THAIDY MORENO RODRIGUEZ

MOLECULAR CHARACTERIZATION OF THE ROLE OF CHROMATIN REMODELING COMPLEXES IN TUMOR PROGRESSION

THESIS

UNIVERSIDAD DE CANTABRIA

FACULTAD DE MEDICINA

DEPARTAMENTO DE BIOLOGÍA MOLECULAR

INSTITUTO DE BIOMEDICINA Y BIOTECNOLOGÍA DE CANTABRIA

(IBBTEC)

Molecular characterization of the role of chromatin remodeling complexes in tumor progression

DIRECTOR: Ignacio Varela Egocheaga

AUTHOR: Thaidy Moreno Rodriguez

Universidad de Cantabria

Octubre 2017

El Dr. Ignacio Varela Egocheaga, Profesor Contratado Doctor del Departamento de Biología Molecular de la Facultad de Medicina de la Universidad de Cantabria

CERTIFICA: que la Lda. Dña. Thaidy Moreno Rodriguez ha realizado bajo su dirección el presente trabajo titulado "Caracterización Molecular del Papel de los Complejos Remodeladores de la Cromatina en la progresión Tumoral" ("Molecular characterization of the role of chromatin remodeling complexes in tumor progression") en el Instituto de Biomedicina y Biotecnología de Cantabria.

Considero que este trabajo reúne los requisitos de originalidad y calidad científica necesarios para su presentación como Memoria de Doctorado por la interesada, al objeto de poder optar al grado de Doctor por la Universidad de Cantabria.

Y para que conste y surta los efectos oportunos, firmo el presente certificado.

Santander, a 24 de Octubre de 2017

Fdo. Ignacio Varela Egocheaga

Esta tesis ha sido realizada en el Instituto de Biomedicina y Biotecnología de Cantabria (Santander) y una parte en el Francis Crick Institute (London).

La financiación necesaria para la realización de esta Tesis doctoral ha sido aportada por:

MInisterio de Economía y Competitividad a traves de los Proyectos del Plan Nacional: SAF2012-31627 y SAF2016-76758-R.

La Fundacion Ramon Areces a traves del Proyecto: Proyecto de investigación de ciencias de la vida 2015: Caracterización molecular del papel de la disfunción mitocondrial en el desarrollo tumoral.

European Research Council a través del proyecto ERC-2014-STG-637904 INTRAHETEROSEQ.

El autor de esta tesis ha disfrutado de una Ayuda para contratos predoctorales para la formación de doctores 2013, referencia BES-2013-062983, concedida por el MInisterio de Economía y Competitividad, al igual que una Ayuda a la movilidad predoctoral para la realización de estancias breves en centro de I+D españoles y extranjeros de la convocatoria 2015, referencia EEBB-I-16-11749.

To my family.

ABBREVIATIONS

ADN	Ácido Desoxirribonucleico
ADNmt	Ácido Desoxirribonucleico mitocondrial
APOBEC	**Apolipoprotein B mRNA editing catalytic polypeptide-like**
ARN	Ácido Ribonucleico
ATAC-Seq	**Assay for Transposase Accessible Chromatin sequencing**
ATCC	**American Type Culture Collection**
ATP	**Adenosine triphosphate**
BAM	**Binary Alignment Map**
BLCA	**Bladder Urothelial Carcinoma**
bp	**Base pairs**
BRCA	**Breast Adenocarcinoma**
BSA	**Bovine Serum Albumin**
cDNA	**Complementary DNA**
CDS	**Coding DNA Sequence**
CFSE	**Carboxyfluorescein diacetate succinimidyl ester**
CGAP	**Cancer Genome Anatomy Project**
CGCI	**Cancer Genome Characterization Initiative**
CGH	**Comparative Genomic Hybridization**
CGP	**The Cancer Genome Project**
CHD	**Chromodomain Helicase DNA-binding**
ChIP-seq	**Chromatin Immunoprecipitation Sequencing**
CLL	**Chronic Lymphocytic Leukemia**
CML	**Chronic Myeloid Leukemia**
CMV	**Cytomegalovirus**
COAD READ	**Colon and Rectal Carcinoma**

Ct	Cycle threshold
CTD2	Cancer Target Discovery and Development program
DAPI	4, 6´-diamino-2 phenylindole dihydrochloride
DEG	Differentially Expressed Genes
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulphoxide
DNA	Deoxyribonucleic Acid
DNAse-Seq	DNase l hypersensitive sites sequencing
DNMT	DNA-methyltransferase
dNTP	Deoxyribonucleoside Triphosphate
Dox	Doxycycline
DTT	Dithiothreitol
EDTA	Ethylenediaminetetracetic acid
EGTA	Ethyleneglycoltetraacetic acid
ENCODE	The Encyclopedia of DNA Elements
FAIRE-Seq	Formaldehyde-assisted isolation of regulatory elements sequencing
FBS	Fetal Bovine Serum
FISSEQ	Fluorescent in situ RNA sequencing
FITC	Fluorescein Isothiocyanate
GBM	Glioblastoma Multiforme
GFP	Green Fluorescent Protein
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HGP	The Human Genome Project
Hi-C	Chromatin Conformation Capture sequencing
HNSC	Head and Neck Squamous Cell Carcinoma
ICGC	International Cancer Genome Consortium

IF	**Immunofluorescence**
INO80	**Inositol auxotroph 80**
ISS	***In situ* sequencing**
ISWI	**Imitation switch**
kb	**Kilobase**
KD	**Knockdown**
kDa	**Kilodalton**
KIRC	**Kidney renal clear cell carcinoma**
LAML	**Acute Myeloid Leukemia**
LOH	**Loss of Heterozygosity**
LUAD	**Lung Adenocarcinoma**
LUSC	**Lung Squamous Cell Carcinoma**
MIDAS	**Mutation Identification and Analysis Software**
miRNA	**Micro-RNA**
MPSS	**Massively parallel signature sequencing**
MspA	***Mycobacterium smegmatis* porin A**
mtDNA	**Mitocondrial DNA**
MW	**Molecular weight**
NCI	**National Cancer Institute**
NGS	**Next-generation sequencing**
NHGRI	**National Human Genome Research Institute**
NP40	**Nonidet-P40 or octyl phenoxylpolyethoxylethanol**
OV	**Ovarian Serous Carcinoma**
OXPHOS	**Oxidative Phosphorylation**
PAGE	**Polyacrylamide Gel Electrophoresis**
PBS	**Phosphate Buffer Saline**
PCR	**Polymerase Chain Reaction**
PDAC	**Pancreatic Ductal Adenocarcinoma**

PE	**Paired End**
PLA	**Proximity Ligation Assay**
qPCR	**Quantitative Polymerase Chain Reaction**
RAMSES	**Realignment Assisted Minimum Evidence Spotter**
RNA	**Ribonucleic Acid**
RNA-seq	**Ribonucleic Acid sequencing**
ROS	**Reactive Oxygen Species**
RPA	**The Replication Protein A complex**
RPKM	**Reads per kilobase per million reads**
rpm	**Revolutions per minute**
RPMI	**Roswell Park Memorial Institute medium**
rRNA	**Ribosomal RNA**
RT-qPCR	**Reverse transcription and quantitative polymerase chain reaction**
SAM	**Sequence Alignment Map**
SBL	**Sequencing by Ligation**
SBS	**Sequencing by Synthesis**
SDS	**Sodium Dodecyl Sulfate**
SDS-PAGE	**Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis**
seq	**Sequencing**
shARID2	**Short hairpin RNA against *ARID2* gene**
shRNA	**Short hairpin RNA**
SOLiD	**Sequencing by Oligo Ligation Detection**
SQCCs	**Lung Squamous Cell Carcinomas**
SWI/SNF	**SWItch/Sucrose Non Fermentable**
TARGET	**Therapeutically Applicable Research to Generate Effective Treatments**
TBS-T	**Tris Buffer Saline –Tween 20**
TCGA	**The Cancer Genome Atlas**

TCR	**Transcription-coupled repair**
TE	**Tris-EDTA buffer**
UCEC	**Uterine Corpus Endometrial Carcinoma**
WGS	**Whole-genome sequencing**
NGS	**Next-generation sequencing**

CONTENTS

SUPPLEMENTARY INFORMATION Digital

INTRODUCTION

“I always have a quotation for everything-it saves original thinking.” Dorothy Sayers

1. NEXT GENERATION SEQUENCING

1.1 The first steps in DNA sequencing

As a natural step after the description in the past century of the double helix structure of the DNA by Watson and Crick; sequencing techniques emerged in 1977 when Maxam and Gilbert developed a method based on chemical cleavage (1) and, in the same year, Sanger and collaborators revealed another technique which was based on a chain-termination method (2). The latter was universally acquired by most of the labs in the world and popularized as Sanger sequencing.

A significant number of modifications were implemented to Sanger sequencing over the years. Thus, a set of four fluorophores allowed the reduction of the number of required independent reactions, and the fragment size discrimination through capillary electrophoresis simplified and automatized the process (3). In the last 80s, the ABI 370A DNA sequencer, the first automatic sequencing machine, allowed for the first time the determination of the encoding sequence of the ß-adrenergic and muscarinic cholinergic receptor genes from rat heart by Craig Venter and colleagues (4,5).

1.2 Second-generation sequencing

The advent of next-generation sequencing (NGS) technology in 2005 revolutionized genomics research. Although there are at present several NGS platforms, all of them are based in the random generation of short DNA fragments and their massively parallel sequencing immobilized into a solid surface. A typical experiment generates thousands of millions of bases of sequence, facilitating the sequence of a complete human genome in a single experiment and for a cost several orders of magnitude lower of the required for traditional technologies. These technologies not only allow the identification of small changes in the genome sequence like base substitutions or small insertions and deletions (indels), but also permit the identification of big genomic rearrangements. For that a paired-end sequencing protocol is typically used. The distance and orientation of the generated paired reads allows the identification of breakpoints in the genome with a few hundred bases of resolution. The big amount of data generated in each experiment requires considerable computer resources and skills but multitude of software is already available for reads alignment and variant identification (6).

The first commercially successful second-generation sequencing system was the 454 Genome Sequencer (GS) of the company 454 Life Sciences in 2005, subsequently acquired by Roche in 2007. Both the 454 and the smaller-scale low cost instrument, 454 Junior, use the approach developed by Ronaghi and colleagues based in a real-time DNA sequencing strategy using detection of the pyrophosphate released in each nucleotide incorporation (commonly called as pyrosequencing) (7). The pyrosequencing process is carried out in fragments immobilized into beads captured inside emulsion droplets. The four di-deoxinucleotides are flushed in rotatory cycles. If there is incorporation, one pyrophosphate per nucleotide is released and converted into ATP by an ATP-sulfurylase emitting light. After capturing the light intensity, the remaining unincorporated nucleotides are washed away and the next nucleotide is provided. This strategy, unlike the one used by other technologies, produces an output of reads of variable length (determined by the template sequence). Additionally, the lack of terminators limits the resolution of polynucleotide tracks (8,9). In 2012, Roche announced the closing of its next-generation sequencing division and the providing of service just through mid-2016 to 454 platforms.

In 2006, the Solexa sequencing platform was commercialized, and acquired by Illumina in early 2007. Called popularly as Illumina sequencing, it is based, like in the case of 454 technologies, on a sequencing-by-synthesis approach using a molecular clustering technique. Nevertheless, in this case, terminator nucleotides are used, each one labelled with a different fluorophore. In each cycle, all four di-deoxinucleotides are flushed into the reaction chamber. After discriminating the nucleotide incorporated in each cluster, the terminator is washed away and the chamber is ready for a new incorporation cycle. In this technology, the template is immobilized in a glass densely coated with oligonucleotides complementary to the library fragment adaptors. In the paired-end sequencing mode after the production of the first read, the sequenced fragment is reversed by the hybridization of the adaptor in the opposite side of the read to a different set of oligonucleotides present in the glass creating a bridge. A new set of sequencing cycles is then performed producing the second read from the opposite extreme of the fragment (10). At present, Illumina offers low throughput cheap benchtop platforms (MiniSeq and MiSeq series) as well as high throughput more expensive production platforms (NextSeq, HiSeq and NovaSeq series), https://www.illumina.com/systems/sequencing-platforms.html.

Sequencing by Oligo Ligation Detection (SOLiD) was the third high-throughput system contemporaneous to the 454 and Illumina technologies (6). Acquired first by Applied Biosystems (ABI) and later by Life Technologies, it was commercially released at the end of 2007. The principal difference in this platform is the use of sequencing by ligation (SBL) strategy, in which a combination of 8-mers labelled with four different colors are used to detect two nucleotides at the same time. As there are more potential di-nucleotide combinations (16 in total) than colors, the redundancy in the use of fluorophores must be resolved by performing several sequencing reactions with an offset of one nucleotide. Consequently, each nucleotide is evaluated twice which improves greatly the error rate of this technology. Nevertheless, the use of SOLiD platforms involves the posterior alignment and analysis of the reads keeping the “color space” system which requires specific software. In addition, the fast increase in the quantity of sequence produced per run together with the decrease in the price obtained by the sequencing-by-synthesis (SBS) platforms, were not maintained by the SOLiD sequencers which reduced greatly the use of this technology by the scientific community at this moment (10-13).

The semiconductor-based machines (Ion Torrent and Ion Proton series) released in late 2010, are the last bet of Life Technologies to compete with Illumina. These instruments do not require fluorescence or chemiluminescence or even a camera for detection which promised to reduce greatly the sequencing reaction time and cost. Similarly to the strategy followed by 454 technology, the IPG system involves the real-time detection of each nucleotide incorporation, in this case directly by sensing changes in the pH produced by the release of a hydrogen ion. Consequently, the lack of terminators limits the resolution of polynucleotide tracks. The average processing time takes one hour per 100 bp (13,14).

1.3 Third-generation sequencing

The generation of new sequencing machines have prompted to some authors to propose the advent of the third-generation sequencing technologies. Although there is no consensus on the definition, there is a general agreement on the characteristics that defines the third generation:

1. Sequencing of single molecules avoiding the need to amplify the DNA fragments (14).
2. Generation of long sequencing reads solving problems related with repetitive regions.
3. Significant improvement of sequencing time and cost (13).

The Helicos sequencing system was the first commercial implementation of single-molecule sequencing, marketed in 2010 by Helicos Biosciences which went bankrupt in 2012 (15). The instrument worked in a similar fashion to Illumina technology by imaging each DNA sequence fixed to a planar surface and extended by a modified polymerase. Opposing to what happens in Illumina instruments, only one single fluorescent nucleotide was used in each cycle. This allowed the increase of sensitivity necessary to work with single molecules (12). The reaction time was high, and the read lengths were sort. However, PCR was not required for sequencing, providing a significant improvement over other technologies (16).

Pacific Biosciences uses a Φ29 DNA polymerase attached to the bottom surface of a plate, which efficiently incorporates phospholinked dNTPs with a high speed allowing very long read sequences (12,17). For the detection in real-time, the sequencer uses a specialized flow cell: a "nanophotonic" structure called zero-mode waveguide (ZMW) that could potentially differentiate a modified base incorporation versus a non-modified one (18). The higher error rates produced by the polymerase are compensated by the application of circular consensus sequencing (CCS) on those bases that are sequenced several times in the same run (19).

Oxford Nanopore Technologies released MinION in 2014, the first commercially available sequencer in portable size (similar to a smartphone) that can be plugged into a USB port. This platform combines the potential for long read lengths (>5 kbp) with high speed (1 bp/10 ns) (20). The technology is based on single-stranded DNA/RNA molecule detection using a change in the electric potential as they pass through a modified α-hemolysin or a *Mycobacterium smegmatis porin A* (MspA pore) (9,21,22). Some reports indicate that MspA-based nanopore sequencing technology is able to detect directly differently modified cytosine nucleotides (5-methylcytosine and 5-hydroxymethylcytosine) (23,24). Among the drawbacks of this sequencer is a median error rate of 12% (13,25). The company announced a new sequencer called "PromethION", still not commercially available, capable of producing ~2 to 4 Tb in 2 days and with read lengths up to 200 kbp (13).

1.4 The future of sequencing technologies

The objective of the future sequencing technologies is to preserve the spatial distribution of the sequencing allowing the reconstruction of the sequences with the original histological localization (26). Several research groups are setting the bases for such technology. For example, Larsson et al, described a method based on padlock probes and *in situ* target-primed rolling-circle amplification for point mutation and individual transcript detection in human and mouse cells and tissues. (27). Additionally, Ke and collaborators developed an *in situ* sequencing (ISS) method to analyze point mutations and gene expression changes in human breast cancer tissue sections (28). Finally, Lee and collaborators described a similar method called fluorescent *in situ* RNA

sequencing (FISSEQ). They analyzed RNA expression from 8102 genes in human primary fibroblasts during a wound-healing assay (29). Although still under the developmental stage, these improvements could be crucial to understanding the context in which these changes at the DNA and RNA levels occurs providing an excellent tool for biological studies.

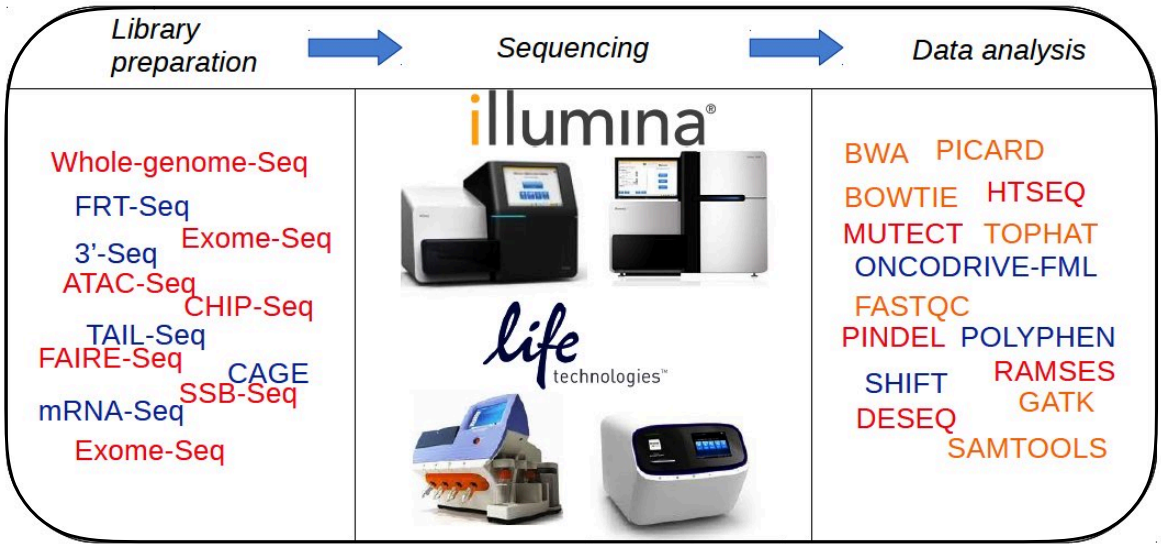


Figure 1. Representation of the different stages in next-generation sequencing. This figure represents the three different stages involved in any next-generation sequencing experiment: Library preparation with a list of potential applications depending on the starting material (red: DNA, blue: RNA), library sequencing with different available platforms and data analysis with a list of available software (Orange: Sequencing processing and alignment, Red: Variant detection, Blue: functional consequence analysis).

1.5 NGS applications

NGS technologies are really flexible and powerful and nowadays more than 200 name-specific applications have been defined. The type of biological information extracted from each one of the applications depends on the concrete combination of the starting biological material as well as the library preparation, sequencing and analysis protocols (30). Here I will summarize only some of those applications that are related to the contents of this thesis.

1.5.1 Identification of genetic and epigenetic mutations

The unbiased sequencing of the total genome (Whole-genome sequencing or WGS) is the most comprehensive method to identify the complete list of nucleotide sequence changes present in the entire genome including single nucleotide variants, indels, copy number variations, and large-scale reorganizations. This can be used to identify differences versus a reference genome (re-sequencing) or to determine the full sequence of a before unknown genome (*de novo* sequencing). Additionally, the application of DNA-Seq technologies to bisulfite-modified DNA allows the identification of methylated nucleotides.

When we are not interested in studying at high coverage the whole genome, the sequencing of specific sequences of the genomic DNA (targeted sequencing), either performed through region capture with designed soluble probes or through PCR-based strategies, constitutes a cheaper alternative. The most extended application is the targeted sequencing of all the protein coding exons of the genome (whole exome) (31). Additionally, several companies offer the customer the opportunity to design a pool of probes to purify any combination of selected genomic regions (32).

1.5.2 Identification of transcriptional alterations

NGS technologies generate a number of reads that is proportional to the initial number of template molecules in the sample. This characteristic opens the door to use RNA-derived cDNA as starting material (RNA-Seq) to perform gene expression studies including all different types of RNA producing genes (mRNAs, non-coding RNAs or small RNAs) with higher sensitivity and reliability than the obtained by array-based technologies. Additionally, as the RNA-Seq strategies are not based in a previously generated gene model, new genes, fusion-genes or splice variants can be identified with the appropriate analysis software (33).

Finally, specific library preparation methods have been developed that retain the molecule strand orientation during the cDNA conversion (directional RNA-Seq) which informs of the DNA molecule that has served as template to generate the RNA (34).

1.5.3 Determination of changes in the functional status of the DNA

As a contrast to the unbiased sequencing of the whole genome, NGS can be combined with the isolation of specific regions of the genome based on its accessibility to specific enzymes, therefore providing information about the structure of the DNA. Thus, DNAse-Seq (regions sensitive to DNAse I digestion), FAIRE-Seq (regions sensitive to cross-link with paraformaldehyde) or ATAC-Seq (regions accessible by Tn5 transposase) are techniques that inform on the open or closed structure of the DNA genome-wide (35).

The specific DNA-DNA contacts can also be elucidated using NGS allowing a higher order of structure resolution. For that Hi-C strategy combines the cross-link of nearby DNA sequences with restriction digestion and random ligation of the generated fragments prior to their genome-wide sequencing in a NGS platform (36).

Finally, ChIP-Seq technique is based in the ultrasequencing of immunoprecipitated DNA using a specific antibody against a DNA-binding protein. In that way, it permits the identification of all protein-DNA binding sites in a concrete moment (37).

1.5.4 Determination of the composition of complex organism communities

The sensitivity and specificity of NGS technologies can also be exploited to deconvolute complex DNA mixtures as the ones produced by the global DNA extraction of different biological materials. This family of techniques, called globally as metagenomics, have become very popular in water or soil ecology studies but have also a high potential for the study of human diseases (38).

In the first place, NGS offers a good opportunity for the identification of the specific microorganism or microorganisms responsible for a human infection with a specificity and speed much higher than the traditional culture-based methods. Additionally, it is estimated that we contain at least as much bacteria as human cells in our body (39) and multiple evidence has been collected about the high impact of the microbiota on human health. Therefore, multiple research groups have focused their efforts in analyzing changes in the microbiota composition as potential players in multiple human diseases (40).

1.6 Data analysis

Accompanying all these advances of laboratory methodologies, a new generation of bioinformatics tools has emerged as a requisite for the analysis of sequencing data. Each application requires specific software designed to extract meaningful information out of the sequencing data. Additionally, it is usually necessary in most of the cases small pieces of code that are generally generated by the proper researchers to adapt the analysis to their specific needs.

It would be very difficult to do a comprehensive review of all available tools so we would focus on those steps that are especially relevant in the context of the present research work.

1.6.1 Sequence quality control and read alignment

Before starting the data analysis, it is convenient to perform an initial quality control of the sequencing data. Basic parameters like the abundance of specific repetitive sequences, the homogeneity of the sequence quality through the read length or the GC content of the sequencing data can be extracted directly from the raw read data with tools like FastQC (https://www.bioinformatics.babraham.ac.uk/publications.html). Other useful parameters like the diversity of the sequenced fragment library (amount of PCR duplicates) or the distribution of insert sizes can be extracted only after read alignment using tools like SAMTools (41) or PICARD (http://broadinstitute.github.io/picard/). Most of the NGS applications require the alignment of the short sequences (reads) generated by these technologies to a reference genome/transcriptome. This step is typically the most resource-consuming process as well as the most critical for the subsequent steps will rely on a correct alignment of the reads. Several aligners have been designed specifically for the use of NGS data. Almost all of them try to identify all potential locations on the genome for a specific part of the read (seed).

Subsequently, this seed is extended and the number of mismatches or gaps required for the alignment of the complete read is computed by the aligner to choose the most probable location of the read. This is finally reported accompanied by a quality score representing the certainty of the alignment. This strategy was first reported by the aligner MAQ (42) and posteriorly improved using Burrows-Wheeler indexing of the reference genome by BWA (43) and Bowtie (44). In the case of RNA-Seq applications, a new tool called TOPHAT (45) works in two stages. First, it identifies all the exons of the genome by a stringent direct alignment of the reads to the reference genome; unaligned reads are subsequently used to find the junctions (splicing events) between the exons.

This alignment is standardized in the SAM/BAM/CRAM format, and several software suites have been developed for its manipulation like sorting, cleaning or indexing. Some examples are SAMTools (41), PICARD (http://broadinstitute.github.io/picard/) or GATK (46). Additionally, some software like IGV allows the graphical visualization of the alignments (47).

1.6.2 Identification of small sequence changes

Identification of base substitutions is probably the most studied topic under the identification of mutations. Many of the tools available are based on a Bayesian model, initially described in MAQ, to compute the different probabilities associated with each potential genotype in a specific genomic position. This approximation works well for germline substitutions. Nevertheless, in the detection of cancer somatic mutations, the normal DNA contamination, copy number alterations and intratumor heterogeneity difficult the construct of an expected frequency model. In this context, modern software have been created specifically to detect somatic single nucleotide variants (sSNVs). Some examples are SomaticSniper (48), JointSNVMix (49), Strelka (50), Varscan 2 (51), Seurat (52) or MuTect (53). Detecting indels have been proved to be a more difficult task and very few tools show good sensitivity and specificity. The main problem relies on the alignment of reads containing these type of mutations. For that reason, some tools like Pindel (54) and Dindel (55) use a second alignment step on singleton reads to identify insertions and deletions. Nevertheless, these tools are not very specific and require generally additional filtering steps.

1.6.3 Detection of genomic structural variants and copy number changes

Paired-end sequencing can be used to identify structural variants like insertions, deletions, duplications, inversions, and translocations. Chromosome mapping, read orientation or insert size of the read pair is used to identify genomic breakpoints. Additionally, breakpoints can also be identified in high-coverage sequencing data in split-apart reads in which different segments of the same read map to different parts of the genome. Some tools that can be used to identify this kind of mutations are Breakdancer (56) and DELLY (57).

The quantitative nature of NGS technologies can also be used to detect copy number changes. Thus, the number of reads coming from a specific genomic region depends on the number of DNA copies for that region. Once the genome has been segmented and the data normalized, some tools allow the identification of deletions or amplifications. Some tools available for this are CODEX (58), CNV-seq (59), Control-FREEC (60) and ExomeCNV (61). The processing of targeted sequencing data offers an extra layer of problems due to potential bias produced during enrichment. To solve this potential problem, the authors of CopywriteR (62) designed their software to use only data off-target and, in theory, not affected by enrichment biases.

1.6.4 Identification of transcriptomic alterations

Different tools for the detection of gene expression changes also rely on the quantitative nature of NGS technologies. In this case, the number of reads generated for a specific transcript is proportional to the number of copies of this transcript in the original sample, the length of the transcript and the total amount of sequence generated from this sample. The expression data is usually normalized in RPKM (reads per transcript kb and per million total reads) or FPKM (fragments/read-pairs per transcript kb and per million total reads) to compare the gene expression across different sequencing reactions. Two of the most broadly used software to detect gene expression changes in RNA-Seq data are DESeq2 (63) and Cufflinks2 (64,65).

RNA-Seq technologies have also the advantage, compared to the previous array-based technologies, of not depending on a previously defined gene model. Therefore, Cufflinks2 suite counts with specific tools that can identify new genes, transcript-specific expression or new splice variants.

Finally, if the two alleles of a specific gene have different nucleotide sequence, RNA-Seq data can be used to detect the specific expression of each one of the different alleles.

1.6.5 Region enrichment identification

The analysis of Chip-Seq, FAIRE-Seq, ATAC-Seq or DNAse-Seq data relies upon the identification of specific regions with a significant increase in sequence coverage compared to the background "noise" of the genome. These regions or peaks can be identified, after a genome segmentation and coverage normalization of the sequencing data, by different statistical methods.

Several tools have appeared in the last years that have been designed for the analysis of this type of data, nevertheless, this is a field of intense research at the moment. The most popular tool at this moment for the analysis of this type of data is MACS (66), but other tools like FindPeaks (67), F-Seq (68) and QuEST (69) are also used.

2. CANCER GENOMICS

2.1 International cancer initiatives: lessons from the human genome project

The Human Genome Project (HGP), launched in the 1990s, constituted the first example of worldwide international collaborative research. This project aimed to determine the complete sequence of the human genome was the birth of the era of team-oriented research in biology, and therefore, the inspiration for many other projects that followed. Thus, The International HapMap Project was launched in 2003 to generate a comprehensive list of the recombination frequencies between different parts of the genome. The 1000 Genomes Project started in 2008 to generate sequencing data from 1000 different healthy individuals. Finally, The Encyclopedia of DNA Elements (ENCODE) project was launched in 2003 to decipher all potential functional non-protein-coding DNA regions of the human genome. This philosophy also arrived at the cancer research field with the launch of several initiatives. The main objective of these projects was the identification of somatic mutations, typically by sequencing both the tumor and corresponding normal tissue DNA, of a sufficient number of tumors of each tumor type to unravel the main genes/ pathways involved in each tumor progression and, subsequently, use this information to improve the diagnosis, prognosis and treatment of cancer patients around the world.

With the advent of high throughput sequencing possibilities, the drastic reduction of sequencing cost catalyzed the initiation of several ambitious cancer sequencing projects. Some examples are The Cancer Genome Project (CGP), The Cancer Genome Atlas (TCGA), The International Cancer Genome Consortium (ICGC) and the St Jude Children's Research Hospital Washington University Pediatric Cancer Genome Project (PCGP).

2.1.1 The Cancer Genome Project (CGP)

The Cancer Genome Project (CGP) was founded at the Wellcome Trust Sanger Institute in 2000 and aims to establish an unbiased catalogue of mutations involved in human tumorigenesis by complete sequencing of candidate genes. The data generated in the CGP was used to produce the list of 616 cancer genes census in the Catalogue of Somatic Mutations in Cancer (COSMIC) (70). Although the discovery of new genes containing driver mutations has slowed down, a huge amount of work is still necessary to explore and understand the specific role of each cancer gene in each cancer type, particularly for the rarer types. Moreover, accurate identification of all driver genes in the cancer of individual patients is of increasing interest for the application of precision medicine.

2.1.2 The International Cancer Genome Consortium (ICGC)

ICGC was organized in 2008 to launch and coordinate a large number of research projects that have the common aim of elucidating in a comprehensive way the genomic changes (somatic mutations, abnormal expression of genes, epigenetic modifications) present in at least 500 tumor samples from 50 different cancer types and/or subtypes which are of clinical and social importance across the globe. In parallel, the project aimed at making the data available to the entire research community as rapidly as possible, and with minimal restrictions, to accelerate research into the causes and control of cancer (http://icgc.org/#about). The Cancer Genome Project (CPG) is englobed in this consortium and TCGA is also a core data and analysis contributor to ICGC, providing about 60 percent of the patient's data.

Currently, the ICGC has received commitments from funding organizations in Asia, Australia, Europe, North America and South America for 88 project teams in 17 jurisdictions to study over 25,000 tumor genomes. Projects that are currently funded are examining tumors affecting: the biliary tract, bladder, blood, bone, brain, breast, cervix, colon, eye, head and neck, kidney, liver, lung, nasopharynx, oral cavity, ovary, pancreas, prostate, rectum, skin, soft tissues, stomach, thyroid and uterus. The genomic analyses of tumors conducted by ICGC members are now available through the Data Coordination Center housed on the ICGC website (https://icgc.org).

2.1.3 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). It leads the effort to study 39 types of human cancers at the genome-scale, collecting hundreds of tumors (http://cancergenome.nih.gov/).

In addition to the released data generated (https://portal.gdc.cancer.gov), for each tumor type TCGA publishes a paper where summarizes the main discoveries and conclusions.

2.2 Molecular insights coming from the cancer genome projects

2.2.1 Mutational Signatures

Somatic mutations are present in all cells of the human body and occur throughout life. They are the consequence of multiple mutational processes, including the intrinsic infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA and defective DNA repair.

In the last decades, several authors have been able to identify mutational patterns associated to different mutagens. Thus, in smoking-associated lung cancer patients, the accumulation of G>T/C>A transversions, enriched at CpG dinucleotides, is a pattern compatible with DNA damage induced by tobacco carcinogens such as benzo[a]pyrene diolepoxide (71). UV-light-associated skin cancers, accumulate C>T and CC>TT transitions (72), and both show transcriptional strand bias due to the action of transcription-coupled repair (TCR) on pyrimidine dimers. Further examples of exposures leading to mutational patterns include G>T transversions in aflatoxin B1-associated hepatocellular carcinomas and A>T transversions in urothelial tumors from patients exposed to aristolochic acid (73-75). Different mutational processes generate unique combinations of mutation types, termed "Mutational Signatures" (http://cancer.sanger.ac.uk/cosmic/signatures) that have been used by several authors to extract useful information about the origin and evolution of particular tumors.

In 2012 Nik-Zainal and collaborators described that many breast cancer genomes have distinctive mutation processes and were able to identify five separate signatures using a non-negative matrix factorization algorithm (75).

One year afterward, Alexandrov and collaborators analyzed 4,938,362 mutations from 7,042 cancers and extracted more than 30 distinct mutational signatures. Some of them are present in many cancer types and others are confined to a single cancer subtype. They revealed that some signatures are associated with the age at diagnosis, mutagenic exposures or defects in DNA maintenance, but many are of cryptic origin (76).

2.2.2 Identification of new mutation accumulation mechanisms

The genome sequencing studies have identified as well in some samples, evidence of the accumulation of mutations that appear to be derived from a single cellular catastrophe event opposing to the conventional gradual acquisition model over time (77).

Thus, Stephens and collaborators described in 2011 the accumulation of a high number of genomic rearrangements restricted to specific genomic areas in at least 2%–3% of all cancers, across many subtypes, and particularly in bone cancers (25% of the cases) that they named chromothripsis. An amount of evidence suggests that this event could occur early in tumor development and could lead to both oncogene activation and tumor suppressor inactivation (78). Chromothripsis is likely to arise through the simultaneous fragmentation of distinct chromosomal regions followed by an imperfect reassembly by DNA repair pathways or aberrant DNA replication mechanisms (79).

Subsequently, Nik-Zainal and collaborators in 2012 observed a remarkable phenomenon of localized hypermutation that they called kataegis. They observed kataegis in 13 of the 21 breast cancers sequenced, with a pattern of C>T and C>G mutations at TpCpX trinucleotides on the same DNA strand, and in some cases associated with rearrangements that had features of chromothripsis, but also colocalized with other rearrangements (75). A subfamily of the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) cytidine deaminases has been implicated as a source of kataegis, in part because aberrant expression in yeast generates a similar C>T substitution mutational signature (80).

The precise molecular mechanism behind all these processes is still under research but Maciejowski and collaborators propose that chromothripsis and kataegis might be the result of DNA repair and APOBEC editing of the fragmented chromatin bridges produced during telomere crisis (81).

2.3 Mutation profile of the main tumor types

The TCGA Pan-Cancer effort was born to compare the data generated from the first twelve tumor types profiled by The Cancer Genome Atlas (TCGA) and published in 2013 by Kandoth y collaborators. These tumors include breast adenocarcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), uterine corpus endometrial carcinoma (UCEC), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), colon and rectal carcinoma (COAD, READ), bladder urothelial carcinoma (BLCA), kidney renal clear cell carcinoma (KIRC), ovarian serous carcinoma (OV) and acute myeloid leukemia (LAML) (82).

An analysis of a total of 617,354 somatic mutations identified in 3,281 tumors allowed the identification of 127 significantly mutated genes both from well-known (like mitogen-activated protein kinase, phosphatidylinositol-3-OH kinase, Wnt/ß-catenin and receptor tyrosine kinase signaling pathways or cell cycle control) as from emerging (like histone, histone modification, chromatin splicing, metabolism and proteolysis) cellular processes in cancer.

According to what has been previously described, *TP53* was the most frequently mutated gene in the Pan-Cancer cohort (42% of samples). Additionally, mutually exclusive mutations in *KRAS* and *NRAS* as well as known activating mutations in *KRAS* (Gly 12 and Gly 13) and *NRAS* (Gln 61), in COAD/READ (30%, 5% and 5%), UCEC (15%, 4% and 2%) and LUAD (24%, 1% and 2%, respectively) were also described. Finally, these analysis identified *EGFR* mutations in GBM (27%) and LUAD (11%) and gain-of-function mutations in *IDH1* and/or *IDH2* in GBM and AML.

Interestingly, mutations in chromatin regulator genes across several cancer types were also observed, in particular histone-lysine N-methyltransferase genes *KMT2D*, *KMT2C* and *KMT2B* in bladder, lung and endometrial cancers, whereas the lysine (K)-specific demethylase *KDM5C* is prevalently mutated in KIRC (7%). Mutations in chromatin remodeling genes like *ARID1A*, were frequent in BLCA, UCEC, LUAD and LUSC, whereas mutations in *ARID5B* predominate in UCEC (10%).

The analysis also uncovered tumor-type-specific mutated genes like is the case of KIRC with mutations in *VHL* (52%) and *PBRM1* (33%) as well as *SETD2* (12%) and *BAP1* (10%) in lower frequency. *CTCF*, *RPL22*, *ARID1A* and *ARID5B* have the highest frequencies in UCEC. Similarly, *APC* (82%) and Wnt/ß-catenin signaling (93% of samples) are the main mutated pathways in COAD/READ. Exclusive mutations in *NPM1* (27%) and *FLT3* (27%) were identified in LAML, but also in *MIR142*, *DNMT3A* and *TET2*. Similarly, *GATA3* and *MAP3K1* are mutated in BRCA. *KEAP1* is frequently mutated in LUAD (17%) and LUSC (12%) and mutations of *EPHA3* (9%), *SETBP1* (13%) and *STK11* (9%) are more characteristic of LUAD (82).

2.3.1 Colon Cancer

A comprehensive genome-scale analysis of 276 human colon and rectal cancer samples was performed as part of the Cancer Genome Atlas, where exome sequencing, DNA copy number, promoter methylation, messenger RNA and microRNA expression data were generated. Approximately 16% of the samples show a hypermutation phenotype resulted from microsatellite instability due to *MLH1* silencing by hypermethylation and *POLE* mutations. Deregulation in the *WNT* signaling pathway was found in more than 90% of tumors. Among the non-hypermutated tumors, 24 genes were significantly mutated being *APC*, *TP53*, *KRAS*, *PIK3CA*, *FBXW7*, *SMAD4*, *TCF7L2* and *NRAS* the most frequently mutated. Finally, the role of new genes as *FAM123B*, *ARID1A* and SOX9 as well as the overexpression of the *WNT* ligand receptor gene *FZD10* were described (83).

2.3.2 Breast Cancer

As part of the ICGC effort, 560 breast cancers were sequenced and 93 genes carrying probable driver mutations were identified. The 10 most frequently mutated genes were *TP53*, *PIK3CA*, *MYC*, *CCND1*, *PTEN*, *ERBB2*, *ZNF703/FGFR1*, *GATA3*, *RB1* and *MAP3K*. Additionally, the study allowed the validation of five new cancer genes that show statically significant accumulation of mutations: *MED23*, *FOXP1*, *MLLT4*, *XBP1*, *ZFP36L1* (84).

The equivalent TCGA effort, were whole-exome sequenced of 510 breast cancers tumors from 507 patients were generated, identified 35 significantly mutated genes, but only three genes with an incidence higher than 10% across all breast cancers (*TP53*, *PIK3CA* and *GATA3*); in luminal A subtype, *GATA3*, *PIK3CA* and *MAP3K1* specific mutations were observed. The study identified nearly all genes previously reported in breast cancer (*PIK3CA*, *PTEN*, *AKT1*, *TP53*, *GATA3*, *CDH1*, *RB1*, *MLL3*, *MAP3K1* and *CDKN1B*) and novel significantly mutated as *TBX3*, *RUNX1*, *CBFB*, *AFF2*, *PIK3R1*, *PTPN22*, *PTPRD*, *NF1*, *SF3B1* and *CCND3*. Statistically significant exclusion mutations in *PIK3R1*, *PIK3CA*, *PTEN* and *AKT1* were also observed (85).

2.3.3 Lung Cancer

As part of The Cancer Genome Atlas, in 2012, Lawrence and collaborators characterized 178 Lung Squamous Cell Carcinomas (SQCCs) to provide a landscape of genomic and epigenomic alterations. The mutation type most commonly observed were CpG transitions and transversions that are associated to the exposition to tobacco smoke. Eleven significantly mutated genes were identified including *TP53* (nearly all the samples), followed by *CDKN2A*, *PTEN*, *PIK3CA*, *KEAP1*, *MLL2*, *HLA*-A, *NFE2L2*, *NOTCH1* and *RB1*. Significantly altered pathways included *CDKN2A/RB1*, *NFE2L2/KEAP1/CUL3*, *PI3K/AKT* and *SOX2/TP63/NOTCH* (86).

Subsequently, in 2014 also as part of TCGA initiative, the molecular profile of 230 untreated lung adenocarcinomas patients (5% lepidic, 33% acinar, 9% papillary, 14% micropapillary, 25% solid, 4% invasive mucinous, 0.4% colloid and 8% unclassifiable) using mRNA, microRNA and DNA sequencing was published. *TP53* was the most commonly mutated gene (46%) followed by *KRAS* (33%) and *EGFR* (14%) which showed mutual exclusivity with *KRAS*. *STK11*, *KEAP1*, *NF1*, *BRAF*, *PIK3CA*, *MET*, *RB1*, *CDKN2A*, and *RIT1* (with frequencies of 17%, 17%, 11%, 10%, 7%, 7%, 4%, 4% and 2%, respectively) were also commonly mutated. Interestingly, the study also found mutations in chromatin remodeling genes such as *SETD2* (9%), *ARID1A* (7%) and *SMARCA4* (6%). *MGA* newly described mutations in 8% of the samples were mutually exclusive with focal *MYC* amplification (87).

2.3.4 Chronic Lymphocytic Leukemia (CLL)

As in others tumor types, the TCGA initiative published the genetic alterations driving tumorigenesis in chronic lymphocytic leukemia (CLL). Whole-exome sequencing of 538 matched DNA samples allowed the identification of 44 putative driver genes, including 18 previously identified, as well as 26 newly putative genes. Previously described genes include *SF3B1 ATM*, *TP53*, *NOTCH1*, *POT1*, *CHD2*, *XPO1*, and *BIRC3*. New driver genes that probably modulate *MYC* activity are *MGA*, *PTPN11*, and *FUBP1*. Genes in *MAPK-ERK* pathway as *NRAS*, *KRAS*, *BRAF* or *MAP2K1* were also identified. Finally, the study identified driver genes belonging to other pathways, including RNA processing and export (*FUBP1*, *XPO4*, *EWSR1* and *NXF1*), DNA damage (*CHEK2*, *BRCC3*, *ELF4* and *DYRK1A)*, chromatin modification (*ASXL1*, *HIST1H1B*, *BAZ2B* and *IKZF3)* and B-cell-activity-related pathways (TRAF2, TRAF3 and *CARD11*) (88).

In the Spanish participation in the ICGC, the DNA from 452 CLL cases plus 54 monoclonal B-lymphocytosis patients were sequenced. *NOTCH1* was the most frequently mutated gene (12.6%), followed by *ATM* (11%), *SF3B1* (8.6%), *BIRC3* (8.8%), *CHD2* (6%), *TP53* (5.3%) and *MYD88* (4%). 12 novel genes without previous association to CLL were identified including *ZNF292*, *ARID1A*, *ZMYM3* and *PTPN11*. Furthermore, the mutations of *BRAF*, *ZMYM3*, *IRF4*, *NFKB2* as well as the 20p deletion, and 2p16 and 5q34 gains showed prognostic value at the time to first treatment and the mutations in *ASXL1*, *POT1* as well as the 14q24 deletion for overall survival. Finally, recurrent mutations in the non-coding region of *NOTCH1* were described to increase *NOTCH1* activity and result in a more aggressive disease (89).

2.3.5 Pancreatic Cancer

The Australian ICGC participation consisted firstly in the sequencing of 142 early (stage I and II) sporadic pancreatic ductal adenocarcinoma cases. The study identified 16 significantly mutated genes. Some had been previously described as *KRAS*, *TP53*, *CDKN2A*, *SMAD4*, *MLL3*, *TGFBR2*, *ARID1A* or *SF3B1*, but they described novel driver genes involved in chromatin modification as

EPC1 and *ARID2*, DNA damage repair as *ATM* and other mechanisms as *ZIM2*, *MAP2K4*, *NALCN*, *SLC16A4* and *MAGEA6*. Additionally, they identified alterations in genes described as embryonic regulators of axon guidance, particularly *SLIT/ROBO* signaling (90).

Subsequently, years later, they performed a deep whole-genome sequencing and copy number variation analysis of 100 pancreatic ductal adenocarcinomas (PDACs), 75 samples with an epithelial cellularity ≥ 40% and, 25 cell lines derived from patients. Genes as *KRAS*, *TP53*, *SMAD4*, *CDKN2A* and *ARID1A* were reaffirmed as drivers and new candidates as *KDM6A* and *PREX2* were identified (91).

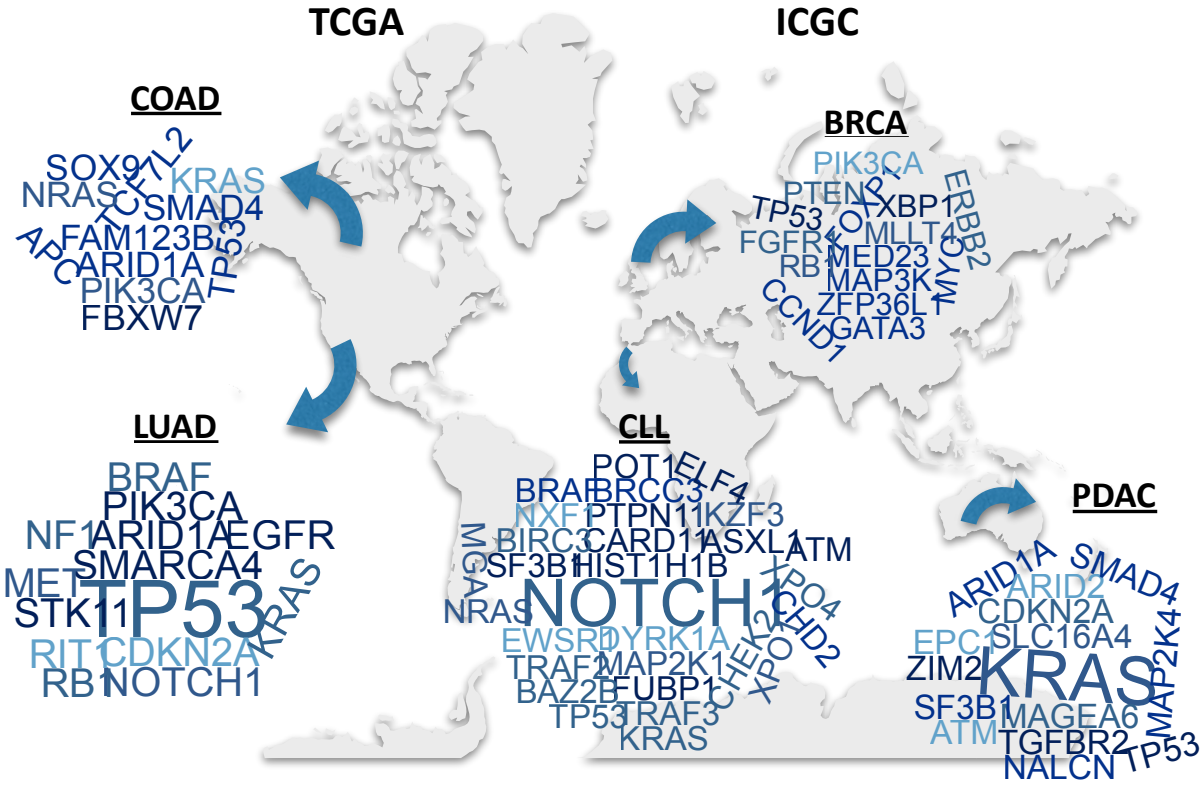


Figure 2. Summary of cancer genome projects. Different wordclouds representing the mutation frequency of the cancer driver genes identified across different tumor types in TCGA and ICGC cancer genome projects.

2.4 Intratumoral heterogeneity (ITH)

The evolutionary mechanisms that could guide cancer progression have become crucial in understanding, predicting, and controlling cancer progression, metastasis, and relapse after surgery or therapy response (92,93). Over the last years it has become evident that there is genetic variation not only between different tumors (intertumor heterogeneity) but also within individual tumors (intratumoral heterogeneity). Was in 1976 when Peter Nowell developed the theory of cancer as a complex cell community that suffers evolutionary processes similar to the ones described in the Darwinian natural selection theory (94). Similarly, in 1999 Cahill and collaborators discussed about the role of genetic instability in tumor formation within the context of Darwinian evolution, and the possible distinct features for instability mechanisms, proposing a model that provides an understanding of how genetic heterogeneity can exist within the 'clonal' process of tumorigenesis (95,96).

The presence of evolutionary independent cell populations inside the tumors challenges the traditional model of lineal progression of cancer, in which a main cell clone is progressively acquiring all capabilities or hallmarks necessary to produce an invasive tumor. Similarly, this new view of tumor progression is difficult to reconcile with the presence of a small population of cancer stem cells, identified in some tumor types and postulated to support the growth of the tumor bulk (97).

According to this, it is plausible that one or other model is more prominent in some tumor types, whereas a combination of both is needed to explain the behavior of others. How this heterogeneity originates is not well understood. It could be either an ongoing process consequence of a general genetic instability or a sporadic event happening in a specific moment during tumor progression.

Taking advantage of the high-sensitivity of next-generation sequencing technologies, the presence of independent cell populations suffering branching evolution has been described in several tumor types (98). This intratumor heterogeneity has important clinical implications. In terms of diagnosis, single

biopsies seem insufficient to capture the whole molecular heterogeneity in the primary tumors, which questions our ability to correctly classify the molecular features of a tumor from a single sample (99,100). Interestingly, Bashashati and collaborators have described the presence, inside the same primary tumor, of clones that could be categorized in different molecular groups (101).

As intratumor heterogeneity has been observed in multiple cancers, this is likely a general problem not restricted to specific tumor types (102).

Treatment of cancer patients is also hampered by the presence of intratumor heterogeneity. As a correct molecular characterization determines nowadays a preferred treatment, it is plausible that targeted therapies only affect some cell clones inside the tumors when others are insensitive to them. Therefore, the treatment could accelerate cancer evolution by removing the dominant clone and relieving interclonal competition (103,104). This could offer an explanation to the frequent cancer relapses seen after targeted therapies. Additionally, if metastases are seeded by some of these minority clones, limited success of the same therapy in metastatic growths can be expected. Consequently, the grade of intratumor heterogeneity has been associated with worse prognosis (105). Finally, specific properties developed by some cell clones, such as the ability to promote angiogenesis or tissue inflammation, could benefit the rest of clones generating a more favorable environment. In that case, it is plausible to hypothesize the existence of collaboration between the different cell clones. In contrast, the presence of highly proliferative but noninvasive clones could impede the proper growth of other more metastatic or dangerous clones. This possibility has been postulated by some authors as a new opportunity for cancer therapy (106).

2.5 Mitochondrial DNA mutations in cancer

Mitochondria are cellular organelles present in all eukaryotic cells that are involved in cellular metabolism, including the ATP generation by oxidative phosphorylation (OXPHOS). The human mitochondrial DNA (mtDNA) strands are differentiated by their nucleotide content, with a guanine-rich strand (Heavy or H-strand) guanine-rich and a cytosine-rich (Light or L-strand). mtDNA is 16 kbp-long and contains the sequence of 13 genes that codify proteins involved in oxidative phosphorylation as well as the whole set of ribosomal and transfer RNA genes. In contrast, all proteins involved in mitochondrial DNA transcription, replication and repair are encoded by nuclear DNA and imported from the cytoplasm.

The presence of abnormal mitochondria in tumor cells has been demonstrated in several studies over the years and a large number of mutations in mtDNA have been described in all tumor types (107,108). These mutations might play a role in tumor development as the presence of dysfunctional mitochondria may offer advantages to tumoral cells. This is the case of the activation of an oxygen independent metabolism in hypoxia conditions or the proliferative signals associated with an increase in reactive oxygen species (ROS) (109). In addition, abnormalities in mitochondrial membrane has been shown to inhibit apoptosis and confer a high resistance to chemotherapeutic treatments (110). Accordingly, mutations in mtDNA have been demonstrated to bear tumorigenic capacity in both murine and human cells, where mutations in mtDNA confer resistance to apoptosis and promote metastasis (111,112).

Most of the mutations described in mtDNA are observed in almost all mitochondrial genomes in the same cell, called homoplasmia. This observation is in line with the hypothesis that cells with dysfunctional mitochondria are positively selected during tumor development. However, Coller and collaborators have proposed that this shift from heteroplasmic to homoplasmic state can be explained without the presence of a selection force (113).

In accordance to a potential role of mtDNA mutations in tumorigenesis, Seok and collaborators identified in 2014, 1907 somatic substitutions in the analysis of the sequencing data of 1675 tumors from the International Cancer Genome Consortium. These substitutions were predominantly C>T and A>G and showed a strong strand bias which lead them to hypothesize that the mutational mechanism responsible for mutation accumulation is fundamentally linked to mtDNA replication (114).

3. CHROMATIN REMODELING COMPLEXES

3.1 Function and families of chromatin remodeling complexes

At the eukaryotic cell nucleus, DNA is associated with proteins (histones), forming packed structures called nucleosomes that constitute the fundamental unit of chromatin. Chromatin structure can be affected by different, but interconnected, processes including covalent modification of histones, exchange of 'generic' core histones by histone variants, disruption of the basic nucleosome structure and histone-DNA contacts, and modification of the DNA itself. This remodeling can play a major part in the multistep process of carcinogenesis (115). Among the described modifications that altered chromatin structure we can find the ATP-dependent chromatin reorganization (116) carried out by chromatin remodeling multi-subunit complexes. These complexes have an important role in the control of cellular processes that use genomic DNA as a template including transcription, replication, recombination, and repair (117).

There are four major families of ATP-dependent chromatin remodeling complexes: SWI/SNF, ISWI, INO80 and CHD (115).

3.1.1 SWI/SNF Family

The SWI/SNF complex is an evolutionarily conserved multi-subunit protein complex in eukaryotes as *S. cerevisiae*, D*rosophila melanogaster* and humans, which uses the energy of ATP hydrolysis to mobilize nucleosomes and remodel chromatin. In mammals, two different complex families have been described: the BAF complex (BRG1 Associated Factors), also called SWI/SNF-A and the PBAF complex (polybromo-BRG1 Associated Factors), also called SWI/SNF-B.

The mammalian complexes of the SWI/SNF family are composed of one of two mutually exclusive catalytic ATPase subunits (either *SMARCA2* or *SMARCA4*); a set of highly conserved "core" subunits (*SMARCB1*, *SMARCC1*, *SMARCC2* and *SMARCE1*); and a variable number of auxiliary subunits that are thought to contribute to the targeting, assembly and regulation of linage specific functions of the complexes (examples of these subunits are *ARID1A*, *ARID1B*, *ARID2*, *PBRM1* or *BRD7*).

Several of these subunits are codified by genes that generate different isoforms produced by alternative splicing. Additionally, most of these genes belong to gene families that often display differential lineage-restricted expression. It is, therefore, likely that a large number of different SWI/SNF complexes probably exist in mammals and contribute to regulate linage and tissue specific gene expression (118).

3.1.2 ISWI Family

The imitation switch (ISWI) family is composed of seven different complexes: *NURF* (nucleosome remodeling factor), *CHRAC* (chromatin accessibility complex), *ACF* (ATP-utilizing chromatin assembly and remodeling factor), *WICH* (WSTF–ISWI chromatin remodeling complex), *NoRC* (nucleolar remodeling complex), *RSF* (remodeling and spacing factor) and CERF (CECR2-containing remodeling factor), each containing one of the two ATPase homologues: *SMARCA5* or *SMARCA1*. The *ACF* and *CHRAC* complexes are

involved in RNA polymerase II transcription and DNA replication and repair, while the *WICH* complexes are implicated in DNA replication and repair. The *NoRC* and *WICH* complexes regulate the transcription of RNA polymerase I and III genes. The RSF complex is implicated in the maintenance of the proper centromere structure. The CERF complex is involved in neurulation and in the regulation of mesenchymal/ectodermal transcription factors also through the regulation of RNA polymerase II transcription (119-122).

3.1.3 INO80 Family

The inositol auxotroph 80 (INO80) family is named for its ability to regulate the expression of inositol-responsive genes and has three complexes: *INO80*, *SRCAP* (Snf2-related *CREBBP* activator protein), and *TIP60/p400* (122,123). The ATPases core of these complexes also have histone acetyltransferase activity. The INO80 complexes are involved in promoting gene transcription, checkpoint regulation, DNA replication, telomere maintenance, chromosome segregation and nucleosome structure but are mostly studied for their role in DNA repair and DNA damage checkpoint responses (124).

3.1.4 CHD Family

The CHD (chromodomain helicase DNA-binding) family includes CHD and NuRD (nucleosome remodeling and deacetylases) complexes. In humans, there are three *CHD* subfamilies complexes. Subfamily 1, which include *CHD1* and CHD2 monomers and, is characterized by the DNA binding domains. They play an important role in the maintenance of embryonic stem cells, DNA damage responses, and tumor suppression. Subfamily 2, that enclose *CHD3*, 4, and 5, and is characterized by two PHD zinc finger domains and the function of the complex depend on the subunit composition. Subfamily 3, contain *CHD6*, 7, 8, and 9, and are characterized by two BRK domains (122,125).

3.2 Chromatin remodelers in cancer

3.2.1 SWI/SNF in cancer

Defects in several subunits of the SWI/SNF family complexes have been associated with tumor progression (126). The 20% of all human malignancies have defects in these complexes, making them the most frequently mutated chromatin remodeling complex in cancer (126,127). Thus, *SMARCB1* is inactivated via biallelic mutations in nearly all malignant rhabdoid tumors, which are aggressive cancers that occur in young children (128). Similarly, *PBRM1* is inactivated in 40 % of clear cell renal carcinoma patients, where lower frequency mutations have been also identified in *KDM6A* and *KDM5C* (histone demethylases) as well as *SETD2* (histone methylase) (126,129). This suggests key roles of chromatin structure in this tumor linage. Additionally, *ARID1A* is mutated in 50% of ovarian cell carcinomas and in 30% of endometrioid carcinomas (130); *ARID1B* single-copy deletions and ARID1A copy number losses have been detected in 74% and 47% of pancreatic cancer patients respectively (131). *ARID2* has been found recurrently mutated in hepatocellular carcinoma, melanoma and non-small cell carcinoma (132-134); and *BRD7* is frequently deleted in breast cancer (135). Finally, *SMARCA2* and *SMARCA4* expression is abrogated in non-small cell lung cancer (136). Interestingly, *SMARCA2* expression is also reduced in prostate cancer where its absence correlates with advanced stages of disease progression and poor prognosis (137). Finally, epigenetic silencing of the *ARID1A* promoter has been observed in breast cancers (138-141).

3.3 Chromatin remodelers in controlling gene expression and development

It is still unclear how impairment of SWI/SNF complexes can lead to cancer development, but essential roles for the complexes have been identified during neurogenesis, myogenesis, adipogenesis, osteogenesis and haematopoiesis. Thus, it is likely that they cooperate with tissue-specific

transcription factors to coordinately balance the activation of lineage-specific genes, the activation of differentiation and the suppression of proliferative programs of many cellular linages (142). In this context, a unique combination of core SWI/SNF components and associated factors is present in ES cells, where it seems to contribute to the maintenance of this linage by regulating pluripotency and self-renewal.

Additionally, direct interactions of SWI/SNF complexes with well-described tumor suppressors like *RB* and *TP53* as well as with the oncogene *MYC* have been described, which suggests that defects in these chromatin remodeling complexes could promote cancer through the activation of canonical cancer pathways (143-145).

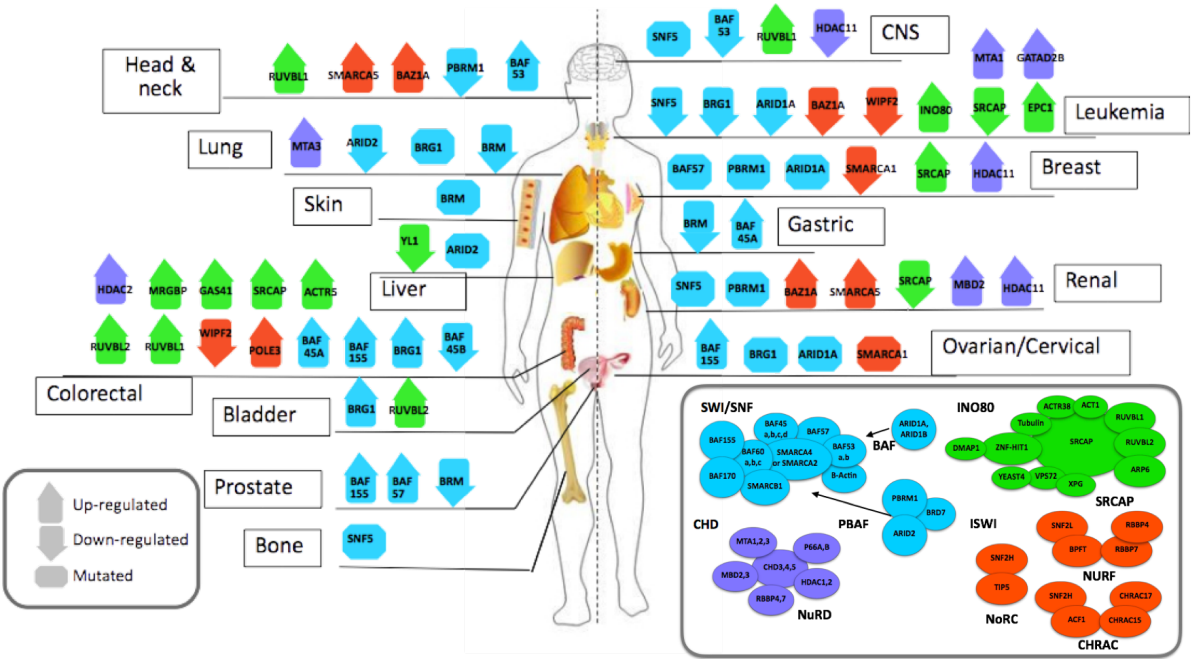


Figure 3. Described alterations on chromatin remodelling genes. Adapted and extended from Weissman and Knudsen (146), with data from Oncomine (www.oncomine.org) and from COSMIC (http://www.sanger.ac.uk/genetics/CGP/cosmic/). Different complexes are represented by colors and type of alteration by different shapes.

3.4 Chromatin remodelers in DNA repair

There is a considerable amount of scientific evidence that links SWI/SNF with DNA repair mechanisms that suggest that a second potential mechanism of tumorigenesis in SWI/SNF deficient cells is through the promotion of genomic instability. In support of this idea, it has been described that mutations in *ARID1A* are mutually exclusive with mutations in *TP53* with could suggest partial redundant functions (127).

There are several mechanisms by which defects in SWI/SNF can promote genomic instability. In first place, SWI/SNF complexes play a role in chromosome segregation in mitosis (147). Secondly, SWI/SNF is described to play essential roles as well in NER (148). Additionally, it has been described that SWI/SNF is recruited to places of DNA damage where they play an essential role in the recruitment and activation of different DNA damage sensors (149,150).

Finally, SWI/SNF is essential for a correct DNA synthesis after DNA damage, a step that is essential for a final damage correction (150,151).

3.5 Therapeutic exploitation of SWI/SNF mutations

The high prevalence of SWI/SNF mutations as well as its broad distribution among different tumor types makes it interesting to study if there is any possibility to exploit potential vulnerabilities derived from SWI/SNF deficiency for cancer treatment. Some of these vulnerabilities have been recently described. Firstly, *MAX* mutations seem to show synthetic lethality properties with mutations in *SMARCA4* (152). Similarly, inhibition of *SMARCA2* protein has been described producing a higher impact on *SMARCA4* mutant tumors (153). Additionally, *ARID1A* mutant cells are more sensitive to the inhibition of *ARID1B* (154).

Another way of action is to exploit a higher genomic instability in SWI/SNF deficient cancer cells to treat the cells with DNA damaging agents. In this line of research, it has been described that *ARID1A* deficiency sensitizes cells to PARP inhibitors (155).

Some SWI/SNF-mutant cancers (*ARID1A*, *SMARCA4* and *PBRM1* mutated) depend on the catalytic and non-enzymatic activity of *EZH2*, perhaps through the stabilization of *PRC2* complex, and are only partially dependent on *EZH2* histone methyltransferase activity (156). Consequently a small molecule-mediated *EZH2* inhibitior has showed promising results in *SMARCB1*-deficient solid tumors (157).

Finally, as several subunits of SWI/SNF complexes contain bromodomains, a specific sensitivity of SWI/SNF deficient cells to bromodomain inhibitors as PFI-3 can be suggested (158,159).

AIMS

“A goal without a plan is just a wish.” Antoine de Saint-Exupéry

The primary aim of this thesis is to determine which chromatin remodeling genes/complexes play an important role in tumor development, as well as, to identify the molecular mechanisms by which defects in these genes play a role and, finally, to determine if these alterations can be used as diagnostic, prognostic or therapeutic tools to improve the management of cancer patients.

This general objective is implemented in the following concrete objectives:

1. Identify those genes of the chromatin remodeling complexes that play an important role in human tumor development.

2. Characterize the molecular mechanisms by which the alterations of chromatin remodeling genes produce their effect in tumor progression and metastasis.

3. Study the possibility of use the alterations in chromatin remodeling complexes to improve the prognosis or treatment of cancer patients.

MATERIALS AND METHODS

“Mathematics is the art of the perfect, physics the art of the optimal and biology the art of the satisfactory.” Sydney Brenner

1. SEQUENCING

1.1 Patient samples

Cancer patient primary samples and, when available, matched corresponding normal samples, were obtained from different tumor Biobanks. In all the cases, we counted with the prior approval of the corresponding ethics committee for each institution. A detailed list of the origin and characteristics of each sample can be found in Supplementary Table 1. In total 479 tumors and 257 matched normal DNAs as well as 45 cell human tumor cell lines (obtained from ATCC and IBBTEC repositories) were used in this thesis (Figure 4).

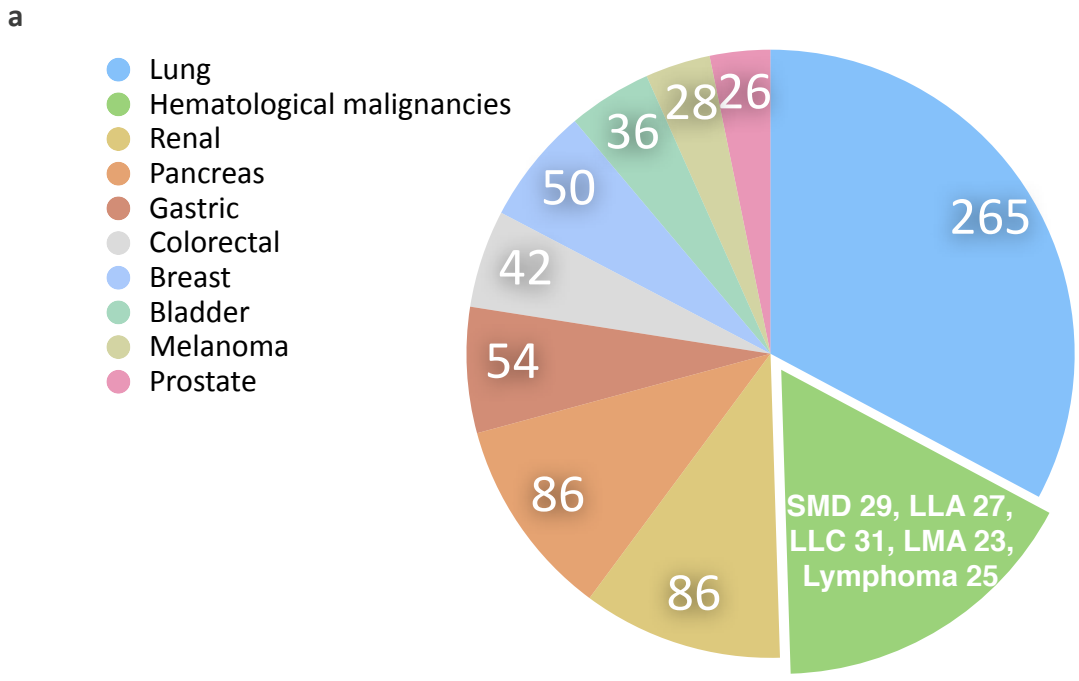


Figure 4. Distribution of sequenced tumors by tissue of origin. Pie chart representation of the number of patient samples that have been sequenced in our screening divided by the tissue of origin (96 of the 265 from lung cancer samples were sequenced only in the codifying region of ARID2).

1.2 Library preparation and sequencing

1.2.1 DNA Extraction

DNA was extracted from fresh frozen tissue or cell lines using the Agencourt DNAdvance Beckman Coulter kit (#A48705, Beckman Coulter, Brea, CA, USA), following manufacturer's instructions. For the formalin-fixed paraffin-embedded sections, the tumor area was micro-dissected, treated with Proteinase K overnight, subjected to a phenol-chloroform organic extraction followed by ethanol precipitation of the DNA.

1.2.2 DNA Libraries

DNA preparations were quantified using the Qubit® dsDNA BR Assay (Q32851, Life Technologies). Normal DNA libraries were performed mixing from 3 to 5 different DNAs. Diagenode Bioruptor® DNA fragmentation was performed with 500 ng of DNA diluted in low TE buffer (#12090015, Thermo Fisher Scientific, UK) to a final volume of 100 µl, and using 30 cycles of 30''/ 30'' (ON/OFF cycles) at 4ºC. For all cleaning steps, we used Agencourt AMPure XP (#082A63881, Beckman Coulter, Brea, CA, USA), following the manufacturer's protocol. Size distribution was analyzed with either the 2100 Bioanalyzer or the 4200 TapeStation using DNA 1000 kit or D1000 ScreenTape Assay (Agilent Technologies, Santa Clara, CA, USA). Sequencing libraries were prepared through a series of enzymatic steps including end-repair and adenylation (DNA Rapid End Repair module, NEXTflex™, #5144-05, Bioo Scientific, Austin, TX, USA), PE adaptor construction through the hybridization of phosphorylated complementary synthetic oligonucleotides, PE adaptor ligation (T4 DNA Ligase, #EP0062, Thermo Fisher Scientific, UK) and PCR indexing amplification (Phusion high fidelity DNA polymerase, # F530L, Thermo Fisher Scientific, UK). Libraries were checked by Nanodrop for chemical contamination, by the 2100 Bioanalyzer for size distribution and finally quantified using the Qubit® and a qPCR reaction with primers designed to target the Illumina adapters. Target capture was performed on pools of 96 libraries using a Sure Select® user-defined probe kit (Agilent Technologies, Palo Alto, CA, USA). The genes contained in each of the designs can be found in Supplementary Table 2. Massively parallel sequencing was carried out in a High-

Seq® machine (Illumina, USA) with a 100 bp paired end (PE) protocol. A single lane was performed for each 96 libraries pool.

In the case of amplicon-based libraries two different strategies where used. When a lot of different products were amplified for the same sample, standard PCR primers were designed, the PCR products generated for each sample were mixed and subjected to the standard library protocol. Alternatively, when a small number of amplicons were designed over a large number of samples, the primers were designed to contain a common adapter sequence that was used, after mixing and purification, to a second PCR to add the barcode and the rest of the Ilumina adapter sequence. These libraries were sequenced in the MiSeq® platform (Illumina, USA) using a 150 or 250 paired-end protocol depending on the amplicon size distribution.

1.2.3 RNA isolation and qRT-PCR

Total RNA was isolated and purified using Extract Me Total RNA Kit (Blirt, DNA Gdansk, Poland) according to the manufacturer's instructions. RNA quality was measured using RNA ScreenTape® (4200 TapeStation Instrument - Agilent Genomics). Reverse transcription was performed using the Takara PrimeScript cDNA Synthesis kit (Takara Bio, Inc., Dalian, Japan) according to the manufacturer's instructions. mRNA expression was measured by qRT-PCR using Luminaris Color HiGreen qPCR Master Mix (Thermo Scientific) with StepOnePlusTM real-time PCR system (Applied Biosystems, Foster City, CA). ß-actin was used as housekeeping gene and the ΔΔCt method was used for quantification and comparison. A list of the primers used for the qRT-PCR experiments can be found in Supplementary table 3.

1.2.4 RNA Libraries

RNA quality and concentration were measured using a RNA Pico chip on a 2100 Agilent Bioanalyzer. For library preparation, mRNA was enriched using NEBNext® Poly(A) mRNA Magnetic Isolation Module. Fragmentation was performed from 1-2 µg mRNA in a buffer containing 4 µL de PrimeScript Buffer and 1 µL random hexamers primers at 94ºC for 15 minutes. The first strand was

synthesized by adding to the previous mix 1 µL of PrimeScript Enzyme and incubating the samples 15 minutes at 37ºC followed by 5 seconds at 85ºC. Second strand was further synthesized by adding to the previous reaction RNAse HI (Thermo) and DNA polymerase I (Thermo) according to manufacturer instructions to a final volume of 100 µL. 2.5 µL of T4 DNA Polymerase (Thermo) was added and the reaction was incubated 5 min at 15ºC. 5 µL of EDTA 0.5 M pH 8.0 was added to stop the reaction. DNA fragments were purified using Agencourt AMPure XP (#082A63881, Beckman Coulter, Brea, CA, USA). Library generation protocol were performed starting from the double-stranded cDNA in a similar way of the DNA libraries.

1.3 Sequencing Data Analysis

1.3.1 First phase

Raw sequence data were subjected to quality control using FastQC v0.11.2 (https://www.bioinformatics.babraham.ac.uk/publications.html) and mapped to the human genome (hg19) using BWA 0.7.3 (43). Samtools 0.1.18 (41) was used for format transformation, sorting and indexing of the bam files. Picard 1.61 (http://broadinstitute.github.io/picard/) was used to fix and clean the alignment and to mark PCR duplicates reads. Finally, GATK 2.2.8 was used to perform local realignment around indels. Bedtools 2.17 (http://bedtools.readthedocs.io/en/latest/#) was used to calculate the enrichment statistics and the target coverage.

For RNA-seq data, paired-end reads from RNA-Seq were aligned using Tophat to the human genome (hg19) (45).

1.3.2 Second phase

Paired tumor/normal bam files were used to identify putative somatic single variants (SVs) using an in-house written algorithm called RAMSES (Realignment Assisted Minimum Evidence Spotter) (217), selecting mutations with a confidence score >2 and mutational frequency higher than 0.05. An in-house Perl script MIDAS (Mutation Identification and Analysis Software) (218). PINDEL 0.2.4 (http://broadinstitute.github.io/picard/) was used to detect indels requiring a minimum of 5 independent reads reporting the indels and with no indels evidence in the control DNA. Potential germline variants were flagged away using 1000 Genomes mutation database with in-house written software.

For transcriptomic analysis, predicted transcripts from Ensembl database were analyzed and transcripts that would lack a CDS start or stop site were filtered out. Differentially expressed genes (DEG) were identified using HTSeq + DESeq (161,162). These R packages for transcriptome expression profile analysis were used according to the manufacturer's instructions to test for differential expression of RNA transcript levels requiring a minimum of 3 counts for a gene in more than two independent samples and using a threshold of fold change >1 and a pvalue <0.05. DEG were manually reviewed and the final list of DEG was created.

1.3.3 Third phase

Functional consequence of the mutations was annotated using ensembl database v.73 through the Perl API. OncodriveFM software was run to detect genes with evidence of selective pressure from the analysis (160).

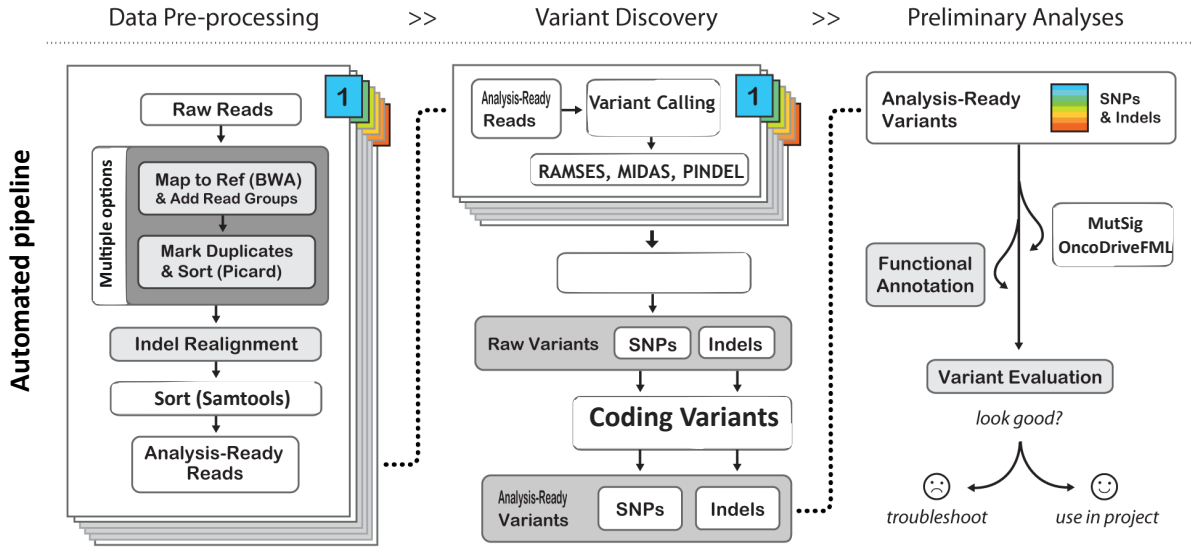


Figure 5. Data analysis workflow. Modified from https://software.broadinstitute.org/gatk/best-practices/, the diagram represents the data analysis stages that have been used in our DNA sequencing data together with the software used.

2. *IN VITRO* E *IN VIVO* EXPERIMENTS

2.1 Cell lines and culture conditions

Both A549 and H460 lung cancer cell lines were sourced from The Francis Crick Institute common repository, authenticated by STR profiling, and tested for mycoplasma. The two cell lines were maintained in DMEM (Lonza, Verviers, Belgium) and RPMI 1640 (Lonza, Verviers, Belgium), respectively, supplemented with 10% FBS (HyClone Victoria, Australia), 1% Gentamycin and 1% Ciprofloxacin at 37°C in a humidified atmosphere containing 5% CO2.

2.2 Generation of stably-transduced cell lines

For stable cell line generation, tetracycline-inducible pTRIPZ constructs V2THS_74399, V3THS_347660 were used for *ARID2* knockdown (Dharmacon/GE Healthcare, Lafayette, CO, USA). The empty vector was used as control. Virus production were performed by transfecting HERK293-T/17clone cells with the pTRIPZ constructs, psPAX2 and pMD2.G plasmids (Addgene) using Fugene HD (Promega Madison, WI, USA). Infected cells were selected with 1 µg/ml puromycin for at least 7 days. Induction of the expression of the shRNAs as well as the turbo-RFP marker was performed with 1 ug/ml of Doxycycline for at least 5 days before analyzing the effect of ARID2 knock-down on the cells.

2.3 FACS sorting of stably- transduced cells

Turbo-RFP expression on cells transduced with shRNAs or empty vectors was induced with 1 µg/ml Doxycycline (Dox) for 16 hrs. Cells were harvested, washed twice in Phosphate-buffered saline (PBS), counted and re-suspended in 0.2 µM filtered sorting buffer containing 250 mL D-PBS (Ca2+ and Mg2+ Free), 3.75 mL of 1 M Hepes Stock solution (final concentration 15 mM), 2.5 gr of BSA in 250 mL (final concentration 1%), 2.5 mL Pen/Strep (100 U/mL of each), 1 mL of 0.5 M stock solution of EDTA (2 mM). Finally, cell preparations were filtered using a nylon mesh with a pore size of 70 µM.

Cells were isolated by FACS based on TurboRFP expression using a FACS-Aria II cell sorter (Becton Dickinson, BD, Franklin Lakes, USA). For proper cell recovery from the sorting process, the cells were collected in tubes containing DMEM supplemented with 50% FBS to prevent the cells from drying out and dying. Cells were seeded in DMEM complete growth medium.

2.4 Proliferation assays

2.4.1 Growth curve

Growth curve analysis was performed over a period of fourteen days. Cells were seeded in 100 mm plates at a density of 500,000 cells per plate. Every two days, cells were trypsinised and the cell number determined by counting using a hemocytometer and re-seeded at a concentration of 500,000 cells per plate. All growth curves were performed in triplicate.

As a complementary approximation, PrestoBlue® (Thermo Fisher Scientific, UK) assay was used to determine cell viability with a colorimetric method. Cells were harvested, washed twice in Phosphate-buffered saline (PBS), counted and re-suspended in DMEM complete growth medium at a

density of 2 * 10^4 cell/mL. Cells were cultured in 96-well plates (BD Falcon, Franklin Lakes, NJ) at a seed density of 2 x 10^3 cells/well. 10 µl of PrestoBlue solution was added to the wells and the plates were incubated at 37 °C for a specified time period. After incubation, absorbance was measured using a Multiskan FC Microplate Photometer (Thermo Fisher Scientific, Waltham, MA) with wavelengths set at 540 and 620 nm.

2.4.2 CFSE

Cell proliferation was analyzed using the carboxyfluorescein diacetate succinimidyl ester labeling method with the CellTrace™ CFSE Cell Proliferation Kit (Invitrogen, CA, USA). The induced cell lines were synchronized by gradual serum deprivation following the protocol described by Lauand and collaborators (163). After 50 h of FBS deprivation, the cells were arrested in G0/G1 phase. Cells were harvested, washed twice in Phosphate-buffered saline (PBS), counted and re-suspended in CellTrace CFSE labelling solution. 1 µL of CellTrace™ stock solution was added to each mL of cell suspension for a final working solution of 5 µM at a density of 10^6 cells/mL and incubated at 37°C for 20 minutes protected from light. DMEM culture media containing FBS was added (10% v/v) to remove any free dye remaining in the solution. After 5 minutes, labelled cells were washed into PBS and pelleted by centrifugation.

Some of these labelled cells were suspended in fresh pre-warmed complete culture medium and were then seeded into 6-well plates at a density of 5 x10^5 cells/well. The remaining labelled cells were suspended in PBS and CFSE fluorescence was measured on a MACSQuant® VYB (Miltenyi Biotec) flow cytometer to ensure the parental population and subsequent division peak tracking during culture. Cells were harvested at defined times and subjected to division peak resolution by flow cytometry. The cell proliferation index was analyzed using ModFit LT™ software (http://www.vsh.com/products/mflt/index.asp). Proliferation index was the sum of the cells in all generations divided by the calculated number of original parent cells.

2.5 Migration assay

In vitro cell migration assays were performed by using 8-µm pore size transwell chambers (Corning™ Transwell™ Multiple Well Plate with Permeable Polycarbonate Membrane Inserts, 3422) in 24-well plates and incubated for 48 h. For the migration assays, 50,000 cells were added into the upper chamber. Cells were plated in medium without serum, and medium containing 10% FBS in the lower chamber served as the chemo-attractant. After 24 h incubation, the cells that did not migrate through the pores were carefully removed using cotton swabs.

The filters were washed with PBS, harvested by treatment with 0.25% trypsin and counted on a hemocytometer. All experiments were performed in triplicate. Filters were fixed in 4% PFA followed by crystal violet staining for microscope visualization.

2.6 Invasion assays

For the invasion assays, 50,000 cells in 50 µL of serum-free DMEM were plated on growth factor-reduced Matrigel (BD Biosciences) pre-coated 8 µm pore transwell chambers, filling the lower chambers with 600 µL DMEM with 10% FBS. After 48 h, non-invading cells were removed from the top of the transwell using cotton swabs. Invasive cells were quantified by fixing chambers in 4% paraformaldehyde for 10 min and staining with crystal violet. To quantify the number of migrated cells, for each Transwell, the cells of 10 random pictures captured under the microscope were counted.

2.7 *In vivo* Tumorigenesis assays

Animal studies were conducted in compliance with guidelines for the care and use of laboratory animals and were approved by the Ethics and Animal Care Committee of Universidad de Cantabria.

2.7.1 Proliferation assays

A549 stably-transduced cell lines were harvested, washed twice in Phosphate-buffered saline (PBS), counted and resuspended in PBS at a density of 10^7 cells/mL.

5 million of cells were subcutaneously injected into the flanks of the 6-8-week-old female mice (Athymic Nude-Foxn1^{nu}). The animals were treated with 1 mM/mL Doxycycline for ~25 days in the drinking water supplemented with 1% sucrose, changed every 2-3 days. After the tumors reached the size of ~0.5 cm^3, mice were euthanizing and tumor tissues were harvested for analyses.

2.7.2 Metastasis assay

A549 and H460 stably-transduced cell lines were harvested, washed twice in Phosphate-buffered saline (PBS), counted and re-suspended in PBS + 0.1% BSA at a density 5 x 10^6 cells/mL.

2.5 million of cells were tail injected into 6-8-week-old female nude mice (Athymic Nude-Foxn1^{nu}). The animals were treated for ~60 days with 1 mM/mL Doxycycline in drinking water for ~60 days supplemented with 1% sucrose, changed every 2-3 days. After two months, mice were euthanizing and tumor tissues were harvested for analyses.

2.8 Western blot analysis

Cells were washed twice in PBS and lysed in RIPA buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1% NP-40, 1 mM Sodium Orthovanadate, 1 mM NaF) containing Halt protease inhibitors Cocktail (Thermo Scientific, 87786), for 30 minutes on ice. Lysates were sonicated using the Bioruptor® (Dia-genode) for ten cycles (30 s on, 30 s off) at high-power and cleared by centrifugation at 16,000 g for 20 min at 4 °C. Protein concentrations were determined by Qubit® Protein Assay (Q33212, Life Technologies). 60-80 µg of total protein lysate was separated by SDS-PAGE in 8% polyacrylamide gels and transferred to nitrocellulose membranes. Subsequently, membranes were washed with TBS-T (50 mM TRIS + 150 mM Sodium chloride + 0,1% Tween 20, pH 7,4) and blocked using 5% non-fat milk solution as blocking agent in TBS (50 mM TRIS + 150 mM Sodium chloride) for 1 h at RT. Membranes were then incubated with primary antibodies anti-ARID2 (E-3, sc-166117, Santa Cruz) and anti-Actin (I-19, sc-1616, Santa Cruz), diluted 1:200 and 1: 1,000 in TBS-T/1% (w/v) BSA at 4°C overnight, respectively. Membranes were washed in TBS-T, three times. The primary antibodies were detected by incubating the membranes in donkey anti-mouse or donkey anti-goat secondary antibodies (LI-COR Biotechnology, Lincoln, USA) conjugated to IRDye 800CW (926-32212) or IRDye 680RD (926-68074) respectively at 1: 15,000 dilutions and incubated for 45 minutes at room temperature. Finally, antibody signals were visualized using Odyssey Clx imager (LI-COR Biotechnology, Lincoln, USA).

2.9 Proliferation inhibition *in vitro* assays

Inhibition assays were performed to determine the half maximal inhibitory concentration (IC50) values for cisplatin, and etoposide in both A549 and H460 stably-transduced cell lines. Briefly, cells were seeded in 96-well plates at 2000 cells per well in 100 µL of complete media, cultured for 24 hours before drug treatment. Drug concentrations ranging from 0.001 µM to 10 mM were prepared in 90 µL of complete media. Cells were treated for 48 hours. Appropriate media and vehicle controls were also added in the media. Viability was determined by adding 10 µL of PrestoBlue® reagent. IC50 value for each drug were determined with GraphPad Prism (https://www.graphpad.com/scientific-software/prism/) software to fit curves to the dose response data.

2.10 Immunohistochemistry analysis

For ARID2 detection on paraffin sections, these were incubated with 1:300-1:500 ARID2 antibody for 32 minutes at 97ºC in citrate buffer pH 6. The sections were developed with HRP-polymer secondary antibodies (Optiview, Roche).

Immunofluorescence was performed in stable cells induced with 1 µg/mL Doxycycline. Cells reach 50–70% confluence on sterile cover slips were rinsed twice with PBS and fixed with 4% paraformaldehyde in PBS for 15 min at room temperature. Cover slips were rinsed tree times with PBS for 5 minutes. Permeabilization was performed with 0.5% Triton X-100 in PBS for 5 minutes at room temperature. The cells were blocked with 3% BSA in PBT (PBS containing 0.05% Triton X-100) and subjected to immunofluorescence staining with ARID2 (1:50) antibody for 30 minutes at room temperature in moist chamber. The cover slips were then washed with PBS three times for 5 minutes. Cells were incubated with Alexa labeled secondary antibodies (1:400) for 30 minutes at room temperature in moist chamber protected from light. Cover slides were mounted in VECTASHIELD Antifade Mounting Medium with DAPI (Vector Labs, Burlingame, CA, USA). The cells were finally examined by fluorescence microscopy (Olympus America Inc, Center Valley, PA). Quantification of fluorescent intensity was performed from randomly selected fields using MetaMorph® (https://www.moleculardevices.com/systems/metamorph-research-imaging/metamorph-microscopy-automation-and-image-analysis-software) and ImageJ software (https://imagej.nih.gov/ij/).

RESULTS

“The greatest obstacle to discovery is not ignorance-it is the illusion of knowledge.”Daniel J. Boorstin

1. OVERALL SEQUENCING RESULTS

1.1 Number of mutations found and mutation profile

In total, we have sequenced 736 samples (479 tumors and 257 matching controls) of eleven different solid tumor types as well as different hematological malignancies (Figure 4). We obtained a mean of 40% of the reads on our region of interest obtaining at least a mean coverage of 50x in 90% of the sequenced samples. The percentage of reads on target is lower than expected and of the one specified for the manufacturer of the enrichment kit. This could be probably the result of the small size of our target region or of the multiplexing strategy that we followed prior enrichment (see Methods). Nevertheless, we obtained very good coverage for most of the sequenced samples.

In total, we have identified 4920 somatic coding mutations with a distribution represented in the Figure 6 (Suppl. Table 4). On average, we were able to call a median of five variants per sample, mainly single nucleotides substitutions, and most of them missense (Figure 6 B). We observed that the mutational profile consists on "C→T" transitions, followed by "C→A" transversions and "T→C" transitions (see Figure 6.c). As expected, the most frequently mutated genes included *TP53*, *KMT2C*, *KMT2D*, *APC* or *KRAS*.

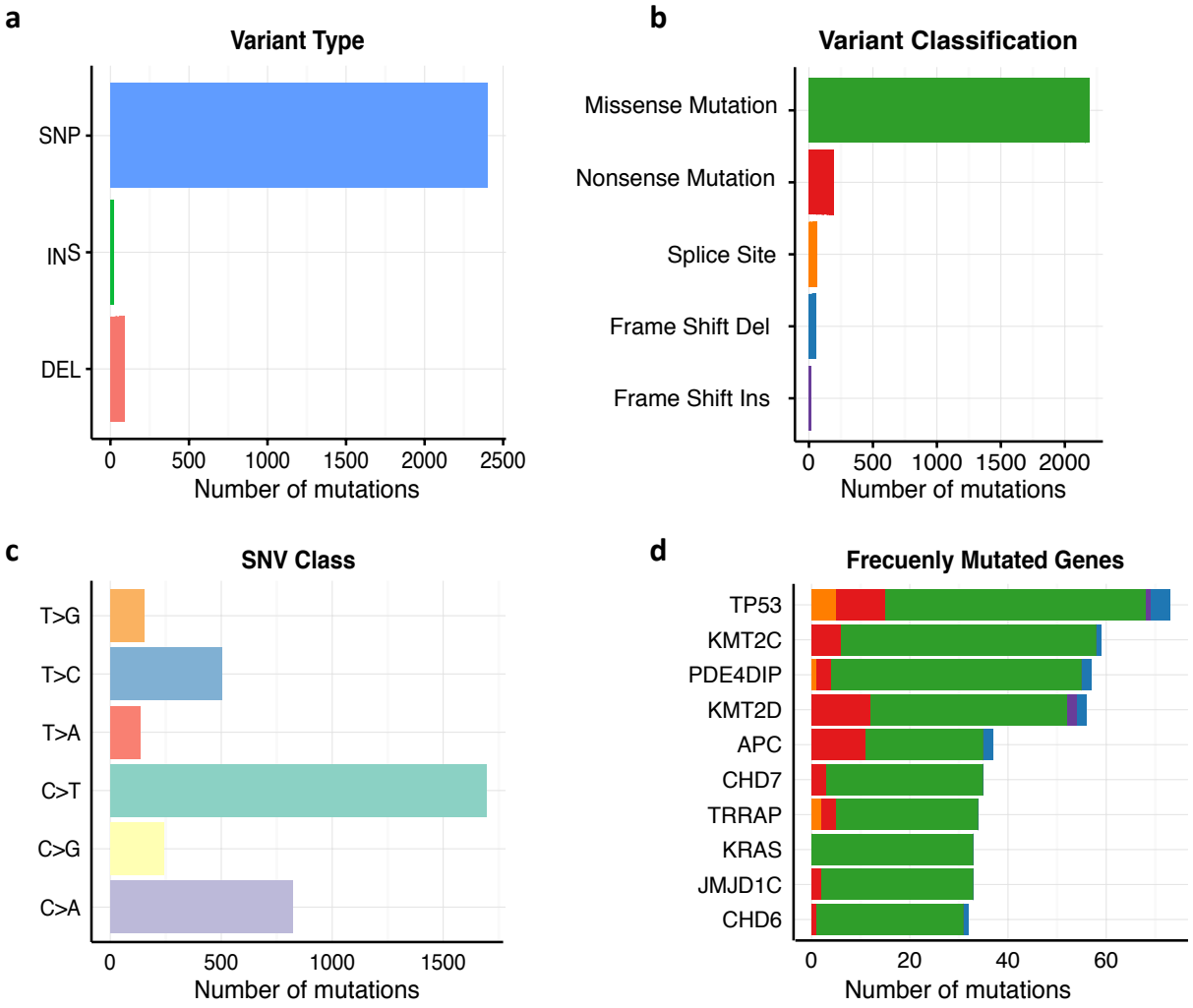


Figure 6. General statistics of the mutations found. (a) Representation of the mutations found according to the mutation type. (b) Distribution of the functional consequences of the mutations. (c) Substitution profile. (d) Most frequently mutated genes found in our screening. For panels b and d, missense (green), nonsense (red), splice site (orange), frameshift deletions (blue), frameshift insertions (lilac) and nonstop (light green) mutations are represented.

1.2 Data specificity

In order to check the validity of our sequencing/analysis strategy, we compared our list of mutations to COSMIC database (http://cancer.sanger.ac.uk/cosmic). Interestingly, 525 of our 4920 mutations had been already reported in this database as is the case of specific mutations of *ARID1A*, *KRAS* or *TP53* (see Figure 7a), which indicates that all these mutations are likely real.

As a second validation strategy, we interrogate the mutation frequency of the known cancer genes in the different tumor types. As it can be seen in Figure 7b, the mutation frequencies of genes like *APC* in colon cancer, *VHL* and *PBRM1* in renal cancer, *KRAS* in Colon and Pancreatic cancer and *ARID1A* in Lung and Colon cancer are very similar to those ones described in the literature by several authors.

Finally, we performed high coverage (3000x) PCR-based ultrasequencing in a randomly selected list of identified mutations (160). We found that near 90% of all the assayed mutations were real (141/160) which demonstrates a high specificity of our strategy. Unfortunately, 29 of them turned out to be present also in the matched normal sample and are, therefore, germline. This is probably the result of a reduced sequencing coverage on that area in the matched normal, likely as a result of the normal pooling strategy that we followed during the original sequencing screening (Figure 7c).

All these results prove that our sequencing/analysis strategy shows a high specificity and supports a high reliability of our mutation list.

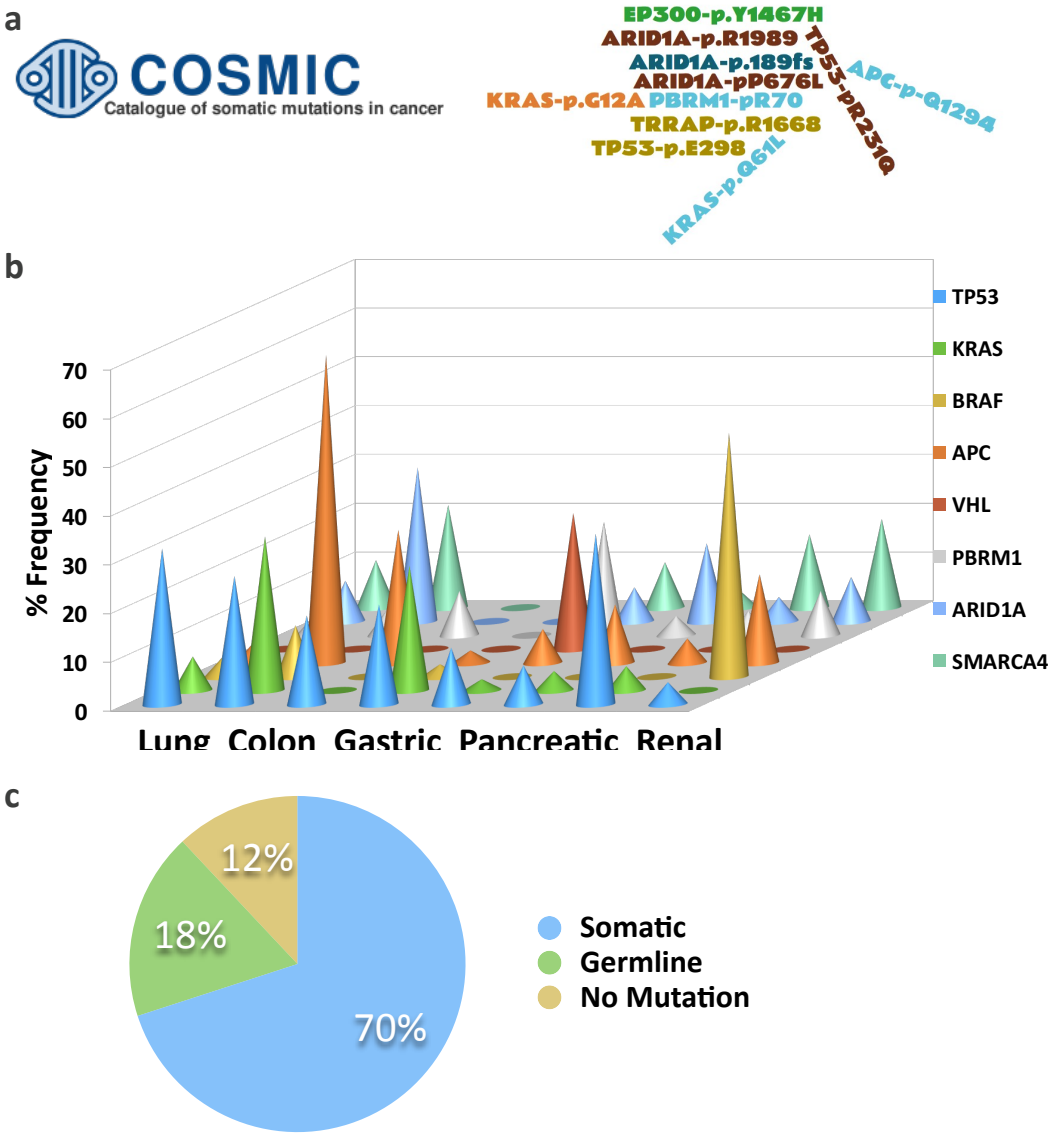


Figure 7. Sequencing/analysis strategy validation. (a) Word-cloud representing some of the identified somatic mutation already described in COSMIC database. (b) Mutated patient frequency according to the different cancer tissues of the main tumor cancer genes included in our screening. (c) Pie chart representing the distribution of results of the orthogonal validation of a random selection of identified mutations.

2. MUTATIONS ACROSS MITOCHONDRIAL GENOME

2.1 High accumulation of mtDNA mutations

A representation of the number of mutations per chromosome, (Figure 8) showed a strikingly high number of mutations in the mitochondrial genome. As it has been broadly discussed in the introduction, several authors have suggested an active role of the mitochondrial mutations in tumor progression.

Counting with a large number of tumor samples, we decided to dedicate part of our efforts in the study of these mutations. In order to achieve better coverage specifically in the mitochondrial genome, we performed a second target probe design, including the mitochondrial genome and genes involved in mtDNA replication and repair, following the same procedure as in the previous design.

2.2 Distribution of mtDNA mutations

In total, we identified 170 mutations in the mtDNA (Figure 8 and Suppl. Table 5).10% of the identified mutations show a mutant read frequency > 80% and therefore we consider them to be in homoplasmia. That means that they are in the majority of the mitochondrial genomes, whereas the rest of the mutations are present in heteroplasmia (only present in a percentage of the mitochondria). This observation is quite informative considering the fact that these mutations are somatic which suggests a fast shift from a heteroplasmic to a homoplasmic state which indicates a selective pressure in the tumor cells to accumulate defective mitochondria. Additionally, we see a homogeneous

distribution of the mutations across all mtDNA genes which indicates that any alteration of any of the genes might have the same effect.

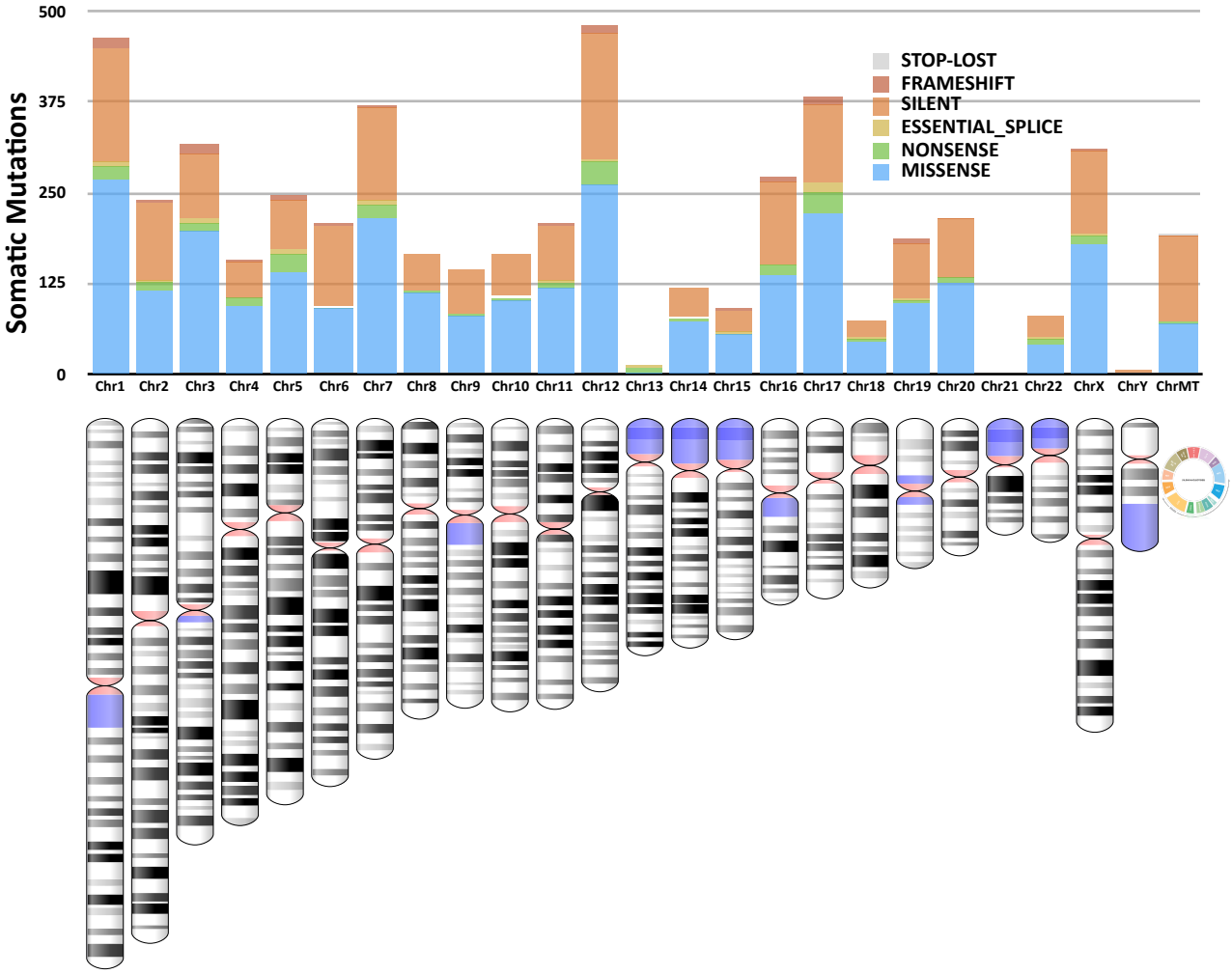


Figure 8. Genomic distribution of somatic mutations. Representation of the number of somatic mutations identified in each chromosome. missense (blue), nonsense (green), splice site (yellow), silent (orange), frameshift (red) and stop lost (gray) mutations are represented.

2.3 Mutational profile and strand bias

A deeper analysis of the mutations found in the mtDNA showed firstly a very striking substitution profile. It is believed that the main cause of mitochondrial DNA damage is the accumulation of Reactive Species of Oxygen (ROS) generated during the oxidative phosphorylation. It is described that the main type of substitutions produced by oxidative stress are G>T/C>A transversions generated by oxidation of guanine to 8-Oxoguanine (8-oxo-G); and T>C/A>G transitions as a result of the oxidation of thymine to thymine-glycol. Interestingly, our data show a large amount of T > C/A>G (Figure 9) due we practically do not detect any G>T/C>A transversions. Conversely, we detect a lot of G>A/C>T transitions that cannot be explained as the result of an excessive oxidative stress.

This effect can be explained by two different mechanisms: either 8-oxo-G is very efficiently repaired in the mitochondria, and therefore does not produce many mutations or, alternatively the 8-oxo-guanine produces a different substitution in the mtDNA than the one produced in the nuclear DNA. It is described that the substitution produced by a DNA damage is highly dependent on the specific DNA polymerase that encounters the DNA damage. According to that, it is plausible that the mtDNA replicating polymerase (POLG) could produce a substitution profile as a result of the accumulation of 8-oxo-G, different from the one produced by the nuclear DNA replicating polymerases.

Additionally, when we plot the ratio of substitutions according to the DNA strand, we see a clear strand bias in the composition of mutated bases in the G>A/C>T transitions between the two strands. Moreover, this strand bias is opposite to what one could expect considering known differences in the nucleotide composition between the heavy and light strand. Thus, more substitutions affecting cytosines in the heavy strand are found whereas this strand is known to contain a lower ratio of that base (Figure 9).

Interestingly, if we continue with the hypothesis that the G>A/C>T mutations are the result of a damaged G, this strand bias would correspond with an accumulation of mutations in the untranscribed versus the transcribed strand. This phenomenon has been observed in the nuclear DNA and is explained by the effect of the transcription coupled repair (TCR) mechanisms. Nevertheless, in nuclear DNA, this DNA-Repair mechanism is carried out by the nucleotide excision repair (NER) machinery and so far, NER proteins are not described to act in the mitochondria. According to all this, we can hypothesize that either the transcription coupled repair is carried out by another complex in the mtDNA or NER proteins do have an effect in the mitochondrial genome.

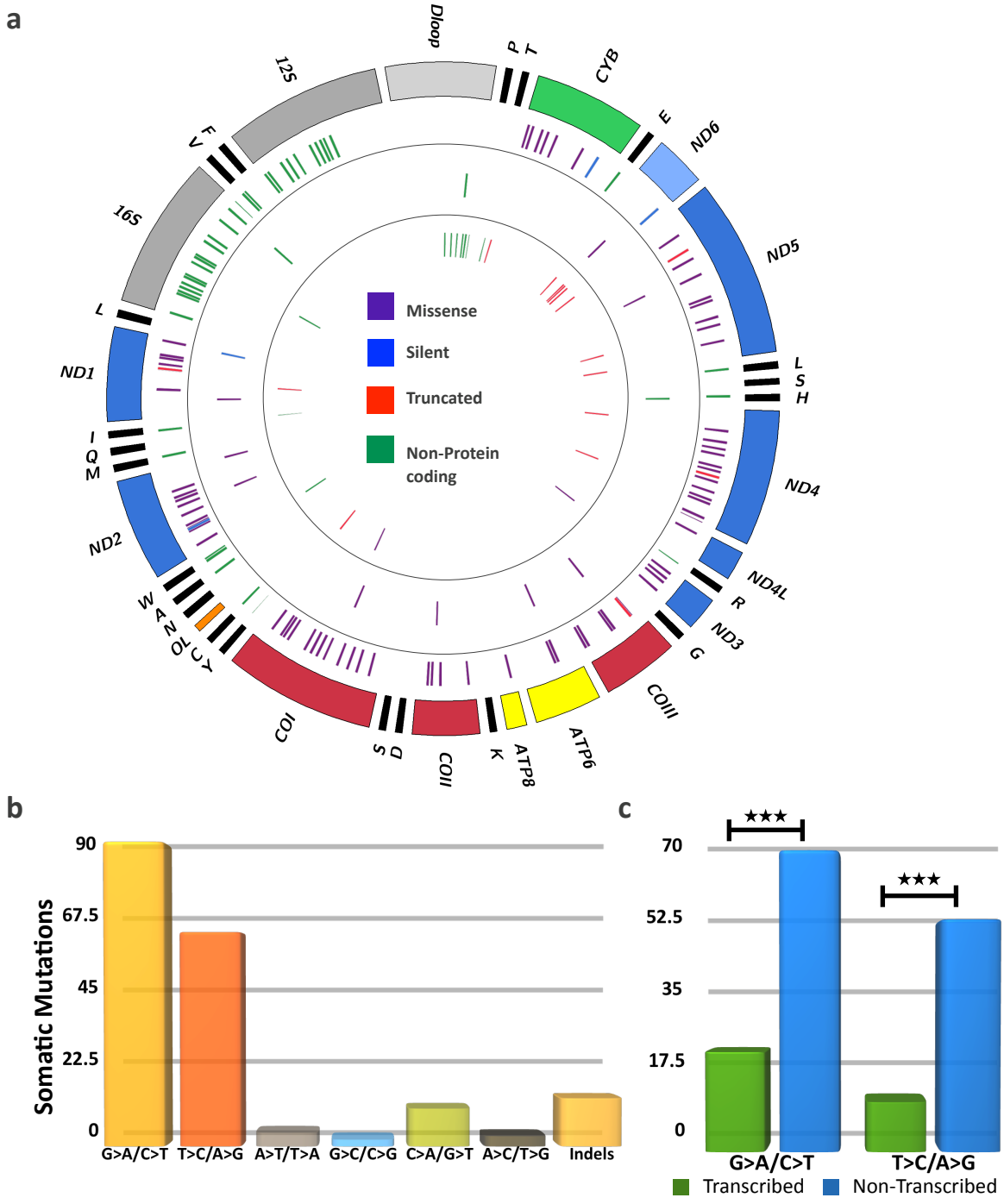


Figure 9. Mutations across mitochondrial genome. (a) Circus plot representing mitochondrial genes organized by color according to the complex to which they belong, NADH dehydrogenase complex (blue), cytochrome oxidase c subunits (red), mitochondrial ATPase complex (yellow) and cytochrome B (green). ND6 is represented in lighter blue to indicate that is in different strand that the rest of the genes. Concentric circles represent the mutation distribution according to functional consequence: missense (purple), silent (orange), truncating protein mutations (red) and the non-codifying mutation (green). From outer to inner tracks, mutations in heteroplasmia (first circle), mutations in homoplasmia (second circle) and small insertions and deletions (last circle) are represented. (b) Substitution spectrum of the mtDNA found mutations. (c) Mutation strand bias represented by mutation counts in transcribed (green) versus untranscribed (blue) strand, (*** $p < 0.001$).

3. DRIVER GENES IN CHROMATIN REMODELING COMPLEXES

Coming back to our complete list of mutations and in order to differentiate those genes that act as drivers of the tumorigenesis (that is, whose mutations confer an advantage to the tumor cells) of those that act as passengers; we run OncodriveFML (164). This software is designed to identify those genes that have evidence of selection showing a profile of mutations significantly different of what expected by chance.

OncodriveFML identified *TP53*, *KRAS* and *APC* as the genes that show stronger evidence of selection (Figure 10a). As the software does not use any functional information, this result argues in favor of the capacity of the strategy to find real cancer driver genes. In addition to known cancer genes, we were able to find in the top of the list several chromatin remodeling genes (Figure 10b and Suppl Table 6). That is the case of *ARID2*, *SRCAP* or *SMARCC1*. Some of these genes, like the case of *ARID2*, have been previously associated with specific tumor types like melanoma. Nevertheless, we see a striking recurrence of mutations in other unreported tumor types like in lung cancer. In the other side, *SRCAP* has not been reported as frequently mutated in any cancer subtype. Interestingly, in our screening we see a high mutation frequency particularly in colon cancer and in pediatric lymphoblastic leukemia. In this thesis, we will study more deeply some of these observations whereas the others are still open research lines that will be followed in the future by our laboratory.

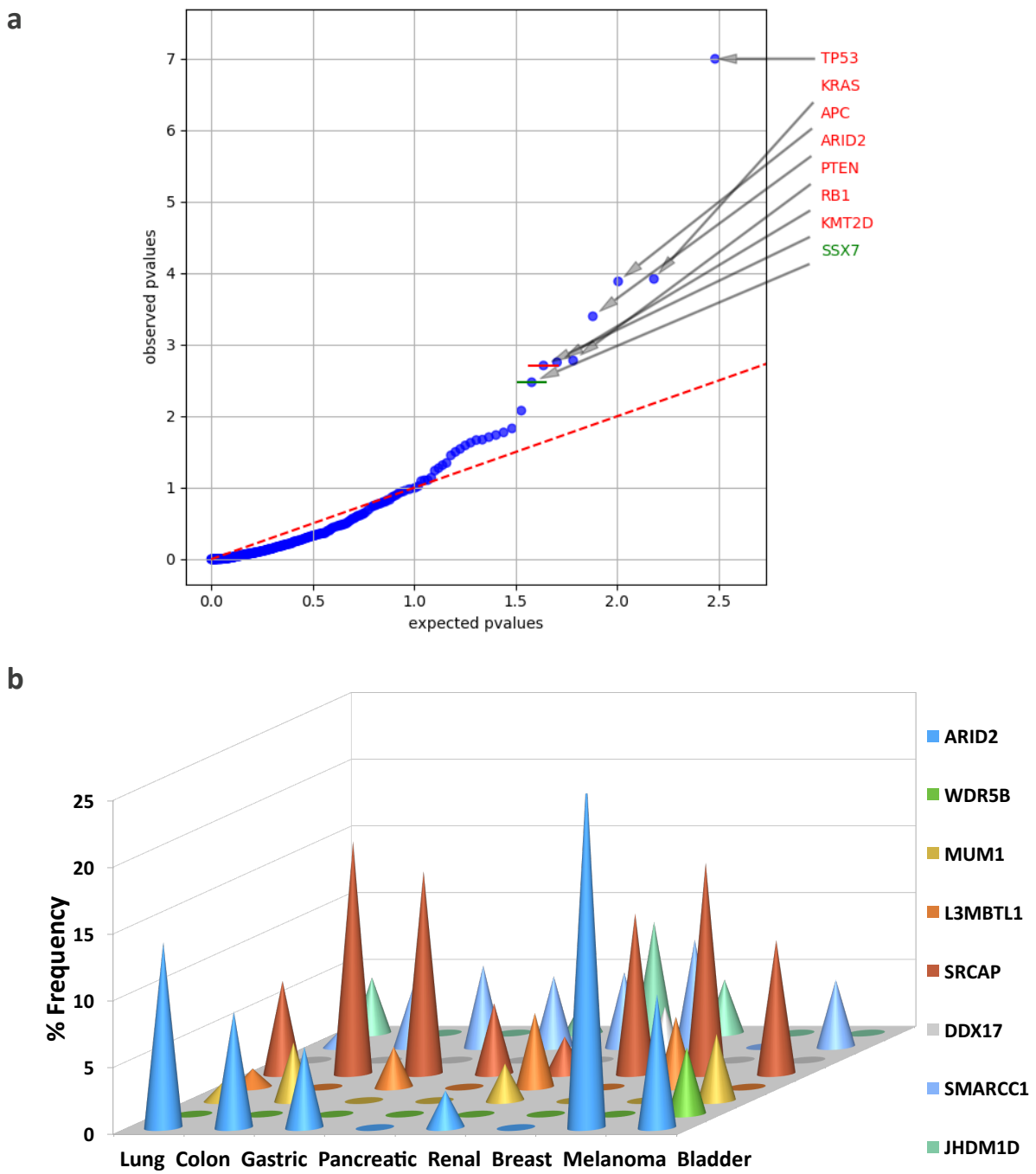


Figure 10. Prediction of cancer driver's genes. (a) Quantile-quantile plot comparing observed and theoretical P values of driver genes among all cancer types in our study predicted by OncodriveFML. (b) Mutated patient frequency divided by tissue of the top raked genes according to OncodriveFML.

3.1 Evidence of selection of driver genes in SWI-SNF complex

67% of our studied samples have at least one mutation in a chromatin remodeling gene which argues in favor of a very important role of these complexes in tumor development. Nevertheless, the abundance in the top of the driver gene list of members of the SWI-SNF complex prompted us to investigate if there is evidence of this complex playing a more important role in tumor progression than the rest.

For that, we performed a Gene Set Enrichment Analysis (GSEA) using the genes belonging to the different complexes as user-defined gene sets, and the significance-ranked driver list from OncodriveFML as input data. As it can be seen in Figure 11, this analysis revealed a statistically significant enrichment of driver genes in the SWI/SNF complex, whereas no enrichment can be seen in the rest of complexes. These results could argue in favor of a more important role of SWI/SNF complex in tumor development compared to the others complexes.

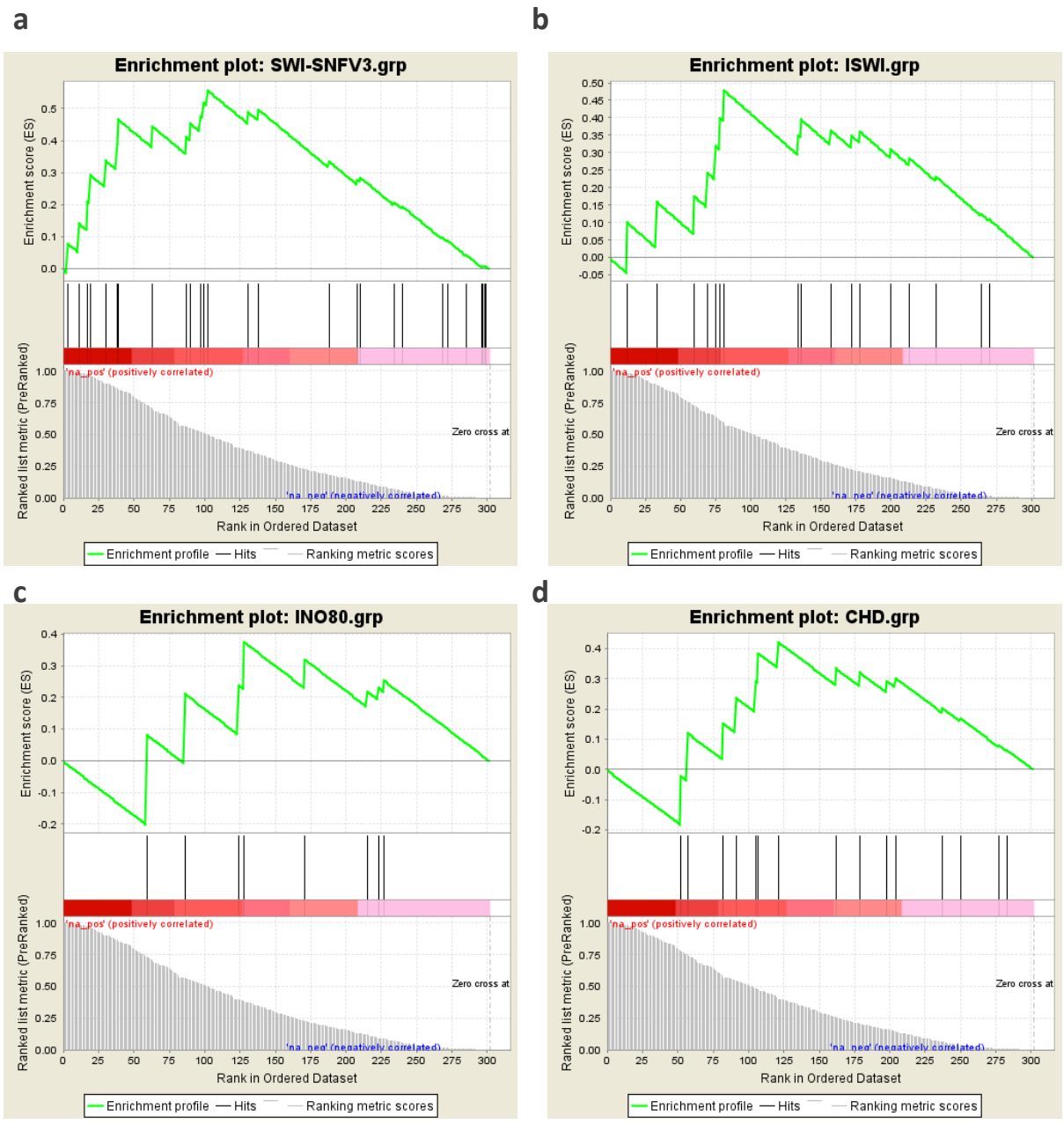


Figure 11. Gene set enrichment analysis (GSEA) results of the driver genes identified by OncodriveFML. The histograms show the gene distribution by complex ranked by p values predicted by OncodriveFML. In each plot, the vertical black lines indicate the position of each of the gene in the pre-ranked chromatin complex data set. The green curve denotes the Enrichment Score (ES). Genes are grouped according to their role in the different chromatin remodeling complexes: SWI/SNF(a), ISWI (b), INO80 (c) and CHD (d).

3.2 SWI-SNF mutations tissue specificity and mutual exclusivity

According to the previous results, we decided to focus our studies in the members of the SWI-SNF complex. When we analyze the distribution of these gene mutations in our sample series, we can see that most of the genes show clear tissue specificity. That is the case of *ARID2* in melanoma, *ARID2* and *SMARCA4* in lung cancer or *PBRM1* in renal cancer. Nevertheless, there are genes, like *ARID1A*, that seems less tissue specific (Figure 12a). These indicate that the mechanisms by which these gene alterations are involved in tumorigenesis are not completely equivalent. This could also be associated to the specific SWI-SNF complex to which they belong: *PBRM1*, along with *ARID2* and *BRD7*, are the defining subunits of the PBAF complex, while *ARID1A* and *ARID1B* define the BAF complex and both components are being mutually exclusive subunits in the BAF complex. Besides this, there is a clear tendency of not accumulating more than one alteration in one chromatin remodeling gene as there is only 1 sample in our cohort with more than one subunit altered (Figure 12b).

Furthermore, when we extend this mutation exclusivity to the highly frequently mutated genes in our screening like *TP53*, *APC* or *KRAS* we see that this exclusivity also involves these genes (Figure 12c). These observations suggest the existence of either incompatible or alternative mechanisms of tumor induction between this commonly mutated cancer genes and the alterations in chromatin remodeling.

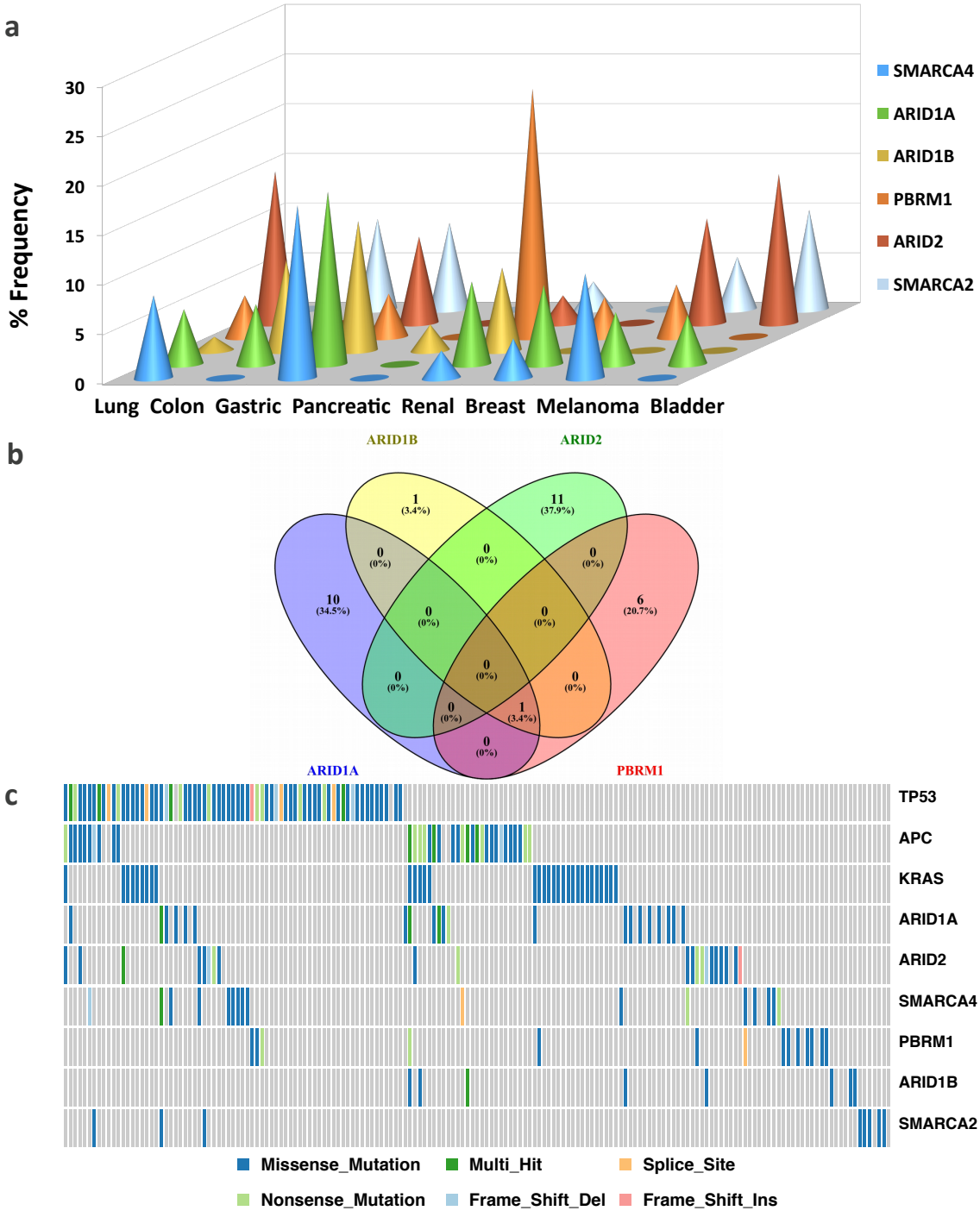


Figure 12. Mutual exclusivity behavior of SWI-SNF mutations. a) Bar graph representing the mutated patient frequency of the main chromatin remodeling genes according to the different cancer types. (b) Venn diagram plot comparing the number of patients mutated in the most frequently mutated SWI-SNF genes (http://bioinfogp.cnb.csic.es/tools/venny/). (c) Representation of the mutated patients for the main SWI-SNF genes including the three most-mutated cancer genes in our screening. Each vertical line represents an independent patient. Colored bars represent a mutated patient for the corresponding gene. Different colors specify the type of mutation.

4. *ARID2* AS A *BONA FIDE* TUMOR SUPPRESSOR GENE IN LUNG CANCER

4.1 Recurrent *ARID2* Mutations in lung cancer patients are associated with loss of protein production and worse prognosis

In our original screen, we observed a high frequency of *ARID2* mutations in our lung cancer patient cohort. Although *ARID2* has been described as a driver gene in melanoma and hepatocarcinoma, its role in lung cancer has not been studied in depth and only one article reports mutations in this gene in 5% of non-small cell lung carcinomas (134). According to that we decided to study deeper the potential role of this gene in lung cancer. When we analyzed our lung cancer cohort, we saw that *ARID2* is the fourth most mutated gene, with 13% frequency, just after genes like *TP53* and *KMT2C* widely reported as mutated in this tumor type (Figure 13). ARID2 is a specific subunit that has been originally identified forming part of the PBAF complex, where it is considered essential to PBAF complex assembly as without it, the complex is disintegrated and PBRM1 degraded. However, the absence of *PBRM1* does not affect the complex formation or produces a diminution in ARID2 levels and PBRM1-deficient complexes have been reported to be capable of regulating transcriptional programs and mediating expression of IFN-α-inducible genes in HeLa cells (165).

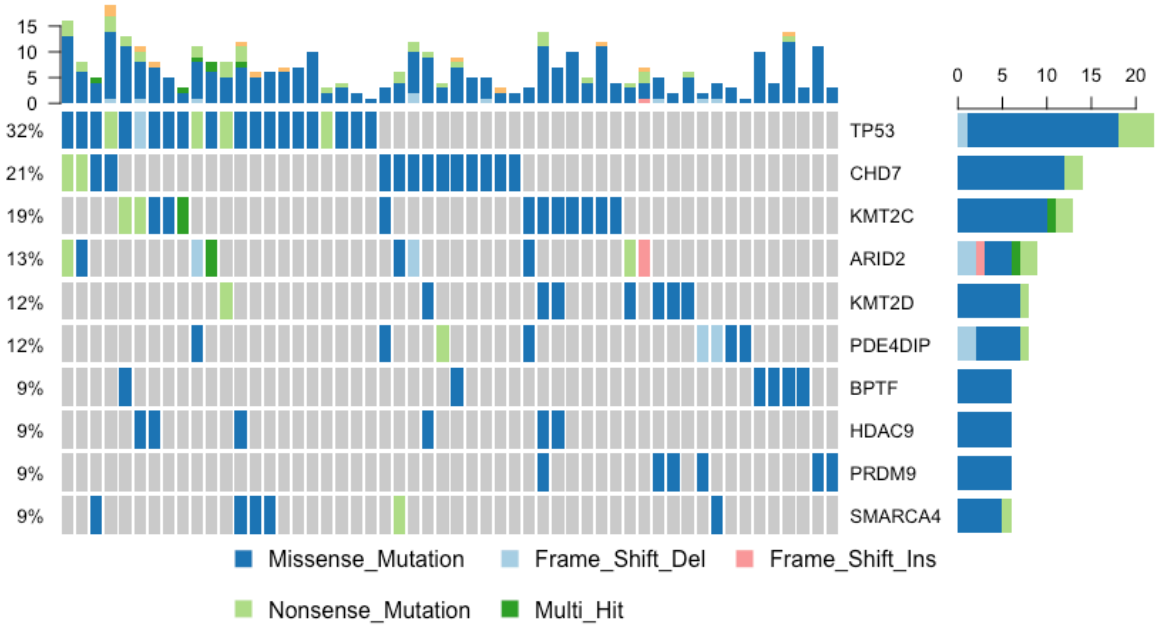


Figure 13. Significantly mutated genes. Box representation of the mutated patients for the most recurrently mutated genes in the lung cancer cohort. The upper panel represents the number of non-silent single nucleotide variants and small insertions or deletions per patient. Each box in the central matrix represent an independent patient. Colored boxes represent mutated patients for the corresponding gene in a color code indicating the type of mutation. Finally, the histogram on the right represents the number of each mutation type found in each individual gene.

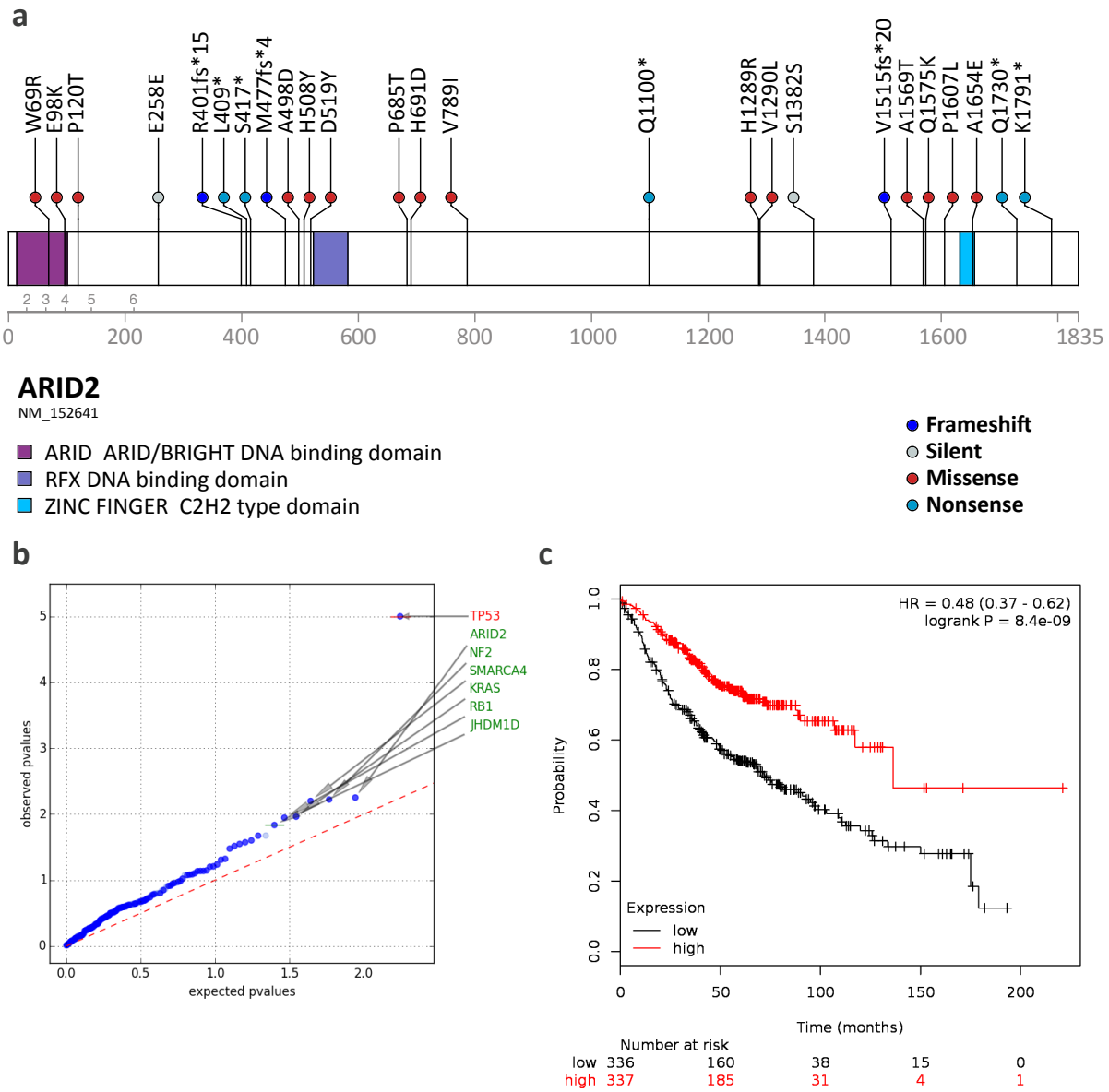


Figure 14. Recurrent mutations in *ARID2* in lung cancer patients. (a) Lollipop graph representing the location of each identified mutation in *ARID2* in relation to the known protein functional domains. Each circle represents a single mutation which annotation at the protein level is indicated. The circle color represents the predicted functional impact of the mutation: missense (red), nonsense (light blue), silent (gray) or frameshift (blue). (b) Quantile-quantile plot comparing observed and theoretical P values of driver genes predicted in lung cancer cohort in our study by OncodriveFML. (c) The overall survival times (OS) extracted from TCGA database using kmplot online software (http://kmplot.com/analysis/) from lung adenocarcinoma patients divided according to *ARID2* low (black) or high (red) expression levels.

Additionally, when OncodriveFML is applied just on the lung samples, it identifes a positive selection in *ARID2*, locating it just below *TP53* which suggest a significant role of *ARID2* mutations in this tumor type progression (Figure 14). In total, we found 13 mutations in *ARID2* that are located all along the coding sequencing without any evident cluster (Figure 14). A high percentage of the identified mutations (8/13) are predicted to produce an inactivation of the protein. All this data suggests a potential role as tumor suppressor of *ARID2* in lung cancer, similarly of what happens in melanoma and hepatocarcinoma.

To investigate if there was already data pointing towards a potential clinical impact on lung cancer of *ARID2* mutations, we checked the public database released from the genome atlas project (http://kmplot.com/analysis/). As it can be seen in Figure 14 there is a clear and statistically significant correlation between low expression of ARID2 and poor prognosis. These results suggest that *ARID2* expression could be potentially used as marker of disease prognosis.

Subsequently, we decided to check if the mutation frequency that we observed in our original screening was representative of the general lung cancer patient population. For that we performed high-coverage, PCR-based sequencing of the coding sequence of *ARID2* of an extra cohort of 96 lung adenocarcinomas from a different clinical source (Hospital Universitario de Canarias).

In this validation cohort, we observed both a similar frequency as well as a similar mutation distribution than the one observed with the original sample cohort (Figure 15 a, b). All the *ARID2* mutations found in both cohorts were validated again by high-coverage PCR-based ultrasequencing in an independent experiment to exclude false positives. All mutations were found real.

Finally, we decided to investigate the effect on the protein production of the identified *ARID2* mutations. For that we performed immunohistochemistry experiments on all the cases in where we counted with stored histological sample left. 85% of the tested *ARID2*-mutated samples (7/8) show reduced or null expression of ARID2 protein whereas all *ARID2*-wildtype samples (32/32) showed consistent high levels of the protein (p=0.029) (Figure 15c).

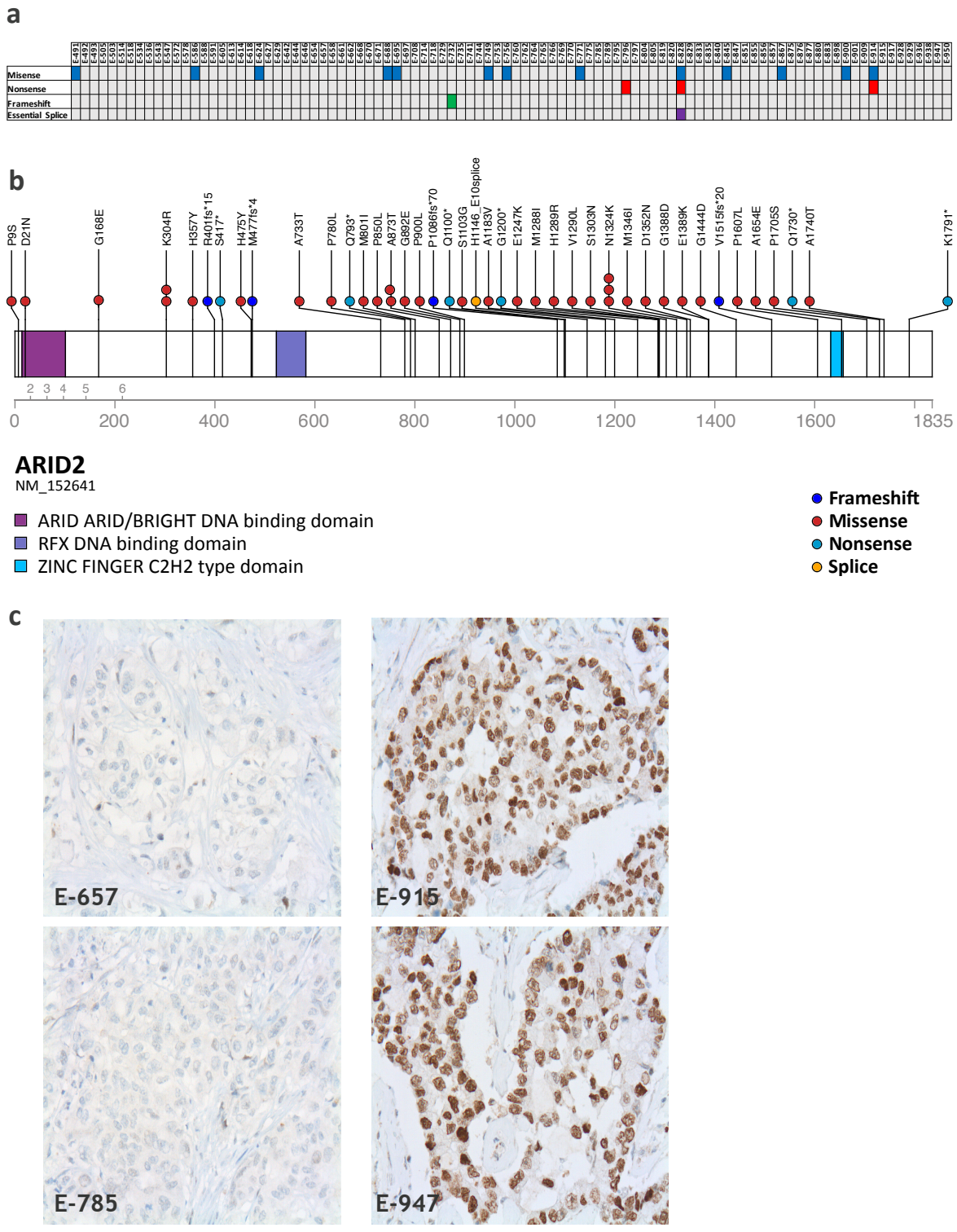


Figure 15. ARID2 mutations in validation cohort. (a) Box representation of the mutated patients in the validation cohort. Each box represents a single patient. Colored boxes represent a mutated patient. Each row/color represent a non-synonymous mutation type: missense substitutions (blue), nonsense substitutions (red), frameshift-inducing indels (green) and substitutions affecting the splice site (purple) . (b) Lollipop graph representing the location of the identified *ARID2* mutations in all lung cancer patients in relation to the functional protein domains. c) Representative images of ARID2 immunohistochemistry experiments on two *ARID2*-mutated lung adenocarcinoma tumors (left panels) as well as two ARID2-nonmutated ones (right panels).

4.2 *ARID2* knock-down produced an increase in proliferation, migration and invasion *in vitro*

In order to understand the effect on cells of *ARID2* alterations, we decided to perform knock-down experiments *in vitro*. For that we used two doxycycline-inducible lentiviral vectors containing two different shRNAs directed against *ARID2* mRNA (V2THS_74399 and V3THS_347660, from now on "v2" and "v3") as well as the empty vector to use as a control (from now on "Empty" or "E"). We generated stably transduced cells in A549 and H460 cell lines derived from human lung adenocarcinoma and which do not contain mutations in *ARID2* according to public databases. As it can be seen in Figure 16. Both ARID2 shRNAs reduced ARID2 mRNA levels at 70% and 85%, respectively. Even greater levels of knockdown were achieved at the protein level with 90% and 100% reduction with V2 and V3 respectively (Figure 16).

As an additional monitoring of the efficiency of *ARID2* knock-down, we performed immunofluorescence experiments using two anti-ARID2 specific antibodies. Expression of ARID2 (in green) is almost undetectable in the ARID2 knock-down cells (positive for the TurboRFP reporter) in contrast with a high expression in stably transfected but non-doxycycline treated cells (Figure 16). ARID2 signal shows homogeneity and a specific nuclear staining. As expected *ARID2* expression is detectable in shEmpty vector stably-transduced cells (Figure 16).

Subsequently, considering that suppression of ARID2 expression has been reported to promote cell proliferation in hepatoma cell lines (166). We decided to investigate in first place the proliferation capacities of our modified cell lines.

As first approximation, we used Carboxyfluorescein succinimidyl ester (CFSE), a fluorescent dye that stain the cells, and is distributed equally in the two daughter cells after cell division, making possible to track the number of cell divisions. As it can be seen in Figure 17, A549 cells showed an increase in cell proliferation after *ARID2* silencing, evidenced for a higher proportion of cells that have suffered more than 5 generations.

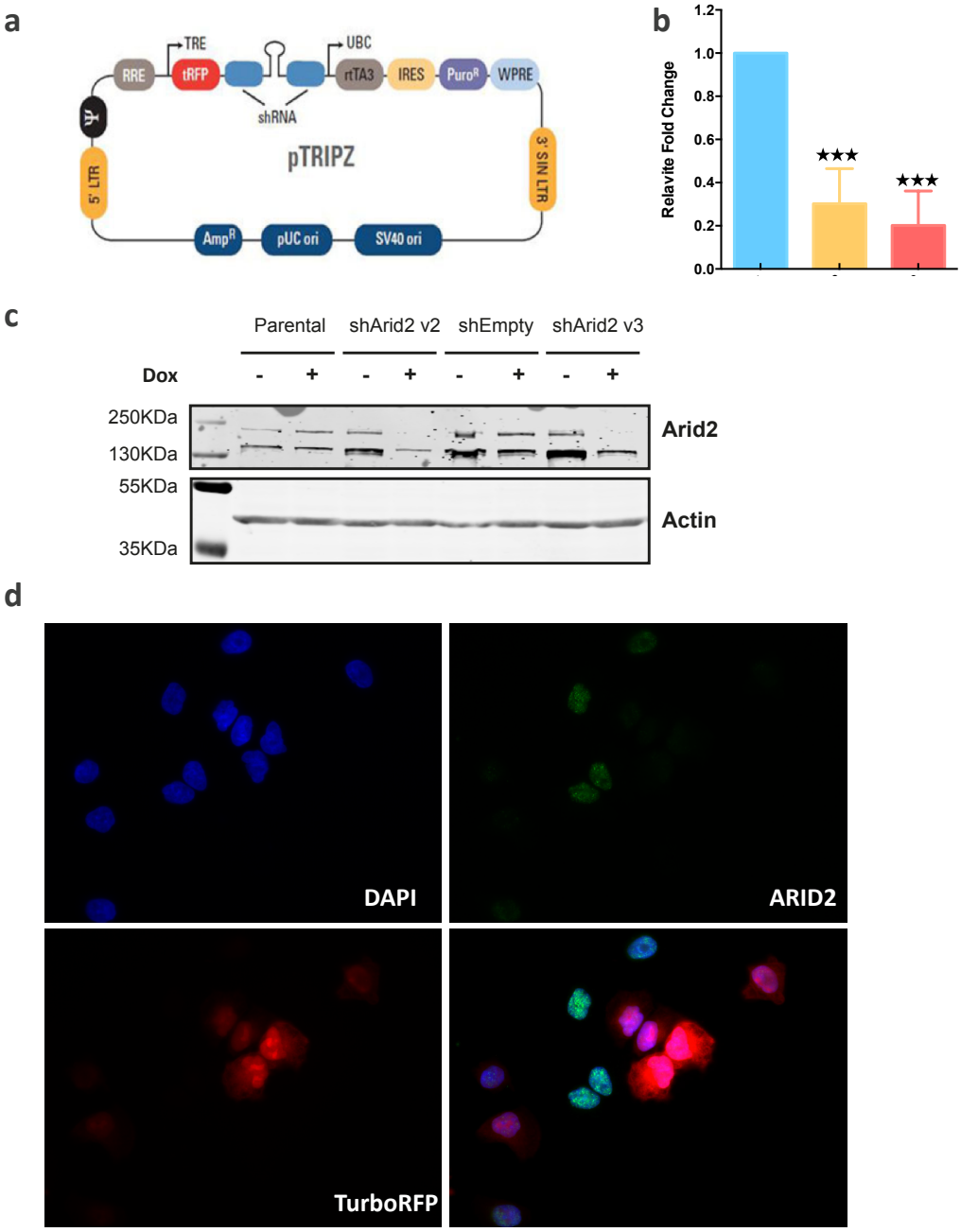


Figure 16. ARID2 knock down experiments. (a) Diagram of pTRIPZ-inducible lentiviral vector (Dharmacon/GE Healthcare, Lafayette, CO, USA). This vector contains a Doxycycline inducible promoter that leads the expression of a turbo-RFP + shRNA fusion RNA. Additionally, it contains puromycin for transfected cell selection. (b) Bar representation of ARID2 expression level fold changes measured by qRT-PCR in A549 cells transduced with shARID2 v2 and v3 as well as the empty vector which is used as control (*** $p < 0.001$). (c) Representative image of a western blot experiment measuring ARID2 protein levels in A549 parental cells as well as those cell lines transduced with ARID2 shRNAs and the empty vector. In all the cases, the results are shown with and without induction of the shRNA expression by doxycycline (Dox) treatment. (d) Representative image of immunofluorescence experiments in H460 cells transduced with shArid2 v3 showing that TurboRFP positives cells (red) efficiently abrogate the expression of ARID2 (green). All cell nuclei are stained with DAPI (blue).

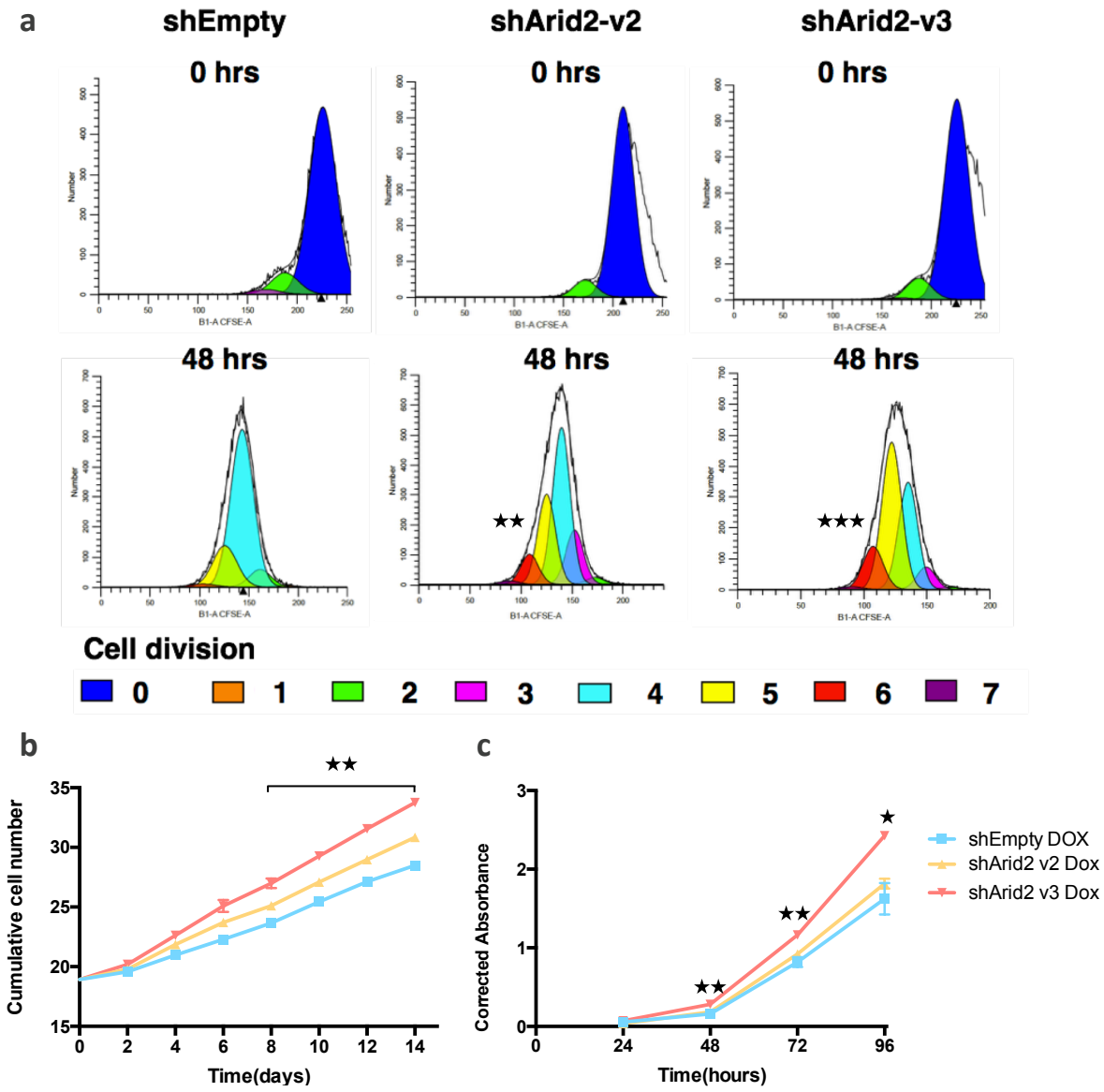


Figure 17. *In vitro* proliferation assays. (a) Representative image of the results of the flow cytometric analysis of cell division by dilution of CFSE in A549 cells. CFSE histograms are shown at 0 (top) and 48 hours (bottom) after CFSE labeling, A549 shEmpty control cells (left), A549 v2 (middle) and A549 v3 cells (right) are shown. Predicted size of population that have suffered different number of cell divisions according to the color legend are represented inside the graph. (b,c) Representation of the cumulative cell number accumulated by serial cell passaging and measured either by direct cell counting (b) (** $p < 0.01$) or by the addition of a metabolic dye(c) (* $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$).

In addition to this approximation, we performed proliferation curves measuring the number of cells after each passage either by direct cell counting, or by use of the metabolic reagent Presto Blue (see methods). In concordance to what was seen in CFSE experiments, *ARID2*-deficient cells showed a higher proliferation rate than control cells (Figure 17b and c).

During the progress of this thesis, Oba and collaborators reported that while an increase in proliferation was not observed, cell migration was increased by *ARID2* knockout in two HCC cell lines (167). According to that, we decided to study the migration and invasion capacities of our modified cell lines.

For that, A549 and H460 cells deficient in ARID2 were starved and seeded in the upper chamber of a Transwell® plate. Transwell migration was efficiently induced by adding FBS in the downer chamber. shRNA-mediated *ARID2* silencing in both NSCLC cells significantly increased migration in response to the serum stimuli (Figure 18). This data demonstrates that the effect of *ARID2* deficiency is stronger in migration (clear in both cell lines) than in proliferation (only clear in A549 cells).

Metastasis not only require a higher migration capacity but also the ability to degrade the extracellular matrix in order to invade distant tissues. In order to evaluate the invasion capacities of our cell lines, we repeated the experiments covering the Transwell with matrigel matrix in order to simulate the extracellular matrix.

Similarly to what happened without matrigel, shRNA-mediated knockdown of *ARID2* significantly increased invasion of A549 and H460 cells. The cells digest Matrigel matrix and survive to the mechanical stress produced by the invasion process, attach to plate and start to grow (Figure 18).

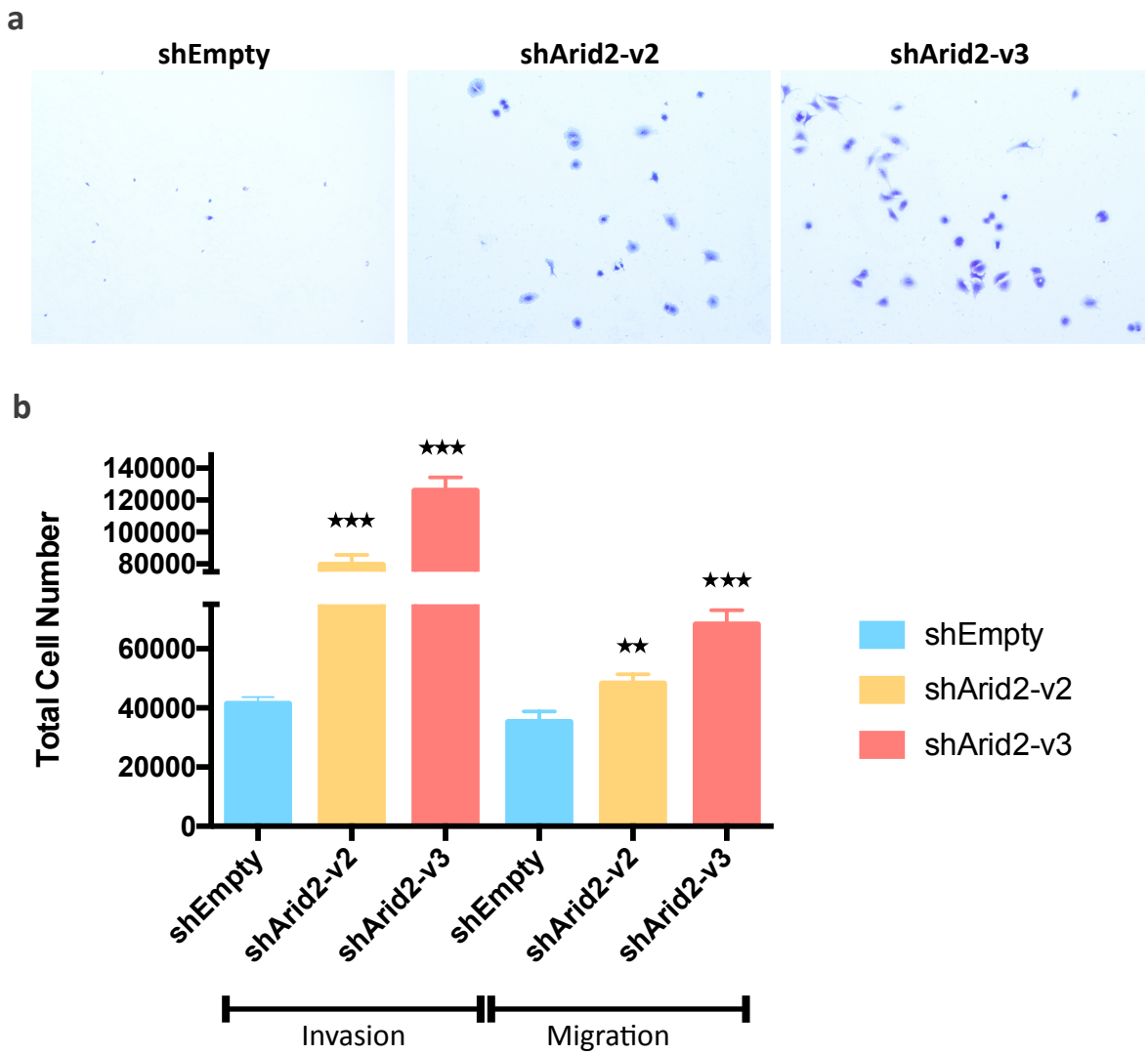


Figure 18. *In vitro* migration and invasion assays. (a) Representative images showing A549 cells subjected to Matrigel invasion and transwell migration that have been stained with crystal violet dye in the downer chamber. (b) Bar representation of the number of cells in the lower chamber in migration and invasion assays of cells transduced with either the empty vector (blue) or the different *ARID2* shRNAs (yellow and red). Data is shown as mean ± SE (standard deviation of the mean) of three independent experiments (** $p < 0.01$ and *** $p < 0.001$).

4.3 *ARID2* deficiency increases proliferation and metastatic potential *in vivo*

In vitro systems do not always recapitulate faithfully tumor progression. For that reason, we decided to test our modified cell lines in an *in vivo* mouse model. In order to investigate the proliferation capacities *in vivo*, we injected subcutaneously 5 million of cells into the flanks of 10 Athymic Nude-Foxn1^{nu} immunocomprised mice. *ARID2*-deficient A549 cells grow significantly faster and form significantly larger tumors in the mice flanks compared to control cells (volume pvalue = 0.032, weight pvalue = 0.047) (Figure 19 b, c).

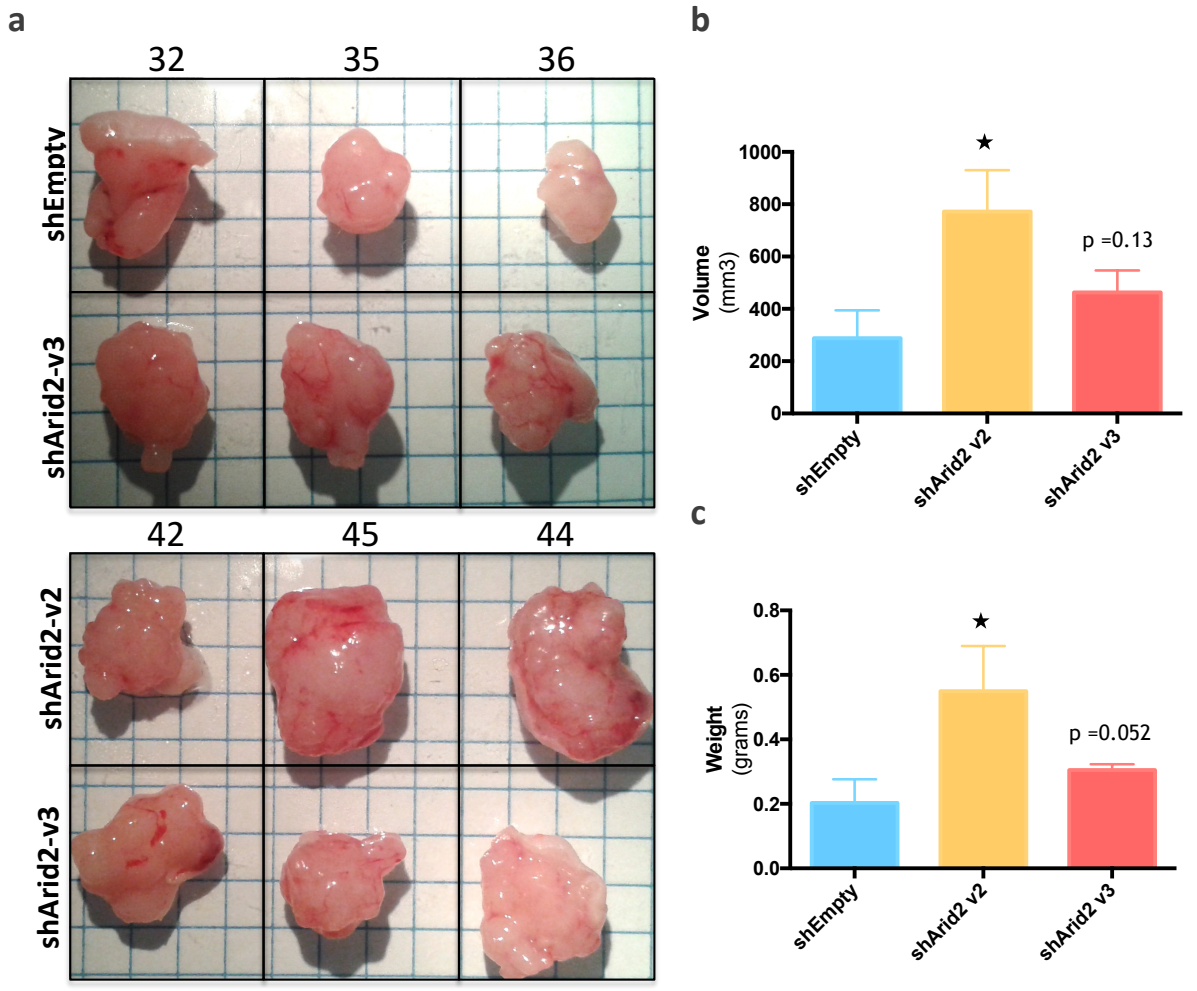


Figure 19. *In vivo* proliferation assay. (a) Images of the tumors generated from 6 nude mice subcutaneously injected with 5 millon A549 cells transduced with shEmpty, shArid2 v2 or v3. (b) Bar representation of the mean volume +/- SE of the tumors found in each group. (c) Bar representation of the mean weight +/- SE of the tumors found in each group (* $p < 0.05$).

4.3.1 Invasion *in vivo*

Subsequently, we decided to test the capacity of ARID2-deficient cells to produce metastasis. For that we tail injected 2,5 million of cells in the tail vein of 40 Athymic Nude-Foxn1^{nu} mice. After 8 weeks, we explored the lungs of these mice to localize potential metastasis that are evidenced by clear nodular masses mainly located randomly in the pulmonary parenchyma although some masses were detected also in the subpleural, hilar, or peribronchiolar regions (Figure 20).

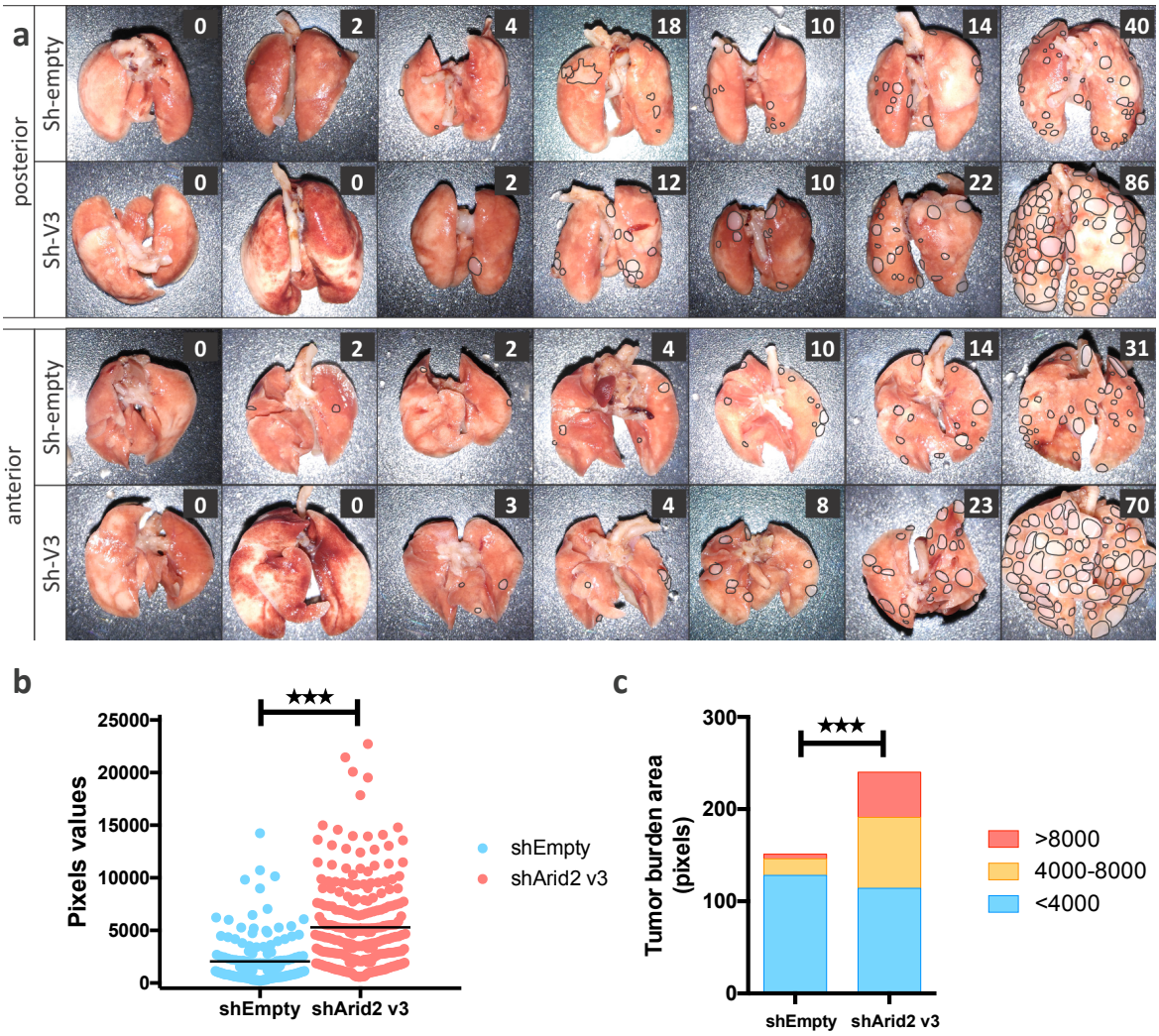


Figure 20. *In vivo* metastasis assay. (a) Images of the lung metastasis generated in intravenously injected mice with A549 cells transduced either with shEmpty, or shARID2 v3. Images of both posterior (top panel) and anterior (bottom panel) face of the lungs are shown. Individual metastasis are delineated in the image and counted (top numbers). (b) Representation of the size (measured in pixels from the images) of each of the identified metastasis divided in the two study mice groups (c) Bar representation of the size of the identified metastasis divided in three groups: small, medium and large (*** $p < 0.001$).

As it can be seen in Figure 20, *ARID2*-deficient A549 and H460 cell lines produced significantly more and larger metastasis than their corresponding control cells (pvalue = 0.0001). We quantified the number and size of the metastases generated by *ARID2* knockdown cells injected in the mice and was significantly greater compared with those of control-shEmpty injected mice (p < 0.0001).

4.4 Transcriptional changes after ARID2 knock-down

As we previously mentioned, ARID2 is capable of regulate transcriptional programs and mediate the expression of IFN-α-inducible genes in HeLa cells (165). In order to understand the molecular mechanisms that could be involved in the differences observed in the cell lines after *ARID2* deficiency, we performed RNA-Seq experiments in three independently generated ARID2-deficient A549 cell lines. After applying several filtering criteria, we found 73 genes upregulated and 107 downregulated (Suppl Table 7).

Interestingly, we observed a significant downregulation in ARID2-deficient cells of genes involved in cellular adhesion such as *CHD6*, *NPNT*, *CNTNAP2*, *FAT3*, *FN1* or *VCAN* (Figure 21). A downregulation of these genes could be associated with a looser connection between the cells and offer maybe an explanation of the increase in metastatic potential of *ARID2*-deficient cells.

Additionally, we observed a down regulation of tumor suppressor genes like *TNFSF10*, *TP63* or *ISMI*, accompanied by an upregulation of pro-tumoral and anti-apoptotic genes like *HOXB1*, *BCL2A1* or *RCVRN*. These changes evidence a change towards a pro-tumoral transcriptional program that probably contributes to the higher tumorigenic properties of *ARID2*-deficient cells. Finally, we observed an upregulation of *GADD45A*, which is known to be involved in DNA damage detection and repair pathways. This observation is very interesting as several of the chromatin remodeling genes have been associated with DNA repair. The validity of all these observations were validated in independently generated *ARID2*-deficient cell lines by qRT-PCR (Figure 21).

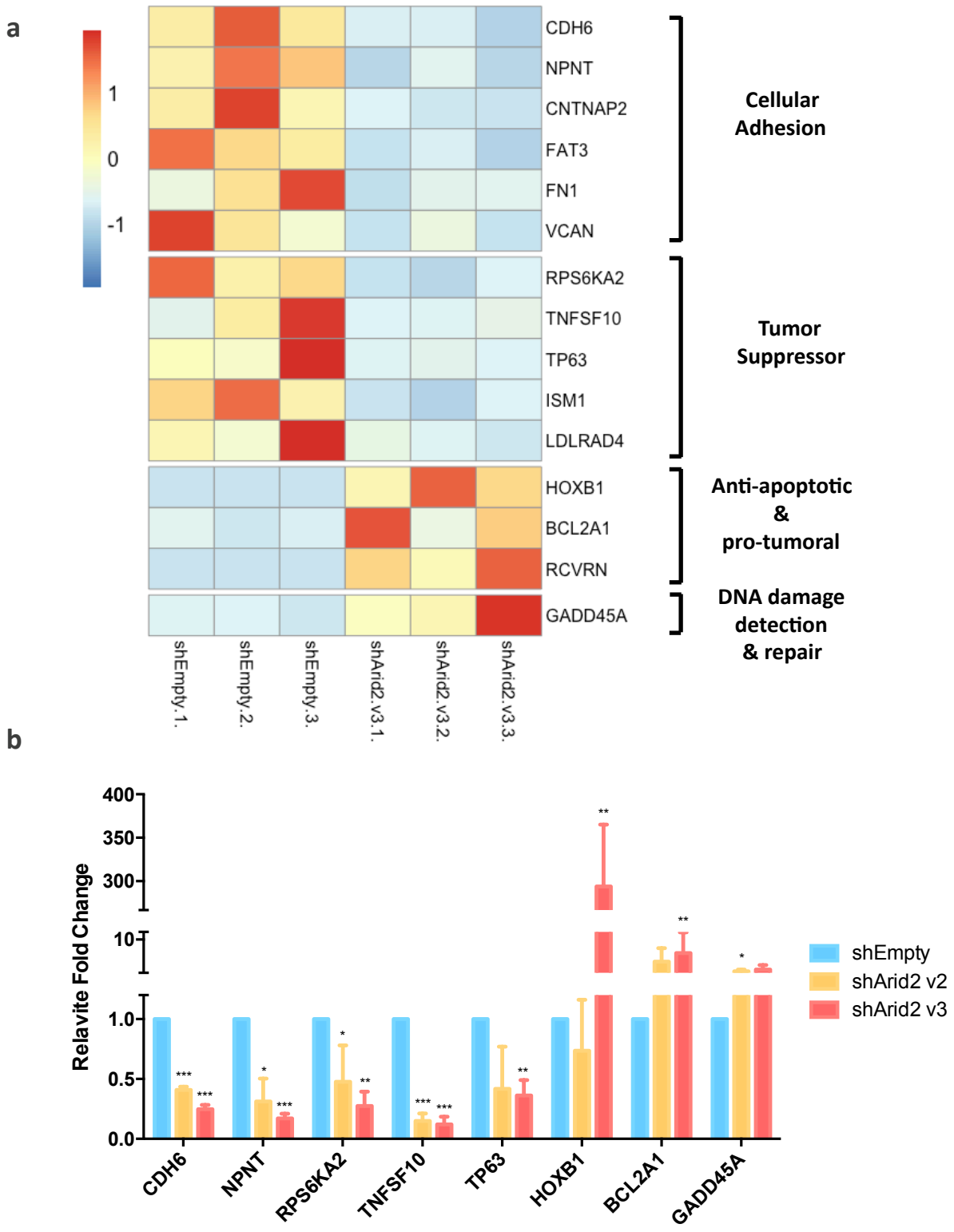


Figure 21. Differentially expressed genes in *ARID2*-deficient cells. (a) Heatmap representation of a selection of differently expressed genes in *ARID2*-deficient cells (N=3) and grouped according to their molecular pathway. Expression differences goes from red (overexpression) to blue (downregulation) according to the log2 of the fold change. (b) Bar representation of the results of the qRT-PCR validation of the expression differences identified in the RNA-Seq experiments. The expression foldchanges are represented as a mean +/- SE of three independent experiments, (*p < 0.05, ** p < 0.01 and *** p < 0.001).

4.5 *ARID2*-deficiency impairs DNA repair and is associated to an accumulation of DNA double strand breaks (DSBs)

Several chromatin remodelers have been proved to play a role in DNA repair (127,148-151). That, together with the upregulation of *GADD45A* seen in the RNA-Seq experiments, prompted us to investigate the potential role of ARID2 in DNA repair.

For that, we decided to perform DNA damage induction experiments with Neocarzinostatin (NCS) and Etoposide in ARID2-deficient cell lines. In the presence of a thiol cofactor, NCS forms a highly reactive biradical species that can induce sequence-specific single and double strand breaks in DNA (169). Neocarzinostatin inhibits DNA synthesis and cellular proliferation by inducing G2 cell cycle arrest and apoptosis (170). On the other hand, Etoposide forms a ternary complex with DNA and the topoisomerase II enzyme inhibiting the ability of the enzyme to ligate cleaved DNA molecules causing DNA strands breaks (171,172). After damage induction, we checked the recruitment of γ*H2AX* and *53BP1* to the damage site. Additionally, we measure the time of the resolution of the foci after DSBs induction.

As it can be seen in Figure 22, ARID2 is recruited to the DNA damage site (foci) and colocalized with both γH2AX and 53BP1 in the DNA damage lesion site, which proves an active role of ARID2 in DNA repair. Additionally, we could see that *ARID2*-deficiency is accompanied to an increase in DSBs as well as to a delay in the foci resolution time (Figure 22).

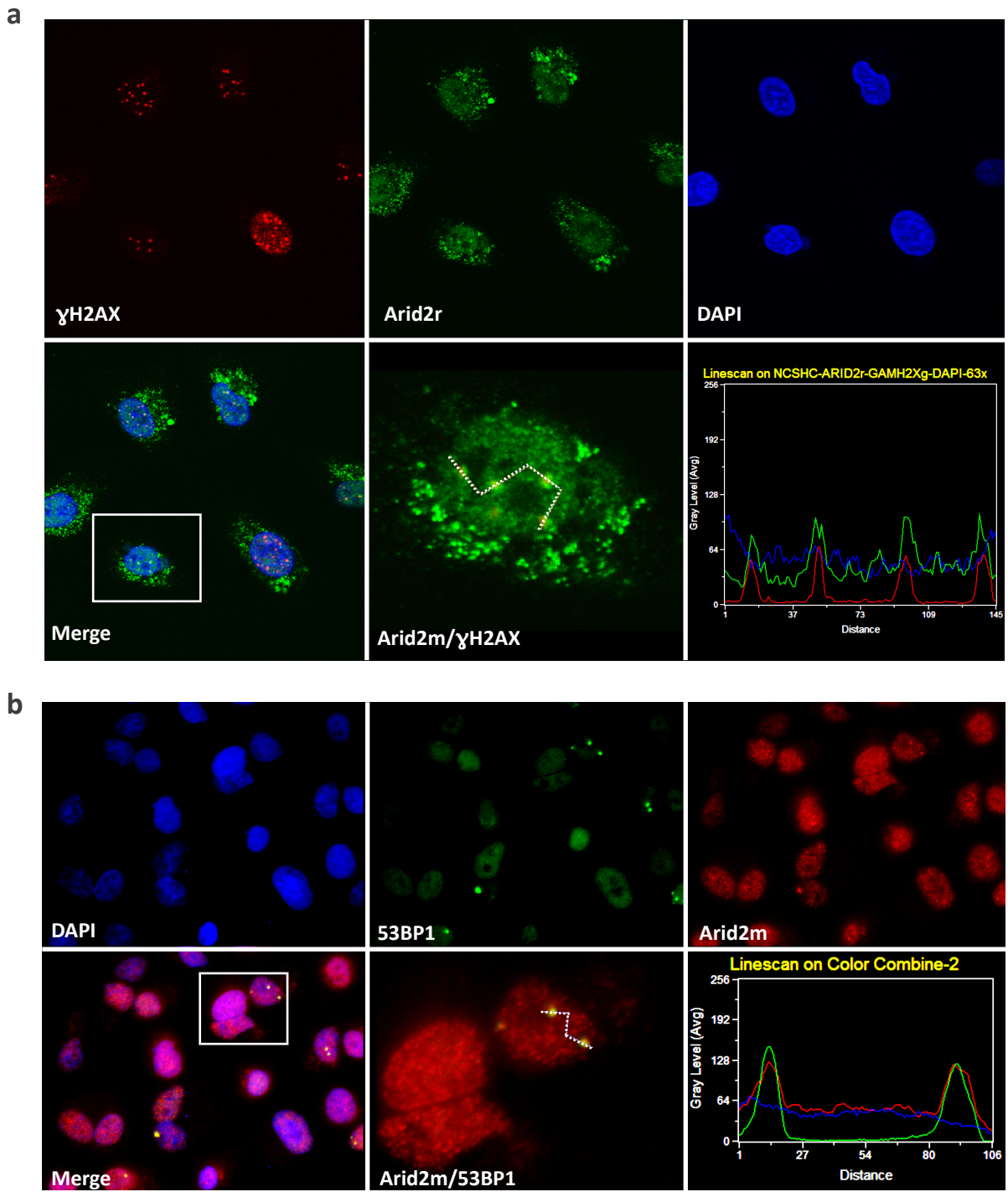


Figure 22. *ARID2* Implication in DNA repair. (a) Representative images of immunofluorescence experiments proving the colocalization of γH2AX (red) and ARID2 (green) in DNA repair foci. Top panels represent the independent signals of each protein together with the cell nuclei labelled with DAPI. The lower panels represent the merged imaged and the analysis carried out by linescan using Metamorph software to prove the co-localization of both fluorescence signals through the analysis pathway (white line). (b) Representative images of immunofluorescence studies that prove the colocalization of 53BP1 (green) and ARID2 (red).

4.6 Chemotherapy sensitivity of ARID2-deficient cells

All our results suggest that, additionally to a change to a pro-tumorigenic transcriptional program, *ARID2*-deficiency could compromise DNA-repair systems and, therefore, promote genomic instability. This genomic instability could suppose a general advantage for the tumor cells but also offers a weakness to resist DNA damaging treatments like cisplatin or etoposide. This is especially interesting in our case because cisplatin is one of the most active cytotoxic agents used for non-small cell lung cancer (NSCLC) treatment.

According to that, we decided to investigate if *ARID2*-deficiency could confer a special sensitivity to these treatments and, therefore, be useful as a marker for patient stratification.

As it can be seen in Figure 23, *ARID2*-deficiency increased cell sensitivity to both cisplatin and etoposide evidenced by a lower IC50 (drug concentration that produced a 50% mortality on the cells). These results are significant and consistent for both treatments and also for both A549 and H460 cell lines.

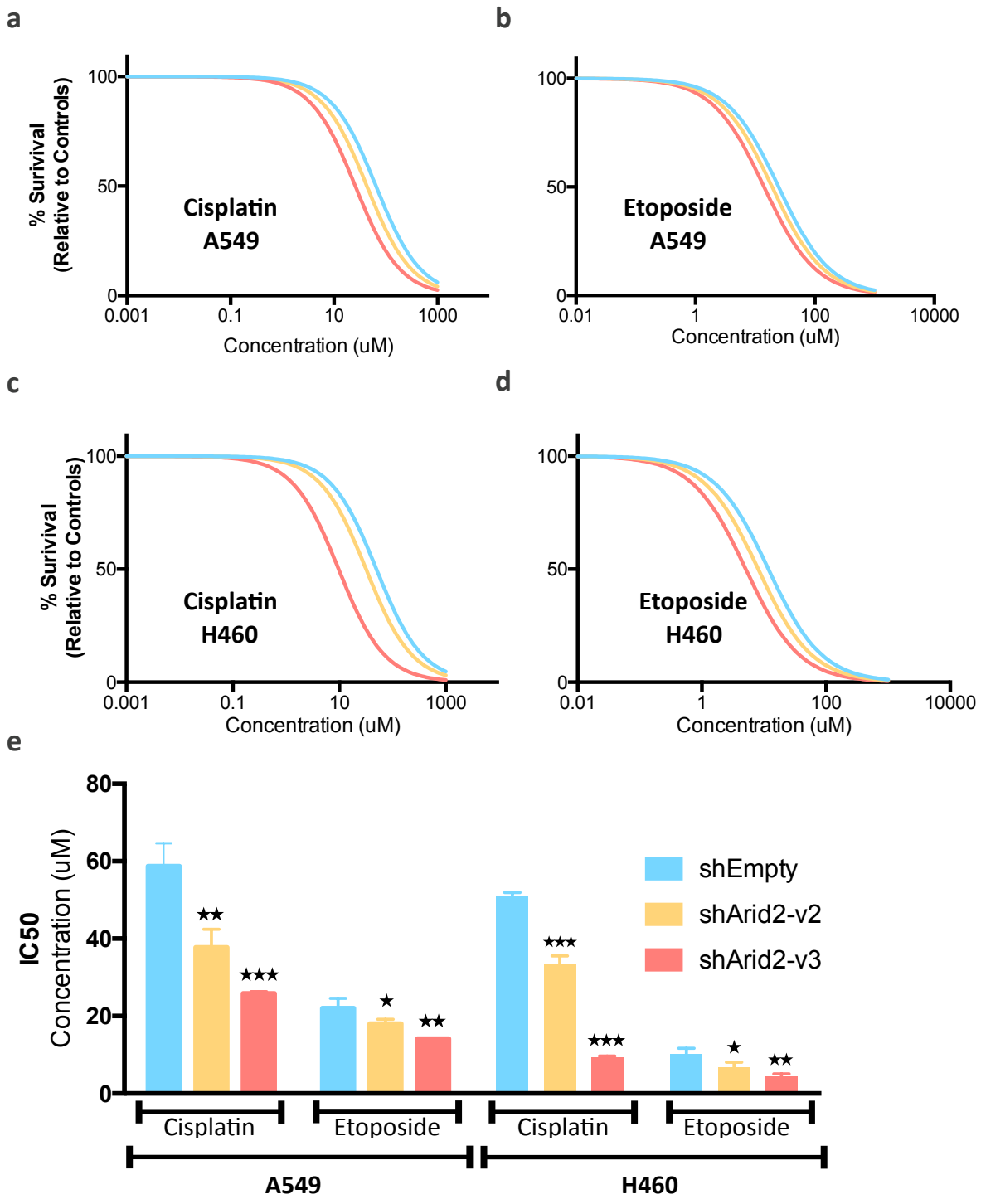


Figure 23. ARID2-deficient cell increased sensitivity to DNA-damaging agents. Graphs of representative experiments measuring cell survival to increasing concentrations of cisplatin (left) and etoposide (right) in A549 (a,b) and H460 (c,d). Transduced cells with shEmpty (blue) or shARID2 v2 (yellow) and v3 (pink) are represented. (e) Bar representation of the calculated IC50 of A549 and H460 cells to cisplatin and etoposide. The results are represented as mean +/- SE of three independent results, (*p < 0.05, ** p < 0.01 and *** p < 0.001).

DISCUSSION

“Knowledge is the small part of ignorance that we arrange and classify.” Ambrose Bierce

1. OVERALL RESULTS OF THE MODIFICATIONS IN LIBRARY PREPARATION AND DATA ANALYSIS PROTOCOLS

The main technical obstacles that were foreseen at the beginning of my PhD project were firstly, the library preparation and sequencing of a large collection of samples with limited human and economical resources. Secondly, the high amount of sequencing data aimed in this project over a limited target sequence of the human genome, the coding sequence of 250 genes, approximately 600 kb, supposed a challenge for the data analysis in order to obtain the highest sensitivity without sacrificing specificity.

To minimize the first obstacle, we have optimized a library generation protocol using individually ordered enzymes outside of commercially available kits and taking advantage of a dual barcoding system with an index of 8 redundant bases. In this protocol variation, we decided to barcode and mix the libraries in groups of 96 samples before enrichment against the manufacturer recommendations. Target sequence enrichment was performed using custom probes using a single enrichment reaction for each 96-multiplexed library group. Finally, each one of this enriched library pools were sequenced in a single lane in a High-Seq instrument. This allows the reduction of the cost of the library preparation from 700€ (price prevailing at the beginning of the project) to an estimated price of less than 5€ per sample. Without these improvements, the goal of sequencing more than seven hundred primary samples would have been completely impossible for a small group like us.

In spite of all these modifications, we obtained an average coverage in the region of interest of at least 50x in 90% of the sequenced samples with an average of 40% of the reads on-target. This is a bit lower of the company specifications that promise at least 50% of the reads on target. We firstly thought that the lower efficiency could be the result of the pooling prior enrichment but conversations with different collaborators using Agilent Technologies allowed us to confirm that the same enrichment ratios were observed when single libraries are used in the enrichment. Thus, it seems that small probe designs tend to be less efficient in capture than larger designs like those used in exome-sequencing. Moreover, in order to finally prove if the number of different libraries pooled before enrichment was affecting efficiency (https://www.agilent.com/cs/library/brochures/5990-3532en_lo%20CMS.pdf) we performed enrichments multiplexing half (48) of the samples and we could not see any improvement in terms of coverage or numbers of reads on-target.

In terms of data analysis, we have performed multiple tests in order to set up the proper pipeline to analyze the high amount of data generated in our lab. For that we had to create quite a lot of different codes that are included as digital files with this thesis. In particular, in order to optimize the analysis with the large number of samples, we created pipelines in perl that automatize the process of alignment, sorting, indexing, cleaning and performing the local indels realignment on the sequencing data. In addition to that, codes to perform the automatic functional annotation of the mutations, the analysis of High-coverage PCR-based NGS data as well as the primer design have been specially designed and used during this thesis and are collected in the digital annex of this thesis.

During the past several years, many variant calling software tools have been developed. Unfortunately, most of them were initially designed to find germline polymorphisms. Cancer variant calling needs higher specificity to detect sequence variants, as a false positive ratio of 1 in 10 kb for example can be acceptable for detecting germline variants (as it is 10 times less frequent that the real variants) but can be fatal in the detection of somatic variants (100 times less frequent than germline polimorfisms). We have tested different freely available software: GATK Unified Genotyper, SomaticSniper, VarScan, Strelka, Mutect, etc. We noted that although in general they have acceptable sensitivity, they tend to have low specificity and they need usually to be complemented by the application of user-defined filters. In many cases, these

filters have a much higher impact in the selection of real variants that any statistical analysis perform during the initial calling.

According to that, and after studying the major causes of false positives in our studies, and with the firm idea that most of them were the results of read alignment errors, we created our own software called RAMSES. The main feature of this software is the performance of a second alignment using a different read aligner (BLAT) to check the correct alignment of the mutant reads. That, plus the implementation inside the software of a series of common filters almost universally used by most authors and aimed at the correction of strand bias; repeats, polymerase delay at specific sequences (173,174) or read position bias, produces a highly reliable mutation list. Additionally, our software is based in a naive approximation of minimum evidence and is therefore not dependent in a previous estimation of the read frequency that must be observed for a real mutation or in complex statistical models. Among the novelties included in our software is a filter to correct false positives produced by the oxidative stress during DNA acoustic shearing(175).

The overall performance of our analysis strategy is supported by the high reliability of our mutation data. In first place, we have identified more than five hundred mutations already reported in the databases and therefore likely real and the observed mutated sample frequencies for the most frequently mutated cancer genes in each tumor type matches with what is reported in the literature. All this indicates that we do not have a serious issue in sensitivity in our software. Secondly, we have done an orthogonal validation of a random selection of mutations and have found that almost 90% of the detected mutations were real. We detected, nevertheless, a problem in our software in order to filter out real germline variants in our data. Preliminary data suggest that this could be the result of a low coverage over those regions in the normal sample and at the moment we are working in RAMSES to improve this aspect.

In summary, in the present thesis we present a series of modifications both for the library preparation as well as for the data analysis that shows a significant improvement over generally used methods in terms of library preparation costs and time as well as in data specificity. These improvements can be used by any laboratory, even for very modest ones, to perform targeted next generation sequencing on large collection of samples to investigate specific biological questions.

2. mtDNA MUTATIONS

Over the years, mutations in mtDNA have been reported in different tumor types. It is broadly assumed that the major cause of the accumulation of these mutations is the exposure of mtDNA to a highly oxidizing environment within the mitochondria. Interestingly, our data does not fit to the main substitution profile expected to be produced by this mechanism and we observe an excess of G>A/C>T transitions instead of the expected G>T/C>A transversions reported by the accumulation of 8-oxo-guanine. As different DNA polymerases have been reported to produce different substitutions as a consequence to the same DNA damage (176), we could hypothesize that DNA polymerase gamma (in charge of replicating mtDNA) could generate G>A/C>T in response to the accumulation of 8-oxo-guanine. We are in our laboritery at the moment performing functional studies to check this hypothesis. Another potential explanation proposed by other authors (114) is that the accumulation of 8-oxo-guanine is very efficiently repaired by the mtDNA repair mechanisms and do not produce mtDNA mutations being the observed G>A/C>T transitions just the result of common errors during mtDNA replication. Although this option is possible, it goes against reports that prove that 8-oxo-guanine repair is much less efficient in the mtDNA than in the nuclear DNA (177).

Another interesting observation that we have found is a bias in the mutation distribution between the two strands. Seok Ju and collaborators have also very recently reported this observation (114) and have suggested that this bias may be a result of the dual mechanism of mtDNA replication that maintain the heavy strand more time exposed to the media and potentially more prone to cytosine and/or adenine deamination. This model, nevertheless, present several limitations in our data. In first place, there is no clear association between the distance of the replication origin and the accumulation of mutations as it could be expected for the exposure of the heavy chain. Additionally, the strand bias is observed also in the T>C/A>G transitions that are

more likely the result of the accumulation of thymine-glycol and not necessarily related with the replication process. Therefore, we propose that the data supports more an effect associated with transcription. It is difficult to differentiate both effects as practically all the genes in the mitochondria are in the same physical strand. If we continue with the hypothesis that the G>A/C>T transitions are the result of the accumulation of 8-oxo-guanine, in both main substitutions, we observe a low reduction of mutations in the transcribed strand, that could be explained by the presence, similarly to what happens in the nuclear DNA, of a transcription-coupled repair mechanism.

Sadly, this hypothesis is not free of inconveniences either, given that this process is carried out by the nucleotide-excision repair machinery (NER) in the nuclear DNA and is broadly reported that mitochondria lack this repair mechanism. Nevertheless, several recent reports put into question these findings. Thus, Liu and collaborators found that Xeroderma pigmentosum group D (*XPD/ERCC2*) gene is located in the mitochondrial inner membrane and play a role protecting and facilitating mtDNA repair (178). On the other hand, Pohjoismäki and collaborators found a specific upregulation of *XPA* and *RAD23A* DNA repair proteins after mitochondrial oxidative stress and also demonstrated the *RAD23A* mitochondrial localization by inmunoflourescence and sucrose gradient purification (179). Finally, there has been several reports that defend the presence of Cockayne syndrome proteins, (*CSA/ ERCC8*) and (*CSB/ ERCC6*) in mitochondria(180-183). In this high controversy, our data also support the presence of such transcription-coupled repair mechanism in the mitochondria but further studies are needed to finally clarify this aspect.

3. SWI/SNF ROLE IN CANCER

Mutations in genes encoding subunits of the chromatin remodeling complexes are continuously reported in literature, but in general it is SWI/SNF complex which has garnered considerable importance. Our data evidences a significant enrichment of driver genes in the SWI/SNF complex, according to OncodriveFML and GSEA. Ours results are in concordance with a recent Gonzalez-Perez and collaborators report, who have performed a massive analysis of chromatin regulators factors in 4623 tumors from datasets of international initiatives like The Cancer Gene Atlas and the International Cancer Genomes Consortium. They found that 650 tumor samples has recurrent mutations in SWI/SNF proteins with a high accumulation in bladder, kidney and uterus tumors(184). Proteomic and bioinformatics analysis demonstrated that SWI/SNF complexes are mutated in 20% of human cancers (127). All these data indicate that SWI/SNF could play a more important role in cancer development that the rest of the complexes.

The mechanism by which this role is exerted is nevertheless not perfectly clear. SWI/SNF complexes mobilize nucleosomes at target promoters and enhancers but also are able to interact with transcription factors, coactivators, and corepressors to modulate gene expression (185), so the deregulation of specific cancer associated pathways could be a first alternative. In accordance to that, members of the BAF complex are able to regulate Wnt/ß-catenin signaling. Specifically, it has been reported that *SMARCA4* interacts with ß-catenin to promote target-gene activation (186), and that the loss of SMARCB1 is sufficient to activate the Wnt/ß-catenin pathway (187). Finally, ARID1B seems to be a repressor of Wnt/ß-catenin signaling (188).

Between other functions, it has been described also that SMARCD2-mediated ATM-p53 activation independent of DNA damage, blocking hepatic cell identity conversion by sensing chromatin opening. The same authors also found by ChIP that SMARCD2, SMARCA4 and PBRM1 bound to the open chromatin regions indicating that SWI/SNF mediates ATM recruitment to these regions (189). In a pathway dependent of DNA damage, PBRM1 is phosphorylated at serine 948 by ATM, enabling the PBAF complex to elicit transcriptional silencing in cis to DSBs, and rapid repair of DSBs within transcriptionally active regions via NHEJ (190). Additionally, alterations in PBRM1 have been reported to upregulated ALDH1A1, increasing tumorigenicity in clear cell renal cell carcinoma (191). Furthermore, loss of *SMARCA4* or *SMARCA2* significantly improved prognosis for overall survival (OS). Interestingly, loss of *PBRM1*, *ARID1A* or *SETD2* had the opposite effect patient survival in ccRCC (192).

How proteins of the same complex can have opposite effects on Wnt/ß-catenin signaling or loss of subunits can improve or aggravate the prognosis depending on the tumor type is not know, however the composition of the BAF complex may provide an answer. It has been describe that BAF complex with different composition exist within the same cell (193).

Interestingly, not all subunits of the SWI/SNF complexes show the same alteration frequency. Some subunits, specially the accessory subunits (*ARID1A*, *ARID1B*, *ARID2* or *PBRM1*), show a higher mutation frequency than the core ones. This observation could indicate that the consequence of these mutations is not the complete abrogation of the function of this complex but instead a change in the regulated targets or in the intensity of these regulation over some targets. Another potential explanation is that these accessory subunits play additional roles in the cell independent of the SWI/SNF complex. One of these functions could be DNA repair as it has been reported the role of several SWI/SNF subunits in different DNA repair mechanisms (148-151).

It is also important to note that the alterations of some subunits show clear preferences for specific tumor types. That is the case of PBRM1 in renal cancer or ARID2 in melanoma, liver and lung cancer. Therefore, whatever the mechanism by which these alterations play a role in cancer development, this role is not completely equivalent in different tumor types.

In accordance with other reports of the synthetic lethality between the alterations of some subunits of the SWI/SNF complex (152-154,194), we could observe in our data a tendency to mutual exclusivity not only among the more frequently mutated SWI/SNF subunits but also with other frequently mutated cancer genes like *RAS* or *TP53*. This observation could lead to the establishment of new synthetic lethality relationships that could be potentially exploited in the treatment of SWI/SNF mutated tumors.

Finally, in addition to the description of already reported SWI/SNF cancer genes, our cancer gene discovery analysis found other genes that show a pattern of selection and that had not been associated before with cancer progression. One of these genes is *SRCAP* which showed positive selection among many tumors types, but with an especial evidence in colorectal cancer. *SRCAP* belongs to the INO80 family which has been directly connected with DNA damage detection and repair. SRCAP has a role catalyzing the deposition of H2A.Z into nucleosomes (195). A couple of articles report mutations in *SRCAP* gene in colorectal (196) and glioblastoma (197) cell lines but in general there is almost no information of its potential role in cancer development. These findings warranty further investigation about the role of this and other less known chromatin remodeling genes in cancer progression.

4. *ARID2* AS A *BONA FIDE* TUMOR SUPPRESSOR GENE IN LUNG CANCER

ARID2 **has been described already as a clear tumor suppressor gene in hepatocellular carcinoma HCC (132) and melanoma (133), but only one single article has suggested a potential implication of this gene in non-small lung carcinoma after having found mutations in 5% of samples of this tumor type (134). Nevertheless, we observed in our screen a much higher mutated patient frequency that the one described, both in a discovery and a validation patient cohort summing up near 200 tumor samples. Moreover, these mutations are associated with a loss of the protein expression and, the low expression of** ***ARID2*** **is in itself associated with a worse prognosis. Therefore, we hypothesized that the role of** ***ARID2*** **in lung cancer has been overlooked until know.**

Previously, it was found that ARID1A knockdown significantly promoted the proliferation, migration and invasion (*MHCC97L*, *MHCC97H*, *SKhep1* and *WRL68*) HCC cell lines (168). Our *in vitro* and *in vivo* experiments demonstrate that *ARID2* silencing promote an increase in proliferation, migration and invasion *in vitro* as well as an increase of the tumorigenic potential of cell lines *in vivo*. These observations are also in accordance to similar experiments performed in the past in other tumor types (166).

ARID2 **deficiency seems to exert this function by two complementary mechanisms. In first place, *ARID2* downregulation produces changes in the cell transcriptional program downregulating genes involved in cellular adhesion and upregulating pro-tumorigenic and anti-apoptotic genes. All together aggregates into a pro-tumoral transcriptional program for the cell. In the other side, we**

report that *ARID2* plays an important role in DNA repair. It is located at the site of DSBs together with other very known DNA sensor proteins like 53BP1 or γH2AX, and its down-regulation is associated in an increase of DSBs and in a delay in their repair.

This dual role of *ARID2* deficiency in cancer is probably extensive to other SWI/SNF subunits which alterations, besides producing changes in the transcriptional program, could also affect other cell mechanisms that are independent of the function of the canonical SWI/SNF complexes. In this line of reasoning, during the realization of this thesis and in accordance to our *in vitro* results, two different reports showed a role of *ARID2* in DNA repair (167,198). Interestingly, these authors proposed that *ARID2* and *PBRM1* participate in a complex or complexes different from the canonical PBAF complex containing *SMARCA4* (198). Thus, it is possible that non-canonical SWI/SNF complexes formed by one or more of the SWI/SNF accessory subunits of the canonical ones, could be behind the observed role of these genes in cancer development which could explain the differences in the tumor specificity or the mutation frequency observed in the samples.

5. USE OF THE SWI/SNF ALTERATIONS TO TREAT CANCER PATIENTS

According to nowadays knowledge, around 20% of all tumors contain alterations in SWI-SNF. Therefore, any advantage in the exploitation of these alterations to improve patient management is very attractive and can be potentially used in a lot of tumor types.

The first potential use of these alterations are as prognostic factors. We have shown that *ARID2* downregulation is associated with worse prognosis in lung adenocarcinoma. Similarly, loss of *ARID1A* expression has been reported to correlate with shorter overall survival (199). Interestingly, tracking breast cancer genomic evolution by sequencing, Yates and collaborators has recently published, that clones seeding relapse have alterations in SWI/SNF genes as *ARID1A*, *ARID1B*, and *ARID2*, and that mutations in *ARID1A* and *ARID2* emerged in three of five patients relapsing after taxane chemotherapy (200). Therefore, it is possible that the screening for SWI/SNF alterations could benefit the management of patients of specific tumor types. We and other groups are starting to analyze this possibility more deeply.

It is worthy to mention, according to this, that we show a high grade of intratumoral heterogeneity in the tumors respecting *ARID2* expression. This could be explained by the presence of different cell subclones inside of the tumors with only a small percentage of the cells harboring the inactivating mutations.

Similarly, it has been described among endometrial cancer that only two-thirds of *ARID1A* mutations were clonal as well as subclonal loss of the expression of *ARID1A* (201,202). Finally, TRACERx Consortium has published that in lung cancer, whereas driver mutations in *EGFR*, *MET*, *BRAF*, and *TP53* were almost always clonal, mutations in SWI-SNF genes were systematically subclonal which indicates a late appearance during tumor evolution. Similarly results were also obtained in breast and hepatocellular cancer (200,203,204).

The other potential use of the SWI/SNF alterations is their exploitation for specific antitumoral therapies. We have reported here that, due to their defect in DNA-repair, *ARID2*-deficient cells are more sensitive to DNA damaging agents like cisplatin or etoposide. Similarly, the SWI/SNF catalytic subunits *SMARCA4* and *SMARCA2* modulate the cellular response to cisplatin through the *ERCC1* recruitment to DNA lesions (205). Finally, some authors have reported the increased sensitivity of SWI-SNF deficient cells to *EZH2* inhibitors (156,157) and the possibility of also a differential behavior against BRD domain inhibitors (158,159).

On the opposite side, it has been described that down-regulation of *PBRM1* attenuates the effect of gefitinib by sustaining *AKT* pathway activation during *EGFR* inhibition (185) and *ARID1A* deficiency has been associated with chemoresistance to platinum-based therapy in ovarian clear cell carcinoma (199).

Finally, the genomic instability associated with the defects in some SWI-SNF subunits (*ARID2*, *PBRM1* or *ARID1A* for example) could suggest a higher generation of neoantigens by the cells containing these defects and, therefore, probably these tumors could be better targets for immunotherapy (206).

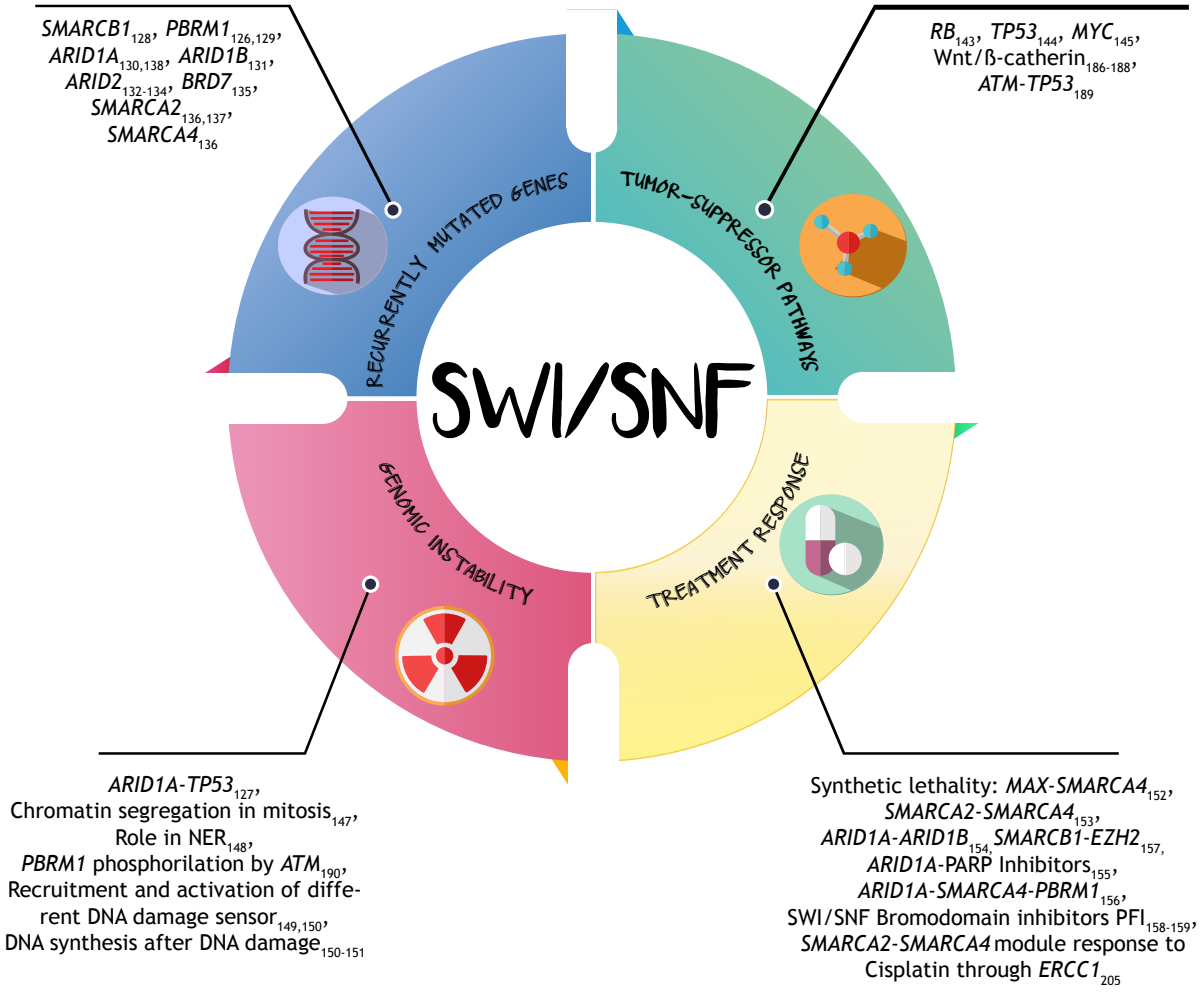


Figure 24. Schematic representation of the mechanisms underlying the tumor-suppressive activity of the SWI/SNF complex. Multiple evidences of an active role of SWI/SNF in tumor suppression are grouped in four "pathways" and the original articles properly referenced.

CONCLUSIONS

“Research is what I'm doing when I do not know what I'm doing." Wernher von Braun

1. We have identified 4900 somatic mutations in the coding region of the 250 genes sequenced in 732 tumor samples.

2. We have identified a total of 170 somatic mutations in the mitochondrial DNA in our collection of samples.

3. The substitution profile and strand bias shown by mtDNA mutations suggest the existence of new mutagenesis and DNA repair mechanisms in this organelle not described until now.

4. 60% of the sequenced tumors have at least one non-synonymous mutation in a chromatin remodeling complex. SWI-SNF complex shows a significant enrichment in cancer driver genes which suggests a more prominent role of this complex in tumor development.

5. The mutations in SWI-SNF complex genes show sample exclusivity among them and also with mutations highly recurrent mutated genes like *KRAS* or *TP53* suggesting some grade of function redundancy.

6. We have found evidence that proves that ARID2 is a *bona fide* tumor suppressor gene in lung cancer which is mutated in 15% of the analyzed samples and is associated with worse prognosis.

7. Downregulation of *ARID2* increases proliferation, migration, invasion and metastatic capabilities in cell lines both *in vitro* and *in vivo*.

8. *ARID2* deficiency is associated with gene expression alterations that produces a pro-oncogenic transcriptional program in the cells.

9. *ARID2* plays an important role in DNA repair and its deficiency is associated with an increase in the DNA damage and a delay in its repair.

10. *ARID2* deficiency increases the sensitivity of cell lines to DNA damaging agents like cisplatin or etoposide, offering the opportunity to use this deficiency to improve cancer patient treatment.

REFERENCES

"Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world." Louis Pasteur

REFERENCES

1. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA. National Academy of Sciences; 1977 Feb;74(2):560–4.

2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. National Academy of Sciences; 1977 Dec;74(12):5463–7.

3. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. NATURE. 1986 Jun;321(6071):674–9.

4. Gocayne J, Robinson DA, FitzGerald MG, Chung FZ, Kerlavage AR, Lentes KU, et al. Primary structure of rat cardiac beta-adrenergic and muscarinic cholinergic receptors obtained by automated DNA sequence analysis: further evidence for a multigene family. Proc Natl Acad Sci USA. National Acad Sciences; 1987 Dec;84(23):8296–300.

5. Hutchison CA. DNA sequencing: bench to bedside and beyond. Nucleic Acids Research. 2007;35(18):6227–37.

6. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008 Oct;26(10):1135–45.

7. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-time DNA sequencing using detection of pyrophosphate release. Analytical biochemistry. 1996 Nov 1;242(1):84–9.

8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. NATURE. 2005 Sep 15;437(7057):376–80.

9. Kircher M, Kelso J. High-throughput DNA sequencing - concepts and limitations. Bioessays. 2010 May 18;32(6):524–36.

10. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008 Nov;92(5): 255–64.

11. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8.

12. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Publishing Group. 2016 May 17;17(6):333–51.

13. Kchouk M, Gibrat JF, Elloumi M. Generations of Sequencing Technologies: From First to Next Generation. Biology and Medicine. 2017;09(03).

14. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.

15. Thompson JF, Steinmann KE. Single molecule sequencing with a HeliScope genetic analysis system. Curr Protoc Mol Biol. 2010 Oct;Chapter 7:Unit7.10.

16. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Human Molecular Genetics. 2010 Oct 15;19(R2):R227–40.

17. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan 2;323(5910):133–8.

18. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics & Bioinformatics. 2015 Oct;13(5):278–89.

19. Hodkinson BP, Grice EA. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Adv Wound Care (New Rochelle). 2015 Jan 1;4(1):50–8.

20. Timp W, Mirsaidov UM, Wang D, Comer J, Aksimentiev A, Timp G. Nanopore Sequencing: Electrical Measurements of the Code of Life. IEEE Trans Nanotechnol. 2010 May 1;9(3):281–94.

21. McGinn S, Gut IG. DNA sequencing – spanning the generations. New Biotechnology. 2013 May 25;30(4):366–72.

22. Wang Y, Yang Q, Wang Z. The evolution of nanopore sequencing. Front Genet. 2014;5:449.

23. Laszlo AH, Derrington IM, Brinkerhoff H, Langford KW, Nova IC, Samson JM, et al. Detection and mapping of 5-methylcytosine and 5-

hydroxymethylcytosine with nanopore MspA. Proc Natl Acad Sci USA. National Acad Sciences; 2013 Nov 19;110(47):18904–9.

24. Schreiber J, Wescoe ZL, Abu-Shumays R, Vivian JT, Baatar B, Karplus K, et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. Proc Natl Acad Sci USA. National Acad Sciences; 2013 Nov 19;110(47):18910–5.

25. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Res. 2015;4:1075.

26. Ke R, Mignardi M, Hauling T, Nilsson M. Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. Hum Mutat. 2016 Dec;37(12):1363–7.

27. Larsson C, Grundberg I, Söderberg O, Nilsson M. In situ detection and genotyping of individual mRNA molecules. Nature Publishing Group. 2010 May;7(5):395–7.

28. Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, et al. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Meth. 2013 Sep;10(9):857–60.

29. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014 Mar 21;343(6177):1360–3.

30. Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics. Nature Publishing Group; 2009 Dec 8;11(1):31–46.

31. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Human Molecular Genetics. 2010 Oct;19(R2):R145–51.

32. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Meth. 2010 Feb;7(2):111–8.

33. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Publishing Group. 2009 Jan;10(1):57–63.

34. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature Publishing Group. 2010 Sep;7(9):709–15.

35. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nature Publishing Group. 2014 Nov;15(11):709–21.

36. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289–93.

37. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007 Jun 8;316(5830):1497–502.

38. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet. 2004;38:525–52.

39. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. Public Library of Science; 2016 Aug;14(8):e1002533.

40. Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, et al. The Microbiome and Human Biology. Annu Rev Genom Human Genet. 2017 Mar 20;18(1):annurev–genom–083115–022438.

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

42. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research. 2008 Nov 1;18(11):1851–8.

43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

44. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

45. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. BioMed Central; 2013 Apr 25;14(4):R36.

46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for

analyzing next-generation DNA sequencing data. Genome Research. 2010 Sep 1;20(9):1297–303.

47. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2012 Apr;14(2):178–92.

48. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012 Feb;28(3):311–7.

49. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012 Apr;28(7):907–13.

50. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012 Jul 15;28(14): 1811–7.

51. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. Cold Spring Harbor Lab; 2012 Mar;22(3):568–76.

52. Christoforides A, Carpten JD, Weiss GJ, Demeure MJ, Hoff Von DD, Craig DW. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. BMC Genomics. 2013 May 4;14(1):302.

53. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013 Mar;31(3): 213–9.

54. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009 Nov 1;25(21):2865–71.

55. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Research. 2011 Jun;21(6):961–73.

56. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. Nat Meth. 2009 Aug 10;6(9):677–81.

57. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep 1;28(18):i333–9.

58. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Research. 2015 Mar 31;43(6):e39.

59. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009 Mar 6;10:80.

60. Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics. 2011 Jan 15;27(2):268–9.

61. Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011 Oct 1;27(19):2648–54.

62. Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, et al. CopywriteR: DNA copy number detection from off-target sequence data. Genome Biol. 2015 Feb 27;16(1):49.

63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

64. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010 May;28(5):511–5.

65. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 2013 Jan;31(1):46–53.

66. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

67. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics. 2008 Aug 1;24(15):1729–30.

68. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008 Nov 1;24(21):2537–8.

69. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Meth. 2008 Sep;5(9):829–34.

70. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004 Mar;4(3):177–83.

71. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002 Oct 21;21(48):7435–51.

72. Pfeifer GP, You Y-H, Besaratinia A. Mutations induced by ultraviolet light. Mutation Research. 2005 Apr 1;571(1-2):19–31.

73. Macé K, Aguilar F, Wang JS, Vautravers P, Gómez-Lechón M, Gonzalez FJ, et al. Aflatoxin B1-induced DNA adduct formation and p53 mutations in CYP450-expressing human liver cell lines. Carcinogenesis. 1997 Jul;18(7):1291–7.

74. Nedelko T, Arlt VM, Phillips DH, Hollstein M. TP53 mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. Int J Cancer. 2009 Feb 15;124(4):987–90.

75. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. CEll. 2012 May 25;149(5):979–93.

76. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. NATURE. 2013 Aug 22;500(7463):415–21.

77. Zhang C-Z, Leibowitz ML, Pellman D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. Genes Dev. 2013 Dec 1;27(23):2513–30.

78. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic

Eventduring Cancer Development. CEll. Elsevier Inc; 2011 Jan 7;144(1):27–40.

79. Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer. 2012 Oct;12(10):663–70.

80. Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. eLife. 2013 Apr 16;2:e00534.

81. Maciejowski J, Li Y, Bosco N, Campbell PJ, de Lange T. Chromothripsis and Kataegis Induced by Telomere Crisis. CEll. 2015 Dec 17;163(7):1641–54.

82. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. NATURE. 2013 Oct 17;502(7471):333–9.

83. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. NATURE. 2012 Jul 18;487(7407):330–7.

84. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. NATURE. 2016 Jun 2;534(7605):47–54.

85. Cancer Genome Atlas Network, Getz G, Saksena G, Park PJ, Chin L, Mills GB, et al. Comprehensive molecular portraits of human breast tumours. NATURE. 2012 Oct 4;490(7418):61–70.

86. Lawrence MS, Voet D, Lawrence MS, Voet D, Jing R, Jing R, et al. Comprehensive genomic characterization of squamous cell lung cancers. NATURE. 2012 Sep 9;489(7417):519–25.

87. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. NATURE. 2014 Jul 31;511(7511):543–50.

88. Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, et al. Mutations driving CLL and their evolution in progression and relapse. NATURE. 2015 Oct 22;526(7574):525–30.

89. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. NATURE. 2015 Oct 22;526(7574):519–24.

90. Biankin AV, Waddell N, Kassahn KS, Gingras M-C, Muthuswamy LB, Johns AL, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. NATURE. 2012 Nov 15;491(7424):399–405.

91. Patch A-M, Bailey P, Johns AL, Miller D, Quek K, Quinn MCJ, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. NATURE. 2015 Feb 26;518(7540):495–501.

92. Burrell RA, Swanton C. ScienceDirectThe evolution of the unstable cancer genome. Curr Opin Genet Dev. Elsevier Ltd; 2014 Feb 1;24:61–7.

93. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. Nature Publishing Group. 2017 Apr; 18(4):213–29.

94. Nowell PC. The clonal evolution of tumor cell populations. Science. American Association for the Advancement of Science; 1976 Oct 1;194(4260):23–8.

95. Cahill DP, Kinzler KW, Vogelstein B, Lengauer C. Genetic instability and darwinian selection in tumours. Trends Cell Biol. 1999 Dec; 9(12):M57–60.

96. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. NATURE. 2013 Sep 19;501(7467):338–45.

97. Magee JA, Piskounova E, Morrison SJ. Cancer Stem Cells:Impact, Heterogeneity, and Uncertainty. Cancer Cell. Elsevier Inc; 2012 Mar 20;21(3):283–96.

98. Swanton C. Intratumor heterogeneity: evolution through space and time. Cancer Research. 2012 Oct 1;72(19):4875–82.

99. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. N Engl J Med. 2012 Mar 8;366(10):883–92.

100. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nature Publishing Group. Nature Publishing Group; 2014 Feb 2;46(3):225–33.

101. Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian

cancers revealed through spatial mutational profiling. J Pathol. 2013 Oct 14;231(1):21–34.

102. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: Implications for targeted therapeutics. Br J Cancer. 2013 Feb 19;108(3): 479–85.

103. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. NATURE. 2012 Jan 11;481(7382):506–10.

104. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. CEll. 2013 Feb 14;152(4):714–26.

105. Almendro V, Cheng Y-K, Randles A, Itzkovitz S, Marusyk A, Ametller E, et al. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. CellReports. 2014 Feb 13;6(3):514–27.

106. Greaves M, Maley CC. Clonal evolution in cancer. NATURE. 2012 Jan 18;481(7381):306–13.

107. Brandon M, Baldi P, Wallace DC. Mitochondrial mutations in cancer. Oncogene. 2006 Aug 7;25(34):4647–62.

108. Lu J, Sharma LK, Bai Y. Implications of mitochondrial DNA mutations and mitochondrial dysfunction in tumorigenesis. Cell Res. 2009 Jul 1;19(7):802–15.

109. Czarnecka AM, Kukwa W, Krawczyk T, Scinska A, Kukwa A, Cappello F. Mitochondrial DNA mutations in cancer - From bench to bedside. Frontiers in Bioscience. 2010 Jan 1;15(2):437–60.

110. Mizutani S, Miyato Y, Shidara Y, Asoh S, Tokunaga A, Tajiri T, et al. Mutations in the mitochondrial genome confer resistance of cancer cells to anticancer drugs. Cancer Science. 2009 Sep;100(9):1680–7.

111. Shidara Y, Yamagata K, Kanamori T, Nakano K, Kwong JQ, Manfredi G, et al. Positive contribution of pathogenic mutations in the mitochondrial genome to the promotion of cancer by prevention from apoptosis. Cancer Research. American Association for Cancer Research; 2005 Mar 1;65(5):1655–63.

112. Kulawiec M, Owens KM, Singh KK. Cancer cell mitochondria confer apoptosis resistance and promote metastasis. Cancer Biol Ther. 2009 Jul 15;8(14):1378–85.

113. Coller HA, Khrapko K, Bodyak ND, Nekhaeva E, Herrero-Jimenez P, Thilly WG. High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. Nat Genet. 2001 Jun 26;28(2):147–50.

114. Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M, et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. eLife. 2014 Oct 1;3:415.

115. Hake SB, Xiao A, Allis CD. Linking the epigenetic “language” of covalent histone modifications to cancer. Br J Cancer. 2004 Feb 23;90(4): 761–9.

116. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011 Mar;21(3):381–95.

117. Roberts CWM, Orkin SH. The SWI/SNF complex--chromatin and cancer. Nat Rev Cancer. 2004 Feb;4(2):133–42.

118. Hargreaves DC, Crabtree GR. ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. Cell Res. 2011 Mar;21(3): 396–420.

119. Erdel F, Rippe K. Chromatin remodelling in mammalian cells by ISWI-type complexes--where, when and why? FEBS J. 2011 Oct;278(19): 3608–18.

120. Bartholomew B. ISWI chromatin remodeling: one primary actor or a coordinated effort? Curr Opin Struct Biol. 2014 Feb;24:150–5.

121. Aydin ÖZ, Vermeulen W, Lans H. ISWI chromatin remodeling complexes in the DNA damage response. Cell Cycle. 2014;13(19):3016–25.

122. Witkowski L, Foulkes WD. In Brief: Picturing the complex world of chromatin remodelling families. J Pathol. 2015 Dec;237(4):403–6.

123. Zhang P, Torres K, Liu X, Liu C-G, Pollock RE. An Overview of Chromatin-Regulating Proteins in Cells. Curr Protein Pept Sci. 2016;17(5): 401–10.

124. Morrison AJ, Shen X. Chromatin remodelling beyond transcription: the INO80 and SWR1 complexes. Nat Rev Mol Cell Biol. 2009 Jun;10(6):373–84.

125. Stanley FKT, Moore S, Goodarzi AA. CHD chromatin remodelling enzymes and the DNA damage response. Mutation Research. 2013 Oct;750(1-2):31–44.

126. Shain AH, Pollack JR. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. Kashanchi F, editor. PLoS ONE. 2013;8(1):e55119.

127. Kadoch C, Hargreaves DC, Hodges C, Elias L, Ho L, Ranish J, et al. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nature Publishing Group. 2013 Jun;45(6):592–601.

128. Versteege I, Sévenet N, Lange J, Rousseau-Merck MF, Ambros P, Handgretinger R, et al. Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. NATURE. 1998 Jul 9;394(6689):203–6.

129. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. NATURE. 2011 Jan 27;469(7331):539–42.

130. Guan B, Wang T-L, Shih I-M. ARID1A, a factor that promotes formation of SWI/SNF-mediated chromatin remodeling, is a tumor suppressor in gynecologic cancers. Cancer Research. American Association for Cancer Research; 2011 Nov 1;71(21):6718–27.

131. Shain AH, Giacomini CP, Matsukuma K, Karikari CA, Bashyam MD, Hidalgo M, et al. Convergent structural alterations define SWItch/Sucrose NonFermentable (SWI/SNF) chromatin remodeler as a central tumor suppressive complex in pancreatic cancer. Proc Natl Acad Sci USA. 2012 Jan 31;109(5):E252–9.

132. Li M, Zhao H, Zhang X, Wood LD, Anders RA, Choti MA, et al. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet. Nature Publishing Group; 2011 Aug 7;43(9):828–9.

133. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat J-P, et al. A Landscape of Driver Mutations in Melanoma. CEll. Elsevier Inc; 2012 Jul 20;150(2):251–63.

134. Manceau G, Letouzé E, Guichard C, Didelot A, Cazes A, Corté H, et al. Recurrent inactivating mutations of ARID2 in non-small cell lung carcinoma. Int J Cancer. Wiley Subscription Services, Inc., A Wiley Company; 2013 May 1;132(9):2217–21.

135. Drost J, Mantovani F, Tocco F, Elkon R, Comel A, Holstege H, et al. BRD7 is a candidate tumour suppressor gene required for p53 function. Nat Cell Biol. 2010 Apr;12(4):380–9.

136. Reisman DN, Sciarrotta J, Wang W, Funkhouser WK, Weissman BE. Loss of BRG1/BRM in human lung cancer cell lines and primary lung cancers: correlation with poor prognosis. Cancer Research. 2003 Feb 1;63(3):560–6.

137. Reisman D, Glaros S, Thompson EA. The SWI/SNF complex and cancer. Oncogene. 2009 Apr 9;28(14):1653–68.

138. Zhang X, Sun Q, Shan M, Niu M, Liu T, Xia B, et al. Promoter hypermethylation of ARID1A gene is responsible for its low mRNA expression in many invasive breast cancers. PLoS ONE. 2013;8(1):e53931.

139. Hodges C, Kirkland JG, Crabtree GR. The Many Roles of BAF (mSWI/SNF) and PBAF Complexes in Cancer. Cold Spring Harb Perspect Med. 2016 Aug 1;6(8):a026930.

140. Dutta A, Sardiu M, Gogol M, Gilmore J, Zhang D, Florens L, et al. Composition and Function of Mutant Swi/Snf Complexes. CellReports. 2017 Feb 28;18(9):2124–34.

141. Pulice JL, Kadoch C. Composition and Function of Mammalian SWI/SNF Chromatin Remodeling Complexes in Human Disease. Cold Spring Harb Symp Quant Biol. 2017 Apr 13;81:53–60.

142. Wilson BG, Wilson BG, Roberts CWM, Roberts CWM. SWI/SNF nucleosome remodellers and cancer. Nat Rev Cancer. 2011 Jun 9;11(7):481–92.

143. Burrows AE, Burrows AE, Smogorzewska A, Smogorzewska A, Elledge SJ, Elledge SJ. Polybromo-associated BRG1-associated factor components BRD7 and BAF180 are critical regulators of p53 required for induction of replicative senescence. Proceedings of the National Academy of Sciences. 2010 Aug 10;107(32):14280–5.

144. Nagl NG, Zweitzig DR, Thimmapaya B, Beck GR, Moran E. The c-myc gene is a direct target of mammalian SWI/SNF-related complexes during differentiation-associated cell cycle arrest. Cancer Research. 2006 Feb 1;66(3):1289–93.

145. Flowers S, Beck GR, Moran E. Transcriptional Activation by pRB and Its Coordination with SWI/SNF Recruitment. Cancer Research. 2010 Nov;70(21):8282–7.

146. Weissman B, Knudsen KE. Hijacking the chromatin remodeling machinery: impact of SWI/SNF perturbations in cancer. Cancer Research. 2009 Nov 1;69(21):8223–30.

147. Dykhuizen EC, Hargreaves DC, Miller EL, Cui K, Korshunov A, Kool M, et al. BAF complexes facilitate decatenation of DNA by topoisomerase IIα. NATURE. 2013 May 30;497(7451):624–7.

148. Ray A, Mir SN, Wani G, Zhao Q, Battu A, Zhu Q, et al. Human SNF5/INI1, a component of the human SWI/SNF chromatin remodeling complex, promotes nucleotide excision repair by influencing ATM recruitment and downstream H2AX phosphorylation. Mol Cell Biol. American Society for Microbiology; 2009 Dec;29(23):6206–19.

149. Lee H-S, Park J-H, Kim S-J, Kwon S-J, Kwon J. A cooperative activation loop among SWI/SNF, gamma-H2AX and H3 acetylation for DNA double-strand break repair. The EMBO Journal. 2010 Apr 21;29(8):1434–45.

150. Niimi A, Chambers AL, Downs JA, Lehmann AR. A role for chromatin remodellers in replication of damaged DNA. Nucleic Acids Research. 2012 Aug;40(15):7393–403.

151. Kakarougkas A, Kakarougkas A, Ismail A, Ismail A, Chambers AL, Chambers AL, et al. Requirement for PBAF in transcriptional repression and repair at DNA breaks in actively transcribed regions of chromatin. Molecular Cell. 2014 Sep 4;55(5):723–32.

152. Romero OA, Torres-Diz M, Pros E, Savola S, Gomez A, Moran S, et al. MAX inactivation in small cell lung cancer disrupts MYC-SWI/SNF programs and is synthetic lethal with BRG1. Cancer Discovery. 2014 Mar; 4(3):292–303.

153. Oike T, Oike T, Ogiwara H, Ogiwara H, Tominaga Y, Tominaga Y, et al. A Synthetic Lethality-Based Strategy to Treat Cancers Harboring a Genetic Deficiency in the Chromatin Remodeling Factor BRG1. Cancer Research. 2013 Sep 2;73(17):5508–18.

154. Helming KC, Helming KC, Wang X, Wang X, Wilson BG, Wilson BG, et al. ARID1B is a specific vulnerability in ARID1A-mutant cancers. Nature Medicine. 2014 Mar;20(3):251–4.

155. Shen J, Peng Y, Wei L, Zhang W, Yang L, Lan L, et al. ARID1A Deficiency Impairs the DNA Damage Checkpoint and Sensitizes Cells to PARP Inhibitors. Cancer Discovery. American Association for Cancer Research; 2015 Jul;5(7):752–67.

156. Kim KH, Kim W, Howard TP, Vazquez F, Tsherniak A, Wu JN, et al. SWI/SNF-mutant cancers depend on catalytic and non-catalytic activity of EZH2. Nature Publishing Group. 2015 Dec;21(12):1491–6.

157. St Pierre R, Kadoch C. Mammalian SWI/SNF complexes in cancer: emerging therapeutic opportunities. Curr Opin Genet Dev. 2017 Feb;42:56–67.

158. Schiaffino-Ortega S, Balinas C, Cuadros M, Medina PP. SWI/SNF proteins as targets in cancer therapy. Journal of Hematology & Oncology. BioMed Central; 2014 Nov 13;7(1):81.

159. Gerstenberger BS, Trzupek JD, Tallant C, Fedorov O, Filippakopoulos P, Brennan PE, et al. Identification of a Chemical Probe for Family VIII Bromodomains through Optimization of a Fragment Hit. J Med Chem. 2016 May 26;59(10):4800–11.

160. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Research. 2012 Nov 24;40(21):e169–9.

161. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166–9.

162. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. BioMed Central; 2010;11(10):R106.

163. Lauand C, Niero EL, Dias VM, Machado-Santelli GM. Cell cycle synchronization and BrdU incorporation as a tool to study the possible selective elimination of ErbB1 gene in the micronuclei in A549 cells. Braz J Med Biol Res. 2015 May;48(5):382–91.

164. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. BioMed Central; 2016 Jun 16;17(1):128.

165. Yan Z, Cui K, Murray DM, Ling C, Xue Y, Gerstein A, et al. PBAF chromatin-remodeling complex requires a novel specificity subunit, BAF200, to regulate expression of selective interferon-responsive genes. Genes Dev. 2005 Jul 15;19(14):1662–7.

166. Duan Y, Tian L, Gao Q, Liang L, Zhang W, Yang Y, et al. Chromatin remodeling gene ARID2 targets cyclin D1 and cyclin E1 to suppress hepatoma cell progression. Oncotarget. Impact Journals; 2016 Jul 19;7(29):45863–75.

167. Oba A, Shimada S, Akiyama Y, Nishikawaji T, Mogushi K, Ito H, et al. ARID2 modulates DNA damage response in human hepatocellular carcinoma cells. J Hepatol. 2017 May;66(5):942–51.

168. Huang J, Deng Q, Wang Q, Li K-Y, Dai J-H, Li N, et al. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet. 2012 Oct;44(10):1117–21.

169. Kim KH, Kwon BM, Myers AG, Rees DC. Crystal structure of neocarzinostatin, an antitumor protein-chromophore complex. Science. 1993 Nov 12;262(5136):1042–6.

170. Bañuelos A, Reyes E, Ocadiz R, Alvarez E, Moreno M, Monroy A, et al. Neocarzinostatin induces an effective p53-dependent response in human papillomavirus-positive cervical cancer cells. J Pharmacol Exp Ther. 2003 Aug;306(2):671–80.

171. Long BH, Musial ST, Brattain MG. Single- and double-strand DNA breakage and repair in human lung adenocarcinoma cells exposed to etoposide and teniposide. Cancer Research. 1985 Jul;45(7):3106–12.

172. Deweese JE, Osheroff N. The DNA cleavage reaction of topoisomerase II: wolf in sheep's clothing. Nucleic Acids Research. 2009 Feb;37(3):738–48.

173. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research. 2011 Jul;39(13):e90–0.

174. Allhoff M, Schönhuth A, Martin M, Costa IG, Rahmann S, Marschall T. Discovering motifs that induce sequencing errors. BMC Bioinformatics. BioMed Central; 2013;14 Suppl 5(Suppl 5):S1.

175. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Research. 2013 Apr 1;41(6):e67–7.

176. Greenberg MM. In vitro and in vivo effects of oxidative damage to deoxyguanosine. Biochm Soc Trans. 2004 Feb;32(Pt 1):46–50.

177. Yakes FM, Van Houten B. Mitochondrial DNA damage is more extensive and persists longer than nuclear DNA damage in human cells following oxidative stress. Proceedings of the National Academy of Sciences. 1997 Jan 21;94(2):514–9.

178. Liu X. In vitro chromatin templates to study nucleotide excision repair. DNA Repair. 2015 Dec;36:68–76.

179. Pohjoismäki JLO, Boettger T, Liu Z, Goffart S, Szibor M, Braun T. Oxidative stress during mitochondrial biogenesis compromises mtDNA

integrity in growing hearts and induces a global DNA repair response. Nucleic Acids Research. 2012 Aug;40(14):6595–607.

180. Cline SD. Mitochondrial DNA damage and its consequences for mitochondrial gene expression. Biochim Biophys Acta. 2012 Sep;1819(9-10): 979–91.

181. Aamann MD, Sorensen MM, Hvitby C, Berquist BR, Muftuoglu M, Tian J, et al. Cockayne syndrome group B protein promotes mitochondrial DNA stability by supporting the DNA repair association with the mitochondrial membrane. FASEB J. 2010 Jul;24(7):2334–46.

182. Kamenisch Y, Fousteri M, Knoch J, Thaler von A-K, Fehrenbacher B, Kato H, et al. Proteins of nucleotide and base excision repair pathways interact in mitochondria to protect from loss of subcutaneous fat, a hallmark of aging. J Exp Med. 2010 Feb 15;207(2):379–90.

183. Lagerwerf S, Vrouwe MG, Overmeer RM, Fousteri MI, Mullenders LHF. DNA damage response and transcription. DNA Repair. 2011 Jul 15;10(7):743–50.

184. Gonzalez-Perez A, Jene-Sanz A, Lopez-Bigas N. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. Genome Biol. 2013;14(9):r106.

185. Liao S, Davoli T, Leng Y, Li MZ, Xu Q, Elledge SJ. A genetic interaction analysis identifies cancer drivers that modify EGFR dependency. Genes Dev. 2017 Jan 15;31(2):184–96.

186. Barker N, Hurlstone A, Musisi H, Miles A, Bienz M, Clevers H. The chromatin remodelling factor Brg-1 interacts with beta-catenin to promote target gene activation. The EMBO Journal. EMBO Press; 2001 Sep 3;20(17):4935–43.

187. Mora-Blanco EL, Mishina Y, Tillman EJ, Cho YJ, Thom CS, Pomeroy SL, et al. Activation of β-catenin/TCF targets following loss of the tumor suppressor SNF5. Oncogene. 2014 Feb 13;33(7):933–8.

188. Vasileiou G, Ekici AB, Uebe S, Zweier C, Hoyer J, Engels H, et al. Chromatin-Remodeling-Factor ARID1B Represses Wnt/β-Catenin Signaling. Am J Hum Genet. 2015 Sep 3;97(3):445–56.

189. Ji S, Zhu L, Gao Y, Zhang X, Yan Y, Cen J, et al. Baf60b-mediated ATM-p53 activation blocks cell identity conversion by sensing chromatin opening. Cell Res. 2017 May;27(5):642–56.

190. Berger ND, Stanley FKT, Moore S, Goodarzi AA. ATM-dependent pathways of chromatin remodelling and oxidative DNA damage responses. Philos Trans R Soc Lond, B, Biol Sci. 2017 Oct 5;372(1731).

191. Schoenfeld D, Su W, Zairis S, Mathur D, Rabadan R, Parsons R. Abstract A24: PBRM1 alteration in clear cell renal cell carcinoma increases tumorigenicity through ALDH1A1 upregulation. Cancer Res. 2016 Jan 14;76(2 Supplement):A24–4.

192. Jiang W, Dulaimi E, Devarajan K, Parsons T, Wang Q, O'Neill R, et al. Intratumoral heterogeneity analysis reveals hidden associations between protein expression losses and patient survival in clear cell renal cell carcinoma. Oncotarget. 2017 Jun 6;8(23):37423–34.

193. Wang X, Nagl NG, Wilsker D, Van Scoy M, Pacchione S, Yaciuk P, et al. Two related ARID family proteins are alternative subunits of human SWI/SNF complexes. Biochem J. Portland Press Limited; 2004 Oct 15;383(Pt 2):319–25.

194. Beijersbergen RL, Wessels LFA, Bernards R. Synthetic Lethality in Cancer Therapeutics. Annu Rev Cancer Biol. 2017;1(1):141–61.

195. Wong MM, Cox LK, Chrivia JC. The chromatin remodeling protein, SRCAP, is critical for deposition of the histone variant H2A.Z at promoters. J Biol Chem. American Society for Biochemistry and Molecular Biology; 2007 Sep 7;282(36):26132–9.

196. Mouradov D, Sloggett C, Jorissen RN, Love CG, Li S, Burgess AW, et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. Cancer Research. 2014 Jun 15;74(12):3238–47.

197. Patil V, Pal J, Somasundaram K. Elucidating the cancer-specific genetic alteration spectrum of glioblastoma derived cell lines from whole exome and RNA sequencing. Oncotarget. Impact Journals; 2015 Dec 22;6(41):43452–71.

198. de Castro RO, Previato L, Goitea V, Felberg A, Guiraldelli MF, Filiberti A, et al. The Chromatin-remodeling Subunit Baf200 Promotes Homology-directed DNA Repair and Regulates Distinct Chromatin-remodeling Complexes. J Biol Chem. American Society for Biochemistry and Molecular Biology; 2017 Apr 5;292(20):jbc.M117.778183–471.

199. Katagiri A, Nakayama K, Rahman MT, Rahman M, Katagiri H, Nakayama N, et al. Loss of ARID1A expression is related to shorter

progression-free survival and chemoresistance in ovarian clear cell carcinoma. Mod Pathol. 2012 Feb;25(2):282–8.

200. Yates LR, Knappskog S, Wedge D, Farmery JHR, González S, Martincorena I, et al. Genomic Evolution of Breast Cancer Metastasis and Relapse. Cancer Cell. 2017 Aug 14;32(2):169–184.e7.

201. Gibson WJ, Hoivik EA, Halle MK, Taylor-Weiner A, Cherniack AD, Berg A, et al. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. Nat Genet. 2016 Aug;48(8):848–55.

202. Mao T-L, Ardighieri L, Ayhan A, Kuo K-T, Wu C-H, Wang T-L, et al. Loss of ARID1A expression correlates with stages of tumor progression in uterine endometrioid carcinoma. Am J Surg Pathol. 2013 Sep;37(9):1342–8.

203. He F, Li J, Xu J, Zhang S, Xu Y, Zhao W, et al. Decreased expression of ARID1A associates with poor prognosis and promotes metastases of hepatocellular carcinoma. Journal of Experimental & Clinical Cancer Research. BioMed Central; 2015 May 15;34(1):47.

204. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature Publishing Group. 2017 May 8;31(6):1806–713.

205. Kothandapani A, Gopalakrishnan K, Kahali B, Reisman D, Patrick SM. Downregulation of SWI/SNF chromatin remodeling factor subunits modulates cisplatin cytotoxicity. Experimental Cell Research. 2012 Oct 1;318(16):1973–86.

206. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. 2017 Aug;18(8):1009–21.

207. Stratton MR, Stratton MR, Campbell PJ, Campbell PJ, Futreal PA, Futreal PA. The cancer genome. NATURE. 2009 Apr 9;458(7239):719–24.

208. Ren R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. Nat Rev Cancer. 2005 Mar;5(3):172–83.

209. Wang Z-Y, Chen Z. Acute promyelocytic leukemia: from highly fatal to highly curable. Blood. 2008 Mar;111(5):2505–15.

210. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth

factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004 May;350(21):2129–39.

211. Bollag G, Hirth P, Tsai J, Zhang J, Ibrahim PN, Cho H, et al. Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. NATURE. 2010 Sep;467(7315):596–9.

212. Martin JA, Wang Z. Next-generation transcriptome assembly. Nature Publishing Group. 2011 Oct;12(10):671–82.

213. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. NATURE. 2010 Jan 21;463(7279):360–3.

214. Osley MA, Shen X. Altering nucleosomes during DNA double-strand break repair in yeast. Trends Genet. 2006 Dec;22(12):671–7.

215. Park J-H, Park E-J, Lee H-S, Kim S-J, Hur S-K, Imbalzano AN, et al. Mammalian SWI/SNF complexes facilitate DNA double-strand break repair by promoting gamma-H2AX induction. The EMBO Journal. 2006 Sep 6;25(17):3986–97.

216. Park J-H, Park E-J, Hur S-K, Kim S, Kwon J. Mammalian SWI/SNF chromatin remodeling complexes are required to prevent apoptosis after DNA damage. DNA Repair. 2009 Jan 1;8(1):29–39.

217. Martínez N, Almaraz C, Vaqué JP, Varela I, Derdak S, Beltran S, et al. Whole-exome sequencing in splenic marginal zone lymphoma reveals mutations in genes involved in marginal zone differentiation. Leukemia. 2014 Jun;28(6):1334–40.

218. Conte N, Varela I, Grove C, Manes N, Yusa K, Moreno T, et al. Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture. Leukemia. 2013 Sep;27(9): 1820–5.

RESUMEN EN ESPAÑOL

“Me he dedicado a investigar la vida y no sé por qué, ni para qué”. Severo Ochoa

1. INTRODUCCIÓN

El cáncer es un grupo de enfermedades caracterizadas por la acumulación progresiva de alteraciones genéticas y epigenéticas a lo largo del genoma, transformando células normales en tumorales. Los cambios genéticos incluyen mutaciones puntuales, pequeñas inserciones o deleciones (indels), cambios en el número de copia y reorganizaciones cromosómicas. Algunos de estos cambios denominados conductores (*drivers*), modifican rutas moleculares importantes para la proliferación y diseminación tumoral, confiriéndoles ventaja selectiva, mientras que otras son una mera consecuencia de la inestabilidad genómica intrínseca del tumor (*passengers*) (207).

En 2000, Douglas Hanahan y Robert Weinberg describieron seis cambios fenotípicos esenciales (*hallmarks*) que adquieren las células durante el desarrollo y progresión tumoral. La identificación de genes o rutas involucradas en la adquisición de estas capacidades, es de vital importancia para entender el proceso biológico del cáncer, y mejorar el diagnóstico, tratamiento y pronostico de los pacientes. Durante los últimos años, la caracterización de algunas de estas alteraciones se ha traducido en grandes avances en el tratamiento de algunos tipos tumorales, como en el caso de la identificación de la translocación *BCR-ABL* y *PML-RARa* en leucemia mieloide crónica y leucemia promielocítica respectivamente (208,209), mutaciones de *EGFR* en el cáncer de pulmón (210), o más recientemente de las mutaciones *BRAF* en el melanoma (211).

La llegada de tecnologías de secuenciación de nueva generación (NGS) en 2005 revolucionó la investigación en genómica del cáncer. Estas plataformas producen más de 100 veces más datos por una fracción minúscula del trabajo y presupuesto necesarios con los secuenciadores capilares más sofisticados basados en el método tradicional de Sanger. Sin embargo, el gran volumen de

datos generados implica un aumento considerable de las habilidades e infraestructura informáticas necesarias para analizar los datos. En consecuencia, una nueva generación de herramientas informáticas se ha ido generado en los últimos años para el acceso y la manipulación de los datos; aunque es todavía un campo que requiere mucha más investigación y desarrollo.

Aunque el coste de la secuenciación de genomas completos ha disminuido drásticamente, constituye todavía una inversión económica muy significativa para secuenciar genomas completos de las grandes colecciones de muestras de cáncer necesarias para identificar genes tumorales con penetrancia moderada. Como sólo un pequeño porcentaje del genoma se transcribe en ARN, una buena alternativa es utilizar ARNm como material de partida. La secuenciación de ARN (RNA-Seq) constituye una oportunidad barata para secuenciar todo el transcriptoma de un gran número de muestras.

Además, en RNA-Seq el número de lecturas obtenido es directamente proporcional a la cantidad de material de partida, lo que permite extraer valores de expresión génica con una sensibilidad y una reproducibilidad difíciles de obtener con los *arrays* de expresión tradicionales. Por último, con el uso de las herramientas correctas, se pueden identificar nuevos genes y transcritos, así como *splicing* alternativo o desequilibrios en expresión alélica (212).

Con todas las ventajas ofrecidas por la tecnología de RNA-Seq, las enormes diferencias en la abundancia de ARNm entre los diferentes genes hacen que la identificación de variantes de secuencia en este tipo de experimento sea una tarea muy compleja. La secuenciación dirigida de DNA, normalmente mediante la captura de regiones específicas del ADN genómico mediante hibridación con sondas; constituye una aproximación intermedia entre la secuenciación del genoma y el transcriptoma.

Los complejos remodeladores de cromatina utilizan la energía del ATP para interrumpir el contacto nucleosoma-ADN, moviendo los nucleosomas a lo largo del ADN. Esto facilita el acceso al ADN de proteínas implicadas en la reparación, el control transcripcional, etc.

Existen cuatro grandes familias de complejos remodeladores de cromatina dependientes de ATP: SWI/SNF, INO80/SWR1, ISWI y CHD/NuRD. En mamíferos, los complejos de la familia SWI/SNF se componen de una de las dos subunidades ATPasas catalíticas mutuamente excluyentes (*ya sea BRM/ SMARCA2 o BRG1/SMARCA4*), un conjunto de subunidades altamente conservadas o "núcleo" (SNF5/SMARCB1, INI1 y BAF47) y subunidades intercambiables que se cree que contribuyen al ensamblaje y regulación de funciones específicas de tejido de los complejos (ejemplos de estas subunidades son ARID1A, ARID1B, PBRM1 o BRD7). Varias de estas subunidades son codificadas por genes con múltiples productos producidos por *splicing* alternativo. Además, la mayoría de estos genes pertenecen a grandes familias génicas que a menudo muestran expresión diferencial dependiente de tejido. Por tanto, es probable que exista un gran número de variantes de complejos SWI/SNF en mamíferos y que esto contribuya a la regulación dependiente de tejido (118).

Defectos en varias subunidades de los complejos de la familia SWI/SNF se han asociado con la progresión tumoral (142). Así, por ejemplo, *SNF5* se inactiva mediante mutaciones bialélicas en casi todos los tumores rabdoides malignos. De manera similar, *PBRM1* se inactiva en el 40% de los carcinomas renales de células claras, donde con menor frecuencia también se han identificado mutaciones en *UTX* y *JARID1C* (histonas demetilasas), así como *SETD2* (histona metilasa) (129,213). Además, *ARID1A* se encuentra mutado en el 50% de los carcinomas de ovario y en el 30% de los carcinomas endometrioides (130); *ARID2* también se ha encontrado mutado de forma recurrente en carcinoma hepatocelular (132) y *BRD7* se encuentra delecionado frecuentemente en cáncer de mama. Por último, la expresión de *SMARCA4* y *SMARCA2* se encuentra reprimida en cáncer de pulmón (137). Curiosamente, la expresión de *SMARCA2* también se encuentra disminuida en cáncer de próstata, donde su ausencia se relaciona con estadios avanzados de progresión de la enfermedad y peor pronóstico (136).

Aún no está claro cómo el deterioro de los complejos SWI/SNF puede contribuir al desarrollo de cáncer, pero han sido identificadas funciones esenciales de los complejos en la diferenciación de varios linajes celulares (142). Además, se han descrito interacciones directas de los complejos SWI/

SNF con supresores tumorales como *RB* y *TP53*, así como con el oncogén *MYC* (143-145).

Por último, varios estudios han identificado que la inactivación de los complejos SWI/SNF conduce a una mayor sensibilidad a daños en el DNA (214-216).

2. OBJETIVOS

De acuerdo a todo lo expuesto anteriormente, el principal objetivo de esta tesis doctoral es determinar qué genes/complejos juegan un papel importante en el desarrollo tumoral, así como intentar identificar los mecanismos moleculares por los cuales los defectos en estos genes juegan un papel importante en el desarrollo tumoral y si, finalmente, estas alteraciones se pueden utilizar con valor diagnóstico, pronóstico o terapéutico para mejorar el manejo de los pacientes de cáncer.

Este objetivo general se implementa en los siguientes objetivos concretos:

1. **Identificar aquellos genes de los complejos remodeladores de cromatina que juegan un papel importante en el desarrollo tumoral en humanos.**

2. **Caracterizar los mecanismos moleculares mediante los cuales las alteraciones de los genes remodeladores de cromatina producen su efecto en la progresión tumoral y la metástasis.**

3. **Estudiar la posibilidad de usar las alteraciones en complejos remodeladores de cromatina como marcador de pronóstico o para mejorar el tratamiento de pacientes con cáncer.**

3. RESULTADOS Y DISCUSIÓN

Durante la realización de esta tesis, hemos secuenciado la región codificante de una lista de 250 genes, en un total de 732 muestras, de ellas, 479 son tumores, 257 normales y 45 líneas celulares tumorales.

En un primer lugar nos llamó la atención una gran cantidad de mutaciones encontradas en el ADN mitocondrial (ADNmt). Debido a los diversos estudios que implican un papel de las alteraciones mitocondriales en la progresión tumoral, decidimos estudiar este fenómeno más a fondo. En total hemos identificado 170 mutaciones somáticas en el ADNmt. Aproximadamente el ~10% de estas mutaciones se encuentran en la mayoría de las mitondrias en las células tumorales (homoplasmia) lo que podría indicar una presión selectiva para acumular mitocondrias defectuosas. Además, observamos que la mayoría de mutaciones corresponden a transiciones G>A/C>T y T>C/A>G. Este perfil mutacional no coincide con el esperado asumiendo que la causa principal de acumulación de mutaciones en el ADNmt es el estrés oxidativo que se caracteriza por una acumulación de transversiones G>T/C>A. Esto nos hace plantear la posibilidad de que el daño producido por el estrés oxidativo genera un perfil mutacional distinto en el ADNmt en comparación con el generado en el ADN nuclear. Además, encontramos un sesgo muy importante entre las mutaciones encontradas en la hebra transcrita frente a la hebra no transcrita. Esto podría indicar que existe un fenómeno de reparación asociada a la transcripción activo en la mitocondria a pesar de que este mecanismo está asociado al sistema NER y no se ha descrito este sistema de reparación en el ADNmt.

En cuanto a los datos de mutaciones en genes implicados en la estructura de la cromatina, identificamos un total de 4920 mutaciones codificantes,

observando en primer lugar la presencia de mutaciones recurrentes en genes del complejo SWI/SNF cuya implicación está descrita en el desarrollo tumoral como *ARID1A*, *ARID1B* o *PBRM1*. Además, identificamos nuevos genes candidatos que no se habían postulado anteriormente como implicados en cáncer o no han sido descritos como importantes en los subtipos en los que han sido encontrados en nuestro estudio. Entre estos genes cabe destacar *ARID2*, *SRCAP*, *SMARCA5*, *CTCF*, *CHD4*, *SSX1*, *JHDM1D* o *SMAD4*. Estos genes suponen nuevos candidatos para analizar su implicación en el desarrollo tumoral. Hemos validados algunas de las mutaciones identificadas con el objetivo de comprobar el grado de falsos negativos generados en el experimento. Para ello, seleccionamos aproximadamente 160 mutaciones que fueron validadas mediante PCR acoplada a secuenciación a alta cobertura obteniendo unos valores de especificidad del 70%.

Uno de los genes que ha llamado nuestro interés es *ARID2*, que encontramos recurrentemente mutado en aproximadamente el 15% de las 82 muestras de cáncer de pulmón secuenciadas. Este gen pertenece a la familia de complejos remodeladores de la cromatina SWI/SNF, y específicamente al complejo PBAF. Varios de los genes de esta familia se han reportado frecuentemente mutados en diferentes tipos de cáncer. *ARID2*, se ha descrito como supresor tumoral en melanoma, y hepatocarcinoma, pero no existen muchos artículos que muestren evidencia de su implicación en cáncer de pulmón. Para comprobar la frecuencia mutacional y para tener una cohorte más representativa, hemos secuenciado la zona codificante del *ARID2* en 96 muestras adicionales de adenocarcinoma de pulmón. De nuevo hemos visto una frecuencia de mutaciones codificantes de aproximadamente el 15% en esta segunda cohorte de muestras. Todas las mutaciones encontradas en *ARID2* se han validado con la misma estrategia explicada anteriormente

Para comprobar el posible efecto de estas mutaciones en las muestras humanas, llevamos a cabo experimentos de immunohistoquímica con un anticuerpo específico anti-ARID2 lo que nos permitió comprobar la ausencia de producción de proteína en la mayoría de las muestras que presentaban mutaciones en este gen mientras que esta proteína se expresaba en altos nieveles en las muestras que no presentan mutaciones en *ARID2*. Además, usando la base de datos del Genome Atlas Consortium, vimos que la pérdida de la expresión de *ARID2* se correlaciona significativamente con peor pronóstico apoyando el papel de *ARID2* como supresor tumoral en cáncer de pulmón humano.

Teniendo en cuenta todo esto y el posible papel de *ARID2* como supresor tumoral en otros tipos de cáncer, quisimos comprobar el efecto de la inactivación de *ARID2* en las capacidades de proliferación, migración/invasión y tumorigénesis en diversas líneas celulares de cáncer de pulmón, así como evaluar su implicación en los mecanismos de reparación del ADN. Para ello, decidimos hacer experimentos de silenciamiento usando líneas celulares de cáncer de Pulmón (A549, H1299, H460) usando vectores inducibles lentivirales de expresión de ARNs interferentes shRNA.

Confirmamos el silenciamiento de *ARID2* tanto a nivel de ARNm mediante qPCR como a nivel de proteína mediante Western-blot en células seleccionadas mediante FACS usando la proteína fluorescente turboRFP contenida en el vector de expresión del shRNA. En los primeros experimentos en donde usamos la línea celular A549, la reducción de expresión de *ARID2* provocó un aumento de la proliferación, evidenciado mediante la generación de curvas de crecimiento y mediante ensayos con el marcaje CFSE que permite identificar el número de divisiones celulares. Además de confirmar estas observaciones por triplicado, comprobamos que las células deficientes en *ARID2* presentaban una capacidad aumentada de invasión y migración *in vitro*. Finalmente, comprobamos que estas células presentaban también una mayor capacidad de producir tumores *in vivo*.

Como posible mecanismo para este papel antitumoral de *ARID2*, evaluamos el nivel de reparación del ADN en estas células tras el silenciamiento de *ARID2* usando 2 agentes que producen daño en el ADN: etopósido y neocarzinostatin. Se cuantificó el nivel de daño mediante el contaje de focos de reparación evidenciados mediante inmunofluorescencia con Anti-phospho-H2A.X (Ser139) y Anti-53BP1. Observamos que aquellas células con *ARID2* silenciado tenían un mayor número focos de reparación del daño celular frente a las células control. Además, vimos que el tiempo que tardaban en resolver los focos también era mucho mayor en las células con *ARID2* silenciado. Todos esto nos indicaría una posible implicación de *ARID2* en la reparación del ADN. Además, observamos que ARID2 colocaliza en los focos de reparación del ADN, detectado con dos anticuerpos distintos.

Por último, para identificar otros posibles mecanismos de acción de *ARID2*, llevamos a cabo experimentos de RNA-Seq en la línea celular estable A549 que generamos previamente, en donde fuimos capaces de identificar diferencias de expresión de genes que podrían contribuir al fenotipo tumoral. Así, entre los genes reprimidos encontramos algunos implicados en la adhesión celular como *CDH6*, *NPNT*, *CNTNAP2*, *FAT3*, *FN1* y *VCAN* que podrían estar asociados con el aumento en la capacidad de migración e invasión. También observamos que genes descritos por tener un papel de supresores tumorales como *RPS6K2*, *TNFSF10*, *TP63*, *ISM1* y *LDLRAD4* también están reprimidos tras la inactivación de *ARID2*. Por otro lado, genes descritos por su actividad pro-tumoral y anti-oncogénica como *HOXB1*, *BCL2A1* y *RCVRN* se sobreexpresan en ausencia de *ARID2*. Además, identificamos que *GADD45A*, que pertenece a la ruta de detección y reparación del ADN, esta sobreexpresado al silenciar *ARID2* lo que podría ser el resultado de la inestabilidad genética inducida por la deficiencia en *ARID2*. Todas estas observaciones se validaron mediante qRT-PCR en diferentes líneas celulares estables.

Por último, debido a las observaciones antes descritas y el posible rol de ARID2 en la detección y/o del daño del ADN, hipotetizamos que las células deficientes en ARID2 podrían mostrar una sensibilidad especial al tratamiento con agentes quimioterápicos utilizados actualmente para el tratamiento del cáncer de pulmón y que se basan en la producción de daños en el ADN. Para comprobar esta hipótesis, llevamos a cabo experimentos de resistencia a cisplatino y etopósido y vimos que las células con ARID2 silenciado eran más sensibles a los agentes inductores de daño en el ADN. Todos esto nos hace pensar que ARID2 no solo es un supresor tumoral, sino que también podría servir como un marcador de pronóstico y para ayudar a estratificar a los pacientes de cáncer de pulmón por su capacidad de predicción de respuesta a terapias específicas.

4. CONCLUSIONES

1. **Hemos identificado 4900 mutaciones somáticas en la región codificante de los 250 genes secuenciados en 732 muestras de tumores.**

2. **Hemos identificado un total de 170 mutaciones somáticas en el ADN mitocondrial en nuestra colección de muestras**

3. **El perfil de sustitución y el sesgo de hebra observado en las mutaciones del ADNmt sugieren la existencia de nuevos mecanismos de mutagénesis y reparación del ADN mitocondrial no descritos hasta ahora.**

4. **El 60% de los tumores secuenciados tienen al menos una mutación no sinónima en alguno de los complejos remodeladores de la cromatina. El complejo SWI/SNF muestra un enriquecimiento significativo en genes *driver* lo que podría indicar un papel más importante de este complejo en el desarrollo tumoral.**

5. **Las mutaciones en los genes del complejo SWI-SNF muestran exclusividad entre ellos y también con genes frecuentemente mutados como *KRAS* o *TP53* sugiriendo funciones parcialmente redundantes.**

6. **Hemos encontrado diversas evidencias que demuestran que *ARID2* es un gen supresor de tumores en cáncer de pulmón con una frecuencia mutacional del 15% en las muestras analizadas y asociado a un peor pronóstico.**

7. **El silenciamiento de *ARID2* aumenta la proliferación, la migración, la invasión y las capacidades metastásicas en líneas celulares tanto *in vitro* como *in vivo*.**

8. **La deficiencia de *ARID2* se asocia con alteraciones en la expresión génica que producen un programa transcripcional pro-oncogénico en las células.**

9. ***ARID2* juega un papel importante en la reparación del ADN y su deficiencia está asociada con un aumento en el daño del ADN y un retraso en su reparación.**

10. **El silenciamiento de *ARID2* aumenta la sensibilidad de las líneas celulares a agentes que dañan el ADN como el cisplatino o el etopósido, y ofrecen una oportunidad de usar esta deficiencia para mejorar el tratamiento de los pacientes con cáncer de pulmón.**

APPENDICES

Appendix 1. Papers published during this thesis

Appendix 2. Manuscripts resulted from this thesis

APPENDIX 1

PAPERS PUBLISHED DURING THIS THESIS

Leukemia (2013) 27, 1820–1825

www.nature.com/leu

OPEN

ORIGINAL ARTICLE

Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture

N Conte[1,2,8], I Varela[3,8], C Grove[1], N Manes[1], K Yusa[1], T Moreno[3], A Segonds-Pichon[4], A Bench[5], E Gudgin[5], B Herman[6], N Bolli[1,5], P Ellis[1], D Haddad[1], P Costeas[7], R Rad[1], M Scott[5], B Huntly[5], A Bradley[1] and GS Vassiliou[1]

Advances in sequencing technologies are giving unprecedented insights into the spectrum of somatic mutations underlying acute myeloid leukaemia with a normal karyotype (AML–NK). It is clear that the prognosis of individual patients is strongly influenced by the combination of mutations in their leukaemia and that many leukaemias are composed of multiple subclones, with differential susceptibilities to treatment. Here, we describe a method, employing targeted capture coupled with next-generation sequencing and tailored bioinformatic analysis, for the simultaneous study of 24 genes recurrently mutated in AML–NK. Mutational analysis was performed using open source software and an in-house script (Mutation Identification and Analysis Software), which identified dominant clone mutations with 100% specificity. In each of seven cases of AML–NK studied, we identified and verified mutations in 2–4 genes in the main leukaemic clone. Additionally, high sequencing depth enabled us to identify putative subclonal mutations and detect leukaemia-specific mutations in DNA from remission marrow. Finally, we used normalised read depths to detect copy number changes and identified and subsequently verified a tandem duplication of exons 2–9 of *MLL* and at least one deletion involving *PTEN*. This methodology reliably detects sequence and copy number mutations, and can thus greatly facilitate the classification, clinical research, diagnosis and management of AML–NK.

Leukemia (2013) **27,** 1820–1825; doi:10.1038/leu.2013.117

Keywords: acute myeloid leukaemia; diagnosis; classification; targeted capture; next generation sequencing; minimal residual disease; MIDAS

INTRODUCTION

Advances in DNA sequencing technologies are revolutionising our understanding of the genetic basis of cancer.[1] One of the first cancers studied by whole-genome sequencing was acute myeloid leukaemia with a normal karyotype (AML–NK),[2,3] a disease whose molecular aetiology was, until recently, poorly understood. As a result, we now know of more than 10 genes mutated in >5% of cases of AML–NK and of several others mutated less often.[4–6] Additionally, it has become clear that mutations other than those affecting *FLT3*,[7] *NPM1*[8] and *CEBPA*[9] have a significant impact on prognosis and can help stratify anti-AML therapy for individual patients.[4] In this light, many are calling for a shift towards a classification system for AML–NK based primarily on mutational profiling.[4]

Currently, many diagnostic laboratories routinely screen for mutations in *NPM1* and *FLT3*, both of which show clustering of somatic mutations in 1–3 exons. However, mutational screening for genes such as *CEBPA* and *TET2*, which do not exhibit mutation clustering, is only employed in specialist laboratories. Furthermore, with the identification of an increasing number of mutant genes in AML, detailed molecular genotyping can no longer be practicably performed using conventional molecular methods such as capillary sequencing or melt curve analyses. Moreover, modern sequencing technologies have demonstrated that many cases of AML are composed of several related subclones, arising through the acquisition of different somatic mutations during clonal evolution from a single-ancestral cell.[6,10] These clones are often invisible to conventional diagnostic methods, yet they commonly represent a significant, if not the main, clone at the time of leukaemia relapse.[10] As relapse is the main vehicle for the poor prognosis of AML, the detection of clones carrying adverse mutations at the time of diagnosis can help identify and stratify high-risk patients.

Given the above, a full molecular diagnostic evaluation of AML requires the identification of all mutations with prognostic or therapeutic significance in the main clone, as well as in subclones when these are present. Here we successfully employ targeted DNA capture with cRNA baits followed by deep sequencing and tailored informatics to simultaneously study 24 genes known to be recurrently mutated in AML–NK and 10 control genes.

MATERIALS AND METHODS

Leukaemic DNA samples

DNA samples from total bone marrow cells, excess to diagnosis, were obtained after informed consent within our ethics-approved study (07/MRE05/44) from seven patients with AML–NK. Remission samples were obtained from two of these patients and a relapse sample from one.

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [2]EMBL-European Bioinformatics Institute, Cambridge, UK; [3]Instituto de Biomedicina y Biotecnología de Cantabria, University of Cantabria, Santander, Spain; [4]Bioinformatic Department, Babraham Institute, Cambridge, UK; [5]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge, UK; [6]Agilent Technologies, IQ Winnersh, Reading, UK and [7]Centre for the Study of Haematological Malignacies, Nicosia, Cyprus. Correspondence: Dr GS Vassiliou, Haematological Cancer Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridgeshire CB10 1SA, UK.
E-mail: gsv20@sanger.ac.uk
[8]These authors contributed equally to this work.
Received 21 October 2012; revised 17 March 2013; accepted 10 April 2013; accepted article preview online 18 April 2013; advance online publication, 24 May 2013

Bait design

We designed a set of Sure Select cRNA biotinylated oligonucleotide baits (Agilent Technologies, Palo Alto, CA, USA) to capture all exons from a set of 24 genes known to be recurrently mutated in AML and 10 control genes, some known to be mutated in solid tumours (Table 1). The custom bait library was designed using eArray software (Agilent Technologies). The exons of the 34 genes were downloaded from Biomart (http://www.ensembl.org/biomart/martview/) and used to create 120 bp baits starting every 24 bp. The masking option used was 'RepeatMasker' and the software allowed baits to overlap by a maximum of 20 bp with repeat masked regions. The centred design strategy was used, which ensured that the tiling level was maintained and baits were not 'squeezed' in the specified interval. As a result, the input region could be expanded by 20 bp at each end. The concentration of individual baits was adjusted manually depending on the target nucleotide composition to optimise DNA capture GC-rich regions ($n=1579$) had 2x and orphan regions ($n=649$), defined as those covered by a single bait, 5x more bait molecules per locus than standard regions ($n=5997$). The total target region size was 24 2051 nucleotides and the library design is available under the unique ELID reference: 0324251.

DNA target selection by 'pull-down'

DNA fragmentation, library preparation and solution phase hybrid capture were performed according to manufacturer's instructions (Agilent Technologies) and modified from previously published protocols.[11]

Sequencing and mapping

We sequenced 10 samples on a single multiplexed lane on an Illumina HiSeq 2000 and aligned the resulting reads to the hg19 reference genome with BWA (Burrows–Wheeler Alignment; http://bio-bwa.sourceforge.net/bwa.shtml).[12]

Table 1. Genes analysed by targeted capture

Gene ID	*Chromosome*	*Position (Mb)*
AML genes		
NRAS	1	115.2
DNMT3A	2	25.5
SF3B1	2	198.3
IDH1	2	209.1
KIT	4	55.5
TET2	4	106.1
CSF1R	5	149.4
NPM1	5	170.8
EZH2	7	148.5
JAK2	9	5.0
PTEN	10	89.6
WT1	11	32.4
MLL	11	118.3
CBL	11	119.1
KRAS	12	25.4
PTPN11	12	112.9
FLT3	13	28.6
IDH2	15	90.6
TP53	17	7.6
NF1	17	29.4
CEBPA	19	33.8
ASXL1	20	30.9
RUNX1	21	36.2
KDM6A	X	44.7
Control genes		
UGT1A1	2	234.7
PIK3CA	3	178.9
IKZF1	7	50.3
EGFR	7	55.1
BRAF	7	140.4
XRCC2	7	152.3
PAX5	9	36.8
TLR4	9	120.5
CYP2D6	22	42.5
HPRT1	X	133.6

Abbreviation: AML, acute myeloid leukaemia.

Coverage and statistical analysis

Coverage histograms and tables describing the coverage distribution for our set of targeted bases were produced using TEQC, 'Target Enrichment Quality Control' Bioconductor package.[13] To validate the ability of our assay to identify copy number changes, we used read numbers of two X-linked genes (*HPRT* and *KDM6A*). First, we generated a list of non-redundant 'amalgamated exons', each representing all overlapping annotated exons. Read count normalisation was done using open-source software and bespoke R scripts: for each sample, read counts per position were calculated using Bedtools 2.12.0 (http://code.google.com/p/bedtools),[14] then normalised read counts were calculated by averaging the exon-specific read counts and dividing by the total number of mapped reads for that sample. As DNA quality can affect capture efficiency and thus read counts, we first wanted to ensure all 10 samples gave comparable standardised read depths for the majority of target regions. For this, we looked at the average read count for each patient at each gene. Sample P5 was an outlier for 23 of the 34 genes and was removed from copy number calculations. All other samples were outliers for three genes or less (Supplementary Figure S1). To identify copy number variation at individual exons, we calculated the coefficient of variability for each exon for the nine patients. We then used the Tukey boxplot approach to identify the outlier exons ($>$ upper quartile $+$ 1.5*IQR, interquartile range). Data from genes with an increased coefficient of variability at more than one exon were examined manually. The mixed-lineage leukaemia (*MLL*) deletion was also detected by analysis with ExomeCopy (http://www.bioconductor.org/packages/2.11/bioc/html/exomeCopy.html).

Mutation calling

Alignment and post-processing. Fastq files were aligned against the human genome (hg19 version) using BWA algorithm (v 0.5.9). Afterwards SAMTOOLS (0.1.18) view, sort, index and fixmate algorithms were used to generate, sort, index and fix co-ordinates of the generated bam files. PICARD (v1.61) java libraries were used to mark PCR duplicates and finally GATK (v. 1.4.20) tools were used to perform local realignment around indels. All these steps were automated using a single in-house written script available upon request.

Variant calling. SAMTOOLS pileup command was used to generate pileup files from the generated bam files (version 0.1.8) (http://samtools.sourceforge.net/).[15] A flexible in-house Perl script (MIDAS, Mutation Identification and Analysis Software; available upon request) was created to parse the pileup file and to take into account in each position only those reads with a sequence quality higher than 25 and a mapping quality higher than 15, and consider only those positions that had a coverage of at least 10 both in the tumour and in the control sample (unless otherwise stated, P1CR was used as control for all comparisons). On those positions, and taking into account the high coverage obtained in this experiment, we reported the possible existence of a substitution whenever there was at least 20 independent reads reporting a different base vs the reference genome in the tumour sample and less than 5% of the reads reporting the same variant in the control sample. We also discarded those positions with at least one-third of this evidence reporting a third allele, as we consider that those regions would probably represent difficult sequences for the aligner and would likely produce false positives. We considered variants present in $>20\%$ of reads as those representing the main/dominant leukaemic clone. In the case of indels (small insertions and deletions), we considered positive those regions with at least 10 independent reads reporting the same indel in the tumour sample and with less than 5 reads in the control sample, and with at least 10 times more reads reporting the indel in the tumour vs the control sample. Similarly to what we did with substitutions, those regions with an evidence of a second indel higher than 40% of the evidence for the primary indel were discarded. Our workflow is shown in Figure 1. Of note, MIDAS allows adjustment of tolerance thresholds to suit the type of control sample used (for example, they can be increased to facilitate the use of a remission sample as a control, which may harbour residual low-level mutant reads).

Comparison with other software/algorithms. The performance of our software was checked using independent variant calling algorithms. In particular, we run SomaticSniper (v. 1.0.2; http://gmt.genome.wustl.edu/

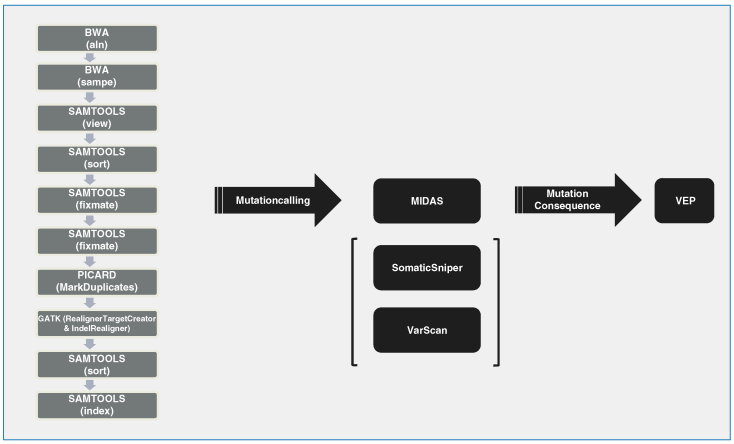


Figure 1. Workflow diagram for data analysis and mutation calling. After initial parsing of sequencing data through a series of open source software tools, mutation calling is performed by our in-house Perl script (MIDAS). Mutational consequences are then determined by Variant Effect Predictor, Ensembl. For the purposes of comparing MIDAS with other callers, SomaticSniper and VarScan were used instead.

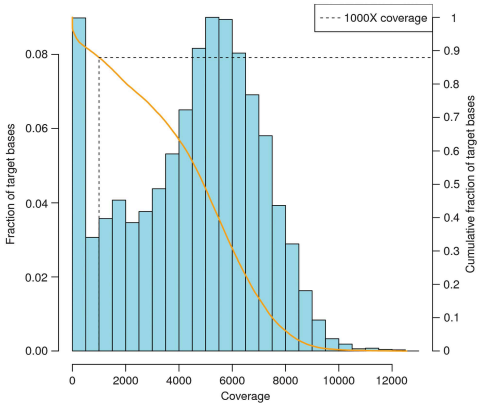


Figure 2. Distribution of the depth of sequencing coverage of the target genes. Representative data from sample P1 showing the fraction of bases covered at incremental depth windows (blue bars and left hand *y* axis) and the cumulative fraction of bases covered at or above the specified coverage (orange line and right hand *y* axis). This shows that ~88% of bases were covered at by at least 1000x sequencing reads.

somatic-sniper/current/)[16] on the bam files using the default parameters and VarScan (v2.3, http://varscan.sourceforge.net/)[17] on the pileup files using both the default mode and a high-sensitivity mode setting a minimum variant frequency of 0.01, a normal purity of 0.95 and a tumour purity of 0.20. In order to be able to compare the results with the calls made by our software, the raw data generated by the other callers were afterwards filtered according to the frequency and ratios criteria specified in the above paragraph.

The predicted protein consequences of variations were derived using Variant Effect Predictor from Ensembl, http://www.ensembl.org/info/docs/variation/vep/index.html.

Validation of mutations and copy number changes identified by next-generation sequencing

All dominant clone mutations were confirmed using PCR and capillary sequencing. PCR was performed with Platinum Taq Polymerase (Invitrogen Corporation, Carlsbad, CA, USA) for 35 cycles at 56 °C annealing and 72 °C extension for 30 s. To amplify across the breakpoint of the MLL-partial tandem duplication, we used LongAmp 2x Taq mastermix (New England Biolabs, Ipswich, MA, USA) for 35 cycles at 57 °C annealing and 65 °C extension for 3 min. PCR for detection of FLT3-internal tandem duplication was performed as described previously.[18] Mutant reads were visualised using IGV (Integrative Genomics Viewer; http://www.broadinstitute.org/igv/bam). To verify the two *PTEN* deletions, we used six known single-nucleotide polymorphisms within introns of the *PTEN* gene. We amplified these by PCR, followed by second-round PCR with barcoded Illumina adapter primers and sequencing on a MiSeq sequencer. We used these results to look for evidence of copy number change for one of the two alleles compared with a reference normal (P6 vs ctrl) or a paired remission sample (P2 vs P2CR). All primer sequences are given in Supplementary Table S1.

RESULTS

Analysis of our sequencing data showed a mean coverage depth of 5136× per nucleotide position within the target region (Figure 2). The 10× and 100× coverage were 96.4% and 94.8%, respectively, for the desired target region (that is, all exons of 34 genes) (Supplementary Table S2), with most of the remaining 3.6–5.2% representing repetitive regions for which baits could not be designed. With regards to substitutions and indels among the seven AMLs studied, our mutation caller, MIDAS, identified 20 exonic and one intronic mutations in the main leukaemic clone (2–4 mutations per AML, Table 2). The same 20 exonic mutations were identified by the VarScan platform and all were successfully validated using Sanger sequencing (Supplementary Figure S2), giving both MIDAS and VarScan 100% specificity for this data set. SomaticSniper, which was designed for the identification of substitutions but not indels, performed slightly less well (Supplementary Table S3). Of the 20 exonic mutations, 11 were single-base non-synonymous substitutions at known sites (9 missense and 2 nonsense) and 9 were small indels (8 associated with premature termination and 1 with a single amino-acid insertion).

Table 2. Diagnostic information and mutations in the dominant leukaemic clone of patient samples

Sample ID	*Age*	*Sex*	*Sample type*	*FAB*	*WCC ($x10^9$/l)*	*BM blasts %*	*CD34+ %*	*CD13+ %*	*CD33+ %*	*CD7+ %*	*CD56+ %*	*Karyotype*	*Mutations in dominant AML clone*		
P1	45	F	P	M5a	140	90	3	56	75	26	0	46XX	*NPM1 L287fs**	*DNMT3A R882C*	
P2	71	M	P	M4	111	85	0	72	92	0	80	46XY	*ASXL1 G642fs**	*TET2 L1119**	*NF1 intron 2*
P3	73	M	P	M2	108	95	34	44	75	0	0	46XY	*CEBPA A111fs**	*KRAS K117N*	*NRAS G12D*
P4	43	F	P	M1	24.4	95	0	12	81	0	2	46XX	*NPM1 L287fs**	*WT1 R145fs**	
P5	47	M	P	M5a	38	80	0	74	33	45	0	46XY	*NPM1 L287fs**	*IDH2 R140Q*	
P6	80	M	P	M1	116	95	0	53	99	6	79	46XY	*ASXL1 C594**	*IDH1 R132H*	
P7	59	F	P	M4	2.6	60	85	80	7	0	0	46XX	*DNMT3A G590fs**	*TET2 P612fs**	
P1CR	45	F	CR	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a		*IDH2 R172K*	
P2CR	71	M	CR	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a			
P4Rel	45	F	Rel1	M1	3	65%	nd	nd	nd	nd	nd	46XX	*NPM1c (TCTG)*	*IDH2 R140Q*	

Abbreviations: BM, bone marrow; CR, complete remission; F, female; M, male; n/a, not applicable; nd, not determined.

In order to determine whether the read depth for our target regions correlated with DNA copy number, we compared standardised read numbers for two X-linked genes, *HPRT* and *KDM6A*, between male and female cases, and contrasted this with the same ratio for the remaining (autosomal) genes. This demonstrated that female cases displayed approximately twice the number of normalised reads of male cases for the two X-linked genes (F:M ratios of 2.0 for *KDM6A* and 1.91 for *HPRT*), signifying that read numbers approximately reflect copy number in the starting DNA. In keeping with this, male and female cases gave similar normalised read numbers (M:F ratios close to 1) for the 32 autosomal genes (Supplementary Figure S3). Samples that deviated from this ratio were later found to harbour copy number variation at the relevant gene locus in one or more samples (for example, *CYP2D6* and *PTEN*). Furthermore, the quantitative nature of the data was evident at the level of individual exons and not just whole genes (for example, Supplementary Figure S5), demonstrating that the data were quantitative even at the level of small independently captured loci.

Given the above, we went on to look for copy number aberrations involving the target exons using the Tukey Box-plot method. The only autosomal gene loci exhibiting a significantly increased coefficient of variability at multiple exons were *CYP2D6*, *MLL* and *PTEN* (Supplementary Figure S4). *CYP2D6* is known to exhibit copy number variation and per exon read numbers were in keeping with one individual (P3) having a lower *CYP2D6* copy number than the others (Supplementary Figure S5b). In the case of *PTEN*, two samples (P2 and P6) had lower read numbers (Supplementary Figure S5c), suggesting that these two cases of AML may harbour deletions involving *PTEN*. To confirm this using our limited material, we looked at differential allelic read counts for six intronic single-nucleotide polymorphisms within the *PTEN* locus, using PCR amplification followed by sequencing on a MiSeq sequencer. Our results confirm copy number change at the *PTEN* locus for P2 by demonstrating a preferential reduction in read counts from one allele of two independent informative single-nucleotide polymorphisms when compared with the matched remission sample (P2CR) (Supplementary Table S4). In the case of P6, only one single-nucleotide polymorphisms was informative and although this was suggestive of copy number loss, we cannot be completely confident this is the case in the absence of a matched normal sample. In the case of *MLL*, one sample (P6) showed an increased number of normalised reads for exons 2–9 only, suggesting the presence of a partial tandem duplication (Figures 3a and b and Supplementary Figure S5d). This was also identified by analysis using the ExomeCopy package[19] (Supplementary Figure S6). The presence of a partial tandem duplication was confirmed using PCR primers to amplify the region spanning the junction (Figure 2c).

We went on to analyse our data to identify single-nucleotide substitutions uniquely present in putative leukaemic subclones representing as few as 1% of cells. We identified putative subclonal mutations representing 3–20% of reads in four leukaemic samples: (i) a *FLT3* internal tandem duplication in sample P3, which was flagged as a series of indels and substitutions and confirmed by PCR (we went on to test all seven AML samples for *FLT3*-internal tandem duplication and only sample P3 was positive—data not shown), (ii) *NRAS*-G12S and *PTPN11*-Q506P mutations in sample P5. The latter two mutations occurred in 4.4% and 4.1% of reads, respectively, in keeping with possible co-occurrence in the same subclone, (iii) *FLT3*-N676K in sample P4 and (iv) *TP53*-G374fs*8 in sample P4Rel (Supplementary Table S5). Finally, we analysed the two paired diagnosis-remission samples (P1 vs P1CR and P2 vs P2CR) to look for evidence of residual mutant reads in each remission sample. Both remission samples were in morphological complete remission, but sample P1CR was taken after four courses and sample P2CR after one course of chemotherapy. At a level of sensitivity of at least 0.1%,

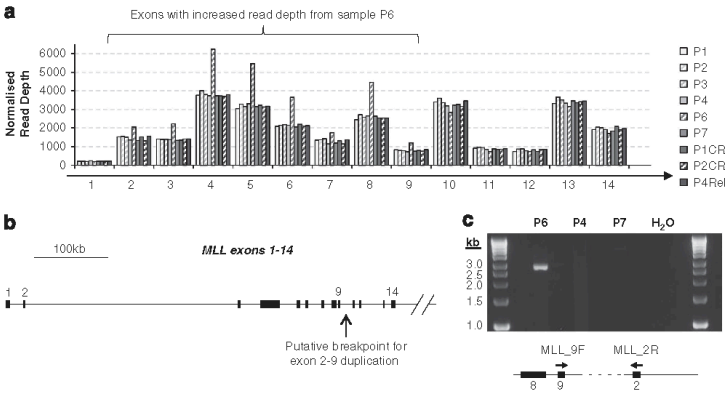


Figure 3. Identification of *MLL* partial tandem duplication (PTD) using sequencing read depth. Normalised per exon sequencing read depths for the first 14 exons of *MLL* show increased depth for exons 2–9 from sample P6 (**a**). This suggested the presence of an exon 2 to exon 9 PTD with a breakpoint in intron 9 (**b**). PCR amplification across the putative breakpoint using an exon 9 forward (MLL_9F) and an exon 2 reverse (MLL_2R) primer confirms the presence of the PTD in this AML sample (**c**).

we found no mutant reads in sample P1CR, while sample P2CR gave residual mutant reads representing 1.5–2.4% of total reads for all three mutations identified at diagnosis (Supplementary Table S6).

DISCUSSION

Advances in sequencing technologies are revolutionising cancer research with somatic mutations underlying most major cancers being avidly identified and characterised in large numbers of cases. Concurrently, clinical and functional studies are defining the diagnostic/prognostic significance of mutations and determining their molecular effects in order to device targeted therapeutic strategies. AML is at the forefront of such progress, and as a result a significant body of information has already been gathered about this leukaemia that can be used to guide clinical practice.

To date, most diagnostic laboratories use allele-specific technologies to identify mutations in genes such as *NPM1* and *FLT3*, which have validated prognostic and therapeutic significance.[7,20,21] Additionally, newer technologies have been shown to reliably identify leukaemia-associated mutations in larger numbers of amplicons using pyrosequencing.[22] However, the increasing number of clinically relevant genes found mutated in AML make conventional amplicon-based approaches impractical, particularly as many such genes can harbour mutations in multiple different locations and exons.[23,24] Additionally, the clear demonstration that many AMLs are composed of multiple subclones that can be differentially susceptible to existing therapies[10] suggests that accurate therapeutic stratification of patients would benefit from the identification of such clones at first presentation.

We describe a method based on targeted DNA capture with cRNA baits followed by deep sequencing that enables the simultaneous identification of mutations in 24 AML genes, including the 10 most frequently mutated in AML–NK, without recourse to normal constitutional DNA from the same individual. Somatic mutations in the dominant leukaemic clone were identified in all cases studied using sequence alignment/configuration with open source software followed by mutation calling using our in-house mutation caller MIDAS (Figure 2). The same mutations were identified by the mutation caller VarScan[17] and all mutations so identified were validated using capillary sequencing, demonstrating 100% specificity for both callers for our data set. By contrast, no novel polymorphisms or mutations were identified in 10 control genes known to be mutated in solid tumours or leukaemias other than AML.

The sequencing depth reached in this study also enabled us to identify putative mutations present in subclones at the time of diagnosis. It is already clear that, compared with the main clone, such subclones may be differentially sensitive to chemotherapy and can expand to become the dominant clone at the time of disease relapse,[10] making their identification at the time of diagnosis important. Nevertheless, at this stage, such subclonal mutations need to be validated using independent methodologies as it remains possible that they represent sequencing or other forms of error.

Additionally, after confirming that our data behaved in a quantitative manner with respect to input DNA copy number in 9 of 10 DNAs studied, we went on to identify copy number variants in leukaemic samples, including an instance of *MLL*-partial tandem duplication and two instances of probable loss of *PTEN*, one of which we were able to validate. In analysing these data it became clear that the lack or copy number information from neighbouring genomic regions made analysis more difficult, and we recommend that future studies of this kind endeavour to capture several features around regions of possible copy number loss to enhance both the power and the reliability of analyses. Finally, we were able to demonstrate evidence of minimal residual disease in a bone marrow DNA sample in morphological complete remission by mining reads from the specific mutations in the remission sample. Quantification of minimal residual disease after induction chemotherapy may have prognostic implications in a heterogeneous disease such as AML–NK and could be employed in interventional studies to determine its significance.

We describe a molecular diagnostic method that enables extensive molecular characterisation of AML–NK at diagnosis and can facilitate clinical management of patients as well as clinical research into this disease. The approach is powerful, reliable and can be introduced into routine clinical practice in order to enhance our ability to identify patients at high risk of relapse as well as those that would benefit from molecularly directed therapies and can also be adapted for minimal residual disease monitoring. The sequencing methodology is modular and target regions can be increased to include any newly discovered gene mutations without significant changes to laboratory

protocols and with only marginal increases in costs. Additionally, we provide a clear analytical workflow employing MIDAS, a novel mutation calling algorithm available on request, which correctly identified 20/20 exonic mutations present in >20% of reads. The blueprint presented here can be used to study other haematological or solid tumours, or groups of tumours with overlapping mutational spectra.

CONFLICT OF INTEREST

Dr Bram Herman is an employee of Agilent Technologies, manufacturer of Sure Select cRNA baits used for targeted capture. His contribution to this work was limited to the bespoke design of our bait set to maximise target DNA capture and the provision of bait sequence files for bioinformatic analyses.

ACKNOWLEDGEMENTS

We acknowledge the use of the National Institute of Health Research (NIHR) Biomedical Research Centre, University of Cambridge. We thank Drs J Craig and C Crawley of Cambridge University NHS Hospitals trust for allowing us to approach their patients for samples. GV is funded by a Wellcome Trust Senior Fellowship in Clinical Science. Work in GV's laboratory is also funded by Leukaemia Lymphoma Research and the Kay Kendal Leukaemia Fund.

REFERENCES

1 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**: 719–724.
2 Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008; **456**: 66–72.
3 Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 2009; **361**: 1058–1066.
4 Patel JP, Gonen M, Figueroa ME, Fernandez H, Sun Z, Racevskis J *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med* 2012; **366**: 1079–1089.
5 Grossmann V, Tiacci E, Holmes AB, Kohlmann A, Martelli MP, Kern W *et al.* Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* 2011; **118**: 6153–6163.
6 Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* 2012; **150**: 264–278.
7 Gale RE, Green C, Allen C, Mead AJ, Burnett AK, Hills RK *et al.* The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. *Blood* 2008; **111**: 2776–2784.
8 Falini B, Mecucci C, Tiacci E, Alcalay M, Rosati R, Pasqualucci L *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med* 2005; **352**: 254–266.
9 Preudhomme C, Sagot C, Boissel N, Cayuela JM, Tigaud I, de Botton S *et al.* Favorable prognostic significance of CEBPA mutations in patients with de novo acute myeloid leukemia: a study from the Acute Leukemia French Association (ALFA). *Blood* 2002; **100**: 2717–2723.
10 Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012; **481**: 506–510.
11 Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009; **27**: 182–189.
12 Bullinger L, Armstrong SA. HELP for AML: methylation profiling opens new avenues. *Cancer Cell* 2010; **17**: 1–3.
13 Hummel M, Bonnin S, Lowy E, Roma G. TEQC: an R package for quality control in target capture experiments. *Bioinformatics* 2011; **27**: 1316–1317.
14 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**: 841–842.
15 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.
16 Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 2012; **28**: 311–317.
17 Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012; **22**: 568–576.
18 Kiyoi H, Naoe T, Nakano Y, Yokota S, Minami S, Miyawaki S *et al.* Prognostic implication of FLT3 and N-RAS gene mutations in acute myeloid leukemia. *Blood* 1999; **93**: 3074–3080.
19 Love MI, Mysickova A, Sun R, Kalscheuer V, Vingron M, Haas SA. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol* 2011; **10.**
20 Smith BD, Levis M, Beran M, Giles F, Kantarjian H, Berg K *et al.* Single-agent CEP-701, a novel FLT3 inhibitor, shows biologic and clinical activity in patients with relapsed or refractory acute myeloid leukemia. *Blood* 2004; **103**: 3669–3676.
21 Sanz M, Burnett A, Lo-Coco F, Lowenberg B. FLT3 inhibition as a targeted therapy for acute myeloid leukemia. *Curr Opin Oncol* 2009; **21**: 594–600.
22 Grossmann V, Kohlmann A, Eder C, Haferlach C, Kern W, Cross NC *et al.* Molecular profiling of chronic myelomonocytic leukemia reveals diverse mutations in >80% of patients with TET2 and EZH2 being of high prognostic relevance. *Leukemia* 2011; **25**: 877–879.
23 Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE *et al.* DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* 2010; **363**: 2424–2433.
24 Delhommeau F, Dupont S, Della Valle V, James C, Trannoy S, Masse A *et al.* Mutation in TET2 in myeloid cancers. *N Engl J Med* 2009; **360**: 2289–2301.

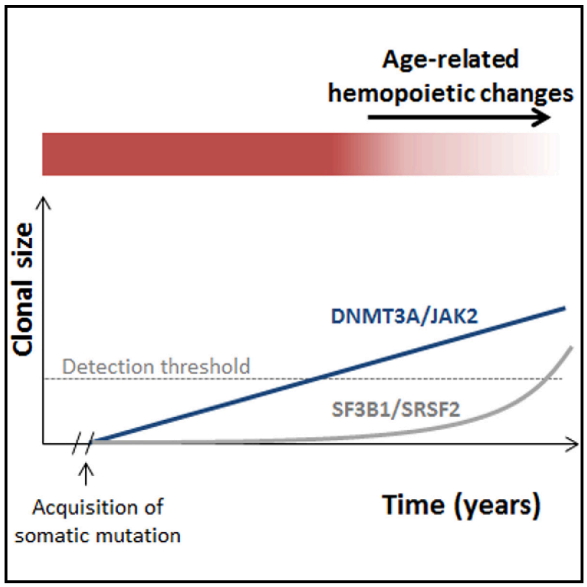
Supplementary Information accompanies this paper on the Leukaemia website (http://www.nature.com/leu)

Cell Reports

Report

Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis

Graphical Abstract



Authors

Thomas McKerrell, Naomi Park, ..., Ignacio Varela, George S. Vassiliou

Correspondence

gsv20@sanger.ac.uk

In Brief

McKerrell et al. employ ultra-deep sequencing to show that age-related clonal hemopoiesis is much more common than previously realized. They find that clonal hemopoiesis, driven by mutations in spliceosome genes *SF3B1* and *SRSF2*, was noted exclusively in individuals aged 70 years or older and that *NPM1* mutations are not seen in association with this phenomenon, endorsing their close association with leukemogenesis.

Highlights

- Clonal hemopoiesis is an almost inevitable consequence of aging in humans
- Spliceosome gene mutations drove clonal hemopoiesis only in persons aged ≥70 years
- NPM1 mutations behave as gatekeepers for leukemogenesis

McKerrell et al., 2015, Cell Reports *10*, 1239–1245
March 3, 2015 ©2015 The Authors
http://dx.doi.org/10.1016/j.celrep.2015.02.005

Cell Reports
Report

Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis

Thomas McKerrell,[1,13] Naomi Park,[2,13] Thaidy Moreno,[3] Carolyn S. Grove,[1] Hannes Ponstingl,[1] Jonathan Stephens,[4,5] Understanding Society Scientific Group,[6] Charles Crawley,[7] Jenny Craig,[7] Mike A. Scott,[7] Clare Hodkinson,[4,8] Joanna Baxter,[4,8] Roland Rad,[9,10] Duncan R. Forsyth,[11] Michael A. Quail,[2] Eleftheria Zeggini,[12] Willem Ouwehand,[4,5,12] Ignacio Varela,[3] and George S. Vassiliou[1,4,7,*]

[1]Haematological Cancer Genetics, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK
[2]Sequencing Research Group, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK
[3]Instituto de Biomedicina y Biotecnología de Cantabria (CSIC-UC-Sodercan), Departamento de Biología Molecular, Universidad de Cantabria, 39011 Santander, Spain
[4]Department of Haematology, Cambridge Biomedical Campus, University of Cambridge, Cambridge CB2 0XY, UK
[5]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK
[6]Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK
[7]Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge CB2 0QQ, UK
[8]Cambridge Blood and Stem Cell Biobank, Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK
[9]Department of Medicine II, Klinikum Rechts der Isar, Technische Universität München, 81675 München, Germany
[10]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[11]Department of Medicine for the Elderly, Cambridge University Hospitals NHS Trust, Cambridge CB2 0QQ, UK
[12]Human Genetics, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK
[13]Co-first author
*Correspondence: gsv20@sanger.ac.uk
http://dx.doi.org/10.1016/j.celrep.2015.02.005

SUMMARY

Clonal hemopoiesis driven by leukemia-associated gene mutations can occur without evidence of a blood disorder. To investigate this phenomenon, we interrogated 15 mutation hot spots in blood DNA from 4,219 individuals using ultra-deep sequencing. Using only the hot spots studied, we identified clonal hemopoiesis in 0.8% of individuals under 60, rising to 19.5% of those ≥90 years, thus predicting that clonal hemopoiesis is much more prevalent than previously realized. *DNMT3A*-R882 mutations were most common and, although their prevalence increased with age, were found in individuals as young as 25 years. By contrast, mutations affecting spliceosome genes *SF3B1* and *SRSF2*, closely associated with the myelodysplastic syndromes, were identified only in those aged >70 years, with several individuals harboring more than one such mutation. This indicates that spliceosome gene mutations drive clonal expansion under selection pressures particular to the aging hemopoietic system and explains the high incidence of clonal disorders associated with these mutations in advanced old age.

INTRODUCTION

Cancers develop through the combined action of multiple mutations that are acquired over time (Nowell, 1976). This paradigm is well established in hematological malignancies, whose clonal history can be traced back for several years or even decades (Ford et al., 1998; Kyle et al., 2002). It is also clear from studies of paired diagnostic-relapsed leukemia samples that recurrent disease can harbor some, but not always all, mutations present at diagnosis, providing evidence for the presence of a clone of ancestral pre-leukemic stem cells that escape therapy and give rise to relapse through the acquisition of new mutations (Ding et al., 2012; Krönke et al., 2013). Studies of such phenomena have defined a hierarchical structure among particular leukemia mutations, with some, such as those affecting the gene *DNMT3A*, displaying the characteristics of leukemia-initiating lesions and driving the expansion of hemopoietic cell clones prior to the onset of leukemia (Ding et al., 2012; Shlush et al., 2014).

These observations suggest that individuals without overt features of a hematological disorder may harbor hemopoietic cell clones carrying leukemia-associated mutations. In fact, such mutations, ranging from large chromosomal changes (Jacobs et al., 2012; Laurie et al., 2012) to nucleotide substitutions (Busque et al., 2012), have been found to drive clonal hemopoiesis in some individuals. Recent reanalyses of large exome-sequencing data sets of blood DNA showed that clonal hemopoiesis is more common than previously realized and increases with age to affect up to 11% of those over 80 and 18.4% of those over 90 years (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). The presence of such clones was associated with an increased risk of developing hematological or other cancers and a higher all-cause mortality, probably due to an increased risk of cardiovascular disease (Genovese et al., 2014; Jaiswal et al., 2014).

Table 1. Mutation Hot Spots Interrogated in This Study

Gene	Target Codon
DNMT3A	R882
JAK2	V617
NPM1	L287
SRSF2	P95
SF3B1	K666
SF3B1	K700
IDH1	R132
IDH2	R140
IDH2	R172
KRAS	G12
NRAS	G12
NRAS	Q61
KIT	D816
FLT3	D835
FLT3	N676

Also see Table S1 for detailed information about numbers of samples screened for each mutation.

The important findings of these studies were based on analysis of exome-sequencing data sets that were generated for the study of constitutional genomes, thus trading genome-wide coverage for reduced sensitivity for detecting small subclonal events. We used the different approach of targeted re-sequencing of selected leukemia-associated mutation hot spots in blood DNA from more than 4,000 individuals unselected for blood disorders. In addition to increasing the sensitivity for detecting subclonal mutations, this approach enabled us to prospectively select and study a large number of elderly individuals. Our results show that clonal hemopoiesis is significantly more common than anticipated, give new insights into the distinct age-distribution and biological behavior of clonal hemopoiesis driven by different mutations, and help explain the increased incidence of myelodysplastic syndromes (MDSs) with advancing age.

RESULTS

To investigate the incidence, target genes, and age distribution of age-related clonal hemopoiesis (ARCH), we performed targeted re-sequencing for hot spot mutations at 15 gene loci recurrently mutated in myeloid malignancies (Table 1) using blood DNA from 3,067 blood donors aged 17–70 (Wellcome Trust Case Control Consortium [WTCCC]) and 1,152 unselected individuals aged 60–98 years (United Kingdom Household Longitudinal Study [UKHLS]; see Figure S1 for detailed age distributions). To do this, we developed and validated a robust methodology, employing barcoded multiplex PCR of mutational hot spots followed by next-generation sequencing (MiSeq) and bioinformatic analysis, to extract read counts and allelic fractions for reference and non-reference nucleotides. This reliably detected mutation-associated circulating blood cell clones with a variant allele fraction (VAF) $\geq$ 0.008 (0.8%; see Supplemental Experimental Procedures and Figure S2).

We obtained adequate coverage ($\geq$1,000 reads at all studied hot spots) from 4,067 blood DNA samples and identified mutation-bearing clones in 105 of these. Of note, not all hot spots were studied in all samples and the derived incidence of mutations in our population as a whole was 3.24% (Table S1). However, the incidence rose significantly with age from 0.2% in the 17–29 to 19.5% in the 90–98 years age group (Figure 1A). We found one or more samples with mutations at 9 of the 15 hot spot codons studied, with VAFs varying widely within and between mutation groups (Table 2).

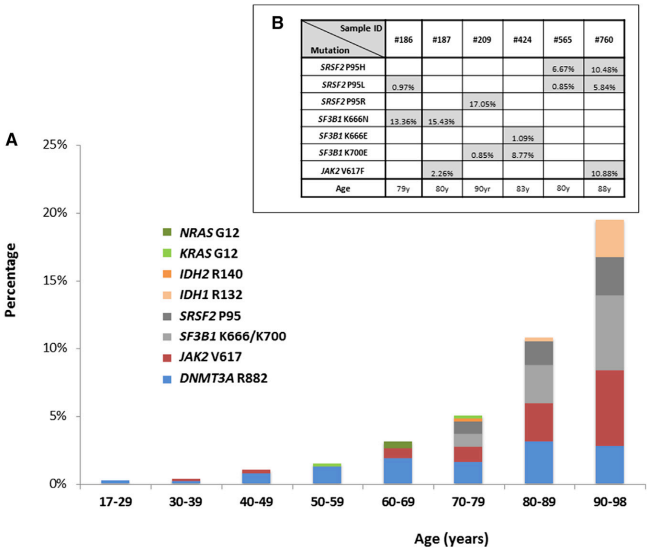
The most-common mutations were those affecting *DNMT3A* R882, whose incidence rose with age from 0.2% (1/489) in the 17–25 to a peak of 3.1% (11/355) in the 80–89 age group. A similar pattern was observed with *JAK2* V617F mutations (Figure 1A). By contrast, spliceosome gene mutations at *SRSF2* P95, *SF3B1* K666, and *SF3B1* K700 were exclusively observed in people aged over 70 years, rising sharply from 1.8% in those aged 70–79 to 8.3% in the 90–98 years age group. Among all samples, we identified only six individuals with more than one mutation; significantly, five of them had two independent spliceosome gene mutations of different VAFs (Figure 1B). Unfortunately, in each of three cases with two mutations at the same or nearby positions, neighboring SNPs were not informative and the variants could not be phased (see Supplemental Experimental Procedures).Occasional mutations in the genes *IDH1*, *IDH2*, *NRAS*, and *KRAS* were also seen. Except for three samples with IDH1/2 mutations, hemoglobin concentrations did not differ significantly between individuals with and without hot spot mutations (Figure S3A). For samples with full blood count results available, *JAK2* V617F mutant cases had a higher platelet count (albeit within the normal range) than "no mutation cases," whereas other results did not differ (Figure S3B). No hot spot mutations were found in the few cord blood (n = 18) and post-transplantation (n = 32) samples studied.

Finally, despite using a very sensitive method and a mutation-calling script written specifically for this purpose, no samples with *NPM1* mutations of VAF $\geq$ 0.008 were identified. In fact, variant reads reporting a canonical *NPM1* mutation (mutation A; TCTG duplication) were detected in only 1 of 4,067 samples at a VAF of 0.0012 (4/3,466 reads).

DISCUSSION

Hematological malignancies develop through the serial acquisition of somatic mutations in a process that can take many years or even decades (Ford et al., 1998; Kyle et al., 2002). Also, it is clear that the presence of hemopoietic cells carrying leukemia-associated mutations is only followed by the onset of hematological malignancies in a minority of cases (Busque et al., 2012; Genovese et al., 2014; Jacobs et al., 2012; Jaiswal et al., 2014; Laurie et al., 2012; Xie et al., 2014). In order to understand the incidence and clonal dynamics of pre-leukemic clonal hemopoiesis, we interrogated 15 leukemia-associated mutation hot spots using a highly sensitive methodology able to detect small clones with mutations.

We show that clonal hemopoiesis is rare in the young but becomes common with advancing age. In particular, we observed that ARCH driven by the mutations studied here doubled in



B

Sample ID / Mutation	#186	#187	#209	#424	#565	#760
SRSF2 P95H					6.67%	10.48%
SRSF2 P95L	0.97%				0.85%	5.84%
SRSF2 P95R			17.05%			
SF3B1 K666N	13.36%	15.43%				
SF3B1 K666E				1.09%		
SF3B1 K700E			0.85%	8.77%		
JAK2 V617F		2.26%				10.88%
Age	79y	80y	90yr	83y	80y	88y

Figure 1. Prevalence and Age Distribution of Hot Spot Mutations Driving Clonal Hemopoiesis

(A) Prevalence of mutations driving clonal hemopoiesis by age.

(B) Samples with more than one mutation, variant allele fraction (VAF) of each mutation present, and age of participant.

Also see Figure S1 for age distribution of all participants.

frequency in successive decades after the age of 50, rising from 1.5% in those aged 50–59 to 19.5% in those aged 90–98 (Figure 1). Of note, 61 of 112 clones identified had a VAF $\leq$ 3% (Table 2), and it is likely that most of these would not have been detected by conventional exome sequencing, which gives lower than 10-fold average coverage compared to the current study (see Table S2 for comparison to such studies), with some recurrently mutated regions giving particularly low coverage (Genovese et al., 2014). Notably, our study did not search for non-hot-spot mutations associated with ARCH such as those affecting genes *TET2* and *ASXL1* or *DNMT3A* codons other than R882 (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). Assuming that the incidence of small clones is similar for such mutations as for the hot spot mutations we studied here, the mean projected true incidence of ARCH driven by leukemia-associated mutations in those older than 90 years is greater than 70% (Figure S4). This makes clonal hemopoiesis an almost inevitable consequence of advanced aging.

Another significant finding of our study is the disparate age distribution of ARCH associated with different mutation types. In particular, we found that, although *DNMT3A* R882 and *JAK2* V617F mutations become more common with age, they were also found in younger individuals. This is in keeping with the increasing cumulative likelihood of their stochastic acquisition with the passage of time. In contrast, spliceosome gene mutations were found exclusively in those aged 70 years or older, replicating the sharp rise beyond this age in the incidence of MDSs driven by these mutations and the fact that, among unselected MDS patients, those with spliceosome mutations are significantly older than those without (Haferlach et al., 2014; Lin et al., 2014; Papaemmanuil et al., 2013; Wu et al., 2012).

Exome-sequencing studies describe a much-lower rate of spliceosome mutations (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014), but this is again likely to reflect their lower sensitivity for detecting small clones, which was a particular limitation at spliceosome mutation hot spots as these were captured/sequenced at lower-than-average depths (Table S2). In our study, 19/33 *SF3B1*- or *SRSF2*-associated clones had a VAF $\leq$ 5%, with 13 of these at VAFs $\leq$ 3% (Table 2), the majority of which would not have been detected by low-coverage sequencing. The identification of ARCH driven by spliceosome gene mutations is in keeping with the fact that these are founding mutations in the clonal evolution of MDS and related hematological malignancies (Cazzola et al., 2013; Haferlach et al., 2014; Papaemmanuil et al., 2013).

We propose that the exclusive identification of spliceosome gene mutations in those aged $\geq$70 years can be explained by differences in the prevailing pressures on clonal selection at different ages, which can in turn explain how different gene mutations can generate detectable clonal expansions at different ages (Figure 2). The alternatives are that spliceosome mutations are associated with slower rates of clonal expansion or that they are detected later because they contribute less to circulating leukocytes. Both of these scenarios are less plausible, given the complete absence of such mutations even at low VAFs in younger age groups. For any somatic mutation imparting a clonal advantage to a stem/progenitor cell and leading to the generation of a steadily expanding clone, one would expect such a clone to be detectable at a smaller size at earlier and a larger size at later time points, as is the case for *DNMT3A* R882 and *JAK2* V617 mutations. Instead, clones (of any size) driven by mutant *SRSF2* and *SF3B1* were observed exclusively in individuals aged 70 years or older, suggesting that these only begin to expand later in life. Furthermore, considerable support for the presence of a different selection milieu comes from the observation that five of six patients with multiple mutations harbored two independent spliceosome gene mutations, indicative of convergent evolution, i.e., evolution to overcome a shared selective pressure or to exploit a shared environment (Greaves and Maley, 2012; Rossi et al., 2008).

It is tempting to consider the nature of age-related changes in normal hemopoiesis that make it permissive to the outgrowth of

Table 2. Amino Acid Consequences and VAFs of the 112 Clonal Mutations Identified in This Study

Mutation Hot Spot	Codon	VAF (%)	Age	Mutation Hot Spot	Codon	VAF (%)	Age	Mutation Hot Spot	Codon	VAF (%)	Age
DNMT3A R882	p.R882H	4.14	25		p.R882H	32.02	81	*IDH1 R132*	p.R132H	42.13	84
	p.R882C	2.33	35		p.R882H	1.14	81		p.R132C	0.92	92
	p.R882H	3.80	42		p.R882H	3.06	81	*IDH2 R140*	p.R140Q	6.67	76
	p.R882H	4.00	42		p.R882H	2.17	81	*SRSF2 P95*	p.P95R	4.46	70
	p.R882H	1.25	43		p.R882H	1.13	82		p.P95L	3.35	72
	p.R882H	19.00	48		p.R882H	1.46	82		p.P95H	0.86	73
	p.R882H	1.18	49		p.R882C	2.62	82		p.P95H	0.84	77
	p.R882S	1.74	49		p.R882C	6.15	89		p.P95L	0.97	79†
	p.R882H	9.87	50		p.R882C	2.00	94		p.P95L	0.85	80††
	p.R882H	0.83	51	*JAK2V617F*	p.V617F	1.56	34		p.P95H	6.67	80††
	p.R882C	1.10	51		p.V617F	4.91	42		p.P95L	0.96	81
	p.R882C	12.50	52		p.V617F	7.72	45		p.P95H	6.40	82
	p.R882C	1.28	53		p.V617F	0.85	62		p.P95L	2.74	85
	p.R882C	2.47	54		p.V617F	25.44	64		p.P95R	7.52	87
	p.R882H	1.95	55		p.V617F	7.41	65		p.P95L	5.84	88**
	p.R882C	30.22	55		p.V617F	1.03	67		p.P95H	10.48	88**
	p.R882C	1.22	56		p.V617F	0.88	71		p.P95R	2.71	88
	p.R882H	0.91	58		p.V617F	3.75	71		p.P95R	17.05	90‡
	p.R882H	4.17	60		p.V617F	1.16	75	*SF3B1* K700	p.K700E	1.04	76
	p.R882H	5.90	60		p.V617F	2.30	77		p.K700E	6.63	81
	p.R882H	9.60	60		p.V617F	1.92	78		p.K700E	0.79	82
	p.R882H	2.73	60		p.V617F	2.26	80*		p.K700E	12.59	83
	p.R882C	9.33	60		p.V617F	4.25	80		p.K700E	8.77	83‡‡
	p.R882H	7.03	61		p.V617F	1.92	80		p.K700E	1.02	84
	p.R882C	1.21	61		p.V617F	3.71	80		p.K700E	0.85	90‡
	p.R882H	0.86	63		p.V617F	15.48	81		p.K700E	1.37	90
	p.R882H	2.54	64		p.V617F	1.21	82	*SF3B1* K666	p.K666N	1.33	70
	p.R882H	3.19	67		p.V617F	1.62	85		p.K666N	5.01	79
	p.R882H	2.74	70		p.V617F	0.83	85		p.K666N	13.36	79†
	p.R882H	4.27	74		p.V617F	1.98	86		p.K666N	15.43	80*
	p.R882H	0.85	74		p.V617F	25.94	88		p.K666N	4.60	81
	p.R882H	0.85	75		p.V617F	10.88	88**		p.K666E	1.09	83‡‡
	p.R882C	1.12	77		p.V617F	2.94	90		p.K666N	35.11	86
	p.R882C	1.15	78		p.V617F	1.23	90		p.K666N	19.70	86
	p.R882H	1.26	79	*KRAS G12*	p.G12 R	0.94	55		p.K666N	16.55	86
	p.R882H	16.66	80		p.G12S	2.78	78		p.K666E	3.34	95
	p.R882C	4.28	80	*NRAS G12*	p.G12S	1.50	61				
	p.R882C	3.66	80		p.G12D	0.96	62				

Mutations identified in the same sample are highlighted with the same symbol (*, **, †, ††, ‡, and ‡‡).

clones driven by spliceosome mutations. HSCs do not operate in isolation; instead, their normal survival and behavior are closely dependent on interactions with the hemopoietic microenvironment (Calvi et al., 2003; Rossi et al., 2008; Zhang et al., 2003). Therefore, both cell-intrinsic and microenvironmental factors influence hemopoietic aging (Rossi et al., 2008; Woolthuis et al., 2011). For example, there is good evidence for age-related changes in cell-intrinsic properties of HSCs in both mice (Chambers et al., 2007; Rossi et al., 2005) and humans (Rübe et al., 2011; Taraldsrud et al., 2009), and it is also clear that aging has a profound effect on the hemopoietic niche, reducing its ability to sustain polyclonal hemopoiesis, favoring oligo- or monoclonality instead (Vas et al., 2012). These and many other observations provide strong evidence that changes in the hemopoietic system subject HSCs to changing pressures during normal aging, driving clonal selection (Rossi et al., 2008).

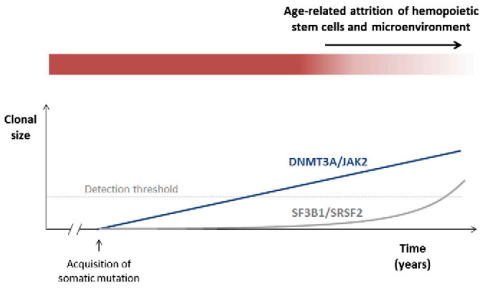


Figure 2. Proposed Kinetics of Hemopoietic Clones Driven by Different Gene Mutations

Mutations such as *DNMT3A* R882H/C or *JAK2* V617F drive a slow but inexorable clonal expansion, leading to the outgrowth of a detectable clone after a certain latency. By contrast, mutations affecting spliceosome genes, such as *SF3B1* and *SRSF2*, and acquired at the same age for the purposes of this model give no proliferative advantage initially but do so later in the context of an aging hemopoietic compartment. Their effects may operate by prolonging stem cell survival and repopulating fitness beyond that of normal stem cells or by exploiting cell-extrinsic changes in the aging microenvironment.

A striking example of such selection was described in a 115-year-old woman whose peripheral white blood cells were shown to be primarily the offspring of only two related HSC clones, whose cargo of approximately 450 somatic mutations did not include known leukemogenic mutations (Holstege et al., 2014). In the absence of somatic driver mutations, it is probable that such selection is driven by well-demonstrated epigenetic differences between individual HSCs (Fraga et al., 2005) or by stochastic events. Furthermore, clonal hemopoiesis in the absence of a known leukemia-driver mutation was also well documented recently (Genovese et al., 2014), and whereas unknown or undetected drivers may be responsible for many cases of this phenomenon, it is also highly plausible that a stochastic process of clonal selection or loss may operate in others. Our study provides evidence that spliceosome gene mutations offer a means to exploit age-related changes in hemopoiesis to drive clonal hemopoiesis in advanced old age, an observation that blurs the boundary between "driver" and "passenger" mutations. Such a context dependency is not a surprising attribute for the effects of spliceosome mutations, which have not, so far, been shown to impart a primary proliferative advantage to normal hemopoietic stem and progenitor cells (Matsunawa et al., 2014; Visconte et al., 2012).

A final important finding of our study was the almost complete absence of canonical *NPM1* mutations in our collection of more than 4,000 people, despite the use of a highly sensitive assay for their detection, designed specifically for this study. Among more than 10 million mapped reads covering this mutation hot spot, we identified only four reads in a single sample reporting a canonical mutation (mutation A; TCTG duplication). Given their frequency in myeloid leukemia (Cancer Genome Atlas Research Network, 2013) and the fact that they are not late mutations (Krönke et al., 2013; Shlush et al., 2014), this observation frames *NPM1* mutations as "gatekeepers" of leukemogenesis, i.e., their acquisition appears to be closely associated with the development of frank leukemia. In this light, the frequent co-occurrence of *DNMT3A* and *NPM1* mutations suggests that the former behave as "rafts" that enable *NPM1* mutant clones to be founded and expanded, thus facilitating onward evolution toward acute myeloid leukemia.

We used a highly sensitive method to search for evidence of clonal hemopoiesis driven by 15 recurrent leukemogenic mutations in more than 4,000 individuals. Our results demonstrate that the incidence of clonal hemopoiesis is much higher than suggested by exome-sequencing studies, that spliceosome gene mutations drive clonal outgrowth primarily in the context of an aging hemopoietic compartment, and that *NPM1* mutations do not drive ARCH, indicating that their acquisition is closely associated with frank leukemia.

EXPERIMENTAL PROCEDURES

Patient Samples

Samples were obtained with written informed consent and in accordance with the Declaration of Helsinki and appropriate ethics committee approvals from all participants (approval reference numbers 10/H0604/02, 07/MRE05/44, and 05/Q0106/74). Maternal consent was obtained for the use of cord blood samples. Samples were obtained from 3,067 blood donors aged 17–70 years (WTCCC; UK Blood Services 1 [UKBS1] and UKBS2 common controls), 1,152 unselected individuals aged 60–98 years (UKHLS; https://www.understandingsociety.ac.uk/), 32 patients that had undergone a hemopoietic stem cell transplant (12 autologous and 20 allogeneic; Tables S3 and S4) 1 month to 14 years previously, and 18 cord blood samples. Age distribution of the WTCCC and UKHLS cohorts/samples is shown in Figure S1. Hemoglobin concentrations were available for a total of 3,587 of the 4,067 samples from which adequate sequencing data were obtained for analysis, including 102 of 105 samples with mutations. Full blood count results were available for 2,952 WTCCC samples. The average blood donation frequency for WTCCC donors was 1.6 donations of one unit per year. Details of donations by individual participants were not available.

Targeted Sequencing

Genomic DNA was used to simultaneously amplify several gene loci using multiplex PCR, in order to capture and analyze 15 mutational hot spots enriched for, but not exclusive to, targets of mutations thought to arise early in leukemogenesis (Table 1). We used three multiplex primer combinations (Plex1–3), guided by our findings, to capture the targeted mutational hot spots (Table S1). Primers were designed using the Hi-Plex PCR-MPS (massively parallel sequencing) strategy (Nguyen-Dumont et al., 2013), except for *JAK2* V617 and "Plex2" primers, which were designed using MPRIMER (Shen et al., 2010). These and additional primer sequences used in each Plex and details of PCR- and DNA-sequencing protocols are detailed in Supplemental Experimental Procedures. Methodological validation experiments are shown in Figure S2.

Bioinformatic Analysis

Sequencing data were aligned to the human reference genome (hg19) using BWA. Subsequently, the SAMTOOLS pileup command was used to generate pileup files from the generated bam files (version 0.1.8; http://samtools.sourceforge.net; Li et al., 2009). A flexible in-house Perl script generated by our group, MIDAS (Conte et al., 2013), was modified in order to interrogate only the hot spot nucleotide positions of interest (those with reported mutations in the COSMIC database; Forbes et al., 2015) on the pileup file, considering only those reads with a sequence quality higher than 25 and a mapping quality higher than 15. For each sample, the numbers of reads reporting the reference and variant alleles at each position were extracted. VAFs were derived by dividing the number of reads reporting the most-frequent variant nucleotide to the total. In order to detect NPM1 mutations with high sensitivity,

we wrote a bespoke Perl script described in Supplemental Experimental Procedures.

Statistical Analyses and Mutation-Calling Threshold

We chose a threshold VAF of ≥0.008 (0.8%) to "call" clones with a heterozygous mutation representing ≥1.6% of blood leukocytes. From validation experiments and data analysis (see Supplemental Experimental Procedures and Figure S2D), we determined that the maximum false-positive error rate for calling a mutation (VAF ≥ 0.008) due to variant allele counts that are solely due to PCR-MiSeq error was negligible ($p < 10^{-5}$). For comparisons of blood cell counts and hemoglobin concentrations, we used non-paired t tests. For summary statistics of read coverage (Table S2) and for the purposes of deriving an estimate of the overall incidence of clonal hemopoiesis (Figure S4), we used published tables of all mutations reported by three recent studies that employed whole-exome-sequencing analyses to identify individuals with clonal hemopoiesis (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014).

ACCESSION NUMBERS

The European Genome-Phenome Archive (EGA) accession number for the sequencing data reported in this paper is EGAS00001000814.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.02.005.

AUTHOR CONTRIBUTIONS

G.S.V. conceived and designed the study. G.S.V. and T. McKerrell supervised the study, analyzed data, and wrote the manuscript. N.P. and T. McKerrell performed experimental procedures. I.V. and T. Moreno wrote scripts and performed bioinformatics analysis. H.P., T. McKerrell, and G.S.V. performed statistical analyses. E.Z., C.S.G., M.A.Q., and R.R. contributed to study strategy and to technical and analytical aspects. U.S.S.G., E.Z., W.O, J.C., C.C., J.B., J.S., C.H., M.A.S., and D.R.F. contributed to sample acquisition and subject recruitment.

ACKNOWLEDGMENTS

This project was funded by a Wellcome Trust Clinician Scientist Fellowship (100678/Z/12/Z; to T. McKerrell) and by the Wellcome Trust Sanger Institute (grant number WT098051). G.S.V. is funded by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA), and work in his laboratory is also funded by Leukaemia Lymphoma Research and the Kay Kendal Leukaemia Fund. I.V. is funded by Spanish Ministerio de Economía y Competitividad subprograma Ramón y Cajal. C.S.G. is funded by a Leukaemia Lymphoma Research Clinical Research Training Fellowship. We thank Servicio Santander Supercomputación for their support. We acknowledge use of DNA from The UK Blood Services Collection of Common Controls (UKBS collection), funded by the Wellcome Trust grant 076113/C/04/Z, by the Juvenile Diabetes Research Foundation grant WT061858, and by the National Institute of Health Research of England. The collection was established as part of the Wellcome Trust Case-Control Consortium. We also gratefully acknowledge use of blood DNA samples and data from participants of the UK Household Longitudinal Study (https://www.understandingsociety.ac.uk/), collected by NatCen and the Institute for Social and Economic Research, University of Essex, and funded by the Economic and Social Research Council, UK. We thank the Cambridge Blood and Stem Cell Biobank and the Cancer Molecular Diagnosis Laboratory, Cambridge Biomedical Research Centre (National Institute for Health Research, UK) for help with sample collection and processing. Finally, we thank Nathalie Smerdon, Richard Rance, Lucy Hildyard, Ben Softly, and Britt Killian for help with sample management, DNA sequencing, and data processing. G.S.V. is a consultant for KYMAB and receives an educational grant from Celgene.

Received: December 14, 2014
Revised: January 19, 2015
Accepted: January 29, 2015
Published: February 26, 2015

REFERENCES

Busque, L., Patel, J.P., Figueroa, M.E., Vasanthakumar, A., Provost, S., Hamilou, Z., Mollica, L., Li, J., Viale, A., Heguy, A., et al. (2012). Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. Nat. Genet. *44*, 1179–1181.

Calvi, L.M., Adams, G.B., Weibrecht, K.W., Weber, J.M., Olson, D.P., Knight, M.C., Martin, R.P., Schipani, E., Divieti, P., Bringhurst, F.R., et al. (2003). Osteoblastic cells regulate the haematopoietic stem cell niche. Nature *425*, 841–846.

Cancer Genome Atlas Research Network (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

Cazzola, M., Della Porta, M.G., and Malcovati, L. (2013). The genetic basis of myelodysplasia and its clinical relevance. Blood *122*, 4021–4034.

Chambers, S.M., Shaw, C.A., Gatza, C., Fisk, C.J., Donehower, L.A., and Goodell, M.A. (2007). Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. PLoS Biol. *5*, e201.

Conte, N., Varela, I., Grove, C., Manes, N., Yusa, K., Moreno, T., Segonds-Pichon, A., Bench, A., Gudgin, E., Herman, B., et al. (2013). Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture. Leukemia *27*, 1820–1825.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. *43*, D805–D811.

Ford, A.M., Bennett, C.A., Price, C.M., Bruin, M.C., Van Wering, E.R., and Greaves, M. (1998). Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia. Proc. Natl. Acad. Sci. USA *95*, 4584–4588.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suñer, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. USA *102*, 10604–10609.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., et al. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia *28*, 241–247.

Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B., et al. (2014). Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res. *24*, 733–742.

Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.J., et al. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nat. Genet. *44*, 651–658.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.

Krönke, J., Bullinger, L., Teleanu, V., Tschürtz, F., Gaidzik, V.I., Kühn, M.W., Rücker, F.G., Holzmann, K., Paschka, P., Kapp-Schwörer, S., et al. (2013). Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. Blood *122*, 100–108.

Kyle, R.A., Therneau, T.M., Rajkumar, S.V., Offord, J.R., Larson, D.R., Plevak, M.F., and Melton, L.J., 3rd. (2002). A long-term study of prognosis in monoclonal gammopathy of undetermined significance. N. Engl. J. Med. *346*, 564–569.

Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. *44*, 642–650.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lin, C.C., Hou, H.A., Chou, W.C., Kuo, Y.Y., Wu, S.J., Liu, C.Y., Chen, C.Y., Tseng, M.H., Huang, C.F., Lee, F.Y., et al. (2014). SF3B1 mutations in patients with myelodysplastic syndromes: the mutation is stable during disease evolution. Am. J. Hematol. *89*, E109–E115.

Matsunawa, M., Yamamoto, R., Sanada, M., Sato-Otsubo, A., Shiozawa, Y., Yoshida, K., Otsu, M., Shiraishi, Y., Miyano, S., Isono, K., et al. (2014). Haploinsufficiency of Sf3b1 leads to compromised stem cell function but not to myelodysplasia. Leukemia *28*, 1844–1850.

Nguyen-Dumont, T., Pope, B.J., Hammet, F., Southey, M.C., and Park, D.J. (2013). A high-plex PCR approach for massively parallel sequencing. Biotechniques *55*, 69–74.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C.J., Ellis, P., Wedge, D.C., Pellagatti, A., et al.; Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood *122*, 3616–3627, quiz 3699.

Rossi, D.J., Bryder, D., Zahn, J.M., Ahlenius, H., Sonu, R., Wagers, A.J., and Weissman, I.L. (2005). Cell intrinsic alterations underlie hematopoietic stem cell aging. Proc. Natl. Acad. Sci. USA *102*, 9194–9199.

Rossi, D.J., Jamieson, C.H., and Weissman, I.L. (2008). Stems cells and the pathways to aging and cancer. Cell *132*, 681–696.

Rübe, C.E., Fricke, A., Widmann, T.A., Fürst, T., Madry, H., Pfreundschuh, M., and Rübe, C. (2011). Accumulation of DNA damage in hematopoietic stem and progenitor cells during human aging. PLoS ONE *6*, e17487.

Shen, Z., Qu, W., Wang, W., Lu, Y., Wu, Y., Li, Z., Hang, X., Wang, X., Zhao, D., and Zhang, C. (2010). MPprimer: a program for reliable multiplex PCR primer design. BMC Bioinformatics *11*, 143.

Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M., Gupta, V., Kennedy, J.A., Schimmer, A.D., Schuh, A.C., Yee, K.W., et al.; HALT Pan-Leukemia Gene Panel Consortium (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature *506*, 328–333.

Taraldsrud, E., Grøgaard, H.K., Solheim, S., Lunde, K., Fløisand, Y., Arnesen, H., Seljeflot, I., and Egeland, T. (2009). Age and stress related phenotypical changes in bone marrow CD34+ cells. Scand. J. Clin. Lab. Invest. *69*, 79–84.

Vas, V., Senger, K., Dörr, K., Niebel, A., and Geiger, H. (2012). Aging of the microenvironment influences clonality in hematopoiesis. PLoS ONE *7*, e42080.

Visconte, V., Rogers, H.J., Singh, J., Barnard, J., Bupathi, M., Traina, F., McMahon, J., Makishima, H., Szpurka, H., Jankowska, A., et al. (2012). SF3B1 haploinsufficiency leads to formation of ring sideroblasts in myelodysplastic syndromes. Blood *120*, 3173–3186.

Woolthuis, C.M., de Haan, G., and Huls, G. (2011). Aging of hematopoietic stem cells: Intrinsic changes or micro-environmental effects? Curr. Opin. Immunol. *23*, 512–517.

Wu, S.J., Kuo, Y.Y., Hou, H.A., Li, L.Y., Tseng, M.H., Huang, C.F., Lee, F.Y., Liu, M.C., Liu, C.W., Lin, C.T., et al. (2012). The clinical implication of SRSF2 mutation in patients with myelodysplastic syndrome and its stability during disease evolution. Blood *120*, 3106–3111.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478.

Zhang, J., Niu, C., Ye, L., Huang, H., He, X., Tong, W.G., Ross, J., Haug, J., Johnson, T., Feng, J.Q., et al. (2003). Identification of the haematopoietic stem cell niche and control of the niche size. Nature *425*, 836–841.

PLOS | ONE

RESEARCH ARTICLE

Colorectal Adenomas Contain Multiple Somatic Mutations That Do Not Coincide with Synchronous Adenocarcinoma Specimens

José P. Vaqué[1]*, Nerea Martínez[1], Ignacio Varela[2], Fidel Fernández[3], Marta Mayorga[3], Sophia Derdak[4], Sergi Beltrán[4], Thaidy Moreno[2], Carmen Almaraz[1], Gonzalo De las Heras[5], Mónica Bayés[4], Ivo Gut[4], Javier Crespo[5,6], Miguel A. Piris[1,3‡]

1 Cancer Genomics Group, IDIVAL, Instituto de Investigación Marqués de Valdecilla, Santander, Spain, **2** IBBTEC-UC-CSIC-SODERCAN Instituto de Biomedicina y Biotecnología de Cantabria, Santander, Spain, **3** Department of Pathology, Hospital Universitario Marqués de Valdecilla, Santander, Spain, **4** Centro Nacional de Análisis Genómico, CNAG, Barcelona, Spain, **5** Gastroenterology and Hepatology Unit, Hospital Universitario Marqués de Valdecilla, Santander, Spain, **6** Infection, Immunity and Digestive Pathology Group, IFIMAV, Santander, Spain

‡ MAP is the senior author on this work.
* jpvaque@idival.org

Abstract

We have performed a comparative ultrasequencing study of multiple colorectal lesions obtained simultaneously from four patients. Our data show that benign lesions (adenomatous or hyperplastic polyps) contain a high mutational load. Additionally multiple synchronous colorectal lesions show non overlapping mutational signatures highlighting the degree of heterogeneity between multiple specimens in the same patient. Observations in these cases imply that considering not only the number of mutations but an effective oncogenic combination of mutations can determine the malignant progression of colorectal lesions.

Citation: Vaqué JP, Martínez N, Varela I, Fernández F, Mayorga M, Derdak S, et al. (2015) Colorectal Adenomas Contain Multiple Somatic Mutations That Do Not Coincide with Synchronous Adenocarcinoma Specimens. PLoS ONE 10(3): e0119946. doi:10.1371/journal.pone.0119946

Academic Editor: Yunli Zhou, Harvard Medical School, UNITED STATES

Received: June 21, 2014

Accepted: January 22, 2015

Published: March 16, 2015

Data Availability Statement: The raw data from this study have been deposited in the NIH Short Read Archive (SRA) under accession number SRP040626.

Funding: Funding provided by Grant PI12/00357 (isciii, FEDER funds) to JPV, Grant Fundación Aecc to MAP, Grant SODERCAN-Gobierno de Cantabria to MAP and Grant from the Spanish "Ministerio de Ciencia e Innovación" (RETICS, SAF2008-03871)to MAP. IV is funded by Spanish Ministerio de Economía y Competitividad subprograma Ramón y Cajal. The authors would like to thank Servicio Santander Supercomputación for their IT support.

Introduction

Our current understanding of colorectal cancer assumes that its pathogenesis includes a progressive accumulation of genomic changes at multiple stages. Thus, initiating events, such as driver mutations affecting APC or KRAS genes, are followed by additional alterations in specific genes such as p16 and p53 [1] and signalling pathways including WNT, MAPK, GNAS or TGFB that, over time, will shape the genomic conditions that drive a pre-malignant lesion towards cancer [2–4]. Thus, premalignant lesions such as colorectal adenomas feature mutational events in APC, BRAF, KRAS and other genes [2, 5]. As the disease progresses, colorectal adenocarcinoma specimens can also accumulate mutations in genes such as p53 and FBXW7 as well as in MAPK, TGFB, PI3K and DNA mismatch-repair pathways [3]. However, the question of whether somatic mutations accumulate in the adenoma-carcinoma sequence in the same patient remains to be investigated.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

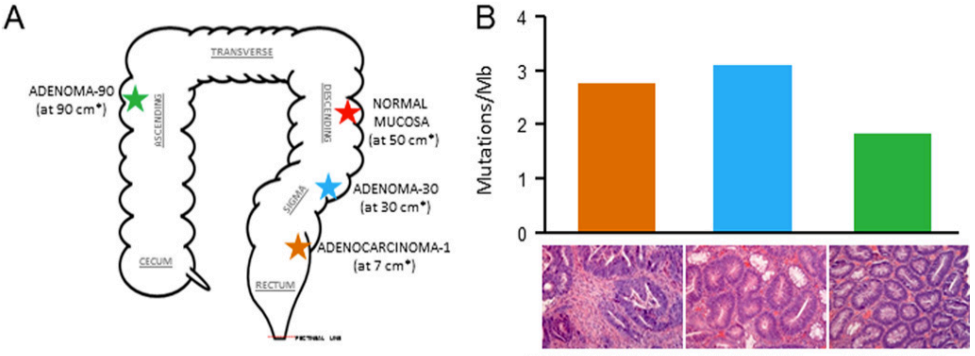
Competing Interests: The authors have declared that no competing interests exist.

Here we have sequenced whole exomes of multiple lesions in four non-MSI colorectal cancer patients corresponding to different adenoma and adenocarcinoma specimen samples taken during the same endoscopic procedure. Our first finding was that adenomas contained a large number of mutations that, in general were reduced but still comparable, with the frequency found in colorectal cancer samples. Additionally, different adenoma lesions within the same patient were strikingly heterogeneous. Analysis of the mutation frequency also showed that a large majority of the mutations found in adenoma samples were subclonal, and probably passenger mutation events.

Results and Discussion

We characterized the genomic variants in a series of untreated colorectal lesions that included adenocarcinomas, adenomas and hyperproliferative polyps taken simultaneously by endoscopic resection, along with normal mucosa, which was used as a control (S1 Table). The topologies of the lesions of each patient are shown in Figs. 1a, 2a, 3a and 3b and the clinical characteristics are summarised in Table 1. We generated two paired-end 75-bp whole exome sequencing libraries and sequenced them using an Illumina HiSeq2000 instrument, which allowed us to map an average of ~102 million reads per sample. Under these conditions, the mean coverage of the target sequenced was 99X (78X-141X) with a mean of 92.1% (89.8–95.9) of targeted bases with at least 15X coverage (S1 Table). Somatic variants were identified using the SAMtools suite. Additionally, we used RAMSES software [6] to call potential mutations showing minimum independent multi-aligner evidence that enabled us to identify subclonal variants present in at least 5% of the reads. We also performed a secondary analysis in a selection of genes with known biological activity that confirmed specific mutations in up to 76.5% of those genes with a mutational percentage above 15% in each sample of our primary analysis (Figs. 1b and 2b and S1 Fig. and S2 Table). Using the data obtained in our primary analysis and aligned with previous observations in colorectal lesions [5], we observed that most mutations were C>T/G>A changes that occurred in CpG in up to 75% of the cases (Fig. 4, and S5 Table). In addition, we reproduced these results using the validated data from the secondary analysis (S2 Fig.). A detailed description of the main findings is included in table 2 and S1–S5 Tables. We decided to focus on those alterations that could potentially induce changes in the expression or activity of the proteins including amino acid changing or truncating mutations. Analysing their incidence, we found that most but not all benign lesions (adenoma or hyperproliferative polyp) contained less genomic alterations than the colorectal cancer specimens (Figs. 1b, 2b, 3a and 3b and table 2); a mutational rate similar to that described by the TCGA network for the non-hypermutated colorectal adenocarcinoma samples [3]. Using this approach we were able to detect one or multiple distinct gene alterations affecting APC in 6 of the 8 adenomas analysed, thereby underlining the relevance of the APC gene inactivation in the genesis of colorectal adenomas. In the same line of evidence, we observed that these benign lesions lacked mutations in genes or pathways considered essential in colorectal cancer [3], with the possible exception of PIK3CG in the adenoma-2 case (Fig. 2c) or KRAS and NRAS mutations found in adenomas-4B and 4C (Fig. 3f). On the other hand, we noticed that a number of mutations found in the adenocarcinomas affected oncogenes such as GHR and INSR (Fig. 1c) or KRAS and ERBB4 (Fig. 2c). These are well known for their ability to activate MAPK signalling. We were able to detect them alongside other somatic mutations affecting SMAD genes (TGFB signalling, Fig. 2c and Fig. 3e) or adenylyl cyclases such as ADCY2 (Fig. 1c) and ADCY1 (Fig. 2c) that participate in the COX2-PGE2-PR-GNAS signalling axis (reviewed in [7]). When comparing the mutational spectrum of the multiple samples from the same patient, we did not find a single recurrent mutation, which in addition to the multiple and non-recurrent alterations

Mutational analysis of Patient 1

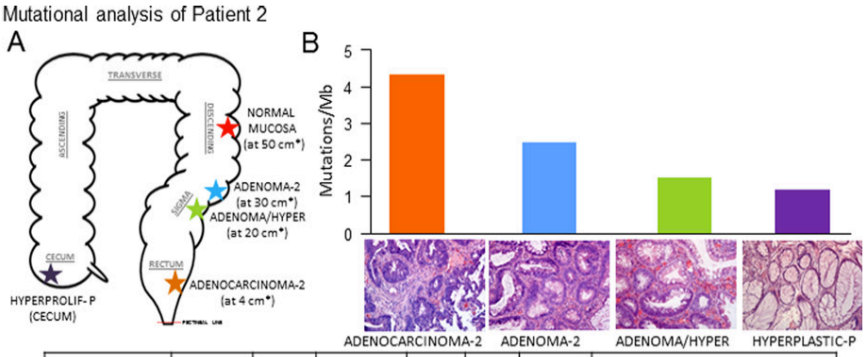


C

Sample	Gene	Chrom	Position	Wt Allele	Mut Allele	% Mutated	Function
ADENOCARCINOMA-1	GHR	Chr5	42695074	G	T	33,8	GROWTH HORMONE RECEPTOR (JAK/STAT)
	ADCY2	Chr5	7804710	C	T	33,7	ADENYLYL CYCLASE-2 (cAMP)
	MAML1	Chr5	179192432	C	T	32,1	NOTCH SIGNALING
	PRDM14	Chr8	70964415	C	T	29,5	HISTONE METHYL-TRANSFERASE (STEMNESS)
	KAT6B	Chr10	76780405	C	T	25,7	HAT (RUNX2-RELATED)
	RAPGEF3	Chr12	48151855	A	G	23,8	EPAC-cAMP SIGNALING
	NBR1	Chr17	41352594	G	A	19,7	PUTATIVE ONCOGENIC
	INSR	Chr19	7119568	A	G	19,3	INSULIN RECEPTOR (RTK)
ADENOMA-30	APC	Chr5	112174061	A	T	50,9	TUMOR SUPPRESSOR/WNT ANTAGONIST
	GPR112	ChrX	135438301	G	T	21,7	GPCR
	LRP1B	Chr2	141115596	G	A	18,9	LDL RECEPTOR FAMILY
	PTGER3	Chr1	71418640	C	T	16,7	GPCR/PROSTAGLANDIN RECEPTOR (CAMP)
	GRM6	Chr5	178419033	G	A	14,3	GPCR-GNAI
	APC	Chr5	112175174	G	T	13,9	TUMOR SUPPRESSOR/WNT ANTAGONIST
	RARA	Chr17	38512455	A	G	9,7	RETOIC ACID RECEPTOR A
	BMP10	Chr2	69093453	G	T	5,3	TGF BETA FAMILY/PROLIFERATIVE
ADENOMA-90	TIAL1	Chr10	121336630	C	A	47,5	mRNA PROCESSING
	NFATC2IP	Chr16	28970176	A	G	46,0	NFAT CO-ACTIVATOR
	TYRO3	Chr15	41865517	A	G	38,4	TRK (PUTATIVE PRO SURVIVAL/GROWTH)
	NYAP2	Chr2	226447383	C	T	37,1	PI3K ACTIVATOR
	BRD4	Chr19	15383837	T	C	33,3	EPIGENETIC REG./MITOSIS
	PLXNB2	Chr22	50724430	G	A	31,9	ACTIVATOR OF RHO/MIGRATION
	APC	Chr5	112164616	C	T	12,3	TUMOR SUPPRESSOR/WNT ANTAGONIST

Fig 1. Mutation analysis of patient 1. A) Scheme showing an approximate representation of the location of each lesion analysed. The distance (*, in cms) from the pectineal line (red dots) is shown. B) Mutational index (number of mutations/Mb) found in the indicated sample from the primary NGS analysis. H&E pictures are representative of each lesion studied by NGS. C) Validated mutations found in a secondary targeted NGS analysis of the indicated samples. Chrom: chromosome; % mutated: percentage of mutant nucleotides found in the corresponding gene within the same sample.

doi:10.1371/journal.pone.0119946.g001



C

Sample	Gene	Chrom	Position	Wt Allele	Mut Allele	% Mutated	Function
ADENOCARCINOMA-2	SMAD3	Chr15	67479795	C	T	30,7	TGF BETA SIGNALING
	DUSP15	Chr20	30450487	C	T	30,0	DUAL PHOSPHATASE
	BAI2	Chr1	32201250	C	G	26,0	GPCR-GNAI
	SNX13	Chr7	17885311	T	C	25	INHIBITOR OF GNAS ACTIVITY
	KRAS	Chr12	25398285	C	T	22,4	ONCOGENIC GTPASE
	AKT2	Chr19	40762843	G	C	19,2	PI3K/MTOR SIGNALING
	ARFGEF1	Chr8	68208781	A	G	18,7	ARF GEF
	SMAD2	Chr18	45368209	T	C	17,9	TGFBETA SIGNALING
	ERBB4	Chr2	212530168	T	A	17,5	RTK EGFR FAMILY
	ADCY1	Chr7	45744165	G	A	17,5	ADENYLYL CYCLASE 1/cAMP
	LRP5	Chr11	68157518	G	A	12,8	WNT CORRECEPTOR
	PLCZ1	Chr12	18837081	A	C	12,6	PHOSPHOLIPASE
	APC	Chr5	112175507	C	T	10,9	TUMOR SUPPRESSOR/WNT ANTAGONIST
	APC	Chr5	112174793	G	T	10,5	TUMOR SUPPRESSOR/WNT ANTAGONIST
	IL31	Chr12	122657162	G	C	10,4	N/D
	ARHGAP19	Chr10	99003739	G	A	9,5	NEGATIVE REGULATOR OF RHO
	FAT4	Chr4	126412567	A	T	9,5	CADHERIN
	CDKN3	Chr14	54878241	T	C	8,6	INHIBITOR OF CDK2
	SMOC2	Chr6	168949856	C	T	5,5	ONCOGENIC ANGIOGENESIS RELATED
ADENOMA-2	APC	Chr5	112175213	A	T	69,4	TUMOR SUPPRESSOR/WNT ANTAGONIST
	PIK3CG	Chr7	106526642	A	T	40,2	PI3K GAMMA
	PTK6	Chr20	62163954	C	T	36,2	PUTATIVE ONCOGENIC KINASE
	LRRC4C	Chr11	40136965	A	G	34,3	NETRIN RELATED
	ACY3	Chr11	67413212	G	A	31,5	N/D
	PTPRF-1	Chr1	44035444	G	A	31,2	PHOSPHATASE RELATED TO WNT FUNCTION
	EGLN1	Chr1	231502192	T	C	27,8	HYPOXIA SENSOR/HIF REGULATOR
	SNX13-2	Chr7	17841230	G	C	25	RGS DOMAIN/GNAS INHIBITOR
	APEX2	ChrX	55032975	C	T	16,1	DNA REPAIR (PUTATIVE)
	HK3	Chr5	176314473	G	A	15,9	METHABOLISM
	RAMP1	Chr2	238785871	C	A	11,3	CALCITONIN RECEPTOR
	DBF4B	Chr17	42827993	G	A	11,1	MITOGENIC S/T KINASE
	TNFRSF8	Chr1	12186098	G	A	8,3	TNF FAMLILY RECEPTOR
	DSCAM	Chr21	41711243	G	A	6,0	CELL ADHESION/SIGNALING
ADENOMA/HYPER	F2	Chr11	46749688	C	A	35,9	PRECURSOR OF THROMBIN
	ZNRF3	Chr22	29383086	A	G	31,8	NEGATIVE WNT REGULATOR
	ARHGAP12	Chr10	32197285	T	A	27,7	NEGATIVE REGULATOR OF RHO
	RASGRP2	Chr11	64507568	A	G	12,9	CALCIUM REGULATED RAS ACTIVATOR

Fig 2. Mutation analysis of patient 2. A) Scheme showing an approximated representation of the location of each lesion analysed. The distance (*, in cms) from the pectineal line (red dots) is shown. B) Mutational index (number of mutations/Mb) found in the indicated sample from the primary NGS analysis. H&E pictures are representative of each lesion studied by NGS. C) Validated mutations found in a secondary targeted NGS analysis of the indicated samples. Chrom: chromosome; % mutated: percentage of mutant nucleotide found in the corresponding gene within the same sample.

doi:10.1371/journal.pone.0119946.g002

found in APC, suggests an independent origin of the multiple adenomas and adenocarcinoma in the same patient. In this respect, we could detect individual lesions like for example adenoma-30 (Fig. 1), carrying different mutations in APC detected at different percentages (14% and 51%). This may reflect a degree of subclonal activity that is not exclusive to adenomas, since adenocarcinoma-2 (Fig. 2) also harboured two distinct APC mutations in 10.9% and 10.4% of the alleles read. Moreover, our observations (aligned with those found in [5]), seem to suggest that colorectal adenomas, independently of their size or degree of dysplasia, and even hyperplastic polyps, (both with reduced potential to make progress towards cancer), still feature a relatively high number of subclonal mutations combined into inefficient non-carcinogenic signatures. Thus, early steps of colorectal cancer could be characterised by highly dynamic genetic changes until an efficient neoplastic signature, giving rise to an infiltrating carcinoma, is generated. Due to the limited number of patients analysed we cannot yet generalize whether all benign lesions carry a high mutational load. This may also apply to the observation that mutations found in adenomas do not coincide with those found in synchronous adenocarcinoma specimens in the same patient, a finding that is supported by data from other laboratories [5]. The individual characterisation of these precise mutational signatures controlling tumour dynamics at specific stages of the disease may serve in the near future as an indicator for the development of specific combination therapies.

Materials and Methods

Ethics statement

All human samples used in this study were collected under a written informed consent form that was appropriately signed and authorized by each patient and the doctor(s) involved and approved by the CEIC (Comité Ético de Investigación Clínica, Cantabria). We kept the original records under specific restrictive conditions to fulfil the current legal requirements. All processes were conducted and approved following the specific recommendations of the CEIC.

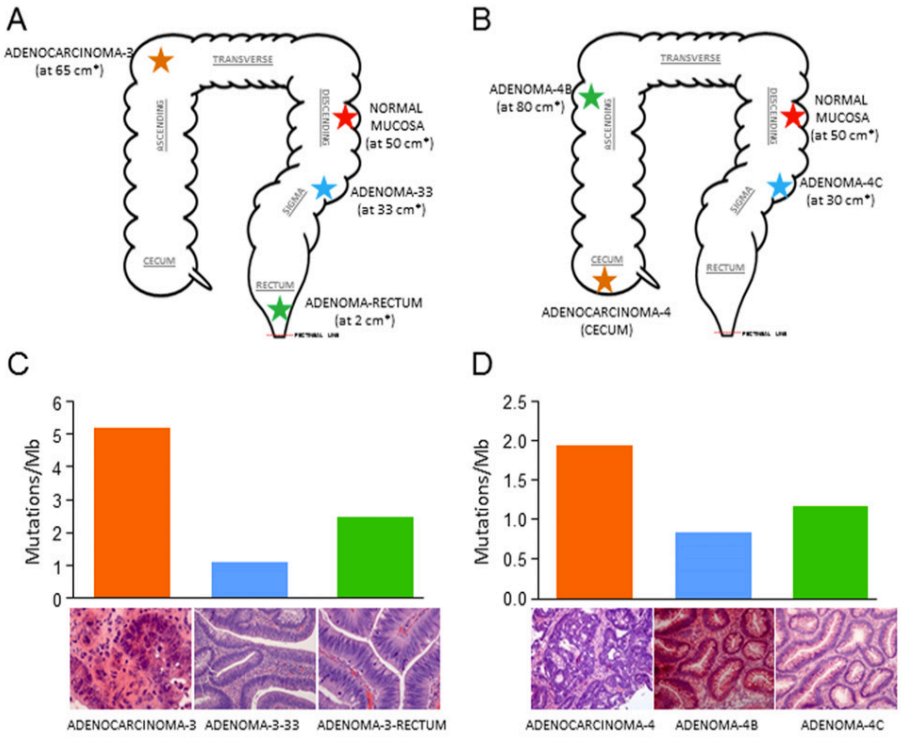
Patients and samples

Nine freshly frozen colorectal samples taken from two previously untreated patients by endoscopic resection were selected for whole exome sequencing. Samples from Patient 1 (Fig. 1) consisted of: a) normal mucosa, b) adenomatous polyp (30 cm), c) adenomatous polyp (90 cm) and d) adenocarcinoma. Samples from Patient 2 (Fig. 2) consisted of: a) normal mucosa, b) hyperplastic polyp, c) adenomatous polyp, d) adenomatous polyp and e) adenocarcinoma. Further information is provided in S1 and S5 Tables. All cases were reviewed by a panel of three pathology specialists and lesions were graded following standard criteria [8].

Genomic DNA extraction, quantification, exome enrichment and sequencing

Purified genomic DNA (3 μg) was extracted from snap-frozen (fresh) samples using standard procedures. Briefly, PBS-washed samples, centrifuged and lysed using "Tissue and cell lysis solution" buffer for the MasterPure kit, complemented by proteinase K (5 μl/100 μl buffer) (Epicenter), shaking overnight at 56°C. DNA was extracted using phenol/chloroform/isoamyl alcohol (in proportions of 25:24:1, respectively) in a fast Lock Gel Light Eppendorf tube (Eppendorf), then washed and precipitated. Genomic DNA was quantified using a Qubit ds DNA BR assay kit and a Qubit 2.0 fluorometer (Invitrogen) following the manufacturer's instructions. Genomic DNA (3 μg) was then enriched in each case for protein coding sequences using the in-solution exome capture SureSelect Human All Exon 50 Mb kit (Agilent

Mutational analysis of Patients 3 and 4



E

Sample	Gene	Chrom	Position	Wt Allele	Mut Allele	% Mutated	Function
ADENOCARCINOMA-3	SMAD2	Chr18	45368208	C	T	51,8	TGF-BETA SIGNALING
	SMAD3	Chr15	67479810	C	T	50	TGF-BETA SIGNALING
	KRAS	Chr12	25398285	C	G	39,2	KRAS ONCOGENE
	MTOR	Chr1	11190804	C	T	24,3	S/T KINASE WITH ONCOGENIC POTENTIAL
	DUSP16	Chr12	12672891	C	T	19,6	MAPK PHOSPHATASE
	DUSP26	Chr8	33451171	A	G	17,1	MAPK PHOSPHATASE
	FAT2	Chr5	150946044	C	T	15,5	PROTO-CADHERIN/CELL ADHESION
AD-3-33	APC	Chr5	112162891	C	T	26,8	TUMOR SUPPRESSOR/WNT ANTAGONIST
	GNAI3	Chr1	110125099	A	G	19,1	ADENYLYL CYCLASE INHIBITOR
	PAK7	Chr20	9525092	G	A	18,2	S/T KINASE (RAC/CDC42 SIGNALING)
AD-3-RECTUM	APC	Chr5	112175171	C	T	67,34	TUMOR SUPPRESSOR/WNT ANTAGONIST
	HTR2C	ChrX	114141413	C	T	69,4	SEROTONIN RECEPTOR (PLC SIGNALING)
	WNT3	Chr17	44846025	C	A	36,8	WNT SIGNALING/PROTO-ONCOGENE
	NMBR	Chr6	142409381	C	T	28,3	BOMBESIN RECEPTOR (PLC SIGNALING)
	CLDN23	Chr8	8560728	G	A	9,2	TIGHT JUNCTIONS/Ca^{++} INDEPENDENT

F

Sample	Gene	Chrom	Position	Wt Allele	Mut Allele	% Mutated	Function
ADENOCARC.-4	TP53	Chr17	7578535	T	C	57,1	P53 ANTI-ONCOGENE
	KRAS	Chr12	25380275	T	G	34,3	KRAS ONCOGENE
	FAT2	Chr5	150933993	C	A	31,7	PROTO-CADHERIN/CELL ADHESION
	APC	Chr5	112173995	G	T	31,1	TUMOR SUPPRESSOR/WNT ANTAGONIST
	PKN1	Chr19	14578550	T	C	22,03	S/T SIGNALING (REGULATED BY RHO)
	APC	Chr5	112175639	C	T	19,04	TUMOR SUPPRESSOR/WNT ANTAGONIST
AD-4B	APC	Chr5	112173917	C	T	33,9	TUMOR SUPPRESSOR/WNT ANTAGONIST
	NRAS	Chr1	115256528	T	G	19,9	NRAS ONCOGENE
	CALCR	Chr7	93065326	G	T	11,7	ADENYLYL CYCLASE ACTIVATOR
A-4C	KRAS	Chr12	25378562	C	T	11,1	KRAS ONCOGENE

Fig 3. Mutation analyses of patients 3 and 4. Scheme showing an approximated representation of the location of each lesion analysed in patient-3 (A) and patient-4 (B). The distance (*, in cms) from the pectineal line (red dots) is shown. C), D) Mutational index (number of mutations/Mb) found in the indicated sample from the primary NGS analysis. H&E pictures are representative of each lesion studied by NGS. Tables below show a selection of genes with oncogenic potential found mutated in the primary analyses of patient-3 (E) and patient 4 (F). Chrom: chromosome; % mutated: percentage of mutant nucleotide found in the corresponding gene within the same sample.

doi:10.1371/journal.pone.0119946.g003

Table 1. Clinical description of the samples analysed.

PATIENT	SEX	AGE	SAMPLE	SAMPLE NAME	DIAGNOSTIC	GRADE OF DYSPLASIA	SIZE (CMS)	MACROSCOPIC DESCRIPTION	LOCALIZATION (FROM PECTINEAL LINE)
1	FEMALE	82	N1E073	normal mucosa-1	normal mucosa	N/A	N/A	healthy mucosa without any macro- or microscopic alteration	50 cm (descendent colon)
1	FEMALE	82	N1E074	Adenoma-30	adenomatous polyp	Tubular adenoma with moderate dysplasia	0,8	Semi-pedunculated polyp	30 cm (sigma)
1	FEMALE	82	N1E075	Adenoma-90	adenomatous polyp	Tubular adenoma showing moderate dysplasia with superficial focal severe dysplasia	0,8	Semi-pedunculated polyp	90 cm (ascendent colon)
1	FEMALE	82	N1E076	Adenocarcinoma-1	adenocarcinoma: poorly differentiated	N/A	5 (length)	stenotic and ulcerated circumferential mass	7 cm (rectum)
2	MALE	82	N2E079	normal mucosa-2	normal mucosa	N/A	N/A	healthy mucosa without any macro- or microscopic alteration	50 cm (descendent colon)
2	MALE	82	N2E069	hyperplastic polyp-2	hyperplastic polyp	N/A	0,3	sessile polyp	cecum
2	MALE	82	N2E092	Adenoma/Hyper-2	adenomatous polyp with hyperplastic mucosa	Tubular adenoma mostly showing mild dysplasia with focal moderate dysplasia	0,2	sessile polyp	20 cm (sigma)
2	MALE	82	N2E070	Adenoma-2	adenomatous polyp	Tubular adenoma mostly showing mild dysplasia with focal moderate dysplasia	0,5	sessile polyp	30 cm (descendent colon)
2	MALE	82	N2E072	Adenocarcinoma-2	adenocarcinoma: Well differentiated	N/A	6 (length)	ulcerated, circumferencial and friable mass (3/4ths of the rectal lumen)	4 cm (rectum)
3	MALE	71	N3J876	normal mucosa-3	normal mucosa	N/A	N/A	healthy mucosa without any macro- or microscopic alteration	50 cm (descendent colon)
3	MALE	71	N3J874	Adenoma-rectum	adenomatous polyp	Tubular adenoma mostly showing focal severe dysplasia	0,5	pedunculated polyp	2 cm (rectum)

(*Continued*)

Table 1. (*Continued*)

PATIENT	SEX	AGE	SAMPLE	SAMPLE NAME	DIAGNOSTIC	GRADE OF DYSPLASIA	SIZE (CMS)	MACROSCOPIC DESCRIPTION	LOCALIZATION (FROM PECTINEAL LINE)
3	MALE	71	N3J873	Adenoma-33	adenomatous polyp	Tubular adenoma mostly showing mild dysplasia	1	pedunculated polyp	33 cm (sigma: from anal margin)
3	MALE	71	N3J872	Adenocarcinoma-3	adenocarcinoma	N/A	5 (length)	stenotic and ulcerated circumferential mass	65 cm (hepatic angle)
4	MALE	62	N4J881	normal mucosa-4	normal mucosa	N/A	N/A	healthy mucosa without any macro- or microscopic alteration	50 cm (descendent colon)
4	MALE	62	N4J878	Adenoma-4B	adenomatous polyp	Tubular adenoma mostly showing mild dysplasia	0,8	Semi-pedunculated polyp	80 cm (ascendent colon)
4	MALE	62	N4J879	Adenoma-4C	adenomatous polyp	Tubular adenoma mostly showing mild dysplasia	0,6	Semi-pedunculated polyp	30 cm
4	MALE	62	N4J877	Adenocarcinoma-4	adenocarcinoma	N/A	8	Ulcerated circumferential mass. Occupies 1/2 of the rectal lumen	cecum

doi:10.1371/journal.pone.0119946.t001

Technologies), following the manufacturer's protocol. The captured targets were subjected to massively parallel sequencing using the Illumina HiSeq 2000 Analyzer (Illumina) with the paired-end 2 × 75 bp read option, in accordance with the manufacturer's instructions. Exome capture and massively parallel sequencing were performed at the Spanish National Genome Analysis Centre (CNAG, Barcelona, Spain). The raw data from this study have been deposited in the NIH Short Read Archive (SRA) under accession number SRP040626.

Sequence mapping and identification of tumour variants

These methods have been described elsewhere [6]. Briefly, base calling and quality control were performed in the Illumina RTA sequence analysis pipeline. Sequence reads were trimmed up to the first base with a quality of more than 10. Mapping to human genome build hg19 (GRCh37) was done with GEM, allowing up to 4 mismatches [9]. Reads not mapped by GEM (~4% of them) were subjected to a final round of mapping with BFAST [10]. Results were merged and only uniquely mapping non-duplicate read pairs were used for subsequent analyses. The SAMtools suite [11] with default settings was used to call SNVs and short INDELS. Variants identified in regions with low mapability [12], with a read depth of < 10 or a strand bias probability of < 0.001 were filtered out. Variants were annotated and effects predicted with ANNOVAR [13] and snpEff [14], including information from dbSNP build 135 [15], the 1000 Genomes Project [16], the Exome Variant Server (NHLBI GO Exome Sequencing Project (ESP), Seattle, WA; http://evs.gs.washington.edu/EVS/) and an internal database of sequence variants identified in a set of > 100 control samples. Tags were added for positions with high strand bias, high tail distance bias, a read depth of < 10 and those in low mapability regions.

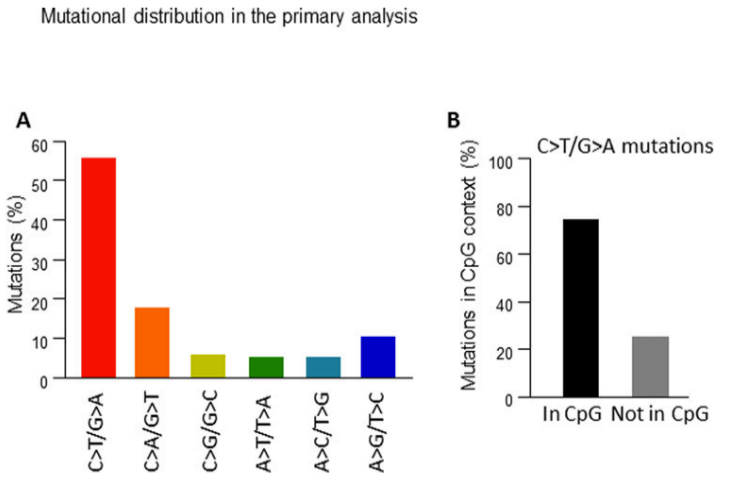


Fig 4. Distribution of mutations detected in the primary analysis. A). Percentage of the indicated mutations detected in the primary analysis. B) Percentage of mutations in CpG dimers.

doi:10.1371/journal.pone.0119946.g004

Table 2. Number of unique amino-acid changing mutations found in the primary analysis.

PATIENT	PATIENT SAMPLE	DATA SOURCE	MUTs*	MUT. INDEX	MUTs INVERSE ANALISYS**	MUT. INDEX (N)
1	N1E074	Adenoma-90 vs. normal mucosa-1	56	1,87	2	0,07
1	N1E075	Adenoma-30 vs. normal mucosa-1	93	3,10	3	0,10
1	N1E076	Adenocarcinoma 1 vs. normal mucosa-1	84	2,80	4	0,13
2	N2E069	Hyperplastic-P vs. normal mucosa-2	35	1,17	2	0,07
2	N2E092	Adenoma/Hyper vs. normal mucosa-2	45	1,50	6	0,20
2	N2E070	Adenoma 2 vs. normal mucosa-2	74	2,47	4	0,13
2	N2E072	Adenocarcinoma 2 vs. normal mucosa-2	130	4,33	4	0,13
3	N3J874	Adenoma-rectum vs normal mucosa-3	74	2,47	9	0,30
3	N3J885	Adenoma-33 vs normal mucosa-3	33	1,10	7	0,23
3	N3J872	Adenocarcinoma-3 vs normal mucosa-3	155	5,17	14	0,47
4	N4J878	Adenoma-4B vs normal mucosa-4	33	1,10	5	0,17
4	N4J879	Adenoma-4C vs normal mucosa-4	25	0,83	8	0,27
4	N4J877	Adenocarcinoma-4 vs normal mucosa-4	58	1,93	9	0,30

MUTs*: Number of amino-acid changing mutations in each lesion vs. normal mucosa, MUT. INDEX: Number of mutations/Mb (Exome) in the colorectal lesion, MUTs INVERSE ANALYSIS**: Number of amino-acid changing mutations found in normal mucosa vs. each lesion, MUT. INDEX (N): Number of mutations/Mb (Exome) in normal mucosa.

doi:10.1371/journal.pone.0119946.t002

For tumour-normal comparison, the probability of a Fisher's exact test was calculated for positions with different genotypes in the two samples.

Detection of subclonal mutations

To identify mutations present in subclonal populations inside the tumours we used a slightly different analysis pipeline. Sequence reads were aligned to the human reference genome (GRCh37) using BWA, and the alignment was consequently cleaned using SAMtools and Picard tools for mating coordinate fixing and PCR duplicate flagging. Finally, GATK indel realigner was used to realign locally around small insertion and deletions (indels). A program specifically written in-house named RAMSES ("Realignment Assisted Minimum Evidence Spotter"; Ignacio Varela, manuscript in preparation) was used to identify coordinates with a minimum value of 2, that were independently aligned with BLAT, and that gave high-quality reads reporting differences from the reference genome in the tumour sample and absolutely no evidence of the same change in the corresponding normal sample. Additionally, mutations near DNA repeats, present in the dbSNP and 1000 Genomes databases, or reported near the end of the reads, were removed. The functional consequence of the mutations was annotated using the Ensembl perl API (Ensembl database, release 69) and only coding mutations were retained.

Secondary analysis by 454 Roche

114 candidate variants from patients 1 and 2 were validated by targeted resequencing using the GS Junior System (Roche). ~300 bp amplicons around the identified mutations were generated, to which specific adaptors were ligated (S3 Table). A pooled, barcoded mixture of amplicons was sequenced using the 454-Junior platform (Roche). The reads were aligned against the human reference genome (GRCh37) using the BWA-SW algorithm. SAMtools was used subsequently to generate bam and pileup files, which were parsed using scripts written in-house. Only those positions with a minimum coverage of 20 in both tumour and normal samples were considered. Mutations with at least 5 independent mutant reads corresponding to a minimum of 1% of the total number of reads at that position in the tumour sample, but with no mutant reads present in the corresponding normal sample, were considered to be validated.

Supporting Information

S1 Fig. Secondary analysis. Percentage of validated mutations in a selection of 92 genes from patients 1 and 2. MP (Mutational percentage): percentage of mutated reads for each mutation. MP>15%: Refers to a mutation found in 15% or more of the total number of reads in the same genomic position. Blue: Confirmed mutations; Red: Not confirmed mutations.
(TIF)

S2 Fig. Distribution of validated mutations. A). Percentage of validated mutations from the secondary analysis. B) Percentage of mutations in CpG. p shows the statistical significance in Fisher´s test.
(TIF)

S1 Table. Mapping and coverage metrics. tROI: Bases that are able to be captured into the genome region that is targeted in the experiment. Specificity: The percentage of non-target bp sequenced among all bases sequenced. Enrichment: Efficiency of recovery for targeted bp in relation to the efficiency of recovery for non-targeted bp, C15: percentage of bases with at least 15X coverage. Mean_cov: mean coverage of the targeted region. Median_cov: median coverage of the targeted region.
(XLS)

S2 Table. Validation panel. Wt Allele: Wild type nucleotide. Mut Allele: Mutated nucleotide. Ref. Reads: Number of reads of the Wt Allele. Mut reads: Number of reads of the Wt Allele. TumorA, C, G or T: Number of reads of each nucleotide.
(XLS)

S3 Table. Oligonucleotides used for validation analysis.
(XLS)

S4 Table. Nucleotide context in validated mutations.
(XLS)

S5 Table. Unique mutations (SNVs) found in this study with potential to provoke amino acid changes. Ref_base: Wild type nucleotide. Mut_base: Mutated nucleotide. Reads_A, C, G or T: Number of reads of each nucleotide. In CpG: the nucleotide is located in a CpG island. Gene ID: Gene name. Transcript ID: Transcript name. c.Annot: Mutation in the cDNA. pAnnot: Mutations in protein. Interpretation: Mutations effect.
(XLS)

Acknowledgments

The authors thank the HUMV-IFIMAV Biobank (Santander, Spain) for providing biological samples.

Author Contributions

Conceived and designed the experiments: JPV MAP JC. Performed the experiments: JPV NM GDLH JC CA FF. Analyzed the data: JPV NM MAP IV TM JC IG SB SD MM FF MB. Contributed reagents/materials/analysis tools: IV TM FF MM SD SB GDLH JC IG MB. Wrote the paper: JPV MAP JC.

References

1. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science. 1991; 253(5015):49–53. Epub 1991/07/05. PMID: 1905840
2. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127):1546–58. Epub 2013/03/30. doi: 10.1126/science.1235122 PMID: 23539594
3. Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487(7407):330–7. Epub 2012/07/20. doi: 10.1038/nature11252 PMID: 22810696
4. Zhou D, Yang L, Zheng L, Ge W, Li D, Zhang Y, et al. Exome capture sequencing of adenoma reveals genetic alterations in multiple cellular pathways at the early stage of colorectal tumorigenesis. PLoS One. 2013; 8(1):e53310. Epub 2013/01/10. doi: 10.1371/journal.pone.0053310 PMID: 23301059
5. Nikolaev SI, Sotiriou SK, Pateras IS, Santoni F, Sougioultzis S, Edgren H, et al. A single-nucleotide substitution mutator phenotype revealed by exome sequencing of human colon adenomas. Cancer Res. 2012; 72(23):6279–89. Epub 2012/12/04. doi: 10.1158/0008-5472.CAN-12-3869 PMID: 23204322
6. Martinez N, Almaraz C, Vaque JP, Varela I, Derdak S, Beltran S, et al. Whole-exome sequencing in splenic marginal zone lymphoma reveals mutations in genes involved in marginal zone differentiation. Leukemia. 2014; 28(6):1334–40. Epub 2013/12/04. doi: 10.1038/leu.2013.365 PMID: 24296945
7. O'Hayre M, Vazquez-Prado J, Kufareva I, Stawiski EW, Handel TM, Seshagiri S, et al. The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. Nat Rev Cancer. 2013; 13(6):412–24. Epub 2013/05/04. doi: 10.1038/nrc3521 PMID: 23640210
8. Konishi F, Morson BC. Pathology of colorectal adenomas: a colonoscopic survey. J Clin Pathol. 1982; 35(8):830–41. Epub 1982/08/01. PMID: 7107955

9. Marco-Sola S, Sammeth M, Guigo R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012; 9(12):1185–8. Epub 2012/10/30. doi: 10.1038/nmeth.2221 PMID: 23103880

10. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS One. 2009; 4(11):e7767. Epub 2009/11/13. doi: 10.1371/journal.pone.0007767 PMID: 19907642

11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

12. Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, et al. Fast computation and applications of genome mappability. PLoS One. 2012; 7(1):e30377. Epub 2012/01/26. doi: 10.1371/journal.pone.0030377 PMID: 22276185

13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. Epub 2010/07/06. doi: 10.1093/nar/gkq603 PMID: 20601685

14. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; 6(2):80–92. Epub 2012/06/26. doi: 10.4161/fly.19695 PMID: 22728672

15. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–11. Epub 2000/01/11. PMID: 11125122

16. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. Epub 2010/10/29. doi: 10.1038/nature09534 PMID: 20981092

Individualized strategies to target specific mechanisms of disease in malignant melanoma patients displaying unique mutational signatures

Soraya Curiel-Olmo[1,*], Almudena García-Castaño[2,*], Rebeca Vidal[3,4,8], Helena Pisonero[1], Ignacio Varela[4], Alicia León-Castillo[1,5], Eugenio Trillo[6], Carmen González-Vela[1,5], Nuria García-Diaz[1], Carmen Almaraz[1], Thaidy Moreno[4], Laura Cereceda[1], Rebeca Madureira[1], Nerea Martinez[1], Pablo Ortiz-Romero[7], Elsa Valdizán[3,4], Miguel Angel Piris[1,5,#], José Pedro Vaqué[1,4,#]

[1]Cancer Genomics Group, IDIVAL, Instituto de Investigación Marqués de Valdecilla, Santander, Spain

[2]Oncology Service, Hospital Universitario Marqués de Valdecilla, Santander, Spain

[3]Department of Pharmacology, University of Cantabria (UC), Santander, Spain, and Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), ISCIII, Madrid, Spain

[4]Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), CSIC, Universidad de Cantabria, Departamento de Biología Molecular, Universidad de Cantabria, Santander, Spain

[5]Pathology Service, Hospital Universitario Marqués de Valdecilla, Santander, Spain

[6]Plastic Surgery Service Hospital Universitario Marqués de Valdecilla, Santander, Spain

[7]Dermatology Service, Instituto I+12, Hospital Universitario 12 de Octubre, Madrid, Spain

[8]Department of Pharmacology, Medicine School, Complutense University, Madrid, Spain

[*]These authors have contributed equally to this work

[#]Senior author

Correspondence to:
José Pedro Vaqué, **e-mail:** vaquej@unican.es

Keywords: *Targeted therapy, melanoma, BRAF, MAPK, somatic mutations*

Received: *March 25, 2015* ***Accepted:*** *July 13, 2015* ***Published:*** *July 25, 2015*

ABSTRACT

Targeted treatment of advanced melanoma could benefit from the precise molecular characterization of melanoma samples. Using a melanoma-specific selection of 217 genes, we performed targeted deep sequencing of a series of biopsies, from advanced melanoma cases, with a Breslow index of ≥4 mm, and/or with a loco-regional infiltration in lymph nodes or presenting distant metastasis, as well of a collection of human cell lines. This approach detected 3–4 mutations per case, constituting unique mutational signatures associated with specific inhibitor sensitivity. Functionally, case-specific combinations of inhibitors that simultaneously targeted MAPK-dependent and MAPK-independent mechanisms were most effective at inhibiting melanoma growth, against each specific mutational background. These observations were challenged by characterizing a freshly resected biopsy from a metastatic lesion located in the skin and soft tissue and by testing its associated therapy *ex vivo* and *in vivo* using melanocytes and patient-derived xenografted mice, respectively.

The results show that upon mutational characterization of advanced melanoma patients, specific mutational profiles can be used for selecting drugs that simultaneously target several deregulated genes/pathways involved in tumor generation or progression.

INTRODUCTION

Melanoma is a form of cancer whose incidence is rising each year in the developed world, and is second to leukemia in terms of loss of years of potential life from cancers [1]. Despite recent improvements in mortality rates, current deaths from melanoma are estimated to comprise 85% of all cancers affecting the skin. This is corroborated by the poor survival associated with melanoma when diagnosed at an advanced stage [2]. Therefore, the development of effective therapies is a major challenge in this field.

Molecular diagnostics of cancer have proved that targeted therapies can be effective in many cancer settings, as measured by the recent improvement in cancer survival statistics (World Cancer Statistics 2008; ICD-10 C18–21). The use of EGFR inhibitors in lung cancer [3, 4] and Imatinib in chronic myeloid leukemia (CML) patients [5] are two relevant examples. Targeted therapies in melanoma are mostly directed towards inhibiting MAPK-ERK1/2 signaling (MAPK hereafter), [6]. Mutational analyses have recently enabled the detection of up to 50% of malignant melanomas carrying an activating mutation in *BRAF* [7], and these can now be treated with specific B-RAF inhibitors [8]. In the clinic, this targeted approach, even when used in combination with MEK inhibitors, is of limited benefit to patient survival and, after a period, the cancer reappears aggressively [9–11].

From a molecular perspective, data from Next Generation Sequencing (NGS) show that more mutated genes than initially expected participate in tumorigenesis, including that of melanoma [12–14]. This involves a dynamic process of subclonal competition that eventually dictates multifactorial clinical resistance to B-RAF inhibitors, which is dependent on reactivation of MAPK signaling or other proliferative and/or pro-survival pathways [15–17].

Taking advantage of available melanoma NGS data, we characterized biopsies from advanced melanoma patients and cell lines by studying the presence of somatic mutations in a selected group of genes. We thereby detected unique signatures of mutated genes that are potentially associated with specific inhibitors, and explored the effects of case-specific combinations of the latter *ex vivo* and *in vivo*. Guided by individual mutational profiles, tailored combinations of inhibitors simultaneously targeting MAPK-dependent and MAPK-independent signaling were very efficient at inhibiting aberrant melanoma growth assessed in multiple cell lines, and xenografted tumors and biopsies grown in mice. Thus, specific mutational signatures could guide the design of personalized therapies based on the use of specific combinations of drugs that target case-specific pathogenic signaling mechanisms.

RESULTS

A targeted approach to characterizing the mutational status of lesions of advanced melanoma patients

To better understand the molecular character of specific melanoma lesions, we set up a targeted mutational study followed by functional analyses (described in Supplementary Figure 1). The genomic design of this study focused on the coding regions of a specific group of 217 genes that had previously been shown to be mutated in melanoma and selected mainly on the basis of their relevance in melanoma and their association with inhibitors of potential clinical use (see Materials and Methods for further explanation). To test this approach our selection of genes was compared *in silico* with the whole genome/exome sequencing (WGS/WES respectively) data already available for 11 advanced melanoma cell lines and 158 human melanomas (see Materials and Methods, [13, 14, 18, 19]). This comparison revealed an average of 3.74 mutated genes that can participate in multiple targetable signaling pathways, including PLC, MAPK, RTKs (receptors with tyrosine kinase activity), PI3K-mTOR and JAK-STAT (Figure 1 and Supplementary Table I). These results prompted us to study advanced melanoma cases (Breslow index ≥4 mm or metastasis) in 18 clinically characterized patients (clinical characteristics summarized in Supplementary Table II) using a targeted primary ultrasequencing approach, followed by secondary validation analysis (see Materials and Methods for further details). By these methods, an average of 3.4 mutated genes were identified in 11 of the 18 patients, enabled the detection of lesion-specific genes such as *BRAF*, *RAC1*, *KRAS*, *HGF* and *MAPK7*, amongst others. Interestingly, there was a wide range of mutation frequencies and combinations, which perhaps reflects the rich and heterogeneous microclonal composition expected in melanoma tumors (400X average depth/mutation; Table I) [20]. Furthermore, actionable mutations such as $BRAF^{V600E}$ that can guide targeted therapy (using B-RAF inhibitors) were detected in the same melanoma alongside other mutated genes that may also guide therapy (Table I). It is significant that mutations in four patients could not be validated due to limitations of the tissue sample (see Materials and Methods), and that no mutations were identified in three other patients. Thus, this targeted approach could be adopted to identify genomic alterations affecting one or several genes. These may be explored as potential targets for therapy in specific cases of melanoma.

Effects of specific targeted therapy guided by mutational signature

To explore how to use mutational data to design targeted therapies based on specific mutational

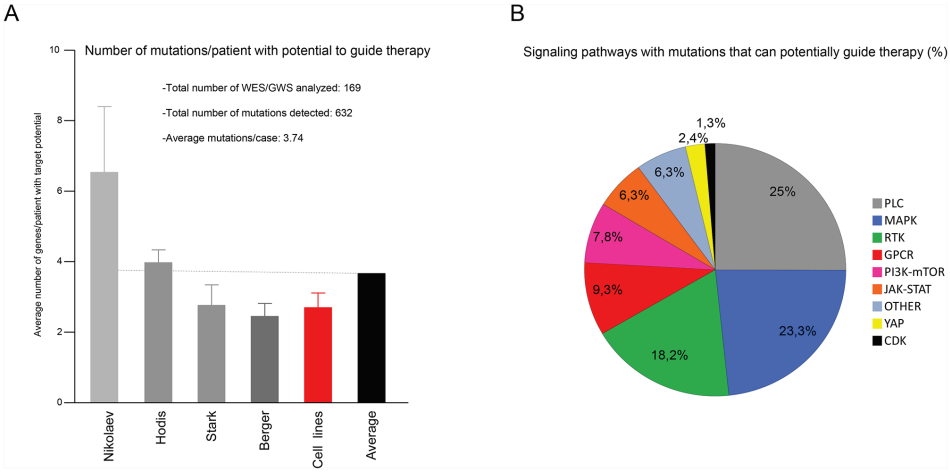


Figure 1: *In silico*-targeted mutational profiling of advanced melanoma patients. A. Meta-analysis showing the average number of mutated genes per case with the potential to guide targeted therapy. Original mutational data from cell lines (red bar) were obtained from the Cancer Cell Line Encyclopedia website (see Material and Methods); mutational data from patients (grey bars) were obtained from Nilolaev [19], Hodis [18], Stark [14] and Berger [13]; Black bar, shows the average frequency of mutations amongst all data sets. **B.** Percentage of hits in A) involved in the indicated signaling pathway.

characterizations, the functional effects of specific combination therapies were studied in advanced melanoma cell lines with known mutational profiles (Supplementary Table III). Taking A375 advanced melanoma cells as an example, we detected and validated mutations in *BRAF*, *FGFR2* and *mTOR* that could reasonably be expected to associate with Vemurafenib (BRAFi (V), hereafter), Vargatef (FGFR2i (Va)) and Everolimus (mTORi (E)). Exponentially growing A375 cells were incubated with increasing concentrations of each inhibitor. This caused a concentration-dependent reduction in cell proliferation from which the IC_{50} of each inhibitor was calculated (Figure 2A and Supplementary Table III). These concentrations were used for subsequent experiments. Next, the mechanistic effects of treatment with each inhibitor (using IC_{50} values in each case) were analyzed in A375 cells that had been serum-starved to provoke the inhibition of the intended mutation-associated downstream signaling. These were assessed by western blot using P-ERK1/2, P-p38 and P-S6 antibodies (Figure 2B).

To discover more about the biological effects of multiple combinations of these inhibitors on proliferation, A375 cells were incubated with IC_{50} concentrations of BRAFi, FGFR2i and mTORi in single, double or triple combinations (blue, green and red lines, respectively, in Figure 2C). The combinatorial treatments were more effective at reducing melanoma cell growth than the monotherapies. The triple combination was the most efficient, and had no non-specific cytotoxic effects (Figure 2C and 2E). These results were confirmed using DNA synthesis as an alternative read-out (Figure 2D and 2E). Thus, under these conditions, a combination of inhibitors guided by a specific mutational signature, simultaneously targeted multiple signaling mechanisms controlling the growth of A375 cells. Analyzing the mechanistic effects of these drug combinations on their associated signaling pathways in this system, showed that treatment with BRAFi inhibited MAPK signaling. However, treatment of A375 cells with the inhibitors mTORi and FGFR2i, alone or in combination (E+Va), had no such effect (Figure 2F and 2G), despite being very effective at inhibiting cell proliferation and DNA synthesis (Figures 2C, and 2E). Thus, using genetically defined inhibitors in this system we can specifically target a combination of MAPK-dependent (V) and MAPK-independent (E+Va) signaling mechanisms that control the malignant growth of A375 melanoma cells. This observation was not confined to these cells and more examples of specific mutational signatures guiding effective combinatorial therapies comprising MAPK-dependent and MAPK-independent mechanisms in other human advanced melanoma cell models are shown in Supplementary Figures 2, 3 and 5.

Increased effects of targeted therapy against an appropriate mutational background

As part of a heterogeneous network of aberrant intracellular signaling, multiple deregulated pathways can participate in the mechanistic control of melanoma growth (Figure 1). We examined whether a combination therapy designed for a specific mutational signature could

Table 1: Validated mutations found in advanced melanoma patients

Patient	Chrom.	Position	Ref_base	Mut_base	Total_Cov	Observed_Freq	Gene_ID	p.Annot
2	Chr4	126370467	G	A	59	0.58	**FAT4**	p.E2766K
2	Chr2	141986902	C	T	83	0.18	**LRP1B**	p.E234K
4	Chr5	167881030	GGA	-	464	0, 53	**WWC1**	p.V861 VE > V
4	Chr5	150923714	T	C	724	0.39	**FAT2**	p.N2325S
4	Chr17	7578490	A	C	884	0.08	**TP53**	p.V147G
5	Chr12	25380269	C	G	177	0, 16	**KRAS**	p.E63D
5	Chr1	9781272	G	C	424	0, 07	**PIK3CD**	p.G593R
6	Chr7	140453136	A	T	39	0, 69	**BRAF**	p.V600E
6	Chr3	3134041	T	A	100	0, 3	**IL5RA**	p.E287D
6	Chr19	15290897	C	T	1200	0, 27	**NOTCH3**	p.G1105S
8	Chr7	6426892	C	T	233	0, 67	**RAC1**	p.P29S
8	Chr7	81346551	G	A	243	0, 61	**HGF**	p.R468C
8	Chr7	140453136	AC	CT	61	0, 54	**BRAF**	p.V600R
8	Chr3	155311800	C	T	145	0, 49	**PLCH1**	p.G104S
8	Chr18	51053024	CC	TT	119	0, 48	**DCC**	p.S1383F
8	Chr5	89910783	C	T	60	0, 4	**GPR98**	p.R52C
8	Chr1	23233289	T	G	76	0, 39	**EPHB2**	p.Y659D
8	Chr13	28886195	C	T	190	0, 32	**FLT1**	p.E1143K
8	Chr18	50432552	C	T	163	0, 37	**DCC**	p.P184L
8	Chr2	170136083	C	T	38	0, 29	**LRP2**	p.G455D
8	Chr11	46406865	G	A	1048	0, 27	**CHRM4**	p.P415S
9	Chr7	31855742	C	T	66	0, 24	**PDE1C**	p.E597K
9	Chr7	140453135	CA	TT	55	0, 11	**BRAF**	p.V600E
12	Chr7	140453136	A	T	142	0, 1	**BRAF**	p.V600E
12	Chr6	32170007	C	T	302	0, 06	**NOTCH4**	p.G1201R
13	Chr2	166905414	C	T	321	0, 15	**SCN1A**	p.G337E
13	Chr5	55247832	G	A	362	0, 12	**IL6ST**	p.L542F
16	Chr16	9858517	C	T	450	0, 06	**GRIN2A**	p.E962K
17	Chr7	140453136	A	T	326	0, 44	**BRAF**	p.V600E
17	Chr4	126239082	C	T	1417	0, 27	**FAT4**	p.L506F
17	Chr17	19285669	C	T	1646	0, 23	**MAPK7**	p.P546S
17	Chr18	50450115	C	T	371	0, 2	**DCC**	p.P246S
18	Chr10	96014751	C	T	651	0, 13	**PLCE1**	p.P1167S
18	Chr2	21233091	G	A	167	0, 08	**APOB**	p.H2217Y
18	Chr13	29001438	C	T	824	0, 08	**FLT1**	p.E432K
18	Chr7	151851165	CC	TT	876	0, 08	**KMT2C**	p.G4069N
18	ChrX	112035176	C	T	420	0, 07	**AMOT**	p.E195K

Patient: Patient number; Chrom: Chromosome number; Position: Genomic location of the mutation in the chromosome; Ref_base: normal nucleotide; Mut_base: mutated nucleotide; Total Cov: Number of reads analyzed at each position; Observed_Freq: Frequency of mutation; Gene_ID: Gene name; pAnnot: Aminoacid change.

be more effective when used against a genetically appropriate background. A group of melanoma cell lines harboring unique mutational signatures (Supplementary Table III) was treated in parallel with the genetically defined inhibitors for A375 cells, BRAFi, FGFR2i and mTORi, alone or in combination. In general, each treatment was more effective in A375 cells than in the other melanoma cell lines, with the possible exception of mTORi used alone (Figure 3A). The triple drug combination treatment (V+E+Va) simultaneously targeting MAPK-dependent and MAPK-independent proliferation mechanisms produced greater inhibition of A375 cell growth than the others (Figure 3A and 3B). Likewise, other specific treatments based on the combination of genetically defined inhibitors in other melanoma cell lines showed a stronger effect than that of A375 cells (Supplementary Figure 4A and 4B). More examples of specific mutational signatures guiding more effective combination therapies when used against an appropriate mutational background are shown in Figure 5E and Supplementary Figure 5C. In summary, based on specific mutational signatures, a specific treatment consisting of a combination of genetically defined inhibitors may have stronger anti-melanoma activity when used against an appropriate genomic background.

In vivo effects of a targeted therapy that combines MAPK-ERK-dependent and MAPK-ERK-independent mechanisms of inhibition

To study the *in vivo* effects of targeted therapy oriented by a specific mutational profile, xenografted tumors from A375 melanoma cells were generated in BALB/C mice ($nu^{-/-}/nu^{-/-}$). Once grown to a volume of approximately 100 mm^3, tumors were assigned to four comparable groups and treated daily with vehicle, a MAPK-dependent inhibitor (BRAFi), a MAPK-independent combination of inhibitors (FGFR2i+ mTORi), or a triple combination of the latter (BRAFi+FGFR2i+mTORi). As shown in Figure 4A and 4B, both treatments used independently reduced tumor growth to a similar extent. However, the triple combination (V+Va+E) proved most effective at reducing melanoma growth in this system. Once the experiment was finished, the remaining growth potential of these tumors was characterized by studying Ki67 and the mitotic index in tumor sections. As might be expected, a marked decrease in both proliferation markers in those tumors treated with the combination of MAPK-dependent and MAPK-independent inhibitors (Figure 4C-4G) was observed, thereby confirming *in vivo* our previous findings in cultured cells (Figures 2D, 2E, and 5).

A pre-clinical example of targeted therapy guided by a specific mutational signature in patient 17

These findings were challenged by integrating the study of a freshly resected biopsy from patient 17 (Table I and Supplementary Table II) in the working pipeline (illustrated in Figure 5A and Supplementary Figure 1). First, a fragment of the biopsy was characterized which enabled the detection of somatic mutations in the *BRAF* and *MAPK7* (*ERK5* hereafter) genes (Table I and Figure 5A), which were associated with specific inhibitors like Vemurafenib (BRAFi (V)) and XMD-8–92 (ERK5i (X)).

Second, freshly isolated melanocytes (MELANOMA17 cells hereafter) were also inspected for the presence of mutations of *BRAF* and *ERK5* (Figure 5B) and for the expression of well-known melanoma markers such as S100A and MCSP (Supplementary Figure 6A and 6B). Once characterized, MELANOMA17 cells were incubated with increasing concentrations of BRAFi and ERK5i, and the IC_{50} of each was calculated (Supplementary Table III). Treatment with BRAFi and ERK5i inhibited B-RAF and ERK5-dependent signaling, assessed by western blot, in starved MELANOMA17 cells (Figure 5C). Simultaneous treatment with both inhibitors was more effective at reducing MELANOMA17 cell proliferation than either inhibitor alone (Figure 5D). Furthermore, and consistent with our previous observations in multiple cell lines, the combination treatment (BRAFi+ERK5i) was also more effective in cells with an appropriate mutational signature (MELANOMA17 cells) when compared with a panel of other melanoma cell lines (Figure 5E), with the possible exception of A375 cells, which were highly sensitive to treatment with the BRAFi dose used (compare the IC_{50} values for MELANOMA17 with those of A375 cells in Supplementary Table III). This combination therapy also consisted of MAPK-dependent (BRAFi) and MAPK-independent (ERK5i) mechanisms, which successfully suppressed the aberrant growth of MELANOMA17 cells (Figure 5F).

Third, another fragment of the biopsy was implanted in NSG mice and allowed to grow until four tumor-comparable groups of mice could be established. The groups of mice were then treated with vehicle, BRAFi, ERK5i, or a combination of the two (V+X). When used separately, the two inhibitors were able to suppress xenografted MELANOMA17 tumor growth to a similar extent, but growth inhibition was more effective when used in combination (Figure 5G-5I).

Thus, it is possible to study the effects of personalized therapies, guided by targeted mutational profiling of advanced melanoma patients, in pre-clinical *ex vivo* and *in vivo* models using freshly resected material from each lesion.

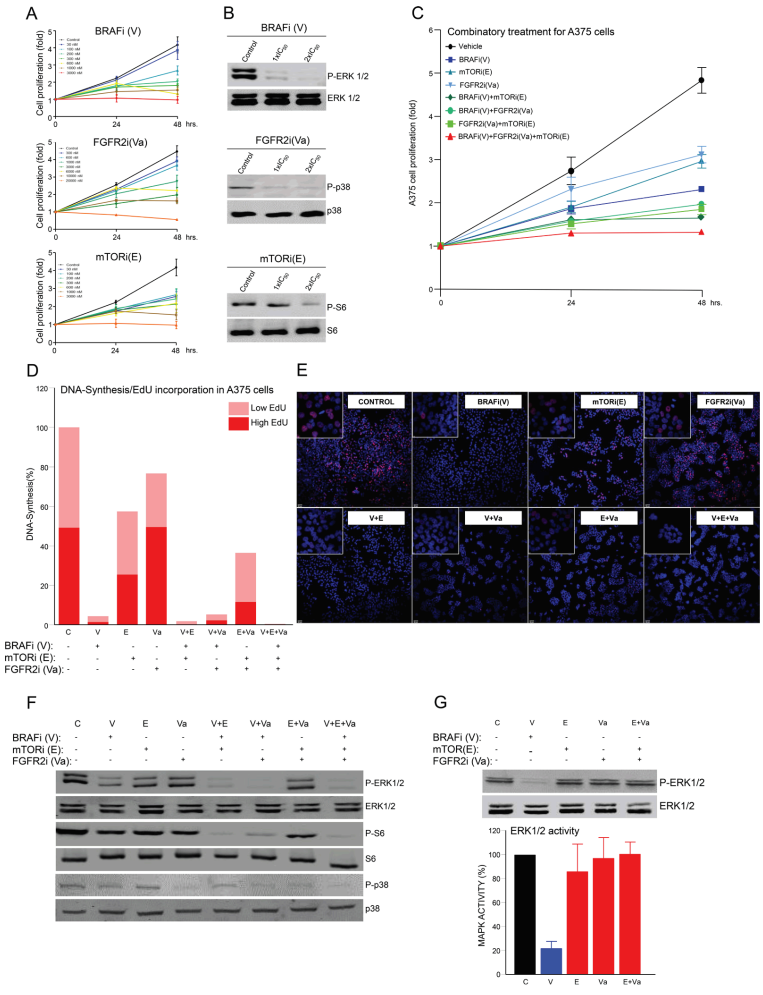


Figure 2: Effects of specific targeted therapy guided by mutational signature. A. Proliferation analysis of A375 cells at 0, 24 and 48 h. Cells were seeded in 96-well plates and treated with the indicated concentrations of each inhibitor: B-RAFi (V: Vemurafenib), FGFR2i (Va: Vargatef), and mTORi (E: Everolimus). **B.** Western blots using whole cell lysates from starved A375 cells incubated for 1 h with control vehicle (DMSO) or the indicated concentration of each inhibitor. The figure shows a representative experiment using P-ERK1/2, ERK1/2, P-p38, p38, P-S6 and S6 antibodies, as indicated. **C.** Proliferation analysis of A375 cells in the same conditions as in A), but incubated with control vehicle (DMSO) or the IC_{50} concentration of the indicated inhibitor alone (blue lines), or in a double (green lines) or triple combination (red line). $N = 6$. Error bars show the SEM. **D.** DNA synthesis using Click-iT® EdU in exponentially growing A375 cells seeded in an 8-well glass and incubated for 48 h with control vehicle (DMSO) or the indicated inhibitor or combination of inhibitors, as in C). Graph bars show percentage of low (clear red) or high (intense red) EdU-stained cells in three photographic fields from a representative experiment. **E.** Representative pictures of each treatment condition showing the nucleus of the total number of cells (blue dots) and EdU-positive cells (red dots). **F** and **G.** Western blots of whole cell lysates of the indicated cells. Cells were starved overnight and incubated for 1 h with control vehicle (DMSO), or the indicated inhibitor, or a combination of inhibitors under the same conditions as in C). Figures show representative experiments using P-ERK1/2, ERK1/2, P-p38, p38, P-S6 and S6 antibodies, as indicated. Bar graphs show the values of three independent experiments in G). Error bars indicate the SEM.

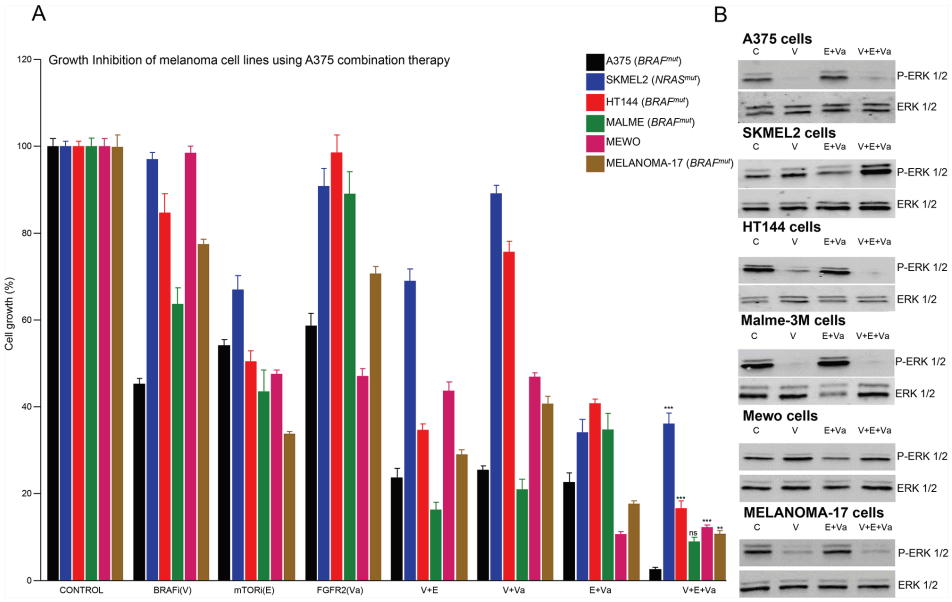


Figure 3: Increased effects of targeted therapy against an appropriate mutational background. A. Proliferation analysis of exponentially growing A375 (BRAF+), SKMEL2 (NRAS+), HT144 (BRAF+), MALME (BRAF+), MEWO and MELANOMA17 (BRAF+) cells treated with vehicle (DMSO) or the IC_{50} concentration (calculated for A375 cells) of the indicated inhibitor alone, in a double or triple combination for 48 h. $N = 6$. Error bars show the SEM. **B.** Western blots of whole cell lysates of the indicated cells. Cells were starved overnight and incubated for 1 h with control vehicle (DMSO), or the indicated inhibitor, or a combination of inhibitors under the same conditions as in A). Figure shows a representative experiment.

DISCUSSION

Metastatic melanoma provides an instructive example of the development of rationalized therapies guided by molecular diagnostics. Mechanistically, targeted therapy mainly involves the inhibition of the MAPK signaling pathway by using BRAF or MEK inhibitors, alone or in combination [21], as suggested by: A) activating mutations in BRAF and NRAS oncogenes in 48% and 15% of all diagnosed melanomas, respectively [22]; and B) multiple MAPK reactivation mechanisms that confer resistance to BRAF inhibitors [23–25]. This has improved the clinical management of those patients with mutated *BRAF*, whereby targeted inhibition of aberrant MAPK signaling can increase their OS by up to 11.4 months, although, from a different perspective, it still offers a limited benefit to these patients [10, 26, 27]. To explain this, we can hypothesize that, as part of an intricate network of transforming mechanisms in melanoma, this disease simultaneously uses multiple oncogenic mechanisms, such as, for example, PI3K, MET and GNAQ [28, 29], that, along with MAPK signaling, act as mechanistic drivers of the disease and promote progression or resistance to therapy. In this regard, data from genetically defined melanoma models now show that a rational combination of MAPK and PI3K inhibitors can improve the effects of therapy when used against specific genomic backgrounds [30]. Moreover, patients with specific mutations gained greater benefits when treated with immunotherapy [31, 32]. Thus, better characterization of advanced melanoma lesions could improve our ability to treat this disease through the use of specific therapies that simultaneously target multiple signaling mechanisms.

From a molecular perspective, melanoma is a very heterogeneous disease in which up to 1, 500 somatic mutations may be harbored in the coding exons of a single lesion [13]. This work studies mutations in 217 genes previously shown to be mutated in melanoma [13, 14, 18, 19, 33, 34] *in silico* by comparing them with mutations in 11 cell lines and 158 human melanomas, and *ex vivo* by characterizing 18 lesions from advanced melanoma patients. Under all conditions, genes like *BRAF*, *RAC1*, *FGFR2* and *IL6R* were mutated at varying frequencies, occurring as part of unique mutational signatures comprising specific combinations of mutated genes that have the potential to participate in multiple signaling

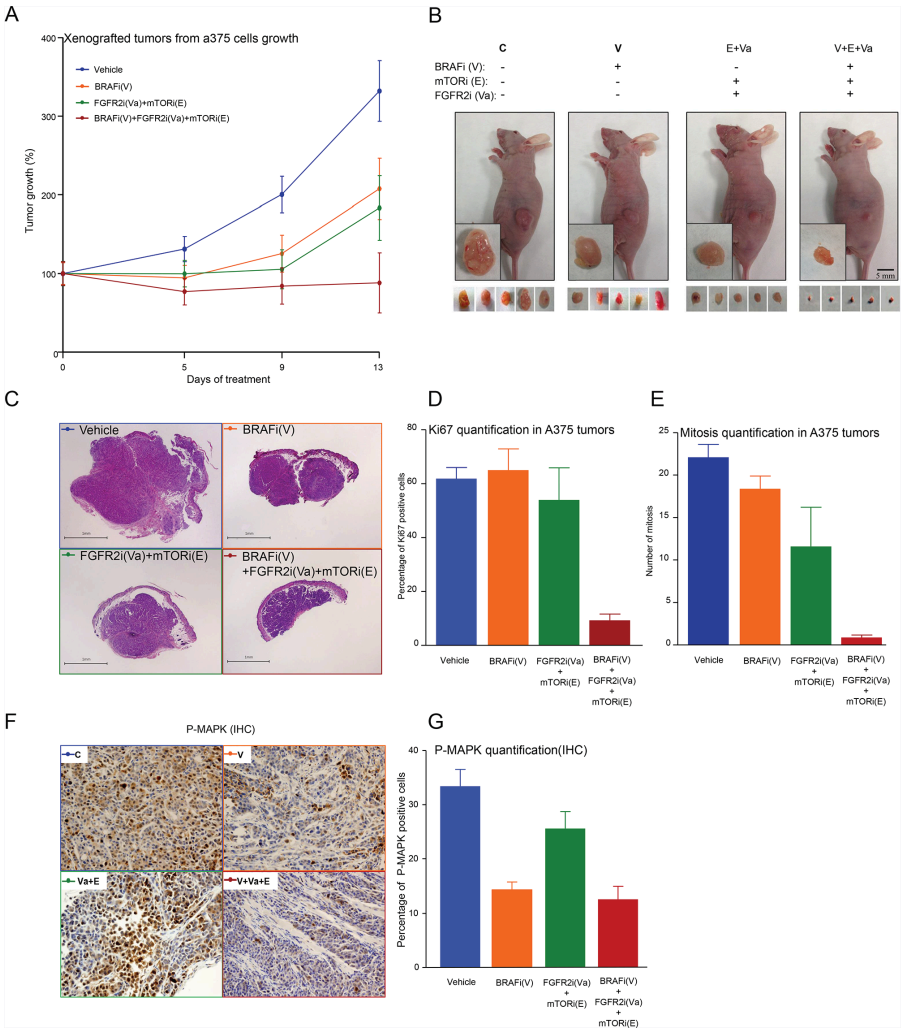


Figure 4: *In vivo* effects of a targeted therapy combining MAPK-ERK-dependent and MAPK-ERK-independent mechanisms of inhibition. A. Xenografted tumor growth-derived A375 cells injected subcutaneously in 48 BALB/C nude mice. Tumor size was monitored until a volume of 100 mm^3 was obtained, whereupon mice were assigned to four treatment groups: 1) Control (DMSO, blue line); 2) BRAFi (V) (orange line); 3) FGFR2i (Va) + mTORi (E) (green line); and 4) BRAFi (V) + FGFRi (Va) + mTORi (E) (red line). Mice were treated daily as indicated (see Materials and Methods for further details) and tumor volumes were measured until day 13, at which point the experiment was ended. Data were obtained from the 12 control, 11 (V), 7 (Va+E) and 10 (V+Va+E) mice that survived the entire process. Error bars indicate the SEM. **B.** Representative pictures illustrating the effects of the indicated treatment on the xenografted tumors that had been resected or were still in the mice. **C.** H&E staining of representative tumor sections from five representative mice for each treatment condition. Tumor sections were analyzed for Ki67-positive staining **D.** or by the number of mitoses **E.** Data are averages of five section cuts in each mouse. Error bars indicate the SEM. **F.** Immunohistochemical (IHC) analysis in tumors corresponding to the indicated treatment, as in C), using an anti-phospho ERK antibody stain. **G.** Tumor sections were analyzed for phosphor-ERK-positive staining. Results are the averages of five section cuts per mouse. Error bars indicate the SEM.

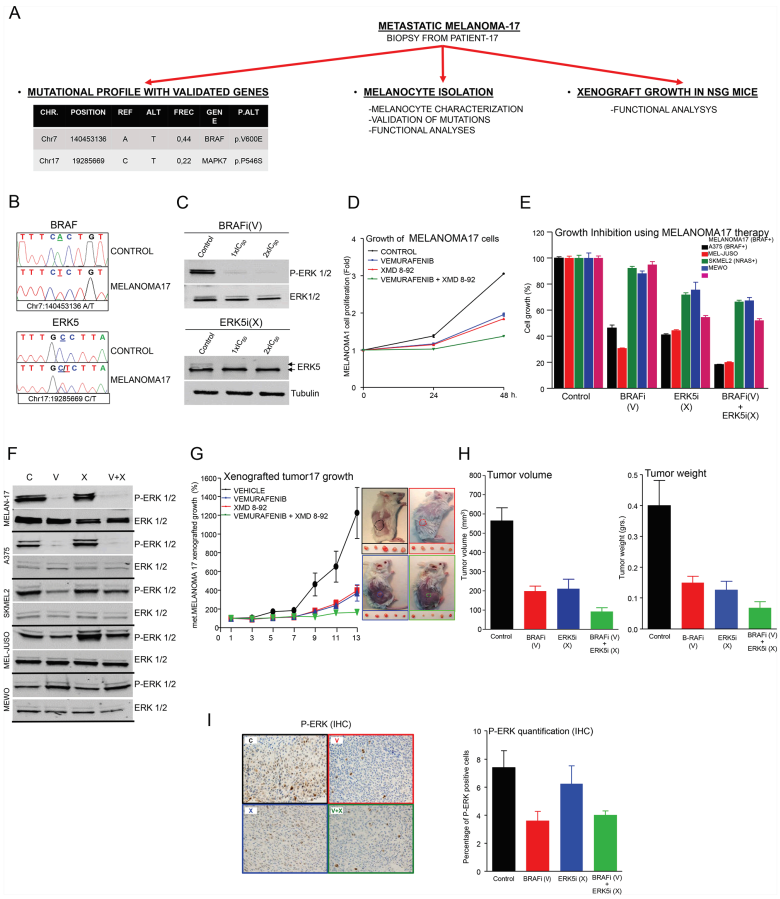


Figure 5: A pre-clinical example of targeted therapy guided by a specific mutational signature in melanoma patient 17. A. Schematic representation of the work performed with a freshly resected biopsy from patient 17. **B.** Sanger sequencing of *BRAF* (above) and *MAPK7* (*ERK5*; below) oncogenes in genomic DNA from control cells or isolated melanocytes from patient 17 (MELANOMA17 cells). **C.** Western blots of whole-cell lysates from starved MELANOMA17 cells incubated for 1 h with control vehicle (DMSO) or the indicated concentration of each inhibitor BRAFi (V: Vemurafenib) or ERK5i (X: XMD-8–92). The figure shows a representative experiment using P-ERK1/2, ERK1/2, ERK5 and tubulin antibodies, as indicated. **D.** Proliferation analysis of MELANOMA17 cells at 0, 24 and 48 h. 3×10^3 cells/well were seeded in 96-well plates and treated with control (DMSO) (black line), or an IC_{50} concentration of B-RAFi (V) (blue line) or ERK5i (X) (red line), alone or in combination (green line). $N = 6$. Error bars show SEM. **E.** Proliferation analysis of MELANOMA17, A375, MEL JUSO, SKMEL2, and MEWO cells at 0, 24 and 48 h, under the same conditions as in D). $N = 6$. Error bars show the SEM. **F.** Western blots using whole cell lysates of the indicated cells. Cells were starved overnight and incubated for 1 h with control vehicle (DMSO), or the indicated inhibitor, or a combination of inhibitors using the same concentrations as in E). Representative experiment using anti-P-ERK1/2 and anti-ERK1/2 antibodies. **G.** Tumor growth derived from 2-mm^3 MELANOMA17-derived tumor fragments implanted subcutaneously in 30 NSG mice (Jackson Laboratories). Tumors were monitored until they attained a volume of 100 mm^3, whereupon mice were assigned to four comparable treatment groups: 1: Control (DMSO, black line), 2: BRAFi (V) (blue line), 3: ERK5i (X) (red line) and 4: BRAFi (V) + ERK5i (X) (green line). Mice were treated daily as indicated (see Materials and Methods for further details) and tumor volumes were monitored until day 13, at which point the experiment was ended. Data were obtained from five survivor mice from each treatment group. Error bars indicate the SEM. The figure shows a representative image from treated tumors that were still in the mice (above) or had been freshly resected (below). **H.** Bar graph of average changes in tumor volume (left) and mass (right) for each treatment condition. $N = 5$. Error bars indicate the SEM. **I.** Examples of IHC analysis of tumors corresponding to the treatment indicated in D) using anti-phospho ERK antibody staining. Bar graphs show results for tumor sections analyzed for phosphor-ERK-positive staining. Data are the averages of five section cuts per mouse. Error bars indicate the SEM.

pathways and to be associated with specific inhibitors. Functionally, the effects of combination therapies guided by specific mutational signatures were analyzed in multiple melanoma cells harboring unique mutational signatures. Those treatments that simultaneously targeted MAPK-dependent and MAPK-independent signaling were most effective at reducing melanoma growth both *ex vivo* and *in vivo*. These observations can be aligned with work from other laboratories, showing that to promote transformation in melanocytes, aberrant MAPK-signaling elicited by BRAF or NRAS oncoproteins requires the active collaboration of other oncogenes, such as *PI3K*, *RAL-GDS*, *GNAS* or *C-MYC*, that can participate in alternative signaling pathways [35–38]. Thus, a combination of genetically defined inhibitors targeting multiple signaling pathways could be more effective against specific cases of malignant melanoma. We might expect targeting well-known melanoma mechanisms to affect the growth of melanoma cells in general, and this can indeed be observed in our data. Nevertheless, combination therapies guided by specific mutational signatures were most effective when used against an appropriate mutational background. Furthermore, different *BRAF*-mutated cell lines, each with an individual mutational signature, had different sensitivities to BRAF inhibition (Supplementary Table III). Finally, our data strongly suggest that combining several mutationally selected inhibitors can specifically block important mechanisms that participate in the control of aberrant melanoma cell growth and in the finely tuned cellular decision to activate DNA synthesis (Figures 2B, 2C, 2E, 4A, 4D and 4E). This rules out the possibility that the results were a consequence of nonspecific cytotoxic activities.

Starting with freshly resected material from a metastatic lesion (patient 17) and trying to match the timing with the clinic, a validated mutational profile was obtained within two weeks of resection. This data enabled the study of a combination therapy based on inhibitors with MAPK-dependent (BRAFi) and MAPK-independent (ERK5i) mechanisms of action in isolated MELANOMA17 cells and in xenografted tumors grown in mice. These gave the best results when combinatorial approaches were used. Of course, this study provides just one example of what targeted characterization of specific lesions might offer by way of diagnostic possibilities for human melanoma in the near future. Considering its potential applicability in routine clinical practice, this approach would require several limitations to be overcome. This would entail: 1) establishing efficient protocols to collect, manipulate and characterize specific lesions that are representative of the various steps of the disease; 2) managing the toxicity due to drug combinations; and 3) dealing with tumor heterogeneity and interactions with the immune system that may be responsible for the eventual resistance acquired after combination treatments. However, there is much scope for studying novel strategies for targeted therapy following a molecular rationale, particularly in a disease like advanced melanoma that offers a limited prospect of survival to patients who are suffering from it (Supplementary Figure 7).

In summary, by adopting targeted approaches we can envisage working with specific signatures of mutated genes that can: 1) help characterize individual lesions in advanced melanoma patients; 2) guide the use of specific inhibitors rationally combined in individualized therapies to target case-specific mechanisms of melanocytic transformation. In this work, a rational combination of genetically defined inhibitors simultaneously targeting MAPK-dependent and MAPK-independent signaling mechanisms showed improved biological outcomes with respect to the malignant growth of specific advanced melanomas.

MATERIALS AND METHODS

Cells and reagents for tissue culture

Eight human advanced melanoma cell lines were used. A375 (CRL-1619™), SK-MEL-28 (HTB-72™), SK-MEL-2 (HTB-68™), MALME-3M (HTB-64™), MEWO (HTB-65™) and HT-144 (HTB-63™) cells were obtained from the American Type Cell Collection (ATCC, Rockville, MD). MEL-JUSO (ACC 74) was obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany). Genomic data from these cells, including those of the somatic mutations detected in this study, are publicly available at the Broad-Novartis Cancer Cell Line Encyclopedia website (CCLE:http://www.broadinstitute.org). MELANOMA17 cells were established from a primary biopsy sample, as explained in the Supplementary Methods. Commercial cell lines were cultured as recommended by ATCC or DSMZ and incubated with inhibitors, as described in the Supplementary Methods.

Cell proliferation and DNA synthesis assays

Cells growing exponentially to approximately 50% confluence in T96 well plates were incubated with the specific inhibitors while keeping the total amount of DMSO (0.5%) constant. Cellular proliferation was evaluated using AlamarBlue reagent (Life Technologies) and colorimetric changes were quantified using the Synergy™ HTX Multi-Mode Microplate Reader (Biotek). To assess the effects on DNA synthesis, cells were grown in a Millicell EZ SLIDE 8-well glass (Merck Millipore, PEZGS0816) and after treatment with specific inhibitors were incubated for a further 2 h with Click-iT® EdU (Alexa Fluor® 594 Imaging Kit; Life Technologies, C10339). Immediately afterwards, cells were prepared for microscopy following the manufacturer's specifications (see Supplementary Methods for further explanation).

Cell images were captured with a Nikon A1R confocal microscope with Plan Apo 10x/0.45NA and Plan ApoVC 60x/1.40NA objectives.

Genomic DNA samples

Matched tumoral and non-tumoral material was obtained from 18 patients diagnosed with advanced melanoma and who were being monitored by the Oncology Department of the Hospital Universitario Marqués de Valdecilla (HUMV; see clinical characteristics in Table I). Tumoral DNA samples were obtained from freshly frozen tissue samples taken at the time of diagnosis, and matched non-tumoral DNA was extracted from saliva or peripheral blood neutrophils. We designed an intra-subject observational study of patients diagnosed with advanced melanoma and with a Breslow index of ≥4 mm, and with a loco-regional infiltration in lymph nodes or presenting distant metastasis. The study, the patient information sheet, and the informed consent form were approved by the Ethics Committee of the HUMV.

Enrichment library design, preparation, sequencing and variant calling

Genomic DNA samples were processed using the Qubit® dsDNA BR Assay Kit (Life Technologies) and quantified using Qubit 2.0 apparatus (Life Technologies). The DNA enrichment library was prepared using a specifically designed HaloPlex Target Enrichment System Kit for this melanoma study (Design ID: 00912-1339502780, Agilent Technologies) following the manufacturer's instructions. The design focused on the coding regions of a group 217 genes known to be mutated in melanoma, and which were selected because they were: A) genes of known relevance in melanoma, including BRAF, NRAS [7], and EGFR [18, 19]; B) genes that may be associated with pharmacological inhibitors of potential clinical use, such as FGFR2, KIT and ERBB4 [18, 20, 21]; and C) genes that may be involved in chromatin architecture (ARID1A and DNMT3A [14]), intracellular signaling (MEK1 [22]), or transcription (NFATC2 [22]). Briefly, 400 ng of genomic DNA was digested with the specific cocktail of restriction enzymes provided in the kit. Digested DNA was then hybridized to a probe for target enrichment, indexed and captured. Each DNA was then amplified by PCR at Tm = 60°C, for 18 cycles, using a Herculase II Fusion Enzyme kit (Agilent Technologies). Next, amplified target libraries were purified using an Agencourt AMPure XP Kit (Beckman Coulter Genomics), following the manufacturer's guidelines, and quantified with Qubit 2.0 apparatus (Life Technologies), using the Qubit® dsDNA HS Assay Kit (Life Technologies). They were also analyzed in parallel by capillary electrophoresis in a 2100 Bioanalyzer (Agilent Technologies), using High Sensitivity DNA reagents and chip Kits (Agilent Technologies). Libraries were sequenced at the Instituto de Medicina Genómica (IMEGEN, Valencia University, Spain) with a MiSeq Personal Sequencer (Illumina). The process of somatic mutation identification described in the Supplementary Methods.

Somatic mutation identification

Sequencing data were aligned against the human reference genome (hg19) using the BWA aligner [39]. The alignment was refined using SAMTOOLS fixmate (PMID: 19505943) and PICARD TOOLS cleanSam tools (http://broadinstitute.github.io/picard.). Local realignment of insertions and deletions (indels) was then performed using the GATK suite [40] before final sorting and indexing. The RAMSES application (PMID: 24296945), written in-house, was used to detect nucleotide substitutions. Small indels were identified using Pindel [41] in paired tumor-normal mode. For greater specificity, only simple insertion and deletion events of fewer than 10 bp were selected. An in-house perl script filter was used to extract high-quality indels: considering the high sequence coverage obtained in these samples, only those indels with a minimum coverage of 20 reads in both tumor and normal samples, and with a minimum frequency of 10% of the reads and a minimum of five independent reads supporting the event in the tumor sample, and with no evidence in the normal sample, were considered. All potential somatic mutations were filtered using the dbSNP132 and 1000 Genomes Project mutation databases and the functional consequence at the protein level was annotated according to the Ensembl database using an in-house perl script based on the Ensembl database API.

Validation analysis

Genomic DNA was amplified using the specific oligonucleotides described in Supplementary Table IV. All amplicons from the same patient were mixed in a tube and each sample was quantified by Qubit 2.0 (Life Technologies), using the Qubit® dsDNA BR Assay Kit (Life Technologies). MELANOMA17 cells were monitored by Sanger sequencing for the presence of mutations in *BRAF* and *MAPK7* (see the supplementary methods for further details).

Statistics

Unless otherwise specified, all experiments were done in independent triplicates and all numerical data were summarized as the average of the values ± the standard error of the mean (SEM) using GraphPad PRISM. Levels of statistical significance are indicated as follows: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

Western blot

Cells growing exponentially at approximately 70% confluence were treated under the desired conditions. Cells were starved overnight (unless otherwise stated), treated with the appropriate inhibitor and lysed as described in [42]. Whole cell lysates were subjected to acrylamide SDS-PAGE, using standard procedures, then transferred onto a nitrocellulose support membrane (Immobilon, Millipore) and western blotted. The primary and secondary antibodies and the data collection method are described in the Supplementary Methods.

Mice and reagents for *in vivo* studies

BALB/c Nude mice CAnN.Cg-Foxn1nu/Crl (Charles River) were injected with 6×10^6 A375 melanoma cells in the subcutaneous dorsal area. Approximately one week later, the tumor reached a volume of about 100 mm^3, at which point mice were assigned to four tumor size-comparable groups of 12 animals and treated as described in the Supplementary Methods.

Fresh tissue from patient 17, who had been diagnosed with metastatic melanoma (Table I and Figure 5), was minced and xeno-injected into NOD. Cg-Prkdcscid Il2rgtm1Wjl/SzJ mice (commonly known as NOD scid gamma (NSG) mice) (Charles River). Briefly, the animals were anesthetized using ketamine (75 mg/kg) and medetomidine (1.0 mg/kg) and a piece of tumor was inserted in the subcutaneous dorsal area through a small incision in the skin and allowed to grow. Next, mice were sacrificed (as described in supplementary methods) and tumors were collected and minced into pieces of about 2 mm^3 and reimplanted into the experimental group of mice. When these mice had grown tumors of an approximate volume of 100 mm^3, they were distributed among four groups of six mice, each with comparable tumor volumes and treatments were started, as described in Figure 5 and in the Supplementary Methods.

ACKNOWLEDGMENTS

We are indebted to the patients who have contributed to this study. We especially thank Dr. Fidel Madrazo, José Revert and Carolina Santa Cruz from IDIVAL, and the staff members of the Biobank and the Pathology Service at HUMV, for their exceptional work in sample collection and organization. We would also like to thank the Santander Super-Computation Service for their support. We would also like to acknowledge Dr. Piero Crespo and Dr. Javier León (IBBTEC, Santander, Spain), and Dr. J. Silvio Gutkind and Dr. Xiaodong Feng (NIH, Bethesda, MD, USA) for helpful discussions and comments about this work.

FUNDING

This work was supported by grants from the ISCIII, co-financed by the European Union (FEDER) (PI12/00357) and a Ramón and Cajal research program from MINECO (RYC-2013-14097) to JPV, and from the ISCIII (RTIC; RD06/0020/0107, RD012/0036/0060) and the Asociación Española Contra el Cáncer (AECC) to MAP. The work was also supported by a research award from LUCHAMOS POR LA VIDA to JPV and AGC. I.V. is supported by the Ramón and Cajal research program.

CONFLICTS OF INTEREST

All authors declare no conflict of interest except MAP.

MAP has the following conflicts of interest: Takeda-advisory board. Novartis, Amgen and Roche: Speaker bureau.

REFERENCES

1. Tran TT, Schulman J, Fisher DE. UV and pigmentation: molecular mechanisms and social controversies. Pigment Cell Melanoma Res. 2008; 21:509–16.
2. Erdei E, Torres SM. A new understanding in the epidemiology of melanoma. Expert Rev Anticancer Ther. 2010; 10:1811–23.
3. Mitsudomi T, Yatabe Y. Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer. FEBS J. 2010; 277:301–8.
4. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, Gemma A, Harada M, Yoshizawa H, Kinoshita I, Fujita Y, Okinaga S, Hirano H, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. N Engl J Med. 2010; 362:2380–8.
5. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med. 2001; 344:1031–7.
6. Kwong LN, Davies MA. Targeted therapy for melanoma: rational combinatorial approaches. Oncogene. 2014; 33:1–9.
7. Curtin JA, Fridlyand J, Kageshita T, Patel HN, Busam KJ, Kutzner H, Cho KH, Aiba S, Brocker EB, LeBoit PE, Pinkel D, Bastian BC. Distinct sets of genetic alterations in melanoma. N Engl J Med. 2005; 353:2135–47.
8. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, Hogg D, Lorigan P, Lebbe C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med. 2011; 364:2507–16.

9. Flaherty KT, Puzanov I, Kim KB, Ribas A, McArthur GA, Sosman JA, O'Dwyer PJ, Lee RJ, Grippo JF, Nolop K, Chapman PB. Inhibition of mutated, activated BRAF in metastatic melanoma. N Engl J Med. 2010; 363:809–19.
10. Sosman JA, Kim KB, Schuchter L, Gonzalez R, Pavlick AC, Weber JS, McArthur GA, Hutson TE, Moschos SJ, Flaherty KT, Hersey P, Kefford R, Lawrence D, et al. Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. N Engl J Med. 2012; 366:707–14.
11. Shi H, Moriceau G, Kong X, Lee MK, Lee H, Koya RC, Ng C, Chodon T, Scolyer RA, Dahlman KB, Sosman JA, Kefford RF, Long GV, et al. Melanoma whole-exome sequencing identifies (V600E)B-RAF amplification-mediated acquired B-RAF inhibitor resistance. Nat Commun. 2012; 3:724.
12. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A, Shefler E, Ramos AH, Stojanov P, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333:1157–60.
13. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, Zhang H, Zeid R, Ren X, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012; 485:502–6.
14. Stark MS, Woods SL, Gartside MG, Bonazzi VF, Dutton-Regester K, Aoude LG, Chow D, Sereduk C, Niemi NM, Tang N, Ellis JJ, Reid J, Zismann V, et al. Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. Nat Genet. 2012; 44:165–9.
15. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008; 40:722–9.
16. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A. 2008; 105:13081–6.
17. Shi H, Hugo W, Kong X, Hong A, Koya RC, Moriceau G, Chodon T, Guo R, Johnson DB, Dahlman KB, Kelley MC, Kefford RF, Chmielowski B, et al. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. Cancer Discov. 2014; 4:80–93.
18. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos AH, Lawrence MS, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150:251–63.
19. Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, Harshman K, Guipponi M, Bukach O, Zoete V, Michielin O, Muehlethaler K, Speiser D, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet. 2012; 44:133–9.
20. Quintana E, Shackleton M, Foster HR, Fullen DR, Sabel MS, Johnson TM, Morrison SJ. Phenotypic heterogeneity among tumorigenic melanoma cells from patients that is reversible and not hierarchically organized. Cancer Cell. 2010; 18:510–23.
21. Salama AK, Flaherty KT. BRAF in melanoma: current strategies and future directions. Clin Cancer Res. 2013; 19:4326–34.
22. Colombino M, Capone M, Lissia A, Cossu A, Rubino C, De Giorgi V, Massi D, Fonsatti E, Staibano S, Nappi O, Pagani E, Casula M, Manca A, et al. BRAF/NRAS mutation frequencies among primary tumors and metastases in patients with melanoma. J Clin Oncol. 2012; 30:2522–9.
23. Wagle N, Van Allen EM, Treacy DJ, Frederick DT, Cooper ZA, Taylor-Weiner A, Rosenberg M, Goetz EM, Sullivan RJ, Farlow DN, Friedrich DC, Anderka K, Perrin D, et al. MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. Cancer Discov. 2014; 4:61–8.
24. Van Allen EM, Wagle N, Sucker A, Treacy DJ, Johannessen CM, Goetz EM, Place CS, Taylor-Weiner A, Whittaker S, Kryukov GV, Hodis E, Rosenberg M, McKenna A, et al. The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. Cancer Discov. 2014; 4:94–109.
25. Holderfield M, Deuker MM, McCormick F, McMahon M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. Nat Rev Cancer. 2014; 14:455–67.
26. Flaherty KT, Infante JR, Daud A, Gonzalez R, Kefford RF, Sosman J, Hamid O, Schuchter L, Cebon J, Ibrahim N, Kudchadkar R, Burris HA 3rd, Falchook G, et al. Combined BRAF, and MEK inhibition in melanoma with BRAF V600 mutations. N Engl J Med. 2012; 367:1694–703.
27. Robert C, Karaszewska B, Schachter J, Rutkowski P, Mackiewicz A, Stroiakovski D, Lichinitser M, Dummer R, Grange F, Mortier L, Chiarion-Sileni V, Drucis K, Krajsova I, et al. Improved Overall Survival in Melanoma with Combined Dabrafenib and Trametinib. N Engl J Med. 2014.
28. Turajlic S, Furney SJ, Stamp G, Rana S, Ricken G, Oduko Y, Saturno G, Springer C, Hayes A, Gore M, Larkin J, Marais R. Whole-genome sequencing reveals complex mechanisms of intrinsic resistance to BRAF inhibition. Ann Oncol. 2014; 25:959–67.
29. Straussman R, Morikawa T, Shee K, Barzily-Rokni M, Qian ZR, Du J, Davis A, Mongare MM, Gould J, Frederick DT, Cooper ZA, Chapman PB, Solit DB, et al. Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion. Nature. 2012; 487:500–4.

30. Deuker MM, Marsh Durban V, Phillips WA, McMahon M. PI3'-Kinase Inhibition Forestalls the Onset of MEK1/2 Inhibitor Resistance in BRAF-Mutated Melanoma. Cancer Discov. 2014.

31. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, Hollmann TJ, Bruggeman C, Kannan K, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med. 2014; 371:2189–99.

32. Hu-Lieskovan S, Mok S, Homet Moreno B, Tsoi J, Robert L, Goedert L, Pinheiro EM, Koya RC, Graeber TG, Comin-Anduix B, Ribas A. Improved antitumor activity of immunotherapy with BRAF and MEK inhibitors in BRAFV600E melanoma. Sci Transl Med. 2015; 7:29ra41.

33. Prickett TD, Agrawal NS, Wei X, Yates KE, Lin JC, Wunderlich JR, Cronin JC, Cruz P, Rosenberg SA, Samuels Y. Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4. Nat Genet. 2009; 41:1127–32.

34. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, Stemke-Hale K, Davies MA, Gershenwald JE, Robinson W, Robinson S, Rosenberg SA, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet. 2011; 43:442–6.

35. Mishra PJ, Ha L, Rieker J, Sviderskaya EV, Bennett DC, Oberst MD, Kelly K, Merlino G. Dissection of RAS downstream pathways in melanomagenesis: a role for Ral in transformation. Oncogene. 2010; 29:2449–56.

36. Michaloglou C, Vredeveld LC, Soengas MS, Denoyelle C, Kuilman T, van der Horst CM, Majoor DM, Shay JW, Mooi WJ, Peeper DS. BRAFE600-associated senescence-like cell cycle arrest of human naevi. Nature. 2005; 436:720–4.

37. Vredeveld LC, Possik PA, Smit MA, Meissl K, Michaloglou C, Horlings HM, Ajouaou A, Kortman PC, Dankort D, McMahon M, Mooi WJ, Peeper DS. Abrogation of BRAFV600E-induced senescence by PI3K pathway activation contributes to melanomagenesis. Genes Dev. 2012; 26:1055–69.

38. Johannessen CM, Johnson LA, Piccioni F, Townes A, Frederick DT, Donahue MK, Narayan R, Flaherty KT, Wargo JA, Root DE, Garraway LA. A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. Nature. 2013; 504:138–42.

39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60.

40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–303.

41. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–71.

42. Chiariello M, Vaque JP, Crespo P, Gutkind JS. Activation of Ras and Rho GTPases and MAP Kinases by G-protein-coupled receptors. Methods Mol Biol. 2010; 661:137–50.

Development and validation of a comprehensive genomic diagnostic tool for myeloid malignancies

Thomas McKerrell, Thaidy Moreno, Hannes Ponstingl, Niccolo Bolli, João M. L. Dias, German Tischler, Vincenza Colonna, Bridget Manasse, Anthony Bench, David Bloxham, Bram Herman, Danielle Fletcher, Naomi Park, Michael A. Quail, Nicla Manes, Clare Hodkinson, Joanna Baxter, Jorge Sierra, Theodora Foukaneli, Alan J. Warren, Jianxiang Chi, Paul Costeas, Roland Rad, Brian Huntly, Carolyn Grove, Zemin Ning, Chris Tyler-Smith, Ignacio Varela, Mike Scott, Josep Nomdedeu, Ville Mustonen and George S. Vassiliou

Blood 2016 128:e1-e9

doi: https://doi.org/10.1182/blood-2015-11-683334

COMMENTARY

blood advances

TO THE EDITOR:

JAK2 V617F hematopoietic clones are present several years prior to MPN diagnosis and follow different expansion kinetics

Thomas McKerrell,[1-3] Naomi Park,[1] Jianxiang Chi,[4,5] Grace Collord,[1] Thaidy Moreno,[6] Hannes Ponstingl,[1] Joao Dias,[7] Petroula Gerasimou,[4,5] Kiki Melanthiou,[8] Chrystalla Prokopiou,[9] Marios Antoniades,[8] Ignacio Varela,[6] Paul A. Costeas,[4,5] and George S. Vassiliou[1-4]

[1]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom; [2]Department of Haematology, University of Cambridge, Cambridge, United Kingdom; [3]Department of Haematology, Cambridge University Hospitals National Health Service Foundation Trust, Cambridge, United Kingdom; [4]The Center for the Study of Haematological Malignancies, Nicosia, Cyprus; [5]The Karaiskakio Foundation, Nicosia, Cyprus; [6]Instituto de Biomedicina y Biotecnologia de Cantabria, Santander, Spain; [7]Cancer Molecular Diagnosis Laboratory, National Institute for Health Research Biomedical Research Centre, University of Cambridge, Cambridge, United Kingdom; [8]Department of Haematology, Nicosia General Hospital, Nicosia, Cyprus; and [9]Department of Haematology, Limassol General Hospital, Limassol, Cyprus

The *JAK2* V617F mutation is the most common somatic mutation in the classical myeloproliferative neoplasms (MPNs), present in >95% of cases of polycythemia vera (PV) and ~50% of essential thrombocythemia (ET) and myelofibrosis (MF).[1-4] It is usually the sole identifiable driver mutation in MPNs[5] and was recently also identified as a driver of age-related clonal hemopoiesis in healthy individuals.[6-9] In order to investigate the preclinical clonal evolution of MPNs, we identified 12 individuals with a *JAK2* V617F mutant MPN, who 4.6 to 15.2 years previously (median 10.2 years) had also donated blood to register with the Cyprus Bone Marrow Donor Registry at the Karaiskakio Foundation (Table 1; Figure 1A).

First, we interrogated all 24 samples for 15 myeloid mutation hot spots including *JAK2* V617 (supplemental Table 1), using a previously described multiplex polymerase chain reaction/MiSeq sequencing protocol that reliably detects nucleotide substitutions present at a VAF ≥0.008.[6] Additionally, for 12 samples with sufficient DNA available, we performed targeted DNA capture for all exons of 41 genes recurrently mutated in myeloid neoplasms (supplemental Table 2) using targeted capture with custom RNA baits (SureSelect, Agilent Technologies ELID ID: 0735431) followed by sequencing on Illumina HiSeq 2500. Substitutions and indels were identified and quantified using Mutation Identification and Analysis Software as described.[6,10] Finally, we genotyped archival registry samples for rs12343867, a single nucleotide polymorphism (C/T) linked to the *JAK2* 46/1 haplotype and polymorphisms at the *TERT*, *SH2B3*, *TET2*, *MECOM*, and *MYB* genes that are associated with a predisposition to MPN.[11,12] The study was approved by the Cyprus National Bioethics Committee (EEBK/ΕΠ/2014/11) and performed in accordance with the Declaration of Helsinki.

Amplicon sequencing returned a median coverage of 6641 reads per nucleotide at the studied hot spots (excluding *NPM1* exon 12). This confirmed the presence of *JAK2* V617F in all 12 diagnostic and 9 of 12 archival samples (supplemental Table 3). The remaining 3 samples were *JAK2* V617F negative at the sensitivity of our assay (VAF ≥0.008). The only other hot spot mutation identified was *SRSF2* P95R in patient P3 (see later). Pull-down sequencing of all exons of 41 genes from 12 samples with sufficient DNA returned an average coverage of 1978 reads per nucleotide and showed a close correlation in VAF quantitation for both *JAK2* V617F and *SRSF2* P95R with amplicon sequencing (supplemental Table 3).

The *JAK2* V617F VAF at MPN diagnosis differed between patients as expected[13]; however, the average rate of clonal growth also varied widely, ranging from 0.36% to 6.2% per annum (Figure 1B). Targeted exon capture from 12 of 24 samples only identified 1 co-mutation with a VAF >0.02, *SRSF2* P95R in patient P3 who had a diagnosis of MF (supplemental Table 3). As this locus was also studied by amplicon sequencing, we were able to quantify the *SRSF2* P95R VAF in both the diagnostic and the archival sample taken 12.6 years earlier. The MPN diagnostic sample VAFs for *JAK2* V617F and *SRSF2* P95R were similar (0.37 and 0.41, respectively; supplemental Tables 3 and 4) indicating that they co-occurred in most cells of the neoplastic clone. However, the archival sample did not harbor *JAK2* V617F (or did so at a level below the sensitivity of our assay) but did harbor the *SRSF2* P95R at a VAF of 0.06 indicating this was the clone-founding mutation.

Table 1. Participant characteristics

Patient ID	Age at MPN diagnosis (y)	Sex	Diagnosis	WBC count ($\times 10^6$/L)	Hb (g/L)	Platelet count ($\times 10^6$/L)	Time between samples (y)
P1	53.6	F	PV	17.3	194	550	10.1
P2	24.2	F	PV	n/a	n/a	n/a	5.5
P3	53.4	F	MF	4.2	112	130	12.6
P4	24.5	M	ET	8.4	15.6	600	4.6
P5	58.7	F	ET	8.94	142	637	14.5
P6	40.6	F	PV	7.7	168	602	6.8
P7	68.2	M	MF	8.1	81	74	15.2
P8	33	M	ET	10.3	149	776	12.1
P9	43.2	M	ET	12.1	148	793	9.4
P10	66.9	F	PV	7.6	154	830	14.7
P11	58.4	M	PV	12.7	220	423	10.3
P12	27	F	ET	5.1	135	750	4.6

F, female; Hb, hemoglobin concentration; M, male; MF, idiopathic myelofibrosis; n/a, not available; WBC, white blood cell.

We and others reported that *JAK2* V617F was a common driver of clonal hematopoiesis[6-9]; however, although *JAK2* V617F clonal hematopoiesis is likely to be the ancestor of most MPNs, our understanding of the latency and rate of clonal expansion of *JAK2* V617F–positive clones are not well understood. Our findings demonstrate that although additional driver mutations can accelerate preclinical clonal expansion of *JAK2* V617F clones, the process is highly variable even between individuals whose clones harbor *JAK2* V617F as the sole identifiable driver and who represent the majority of MPN patients.[14]

A recent study used ultrasensitive sequencing to show that clonal hematopoiesis driven by acute myeloid leukemia (AML)–associated mutations was ubiquitous among healthy adults aged 50 to 60 years.[15] Interestingly, most clones identified in this study were associated with loss-of-function mutations in *DNMT3A* and *TET2,* and although most persisted over long periods (>10 years), they usually exhibited only modest expansion if any. However, in contrast to studies looking at larger numbers of people with lower assay sensitivities,[6-9] this study did not identify JAK2 mutant clones in any of its 20 participants.[15] It is therefore possible that *JAK2* V617F, a hot spot mutation that relies on a G>T transversion that is uncommon in myeloid malignancies,[16] is acquired less often but has a more pronounced effect on clonal growth than many other mutations. Nevertheless, it should be noted that the high rates of clonal growth in our study were observed in individuals that (1) were younger than the average MPN patient (as they were identified because they were registered stem cell donors) and (2) did actually go on to develop MPN and are therefore unlikely to be characteristic of the general behavior of *JAK2* V617F clones. Additionally, there is a possibility that some of these clones may not have been the same from clonal hematopoiesis to MPN diagnosis, as MPN patients can sometimes harbor >1 *JAK2* V617F clone.[17] Nevertheless, clonal expansion rates varied dramatically even among the 12 individuals studied here. As certain germ line polymorphisms are associated with an increased risk of developing both *JAK2* V617F–driven clonal hematopoiesis and MPN,[11,12,17,18] we genotyped our patients for these (supplementary Table 4). In our small group of patients, we observed a possible association between the *JAK2* 46/1 risk haplotype (C) and the average annual rise in *JAK2* V617F VAF. In fact, the 4 fastest-expanding clones were either CC homozygous in the germ line (P2 and P4) or through somatic loss of heterozygosity (P1), or had a second driver mutation (P3). These observations suggest that the 46/1 haplotype may influence the rate of expansion of established *JAK2* V617F mutant clones and that loss of heterozygosity involving the risk haplotype may further expedite clonal expansion. These findings need to be verified in larger studies.

Interestingly, 1 of our patients acquired *JAK2* V617F after *SRSF2* p95R, and this was followed by a clonal sweep and the development of MF >12 years later. Clonal hematopoiesis driven by mutant *SRSF2* is very rare before the age of 70 years,[6] yet this individual harbored a significant *SRSF2* P95R clone (VAF 0.06) aged only 40 years old. A recent study of 182 MF patients identified many with *JAK2* V617F and *SRSF2* P95 co-mutation, including several aged <70 years (n = 21, age range 39-84 years).[19] Taken together, these observations suggest that younger individuals with *SRSF2*-mutant clonal hematopoiesis may be at a high risk of progression to a hematological neoplasm.

Our study demonstrates that *JAK2* V617F neoplasms develop from clonal hematopoiesis over many years, sometimes over more than a decade. Although co-mutations in other myeloid genes may accelerate the rate of *JAK2* V617F–driven clonal expansion, the rate can be highly variable even among those without co-mutations as demonstrated at least by the 7 of our 12 patients studied using targeted capture of 41 myeloid genes. Our findings suggest that heritable polymorphisms such as the *JAK2* 46/1 haplotype may have a role,[20] but this will need to be confirmed in larger future studies and it is probable that nongenetic factors may also be at play. Our ability to predict clonal growth and by extension forecast the likelihood of progression of *JAK2* V617F clonal hematopoiesis to MPN remains limited, and our study contributes to the understanding of this process. In order to better advise individuals with clonal hematopoiesis of their prognosis and to identify those that could benefit from current or future therapies, larger studies are required to define the variables influencing clonal expansion whether they are heritable vs environmental or cell-intrinsic vs cell-extrinsic and related to the hematopoietic microenvironment.[21,22]

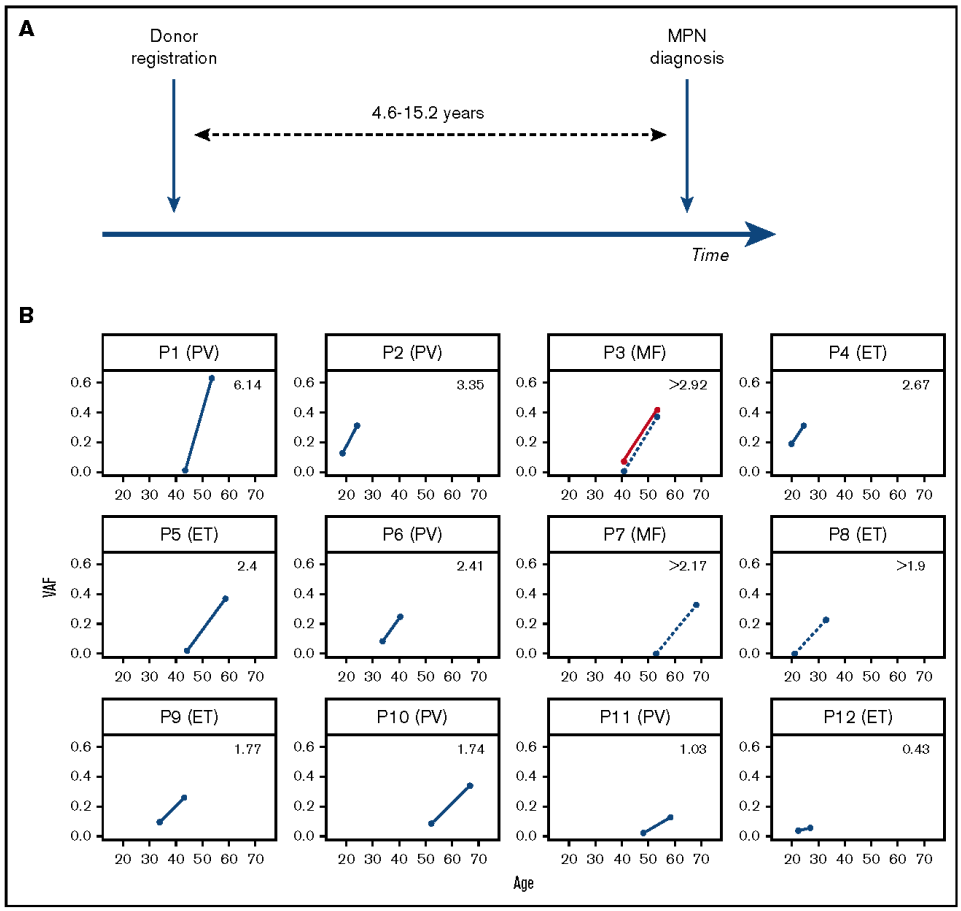


Figure 1. Preclinical expansion of *JAK2* V617F clones in 12 MPN patients. (A) Schema of blood sample collection from 12 individuals at the time of registration as stem cell donors at the Cyprus Bone Marrow Donor Registry and at the time they were diagnosed with MPN 4.6 to 15.2 years later. (B) Variant allele fraction (VAF) sizes of *JAK2* V617F–positive clones at the 2 time points against the age of participants at the time. The specific diagnosis is indicated in brackets next to each patient's ID, and the average annual rise in *JAK2* V617F VAF is indicated in the upper right quadrant of each plot. Samples P3, P7, and P8 had no detectable *JAK2* V617F at donor registration. The VAF rise for *SRSF2* P95R in patient P3 is show in red.

The full-text version of this article contains a data supplement.

Acknowledgements: The authors thank Ayalew Tefferi for providing the numbers and age range of individuals with *JAK2* V617F and *SRSF2* P95 co-mutation from his team's recent study of myelofibrosis.[19]

This work was supported by the Wellcome Trust Sanger Institute (WT098051). T. McKerrell is funded by a Wellcome Trust Clinician Scientist Fellowship (100678/Z/12/Z). G.S.V. is funded by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA), and work in his laboratory is also funded by Cancer Research UK, Bloodwise, the Kay Kendall Leukaemia Fund, and Celgene. I.V. is supported by the Spanish Ministerio de Economía y Competitividad, Programa Ramón y Cajal.

Contribution: G.S.V. and P.A.C. designed the study; G.S.V., P.A.C., and T. McKerrell supervised the study, analyzed data, and wrote the manuscript with contributions from all authors; N.P., T.M., J.C., P.G., and G.C. performed experimental procedures; I.V., T. Moreno, H.P., and J.D. performed bioinformatics analyses; and K.M., C.P., M.A., and P.A.C. contributed to sample and data acquisition.

Conflict-of-interest disclosure: G.S.V. is a consultant for KYMAB and receives an educational grant from Celgene. The remaining authors declare no competing financial interests.

ORCID profiles: T. McKerrell, 0000-0003-4235-0650; G.S.V., 0000-0003-4337-8022.

Correspondence: George S. Vassiliou, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom; e-mail: gsv20@sanger.ac.uk; and Paul A. Costeas, The Karaiskakio Foundation, Nicosia, Cyprus; e-mail: paul.costeas@karaiskakio.org.cy.

References

1. Baxter EJ, Scott LM, Campbell PJ, et al; Cancer Genome Project. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *Lancet*. 2005;365(9464):1054-1061.
2. Levine RL, Wadleigh M, Cools J, et al. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell*. 2005;7(4):387-397.
3. Kralovics R, Passamonti F, Buser AS, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N Engl J Med*. 2005;352(17):1779-1790.
4. James C, Ugo V, Le Couédic JP, et al. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature*. 2005;434(7037):1144-1148.
5. Nangalia J, Green TR. The evolving genomic landscape of myeloproliferative neoplasms. *Hematology Am Soc Hematol Educ Program*. 2014;2014(1):287-296.
6. McKerrell T, Park N, Moreno T, et al; Understanding Society Scientific Group. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Reports*. 2015;10(8):1239-1245.
7. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med*. 2014;371(26):2488-2498.
8. Genovese G, Kähler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med*. 2014;371(26):2477-2487.
9. Xie M, Lu C, Wang J, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med*. 2014;20(12):1472-1478.
10. Conte N, Varela I, Grove C, et al. Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture. *Leukemia*. 2013;27(9):1820-1825.
11. Jones AV, Chase A, Silver RT, et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):446-449.
12. Hinds DA, Barnholt KE, Mesa RA, et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood*. 2016;128(8):1121-1128.
13. Nielsen C, Bojesen SE, Nordestgaard BG, Kofoed KF, Birgens HS. JAK2V617F somatic mutation in the general population: myeloproliferative neoplasm development and progression rate. *Haematologica*. 2014;99(9):1448-1455.
14. Nangalia J, Massie CE, Baxter EJ, et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N Engl J Med*. 2013;369(25):2391-2405.
15. Young AL, Challen GA, Birmann BM, Druley TE. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat Commun*. 2016;7:12484.
16. Alexandrov LB, Nik-Zainal S, Wedge DC, et al; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain. Signatures of mutational processes in human cancer [published correction appears in *Nature*. 2013;502(7470):258]. *Nature*. 2013;500(7463):415-421.
17. Olcaydu D, Harutyunyan A, Jäger R, et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):450-454.
18. Tapper W, Jones AV, Kralovics R, et al. Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. *Nat Commun*. 2015;6:6691.
19. Tefferi A, Lasho TL, Finke CM, et al. Targeted deep sequencing in primary myelofibrosis. *Blood Adv*. 2016;1(2):105-111.
20. Vassiliou GS. JAK2 V617F clonal disorders: fate or chance? *Blood*. 2016;128(8):1032-1033.
21. Arranz L, Sánchez-Aguilera A, Martín-Pérez D, et al. Neuropathy of haematopoietic stem cell niche is essential for myeloproliferative neoplasms. *Nature*. 2014;512(7512):78-81.
22. McKerrell T, Vassiliou GS. Aging as a driver of leukemogenesis. *Sci Transl Med*. 2015;7(306):306fs38.

DOI 10.1182/bloodadvances.2017007047

APPENDIX 2

MANUSCRIPTS RESULTED FROM THIS THESIS

Recurrent somatic mutations in ARID2 in lung cancer are associated with poor prognosis, increased metastatic potential and increased sensitivity to DNA-damaging agents.

Thaidy Moreno(1), Laura González-Silva(1), Carlos Revilla(1), Antonio Agraz-Doblas(1,2), Laura Quevedo(1), Isabel Betancor(2), Javier Freire(3), Santiago Montes-Moreno(3), Laura Cereceda(3), Pablo Isidro(4), Aurora Astudillo(4), Javier Gomez-Roman(3), Eduardo Salido(2), Paola Scaffidi(5) and Ignacio Varela(1)

1) Instituto de Biomedicina y Biotecnología de Cantabria. Universidad de Cantabria-CSIC. Santander, Spain.
2) Departamento de Patología, Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER). Tenerife, Spain.
3) Departamentos de Patología.Biobanco Valdecilla. HUMV/IDIVAL. Santander, Spain.
4) Biobanco del Principado de Asturias (BBPA). Hospital Universitario Central de Asturias. Oviedo, Spain.
5) Cancer Epigenetics Laboratory, The Francis Crick Institute, 1 Midland Road London NW1 1AT UK.

Abstract

The potential role of chromatin structure in cancer development is still a topic of big debate. In the last years, several members of the SWI/SNF chromatin remodeling complexes have been described altered in different tumor types. In the present work, we have identified *ARID2* as a new potential tumor suppressor gene in lung cancer, the major cause of cancer-related deaths worldwide. We observed that approximately 15% of our lung cancer patients present mutations in this gene that are associated with a loss of production of the protein and a worse prognosis. Our results suggest that ARID2 plays a dual role in this type of tumor; firstly, it produces expression changes in several genes to generate a pro-invasive phenotype in the cells. Secondly, it generates genetic instability which increases the capacity of tumor cells to adapt to the environment. All these changes increase the proliferation and metastatic potential of the cells *in vitro* and *in vivo*. Moreover, we have proved that *ARID2* deficiency can be exploited therapeutically as it confers sensitivity to DNA-damaging agents frequently used for the treatment of lung cancer patients like cisplatin or etoposide. All these results support the use of *ARID2* deficiency as prognostic factor for patient stratification and for the selection of specific treatment regimens.

Main text

Lung cancer is the major cause of cancer-related deaths worldwide. The average 5-year survival rate is below 20% irrespective of the subtype, a number that has only marginally improved in the last decades[1]. Consequently, any new knowledge about the molecular mechanisms that drive this disease could have a great impact on the treatment of lung cancer patients. In the last years, large genomic characterization projects have facilitated the identification of major players in this tumor type. Thus, small cell lung cancer (SCLC) that constitutes around 15% of all cases is mainly driven by mutations in *TP53* and *RB1* although the role of other genes like *PTEN*, *SLIT2* or *CREBBP* has been also described[2]. Respecting non-small cell lung cancer cases, constituted mainly by adenocarcinomas (50%) and squamous cell carcinomas (40%), *KRAS* and *BRAF* were the first genes found recurrently altered. Subsequently, mutations in *EGFR* or *HER2* among others have been also identified and used as markers to detect sensitivity to specific anti-tumor therapies[3].

More recently, several members of the SWI/SNF family of chromatin remodeling complexes have been identified recurrently altered in human cancer which sums to the accumulated compelling evidence of the role of SWI/SNF family complexes in cancer development. It is estimated that approximately 20% of all tumors contain alterations in these complexes which means that SWI/SNF subunits are more broadly mutated that any other cancer gene excepting *TP53*[4]. In the case of lung cancer, the production of the two mutually exclusive catalytic ATPase subunits (BRM or BRG1) is lost in 30% of non-small cell lung cases where it is associated with worse prognosis[5]. Additionally, *ARID1A* which produces one of the auxiliary subunits of the complex, has been found recurrently mutated in lung adenocarcionoma[6]. Finally, mutations in another auxiliary subunit producing gene, *ARID2*, have been reported in 5% of non-small cell lung cases in a recent report by Blons and cols[7].

In order to understand better the role of these chromatin remodeling complexes in lung cancer development, we have performed a genetic screening on the coding sequences of a list of selected genes in a collection of 576 cancer cases using targeted next-generation sequencing technologies (Suppl. Table 1). This list of genes includes, besides known cancer genes, those that codify for the subunits of the main chromatin remodeling complexes (Suppl. Table 2). Surprisingly, we found mutations in *ARID2* in approximately 15% of our lung cancer cases irrespective of the cancer subtype (Figure 1a and Suppl. Table 3). This recurrence of ARID2 mutations is significantly higher than reported previously[7] and locates *ARID2* as the forth gene more recurrently mutated in lung cancer after *TP53*, *EGFR* and *KRAS*. To validate these results, we performed targeted sequencing on the coding regions of *ARID2* in a second validation cohort of extra 96 lung adenocarcinoma cases finding the same frequency of *ARID2* mutations (Figure 1b and Supplementary Table 3). In

concordance with a previously described role of *ARID2* as a potential tumor suppressor gene, several of the identified mutations are predicted to produce a premature truncation of the protein and these mutations are located through the whole protein sequence (Figure 1c and Suppl. Table 3). Subsequently, to check the effects of these mutations on the ARID2 protein production, we performed immunohistochemistry analysis on those samples from which we count with an histological sample. These experiments revealed that approximately 15% of the samples analyzed showed loss of expression of ARID2 including more than the 85% of the analyzed *ARID2*-mutated samples (7/8). In contrast all the assayed *ARID2* wild-type samples (32/32) show clear ARID2 production (Figure 1d). Interestingly, when we checked the data published by the Genome Atlas Consortium, the loss of expression of ARID2 correlates significantly with a worse prognosis of the patients which supports the potential role of ARID2 as a *bona fide* tumor suppressor gene and prognostic marker in lung cancer.

ARID2 has been described as tumor suppressor gene in melanoma and liver cancer[8,9]. To check if alterations in this could also promote lung cancer development, we performed knock-down experiments *in vitro* using shRNAs. As it can be observed in Figure 2, *ARID2* mRNA and protein production is efficiently reduced by two different shRNAs. This reduction is accompanied by an increase of the proliferation, invasion and migration capacities of A549 and H460 lung cancer cell lines compared with those cells transduced with empty vector. Moreover, when these cells are injected on immunocompromised mice, they show a greater capacity to produce tumors *in vivo* (Figure 2). These results prove that, similarly to its role in other tumor types, *ARID2* plays a tumor suppressor function in lung cancer.

The precise molecular mechanisms by which alterations in chromatin remodeling complexes promote cancer development are not perfectly known. Interactions with well-described tumor suppressors like TP53, RB or MYC have been described[10–12]. Additionally, they play essential roles in the activation of differentiation and the suppression of proliferative programs of many cellular lineages[13]. In order to understand better the potential impact at the transcriptional level of the deficiency in ARID2, we performed RNA-Seq experiments in the transduced cell lines. Loss of ARID2 is accompanied with changes in gene expression that could offer an explanation to the phenotype observed in the cells (Figure 3). Thus, we observed a downregulation of genes involved in cellular adhesion as NPNT, CNTNAP2, FAT3, FN1 or VCAN which could be associated with the increase in migration and invasion capacities of ARID2-deficient cells. Additionally, we observed downregulation of other tumor suppression genes like RPS6K2, TNFSF10, TP63, ISM1 and LDLRAD4 together with upregulation of protumoral and antioncogenic genes like HOXB1, BCL2A1 or RCVRN. These combined transcriptional program change could in itself explain the tumoral capacities adquired by the cells after ARID2 knock-down. All these changes were further validated by qRT-PCR in cells transduced with two different shRNAs (Figure 3).

Another potential mechanism that has been suggested for the role of SWI/SNF alterations in cancer progression is through the promotion of genomic instability. Firstly, SWI/SNF is described to play essential role in chromosome segregation and in NER[14]. Secondly, SWI/SNF is recruited to places of DNA damage and play an essential role in activate DNA damage sensors[15]. Finally, SWI/SNF is essential as well for a correct DNA synthesis after DNA damage, a step that is essential for a final damage correction[16]. Consequently, we hypothesized that *ARID2* mutations could be associated with a higher genomic instability. In accordance to that, we observe an upregulation of *GADD45A* in our ARID2-deficient cells which could indicate an activation of DNA detection and repair mechanisms in these cells. In order to check this possibility, we performed immunofluorescence experiments to check the nuclear location of ARID2 in cells submitted to DNA damage. As it can be seen in Figure 4, ARID2 co-localizes with H2AX and 53BP1 at the DNA repair foci. Moreover, we observed that ARID2-deficient cells show an increase in the number of DNA damage foci. All this demonstrates that a second tumor promoting mechanism of *ARID2* deficiency in exerted by genomic instability.

Finally, we decide to explore the possibility of exploiting this genomic instability as a therapeutic target for ARID2-deficient patients. For that we performed *in vitro* treatment of ARID2-deficient and control cell lines with cisplatin and etoposide, two widely used treatments in lung cancer patients that produce DNA damage. As it can be seen in Figure 4, ARID2-deficient cells show a higher sensitivity to these drugs that control cell lines. Therefore, ARID2 status can be used as a marker to stratify patients for a proper treatment.

In summary, here we present compelling evidence of the role of *ARID2* as tumor suppressor gene in lung cancer. Although ARID2 has been proposed as tumor suppressor gene in other tumor types, little is known about the molecular mechanisms behind. In this work, we have described that this role in lung cancer is exerted in two ways, firstly by producing the activation of a specific pro-oncogenic transcriptomic program and secondly by the promotion of general genomic instability. Which is most important, we show results that suggest that this genomic instability can be exploited for lung cancer patient treatment.

Methods

Patient Samples

Cancer patient primary samples and, when available, matched corresponding normal samples, were obtained from different tumor Biobanks. In all the cases, we counted with the prior approval of the corresponding ethics committee for each institution. A detailed list of the origin and characteristics of each sample can be found in Supplementary Table 1. In total 177 tumors and 88 matched normal DNAs were used in this project.

DNA Extraction and DNA libraries

DNA was extracted from fresh frozen tissue or cell lines using the Agencourt DNAdvance Beckman Coulter kit (#A48705, Beckman Coulter, Brea, CA, USA), following manufacturer's instructions. For the formalin-fixed paraffin-embedded sections, the tumor area was micro-dissected, treated with Proteinase K overnight, subjected to a phenol-chloroform organic extraction followed by ethanol precipitation of the DNA. DNA preparations were quantified using the Qubit® dsDNA BR Assay (Q32851, Life Technologies). Normal DNA libraries were performed mixing from 3 to 5 different DNAs. Diagenode Bioruptor® DNA fragmentation was performed with 500 ng of DNA diluted in low TE buffer (#12090015, Thermo Fisher Scientific, UK) to a final volume of 100 μl, and using 30 cycles of 30''/30'' (ON/OFF cycles) at 4ºC. For all cleaning steps, we used Agencourt AMPure XP (#082A63881, Beckman Coulter, Brea, CA, USA), following the manufacturer's protocol. Size distribution was analyzed with either the 2100 Bioanalyzer or the 4200 TapeStation using DNA 1000 kit or D1000 ScreenTape Assay (Agilent Technologies, Santa Clara, CA, USA). Sequencing libraries were prepared through a series of enzymatic steps including End-repair and Adenylation (DNA Rapid End Repair module, NEXTflex™, #5144-05, Bioo Scientific, Austin, TX, USA), PE adaptor construction through the hybridization of phosphorylated complementary synthetic oligonucleotides, PE adaptor ligation (T4 DNA Ligase, #EP0062, Thermo Fisher Scientific, UK) and PCR indexing amplification (Phusion high fidelity DNA polymerase, # F530L, Thermo Fisher Scientific, UK). Libraries were checked by Nanodrop for chemical contamination, by the 2100 Bioanalyzer for size distribution and finally quantified using the Qubit® and a qPCR reaction using primers designed to target the Illumina adapters. Target capture was performed on pools of 96 libraries using a Sure Select® user-defined probe kit (Agilent Technologies, Palo Alto, CA, USA). Massively parallel sequencing was carried out in a High-Seq® machine (Illumina, USA) with a 100bp paired end (PE) protocol. A single lane was performed for each 96-library pool.

In the case of amplicon-based libraries two different strategies where used. When a lot of different products were amplified for the same sample, standard PCR primers were designed, the PCR products generated for each sample were mixed and subjected to the standard library protocol. Alternatively, when a small number of amplicons were designed over a large number of samples, the primers were designed to contain a common adapter sequence that was used, after mixing and purification, to a second PCR to add the barcode and the rest of the Ilumina adapter sequence. These libraries were sequenced in the MiSeq® platform (Illumina, USA) using a 150 or 250 paired-end protocol depending on the amplicon size distribution.

RNA isolation and qRT-PCR.

Total RNA was isolated and purified using Extract Me Total RNA Kit (Blirt, DNA Gdansk, Poland) according to the manufacturer's instructions. RNA quality was measured using RNA ScreenTape® (4200 TapeStation Instrument - Agilent Genomics). Reverse transcription was performed using the Takara PrimeScript cDNA Synthesis kit (Takara Bio, Inc., Dalian, Japan) according to the manufacturer's instructions. mRNA expression was measured by qRT-PCR using Luminaris Color HiGreen qPCR Master Mix (Thermo Scientific) with StepOnePlusTM real-time PCR system (Applied Biosystems, Foster City, CA). β-actin was used as housekeeping gene and the ΔΔCt method was used for quantification and comparison. A list of the primers used for the qRT-PCR experiments can be found in Supplementary Table 2.

RNA Libraries

RNA quality and concentration were measured using a RNA Pico chip on a 2100 Agilent Bioanalyzer. For library preparation, mRNA was enriched using NEBNext® Poly(A) mRNA Magnetic Isolation Module. Fragmentation were performed from 1-2 µg mRNA in a buffer containing 4 µL de PrimeScript Buffer and 1 µL random hexamers primers at 94ºC for 15 minutes. The first strand was synthesized by adding to the previous mix 1 µL of PrimeScript Enzyme and incubating the samples 15 minutes at 37ºC followed by 5 seconds at 85ºC. Second strand was further synthesized by adding to the previous reaction RNAse HI (Thermo) and DNA polymerase I (Thermo) according to manufacturer instructions to a final volume of 100 µL. 2.5 µL de T4 DNA Polymerase (Thermo) and incubated 5 min at 15ºC. 5 µl of EDTA 0.5 M pH 8.0 were added. Fragments were purified using Agencourt AMPure XP (#082A63881, Beckman Coulter, Brea, CA, USA). Libraries generation protocol were performed starting from the double-stranded cDNA in a similar way of the DNA libraries.

DNA-Seq Analysis

Raw sequence data wasere subjected to quality control using FastQC v0.11.2 (https://www.bioinformatics.babraham.ac.uk/publications.html) and mapped to the human genome (hg19) using BWA 0.7.3[17]. Samtools 0.1.18[18] was used for format transformation, sorting and indexing of the bam files. Picard 1.61 (http://broadinstitute.github.io/picard/) was used to fix and clean the alignment and to mark PCR duplicates reads. Finally, GATK 2.2.8 was used to perform local realignment around indels. Bedtools 2.17 (http://bedtools.readthedocs.io/en/latest/#) was used to calculate the enrichment statistics and the target coverage. Paired tumor/normal bam files were used to identify putative somatic single variants (SVs) using an in-house written algorithm called RAMSES selecting mutations with a confidence score >2 and mutational frequency higher than 0.05. PINDEL 0.2.4d (http://

broadinstitute.github.io/picard/) was used to detect indels requiring a minimum of 5 independent reads reporting the indel and with no evidence in the control DNA. Potential germline variants were flagged away using 1000 Genomes mutation database with in-house written software. Functional consequence of the mutations was annotated using ensembl database v.73 through the Perl API. OncodriveFM software was run to detect genes with evidence of selective pressure from the analysis[19].

RNA-Seq Analysis

Paired-end reads from RNA-Seq were aligned using Tophat[20] to the human genome (hg19). Predicted transcripts from Ensembl database were analyzed and transcripts that would lack a CDS start or stop site were filtered out. Differentially expressed genes (DEG) were identified using HTSeq + DESeq[21,22]. These R packages for transcriptome expression profile analysis were used according to the manufacturer's instructions to test for differential expression of RNA transcript levels requiring a minimum of 3 counts for a gene in more than two independent samples and using a threshold of fold change >1 and a pvalue <0.05. DEG were manually reviewed and the final list of DEG was created.

Cell Culture

Both A549 and H460 lung cancer cell lines were sourcedobtained from The Francis Crick Institute common repository, authenticated by STR profiling, and tested for mycoplasma. The two cell lines were maintained in DMEM (Lonza, Verviers, Belgium) and RPMI 1640 (Lonza, Verviers, Belgium), respectively, supplemented with 10% FBS (HyClone Victoria, Australia), 1% Gentamycin and 1% Ciprofloxacin at 37°C in a humidified atmosphere containing 5% CO2.

Generation of stably-transduced cell lines

For stable cell line generation, tetracycline-inducible pTRIPZ constructs V2THS_74399, V3THS_347660 were used for ARID2 knockdown (Dharmacon/GE Healthcare, Lafayette, CO, USA). The empty vector was used as control. Virus production were performed by transfecting HERK293-T/17clone cells with the pTRIPZ constructs, psPAX2 and pMD2.G plasmids (Addgene) using Fugene HD (Promega Madison, WI, USA). Infected cells were selected with 1µg/ml puromycin for at least 7days. Induction of the expression of the shRNAs as well as the turbo-RFP marker was performed with 1ug/ml of Doxycycline for at least 5 days before analyzing the effect of ARID2 knock-down on the cells. Turbo-RFP expression on cells transduced with shRNAs or empty vectors was induced with 1µg/ml Doxycycline (Dox) for 16h. Cells were isolated by FACS based on TurboRFP expression using the FACS-Aria II cell sorter (Becton Dickinson, BD, Franklin Lakes,

USA). For proper cell recovery from the sorting process, the cells were collected in tubes containing DMEM supplemented with 50% FBS to prevent the cells from drying out and dying. Cells were seeded in DMEM complete growth medium.

Proliferation assays

Growth curve analysis was performed over a period of fourteen days. Cells were seeded in 100 mm plates at a density of 500,000 cells per plate. Every two days, cells were trypsinised and the cell number determined by counting using a hemocytometer and seeded 500,000 cells per plate. All growth curves were performed in triplicate.

Cell proliferation was also analyzed using the carboxyfluorescein diacetate succinimidyl ester labeling method with the CellTrace™ CFSE Cell Proliferation Kit (Invitrogen, CA, USA). Induced cells line were synchronized by gradual serum deprivation following the protocol described by Lauand and collaborators (163). After 50 h of FBS deprivation, the cells were arrested in G0/G1 phase. Cells were harvested, washed twice in Phosphate-buffered saline (PBS), counted and re-suspended in CellTrace CFSE labelling solution. 1 µL of CellTrace™ stock solution was added to each mL of cell suspension for a final working solution of 5 µM at a density of 106 cells/mL and incubated at 37ºC for 20 minutes protected from light. DMEM culture media containing FBS was added (10% v/v) to removes any free dye remaining in the solution. After 5 minutes, labelled cells were washed into PBS and pelleted by centrifugation. Some of these labelled cells were suspended in fresh pre-warmed complete culture medium and were then seeded into 6-well plates at a density of 5 *105 cells/well. The remaining labelled cells were suspended in PBS and CFSE fluorescence was measured on a MACSQuant® VYB (Miltenyi Biotec) flow cytometer to ensure the parental population and subsequent division peak tracking during culture. Cells were harvested at defined times and subjected to division peak resolution by flow cytometry. The cell proliferation index was analyzed using MODFIT software. Proliferation index was the sum of the cells in all generations divided by the calculated number of original parent cells.

Migration assay

In vitro cell migration assays were performed by using 8-µm pore size transwell chambers (Corning™ Transwell™ Multiple Well Plate with Permeable Polycarbonate Membrane Inserts, 3422) in 24-well plates and incubated for 48 h. For the migration assays, 50,000 cells were added into the upper chamber. Cells were plated in medium without serum, and medium containing 10% FBS in the lower chamber served as the chemo-attractant. After 24 h incubation, the cells that did not migrate through the pores were carefully removed using cotton swabs. Filters were washed with PBS, harvested by treatment with 0.25% trypsin and counted on a

hemocytometer. All experiments were performed in triplicate. Filters were fixed in 4% PFA followed by crystal violet staining for microscope visualization.

Invasion assays.

For the invasion assays, 50,000 cells in 50 µL of serum-free DMEM were plated on growth factor-reduced Matrigel (BD Biosciences) pre-coated 8 µm pore Transwell chamber and the lower chamber was filled with 600 µL DMEM with 10% FBS. After 48 h, non-invading cells were removed from the top of the Transwells using cotton swabs. Invasive cells were quantified by fixing chambers in 4% paraformaldehyde for 10 min and staining with crystal violet. For each Transwell, 10 fields were imaged and counted.

***In vivo* Tumorigenesis assays**

Animal studies were conducted in compliance with guidelines for the care and use of laboratory animals and were approved by the Ethics and Animal Care Committee of Universidad de Cantabria University. For proliferation assays, A549 stably-transduced cell lines c were harvested, washed twice in Phosphate-buffered saline (PBS), counted and re-suspended in PBS at a density of 107 cells/mL. Five million of cells were subcutaneously injected into the flanks of the 6-8-week-old female nude mice (Athymic Nude-Foxn1^{nu}). The animals were treated with 1 mM/mL Doxycycline for ~25 days in the drinking water supplemented with 1% sucrose, changed every 2-3 days. After the tumors reached the size of ~0.5 cm3, mice were euthanizing and tumor tissues were harvested for analyses. For metastasis assays, A549 and H460 stably-transfected cell lines were harvested, washed twice in Phosphate-buffered saline (PBS), counted and re-suspended in PBS + 0.1% BSA at a density 5 * 106 cells/mL. 2.5 million of cells were tail injected into 6-8-week-old female nude mice. The animals were treated for ~60 days with 1 mM/mL Doxycycline in drinking water for ~60 days supplemented with 1% sucrose, changed every 2-3 days. After two months, mice were euthanizing and tumor tissues were harvested for analyses.

Western blot analysis

Cells were washed twice in PBS and lysed in RIPA buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1% NP-40, 1 mM Sodium Orthovanadate, 1 mM NaF) containing Halt protease inhibitors Cocktail (Thermo Scientific, 87786), for 30 minutes on ice. Lysates were sonicated using the Bioruptor® (Dia-genode) for ten cycles (30 s on, 30 s off) at high-power and cleared by centrifugation at 16,000g for 20 min at 4 °C. Protein concentrations were determined by Qubit® Protein Assay (Q33212, Life Technologies). 60-80 µg of total protein lysate was separated by SDS-PAGE in 8% polyacrylamide gels and transferred to nitrocellulose membranes.

Subsequently, membranes were washed with TBS-T (50 mM TRIS + 150 mM Sodium chloride + 0,1% Tween 20, pH 7,4) and blocked using 5% non-fat milk solution as blocking agent in TBS (50 mM TRIS + 150 mM Sodium chloride) for 1 h at RT. Membranes were then incubated with primary antibodies anti-ARID2 (E-3, sc-166117, Santa Cruz) and anti-Actin (I-19, sc-1616, Santa Cruz), diluted 1:200 and 1: 1,000 in TBS-T/5% (w/v) BSA at 4°C overnight, respectively. Membranes were washed in TBS-T, three times.After careful washing with TBS-T, t The primary antibodies were detected by incubating the membranes with donkey anti-mouse or donkey anti-goat secondary antibodies (LI-COR Biotechnology, Lincoln, USA) conjugated to IRDye 800CW (926-32212) or IRDye 680RD (926-68074) respectively at 1: 15,000 dilutions and incubated for 45 minutes at room temperature. Finally, antibody signals were visualized using Odyssey Clx imager (LI-COR Biotechnology, Lincoln, USA).

Proliferation inhibition *in vitro* assays.

Inhibition assays were performed to determine the half maximal inhibitory concentration (IC50) values for cisplatin, and etoposide in both A549 and H460 stably-transduced cell lines. Briefly, cells were seeded in 96-well plates at 2000 cells per well in 100 µL of complete media, cultured for 24 hours before drug treatment. Drug concentrations range (0.001 µM – 10 mM) were prepared in 90 µL of complete media. Cells were treated for 48 hours. Appropriate media and vehicle controls were also added in the media. Viability was determined by adding 10 µL of PrestoBlue® reagent (Thermo Fisher Scientific, UK). After incubation, absorbance was measured using a Multiskan FC Microplate Photometer (Thermo Fisher Scientific, Waltham, MA) with wavelengths set at 540 and 620 nm. IC50 value for each drug were determined with Prism software to fit curves to the dose response data.

Immunohistochemistry analysis

For ARID2 detection on paraffin sections, these were incubated with 1:300-1:500 ARID2 antibody for 32 minutes at 97ºC in citrate buffer pH 6. The sections were developed with HRP-polymer secondary antibodies (Optiview, Roche).

Immunofluorescence was performed in stable cells induced with 1 µg/mL doxycycline. Cells reach 50–70% confluence on sterile cover slips were rinsed twice with PBS and fixed with 4% paraformaldehyde in PBS for 15 min at room temperature. Cover slips were rinsed tree times with PBS for 5 minutes. Permeabilization were performed with 0.5% Triton X-100 in PBS for 5 minutes at room temperature. The cells were blocked with 3% BSA in PBT (PBS containing 0.05% Triton X-100) and subjected to immunofluorescence staining with ARID2 (1:50) antibody for 30 minutes at room temperature in moist chamber. The cover slips were then washed with PBS three times for 5 minutes. Cells were incubated with Alexa labeled secondary antibodies (1:400) for 30 minutes at room temperature in moist chamber protected from light. Cover slides were mounted in VECTASHIELD Antifade Mounting Medium with DAPI (Vector Labs, Burlingame, CA, USA). The cells

were finally examined by fluorescence microscopy (Olympus America Inc, Center Valley, PA). Quantification of fluorescent intensity was performed from randomly selected fields using Metamorph and ImageJ software.

References

1. Lovly, C. M. & Carbone, D. P. Lung cancer in 2010: One size does not fit all. *Nat. Rev. Clin. Oncol.* **8,** 68–70 (2011).
2. Gelsomino, F., Rossi, G. & Tiseo, M. MET and Small-Cell Lung Cancer. *Cancers* **6,** 2100–2115 (2014).
3. Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F. & Wong, K.-K. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat. Rev. Cancer* **14,** 535–546 (2014).
4. Masliah-Planchon, J., Bi?che, I., Guinebreti?re, J.-M., Bourdeaut, F. & Delattre, O. SWI/SNF Chromatin Remodeling and Human Malignancies. *Annu. Rev. Pathol. Mech. Dis.* **10,** 145–171 (2015).
5. Reisman, D. N., Sciarrotta, J., Wang, W., Funkhouser, W. K. & Weissman, B. E. Loss of BRG1/BRM in human lung cancer cell lines and primary lung cancers: correlation with poor prognosis. *Cancer Res.* **63,** 560–566 (2003).
6. Imielinski, M. *et al*. Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. *Cell* **150,** 1107–1120 (2012).
7. Manceau, G. *et al*. Recurrent inactivating mutations of *ARID2* in non-small cell lung carcinoma. *Int. J. Cancer* **132,** 2217–2221 (2013).
8. Hodis, E. *et al*. A Landscape of Driver Mutations in Melanoma. *Cell* **150,** 251–263 (2012).
9. Li, M. *et al*. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nat. Genet.* **43,** 828–829 (2011).
10. Burrows, A. E., Smogorzewska, A. & Elledge, S. J. Polybromo-associated BRG1-associated factor components BRD7 and BAF180 are critical regulators of p53 required for induction of replicative senescence. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14280–14285 (2010).

11. Nagl, N. G. The c-myc Gene Is a Direct Target of Mammalian SWI/SNF-Related Complexes during Differentiation-Associated Cell Cycle Arrest. *Cancer Res*. **66,** 1289–1293 (2006).
12. Flowers, S., Beck, G. R. & Moran, E. Transcriptional Activation by pRB and Its Coordination with SWI/SNF Recruitment. *Cancer Res*. **70,** 8282–8287 (2010).
13. Wilson, B. G. & Roberts, C. W. M. SWI/SNF nucleosome remodellers and cancer. *Nat. Rev. Cancer* **11,** 481–492 (2011).
14. Ray, A. *et al*. Human SNF5/INI1, a component of the human SWI/SNF chromatin remodeling complex, promotes nucleotide excision repair by influencing ATM recruitment and downstream H2AX phosphorylation. *Mol. Cell. Biol*. **29,** 6206–6219 (2009).
15. Lee, H.-S., Park, J.-H., Kim, S.-J., Kwon, S.-J. & Kwon, J. A cooperative activation loop among SWI/SNF, gamma-H2AX and H3 acetylation for DNA double-strand break repair. *EMBO J*. **29,** 1434–1445 (2010).
16. Niimi, A., Chambers, A. L., Downs, J. A. & Lehmann, A. R. A role for chromatin remodellers in replication of damaged DNA. *Nucleic Acids Res*. **40,** 7393–7403 (2012).
17. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl*. **25,** 1754–1760 (2009).
18. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl*. **25,** 2078–2079 (2009).
19. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol*. **17,** (2016).
20. Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. **14,** R36 (2013).
21. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol*. **11,** R106 (2010).

22. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

Acknowledgements

This work has been supported by two grants (SAF2012-31627 and SAF2016-76758-R) of the Spanish Ministerio de Economía y Competitividad. I.V has also been supported by the Spanish Ministerio de Economía y Competitivida, Programa Ramón y Cajal. We would like as well to acknowledge the support of the Servicio Santander Supercomputación. Finally, we would like to thank also the work of the staff members of the different tumor biobanks that suministrated us with patient samples, specially to the IDIVAL and HUCA biobanks, for their exceptional work in sample collection and organization.

Figure Legends

Figure 1. (a) Box representation of the mutated patients for the most recurrently mutated genes in the lung cancer cohort. The upper panel represents the number of non-silent single nucleotide variants and small insertions or deletions per patient. Each box in the central matrix represents an independent patient. Colored boxes represent mutated patients for the corresponding gene in a color code indicating the type of mutation. The histogram on the right represents the number of each mutation type found in each individual gene. (b) Box representation of the mutated patients in the validation cohort. Each box represents a single patient. Colored boxes represent a mutated patient. Each row/color represents a non-synonymous mutation type: missense substitutions (blue), nonsense substitutions (light green), frameshift-inducing deletions (light blue) and multiple substitutions affecting the same patient (green). (c) Lollipop graph representing the location of the identified ARID2 mutations in all lung cancer patients in relation to the functional protein domains. (d) Representative images of ARID2 immunohistochemistry experiments in a ARID2-mutated lung adenocarcinoma tumor (left) as well as a ARID2-nonmutated (right).

Figure 2. (a) Bar representation of ARID2 expression level fold changes measured by qRT-PCR in A549 cells transduced with shARID2 v2 and v3 as well as the empty vector which is used as control. All cells were submitted to doxycycline treatment for at least 5 days. (b) Representative image of a western blot measuring ARID2 protein levels in A549 parental cells as well as those cell lines transduced with ARID2 shRNAs and the empty vector. In all the cases, the results are shown with and without induction of the shRNA expression by doxycycline (Dox) treatment. (c) Representation proliferation quantified the cumulative cell number accumulated by serial cell passaging (d) Bar representation of the number of cells in the lower chamber in migration and invasion assays of cells transduced with either the empty vector (blue) or the different ARID2 shRNAs (yellow and red). Data is shown as mean ± SE (standard deviation of the mean) of three independent experiments. (e) Representative image of the results of the flow cytometric analysis of cell division by dilution of CFSE in A549 cells. CFSE histograms are shown at 48 hours after CFSE labeling, A549 shEmpty control cells (left), A549 v2 (middle) and A549 v3 cells (right) are shown. Predicted size of population that have suffered different number of cell divisions according to the color legend are represented inside the graph. (f) Representative images showing A549 cells subjected to Matrigel invasion and transwell migration that have been stained with crystal violet dye in the downer chambers. (g) Representative images of the lung metastasis generated in intravenously injected mice with A549 cells transduced either with shEmpty, or shARID2 v3. Images of both posterior (top panel) and anterior (bottom panel) face of

the lungs are shown. Individual metastasis are delineated in the image and counted (top numbers). (h) The top graph represents the size of the identified metastasis divided in three groups: small, medium and large. At the botton dot plot represents the size (measured in pixels) of each of the identified metastasis divided in the two study mice groups.

Figure 3. Differentially expressed genes in ARID2-deficient cells. (a) Heatmap representation of a selection of differently expressed genes in ARID2-deficient cells (n=3) and grouped according to their molecular pathway. Expression differences goes from red(overexpression) to blue (downregulation) according to the log2 of the fold change. (b) Bar representation of the results of the qRT-PCR validation of the expression differences identified in the RNA-Seq experiments. The expression foldchanges are represented as a mean +/- SE of three independent experiments.

Figure 4. (a). Representative images of immunofluorescence experiments proving the colocalization of γH2AX (green) and TurboRFP (red) in DNA repair foci in H460 cell line after Neocarzinoztatin treatment during 24 hours. At the right, the plot represents the number of DNA repair foci generated measured by the intensity of γH2AX green signal in cells transduced with control or ARID2sh vectors. (b) Graphs of representative experiments measuring cell survival to increasing concentrations of cisplatin (left) and etoposide (right) in A549: Cells transduced with shEmpty (blue) or shARID2 v2 (yellow) and v3 (pink) are represented. Bellow, bars represent the calculated IC50 of A549 cells to cisplatin and etoposide. The results are represented as mean +/- SE of three independent results, (*$p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$).

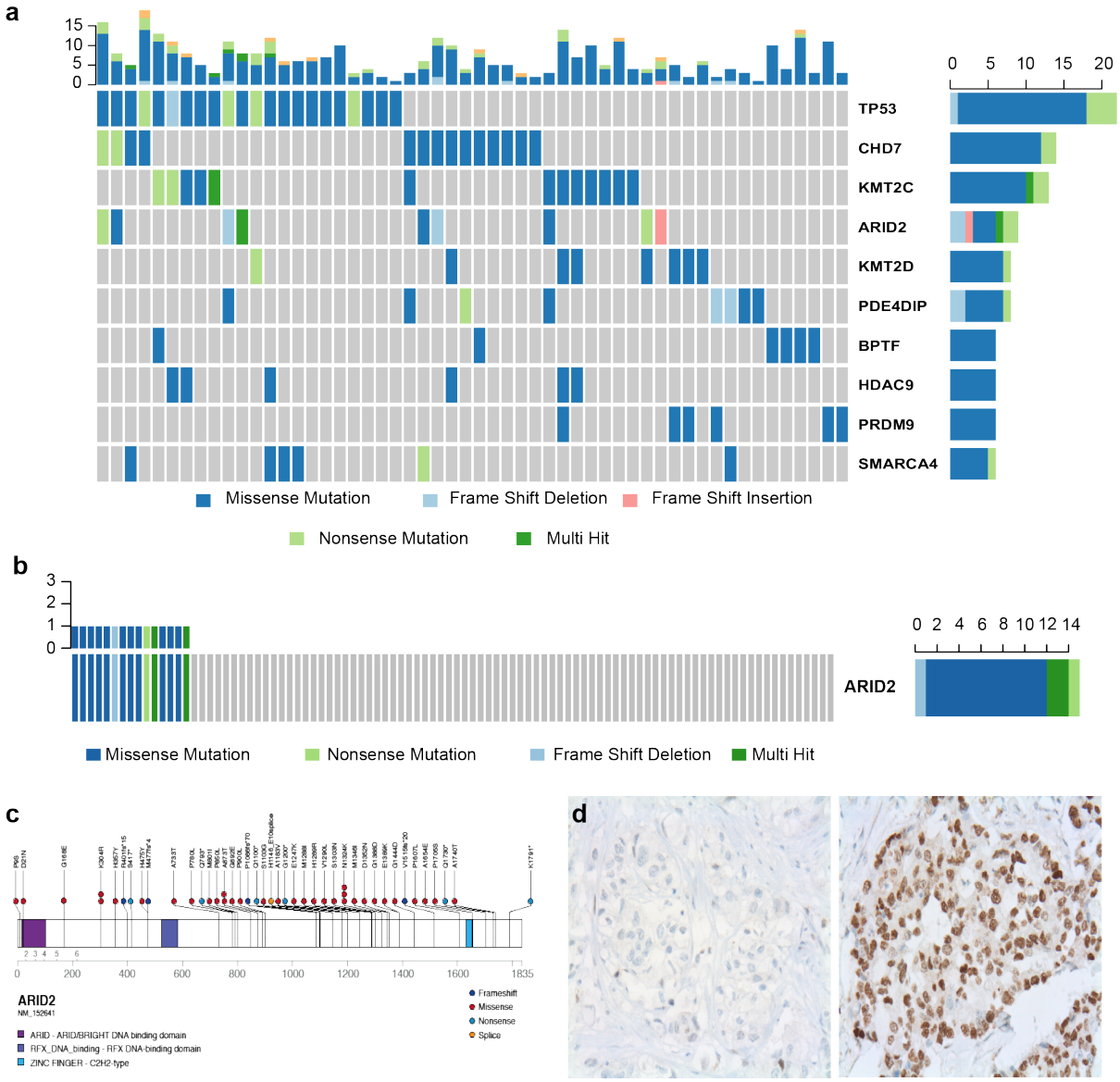

a
15
10
5
0
0 5 10 15 20
TP53
CHD7
KMT2C
ARID2
KMT2D
PDE4DIP
BPTF
HDAC9
PRDM9
SMARCA4
Missense Mutation
Frame Shift Deletion
Frame Shift Insertion
Nonsense Mutation
Multi Hit
b
3
2
1
0
0 2 4 6 8 10 12 14
ARID2
Missense Mutation
Nonsense Mutation
Frame Shift Deletion
Multi Hit
c
0 200 400 600 800 1000 1200 1400 1600 1835
ARID2
NM_152641
ARID - ARID/BRIGHT DNA binding domain
RFX_DNA_binding - RFX DNA-binding domain
ZINC FINGER - C2H2-type
Frameshift
Missense
Nonsense
Splice
d

Figure 1.

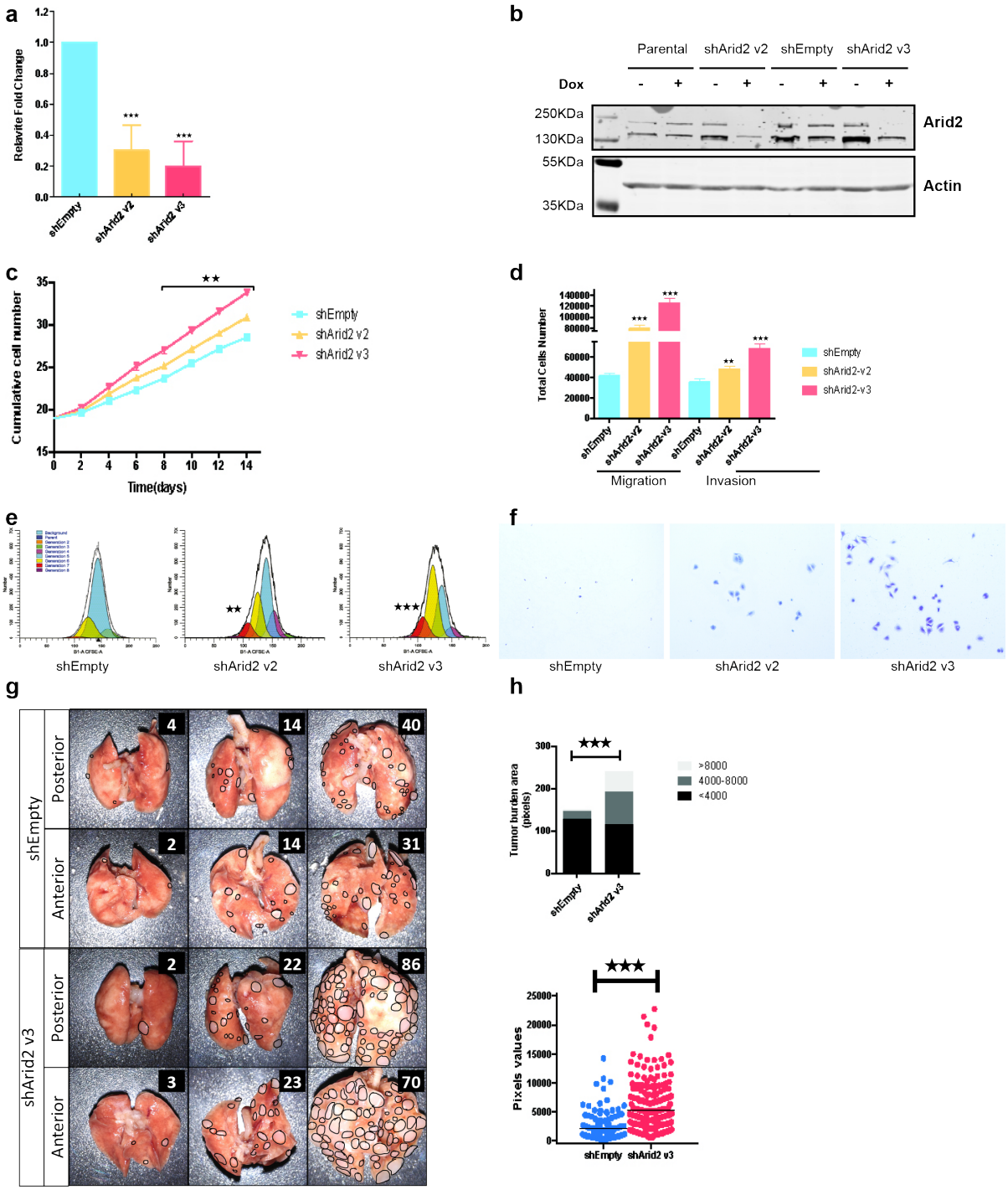
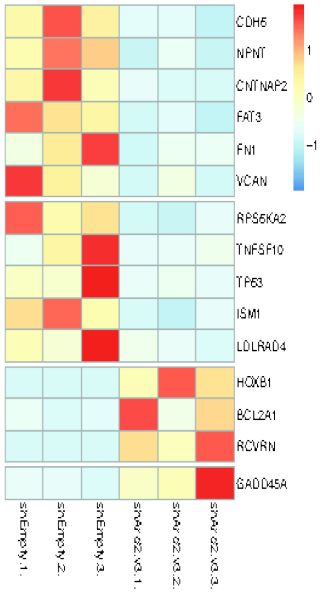

a
Relavite Fold Change
shEmpty
shArid2 v2
shArid2 v3
b
Parental
shArid2 v2
shEmpty
shArid2 v3
Dox
250KDa
130KDa
55KDa
35KDa
Arid2
Actin
c
Cumulative cell number
Time(days)
shEmpty
shArid2 v2
shArid2 v3
d
Total Cells Number
shEmpty
shArid2-v2
shArid2-v3
Migration
Invasion
e
shEmpty
shArid2 v2
shArid2 v3
f
shEmpty
shArid2 v2
shArid2 v3
g
shEmpty
shArid2 v3
Posterior
Anterior
4
14
40
2
14
31
2
22
86
3
23
70
h
Tumor burden area (pixels)
>8000
4000-8000
<4000
shEmpty
shArid2 v3
Pixels values
shEmpty
shArid2 v3

Figure 2.

a


CDH6
NPNT
CNTNAP2
FAT3
FN1
VCAN
RPS6KA2
TNFSF10
TP63
ISM1
LDLRAD4
HOXB1
BCL2A1
RCVRN
GADD45A
shEmpty.1
shEmpty.2
shEmpty.3
shAr.62.v3.1
shAr.62.v3.2
shAr.62.v3.3
1
0
-1

b

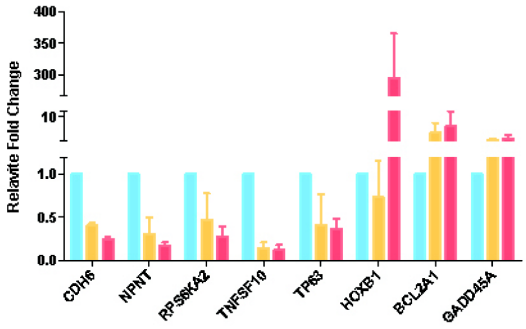

Relavite Fold Change
400
350
300
10
1.0
0.5
0.0
CDH6
NPNT
RPS6KA2
TNFSF10
TP63
HOXB1
BCL2A1
GADD45A

Figure 3.

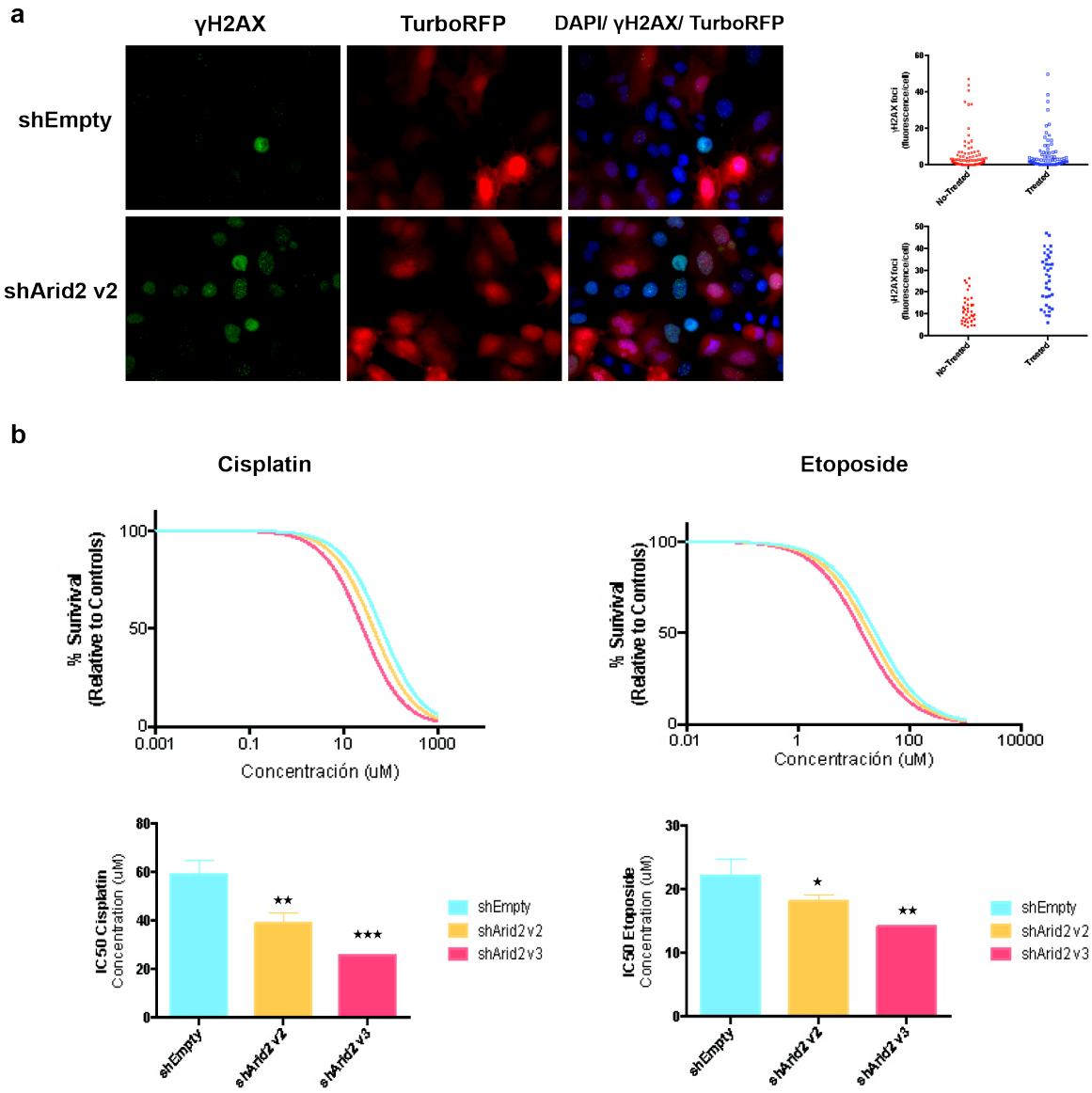

a
γH2AX
TurboRFP
DAPI/ γH2AX/ TurboRFP
shEmpty
shArid2 v2
γH2AX foci (fluorescence/cell)
No-Treated
Treated
b
Cisplatin
Etoposide
% Survival (Relative to Controls)
Concentración (uM)
IC50 Cisplatin Concentration (uM)
IC50 Etoposide Concentration (uM)
shEmpty
shArid2 v2
shArid2 v3

Figure 4.

mtDNA next-generation sequencing in human cancer identifies novel mutational and repair mechanisms

Thaidy Moreno(1), Laura Quevedo(1), Laura González-Silva(1), Carlos Revilla(1), Antonio Agraz-Doblas(1,2), Javier Freire(3), Santiago Montes-Moreno(3), Laura Cereceda(3), Pablo Isidro(4), Isabel Betancor(5), Aurora Astudillo(4), Irene Esposito(6), Manuel Ramírez(7), Francesc Solê(2), Ana Battle-López(8), Milagros Balbín(9), Javier Gomez-Roman(3), Eduardo Salido(5) and Ignacio Varela(1)

1) Instituto de Biomedicina y Biotecnología de Cantabria. Universidad de Cantabria-CSIC. Santander, Spain.
2) Instituto de Investigación contra la Leucemia Josep Carreras(IJC), Barcelona, Spain.
3) Departamentos de Patología.Biobanco Valdecilla. HUMV/IDIVAL. Santander, Spain.
4) Biobanco del Principado de Asturias (BBPA). Hospital Universitario Central de Asturias. Oviedo, Spain.
5) Departamento de Patología, Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER). Tenerife, Spain.
6) Institute of Pathology, Heinrich-Heine-University. Duesseldorf, Germany.
7) Hospital Niño Jesús. Instituto de Investigación Sanitaria de La Princesa. Madrid, Spain.
8) Departamento de Hematología. HUMV/IDIVAL. Santander, Spain.
9) Laboratorio de Oncología Molecular, Instituto Universitario de Oncología del Principado de Asturias (IUOPA). AGC Laboratorio de Medicina. Hospital Universitario Central de Asturias (HUCA). Oviedo, Spain.

Abstract

Mutations in mitochondrial DNA (mtDNA) have been described in almost all tumor types. Nevertheless, its causal involvement in tumorigenesis is still uncertain. Here, by using next-generation sequencing technologies, we report on of the largest mtDNA mutation set described in tumor samples in a single experiment. We have found that many of the mutations in mtDNA are homoplasmic and likely confer a selective advantage to tumor cells. The mutation profile is consistent with oxidative stress constituting the major cause of DNA damage in the mitochondria and producing a unique kind of mutation when recognized by the DNA polymerase gamma. Finally, we provide evidence for the occurrence of a mechanism of transcription-coupled repair in the mitochondria similar to that acting on nuclear DNA. Collectively, these findings offer new insights on the biology of the mitochondria and support its functional involvement in tumorigenesis.

Mitochondria are DNA containing organelles ubiquitous in eukaryotic cells and involved in different aspects of cellular metabolism, including ATP generation through oxidative phosphorylation. The human mitochondrial genome is a ~16 kb circular double-stranded DNA which contains 13 protein-coding genes. These proteins, together with additional ones encoded by the nuclear DNA, form the five complexes involved in oxidative phosphorylation. In contrast, all proteins involved in the transcription and replication of the mitochondrial DNA are imported from the cellular cytoplasm. Many previous studies have reported the presence of abnormal mitochondria in tumor cells and numerous mutations in the mtDNA have been described in virtually all known tumor types (M Brandon et al. 2006; Lu et al. 2009). The role of these mutations in tumorigenesis is unclear but some authors have proposed an oxygen-independent metabolism together with an increase in the ROS-dependent proliferative signals as the main advantages of dysfunctional mitochondrial activity for tumor cells (Czarnecka et al. 2010). Additionally, it has been shown that altered mitochondrial membrane potential impairs apoptosis and confers a higher resistance to chemotherapy (Mizutani et al. 2009). The tumorigenic potential of these alterations have been proved both in murine and human cells, in which pathogenic mtDNA mutations confer apoptosis resistance and promote metastasis (Shidara et al. 2005; Kulawiec et al. 2009). Most of the described mtDNA mutations are observed in the majority if not all the mitochondrial genomes present in each cell (between 10^3 and 10^4), a condition that is known as homoplasmy. This observation is consistent with the proposal that a mitochondria dysfunction is selected during tumor growth. However, other authors have proposed that the mutation shift from a heteroplasmic to a homoplasmic state could occur without selection (Coller et al. 2001).

Despite the growing evidence of the role of mitochondria in tumor development and the recent design of new techniques to identify mtDNA mutations (Maitra et al. 2004), the high number of mitochondrial genomes per cell and the existence of nuclear genome sequence of mitochondrial origin (numts) (Hazkani-Covo et al. 2010) make the identification of mitochondrial mutations a complex task. These difficulties have forced most of the researchers to focus their studies on a very limited number of samples or on certain variable mtDNA sequences like the regulatory region known as the D-loop. Next-generation sequencing technologies offer the opportunity to sequence the whole mitochondrial genome in a large collection of tumor samples, with enough sequence coverage to overcome most of the difficulties inherent to the heterogeneity of mitochondrial genomes (Yan Guo et al. 2012; He et al. 2010).

In the present study, through the use of next-generation sequencing technologies, we have generated a comprehensive list of mtDNA mutations identified in a large collection of cancer samples. All these mutational data provide evidence on the existence of new DNA repair and mutation mechanisms in the mitochondrial genome and support an active role of mitochondria alterations in human cancer.

Results

Mitochondrial genome sequencing and mutation identification

We generated on average 1 million reads of 100 bp aligned to the mitochondrial genome for each one of more than 500 different cancer samples of several tumor types using Illumina technology obtained an average of 10,000x for the mitochondrial genome. A detailed list of samples can be found in Supplementary Table 1.

Bioinformatic algorithms were developed to identify differences between the analysed tumor samples and the reference mitochondrial genome (Ensembl GRCh37). Overall, 170 potential somatic mutations were called by our algorithms (Supplementary Table 2). Among them, 110 were present in the 13 protein coding genes. Interestingly, 93% of the identified mutations (102/110) were predicted to produce a missense change which is a 40% more of what it is expected by chance ($p<0.001$). This could indicate a strong positive selection to accumulate potentially deleterious mutations in the mtDNA. Most of the identified mutations where in different degrees of heteroplamia but we found that a bit less than 10% of the mutations (14/170) were present in at least 80% of the reads and therefore were considered as homoplasmic. As they are somatic mutations, it indicates a very fast shift from an heteroplasmic to an homoplasmic state during the tumor progression.

Notably, we found mutations in almost all mitochondrial genes (Fig. 1). This finding is in agreement with the assumption that most if not all these genes are essential for a correct function of the mitochondria and any disruption in any of them would produce the same deleterious effect.

Mutation profile of mitochondrial DNA in human cancer

The majority of mutations found in our study were G>A/C>T and T>C/A>G transitions (Fig. 2B). Strikingly, these mutations are not evenly distributed between the two DNA strands. Thus, we observed a three-fold enrichment of G>A mutations on the untranscribed strand compared to that expected by chance ($P = 1.06 \times 10^{-9}$) (Fig. 2C). This observation evidenced firstly that this kind of mutations are likely the result of lesions on the G nucleotide and, secondly, that many of the same lesions occurring on the transcribed strand are correctly identified and removed by the cell. This process, known as transcription-coupled repair, is produced when the stalled RNA recruits the nucleotide excision repair (NER) repair machinery and has been well described in the nuclear DNA (Nouspikel 2009). Our data suggest that a similar process takes place in the mitochondria, even when an equivalent NER repair system has not been described yet on this organelle (Larsen et al. 2005) , and is responsible for repairing most of the modified guanines. This

strand bias is also present significantly in the T>C/A>G transitions with a near two-fold enrichment of T>C mutations on the untranscribed strand ($P = 5.56 \times 10^{-6}$).

Mitochondrial mutations are thought to be produced mainly as a result of the higher oxidative stress affecting this organelle as a consequence of ATP oxidative phosphorylation. The major by-product of this process is 7,8-dihydro-8-oxo-deoxyguanosine (8-OXO-dG), and higher levels of this compound have been reported in the mitochondria (Barja and Herrero 2000). Interestingly, no G>T/C>A transversions, the principal mutation described in the nuclear DNA as a result of this modified base, have been observed in our mutation set. In contrast, and as discussed above, the majority of mutations found in our study were G>A/C>T transitions (Fig. 2B). This observation suggests that either all the 8-OXO-dG produced in the mitochondria is efficiently repaired by the mitochondria and does not result in mutations; or, most likely, that this lesion on the guanine produce G>A transitions instead of G>T transversions on mtDNA. This can be explained if the POLG polymerase involved in mtDNA replication mainly introduces T to pair with 8-OXO-dG instead of A, as it happens in the nuclear DNA. Differences in the behaviour of different polymerases in response to the same oxidative DNA damage have been previously described (Greenberg 2004). The second more common mutation observed in our study is T>C/A>G transitions, which is likely the result of the oxidation of thymine to thymidine glycol. This by-product has also been described as a consequence of a high oxidative stress and can pair with guanine during DNA replication (Iijima et al. 2009).

Finally, to check if our mutational data was consistent with that described before, we analyzed the manually curated mtDNA variations database MITOMAP (http://www.mitomap.org) (Ruiz-Pesini et al. 2007). The main aim of this resource is to report published and unpublished data on human mitochondrial DNA variation. We focused our study on those mutations annotated in the database as somatic. This set is expected to be biased towards moderately homoplasmic mutations (as most of the mutations have been identified by capillary sequencing) and towards specific, well studied regions (more than one third of the mutations are reported in the regulatory region known as the D-loop). Nevertheless, we observed the same mutation profile and strand bias described in our data (Supplementary Figure 1). This finding indicates that our mutation profile is not a unique feature of our sample set but the result of a general mutagenesis process affecting all tumor mitochondrial genomes.

Discussion

Several works support the role of mitochondrial dysfunction in human cancer. Nevertheless, the study of mtDNA has been limited by the large number of potentially different mtDNA molecules per cell, the high similarity between the mtDNA sequence and some nuclear regions, and the low sensitivity of traditional

nucleotide sequencing technologies. Accordingly, most previous studies have analyzed low number of samples or have focused on specific mtDNA regions (Chatterjee et al. 2006). The recent advent of the so-called next-generation sequencing technologies has offered a new opportunity to overcome these difficulties (Shendure and Ji 2008) .

In the present study, massively parallel sequencing has allowed us to perform a detailed analysis of the complete mtDNA genome sequence of a large set of tumor samples. The large amount of generated data offers several insights into the mitochondrial biology and its potential involvement in the tumor process. Thus, although mitochondrial heteroplasmy inheritance has been described (Yan Guo et al. 2012) , we have provided evidence that most of the heteroplasmic mutations observed in tumor mitochondria are mainly the result of recently acquired de novo mutations. These mutations are rapidly purified into an homoplasmic state. It has been described that this shift does not require active positive selection (Coller et al. 2001), nevertheless, the high amount of potentially deleterious mutations suggest that these confer selective advantage to the tumor DNA, likely due to a shift to a oxygen-independent energy production. In agreement with this observation, during the preparation of this manuscript, Larman and cols have found independent evidence of selective advantage conferred by mitochondrial mutations in human tumors (Larman et al. 2012). The active role of mitochondria dysfunction in tumorigenesis has been reinforced recently. Thus positive correlation between presence of somatic mtDNA mutations and tumor progression has been described in human breast cancer (Tseng et al. 2011) , and mitochondrial genome instability has been observed to lead to tumorigenesis in mice (P-L Chen et al. 2012). Besides the shift to an oxygen-independent metabolisms, mitochondrial dysfunction could interfere with the function of tumor suppressor genes codified in the nuclear DNA. Thus, SIRT3, a newly identified tumor suppressor is localized in the mitochondria (Kim et al. 2010) and the alteration of the redox balance can alter the phosphorilation of proteins like ATK or PTEN (Kulawiec et al. 2009). Further studies will be necessary to fully characterize the precise mechanism underling the role of mitochondrial dysfunction in tumorigenesis. In this respect, the recent description of a method to correct human mitochondrial mutations *in vivo* would be offer new research possibilities to the field (Wang et al. 2012).

In this study we report a considerable strand bias on the distribution of the mutations. Similar strand bias towards the heavy strand (rich in guanines and that corresponds to the non-transcribed strand in all the genes but ND6) has been explained in the past due to the longest single-stranded time that are spent by this strand on a asymmetrically replication model of the mtDNA. Nevertheless, this explanation would implicate a higher frequency of mutations toward the origin of replication that we do not observe in our study (Figure 1). Additionally, this wo not explain why this effect does not affect equally to all type of mutations. We suggest instead that this strand bias is the consequence of the existence of a mechanism of transcription-coupled repair active on mtDNA similar to the one described in nuclear DNA (Hanawalt and Spivak 2008) and that this mechanism is more active

on G>A/C>T than on T>C/A>G transitions. This hypothesis is especially risky considering that a mitochondrial equivalent of the NER repair mechanism, the one responsible for the transcription-coupled repair on nuclear DNA, has not been described on mtDNA. Nevertheless, it is remarkable that over recent years, several studies have described the occurrence of mtDNA repair mechanisms similar to those taking place on nuclear DNA, including BER, MMR, homologous recombination and non-homologous end joining (Boesch et al. 2011) . Therefore, further studies will be necessary to clarify the detailed molecular mechanism by which transcription-coupled repair occurs in mitochondria. Finally, we must emphasize that the mutational profile described here is consistent with mtDNA mutations being the result of the DNA oxidation produced by oxidative phosphorylation. Interestingly, this oxidation produces a unique mutagenesis effect in the mitochondria when the 8-OXO-dG is recognized by the POLG DNA polymerase, generating as consequence G>A/C>T transitions instead of the G>T/C>A transversions observed on nuclear DNA (Delaney et al. 2012).

In summary, this study represents one of the largest DNA mitochondrial sequencing effort performed until now in tumor samples and offers new insights into the mitochondrial biology, including its functional involvement in human cancer.

Methods

Samples

576 different cancer samples were analysed (see supplementary Table 1). All the patients gave written informed consent for sample collection and analysis. All samples were anonymised at sample collection so none of the researchers involved in the process could have access to the patient personal data. In the case of solid tumors, representative histological samples were reviewed by a pathology committee to verify a minimum of 70% of tumor content.

Massive parallel sequencing and computer analysis

Genomic libraries from the different tumor samples were generated using 500 μg of total genomic DNA. Briefly, genomic DNA was randomly fragmented to between 200 and 300 bp by focused acoustic shearing. These fragments were end repaired and ligated with the Illumina paired-end adaptors. After a PCR enrichment step, we mix groups of 96 samples for target enrichment with a SureSelect user-defined probe set (Agilent) containing the complete mitochondrial genome sequence. After sequencing in a High-Seq instrument, the resultingFASTQ sequence files were aligned to the human genome (Ensembl GRCh37) using the BWA algorithm. The resulting alignment file was processed for our own in-house written software RAMSES. Finally, known polymorphisms recorded in human

mitochondrial genome database (http://www.genpat.uu.se/mtDB) (Ingman and Gyllensten 2006) and mitomap database (http://www.mitomap.org) (Ruiz-Pesini et al. 2007) were removed from the list of potential mutations. Additionally, we remove of our list those mutations present in more than two different samples considering them likely germline variants particular to our sample collection and, therefore, not present in the databases. Considering the sequencing and alignment errors occurring as a consequence of the existence of nuclear DNA sequences homologous to mtDNA, those mutations in which the mutant allele was observed in more than 80% of the reads were considered as homoplasmic.

Figure legends

Figure 1. List of somatic mutations found in mitochondrial DNA. Figurative representation of all the mutations found in mtDNA in different tumor samples. Gene ideograms are shown around the outer ring and are oriented anticlockwise. The different genes belonging to the same protein complex are represented in the same color. ND6 is represented differently to indicate that is located in a different strand that the rest of genes. From outside to inside, the next ring represents the position of the different homoplasmic mutations: non-synonymous mutations (purple rectangles) and synonymous mutations (red rectangles). The next ring represents heteroplasmic mutations and the most inner ring represents small insertions or deletions.

Figure 2. Mitochondrial DNA mutation profile in tumor cells. a) Number of mutations observed in each of the six possible mutation classes. b) Fraction of G>A and T>C transitions observed in transcribed and untranscribed strand. The ratio of the observed versus expected mutations is represented considering the different base composition of each strand. ***P<0.001

Suppl. Figure 1. Mitochondrial DNA mutation profile in mitomap database. a) Number of mutations observed in each of the six possible mutation classes. b) Fraction of G>A and T>C transitions observed in transcribed and untranscribed strand. The ratio of the observed versus expected mutations is represented considering the different base composition of each strand. ***P<0.001

Acknowledgements

This work has been supported by two grants (SAF2012-31627 and SAF2016-76758-R)of the Spanish Ministerio de Economía y Competitividad. I.V has also been supported by the Spanish Ministerio de Economía y Competitivida, Programa Ramón y Cajal. We would like as well to acknowledge the support of the Servicio Santander Supercomputación. Finally, we would like to thank also the

work of the staff members of the different tumor biobanks that suministrated us with patient samples, specially to the IDIVAL and HUCA biobanks, for their exceptional work in sample collection and organization.

References

Barja G, and Herrero A. 2000. Oxidative damage to mitochondrial DNA is inversely related to maximum life span in the heart and brain of mammals. *FASEB J.* **14**: 312–318.

Boesch P, Weber-Lotfi F, Ibrahim N, Tarasenko V, Cosset A, Paulus F, Lightowlers RN, and Dietrich A. 2011. DNA repair in organelles: Pathways, organization, regulation, relevance in disease and aging. *Biochim. Biophys. Acta* **1813**: 186–200.

Brandon M, Baldi P, and Wallace D C. 2006. Mitochondrial mutations in cancer. *Oncogene* **25**: 4647–4662.

Chatterjee A, Mambo E, and Sidransky D. 2006. Mitochondrial DNA mutations in human cancer. *Oncogene* **25**: 4663–4674.

Chen P-L, Chen C-F, Chen Y, Guo XE, Huang C-K, Shew J-Y, Reddick RL, Wallace D C, and Lee W-H. 2012. Mitochondrial genome instability resulting from SUV3 haploinsufficiency leads to tumorigenesis and shortened lifespan. *Oncogene*. http://www.ncbi.nlm.nih.gov/pubmed/22562243 (Accessed August 23, 2012).

Coller HA, Khrapko K, Bodyak ND, Nekhaeva E, Herrero-Jimenez P, and Thilly WG. 2001. High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.* **28**: 147–150.

Czarnecka AM, Kukwa W, Krawczyk T, Scinska A, Kukwa A, and Cappello F. 2010. Mitochondrial DNA mutations in cancer--from bench to bedside. *Front. Biosci.* **15**: 437–460.

Delaney S, Jarem DA, Volle CB, and Yennie CJ. 2012. Chemical and biological consequences of oxidatively damaged guanine in DNA. *Free Radic. Res.* **46**: 420–441.

Guo Y, Cai Q, Samuels DC, Ye F, Long J, Li C-I, Winther JF, Tawn EJ, Stovall M, Lähteenmäki P, et al. 2012. The use of next generation sequencing technology to study

the effect of radiation therapy on mitochondrial DNA mutation. *Mutation Research*. http://www.ncbi.nlm.nih.gov/pubmed/22387842 (Accessed March 14, 2012).

Hanawalt PC, and Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* **9**: 958–970.

Hazkani-Covo E, Zeller RM, and Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* **6**: e1000834.

He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, Diaz LA Jr, Kinzler KW, Vogelstein B, and Papadopoulos N. 2010. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* **464**: 610–614.

Iijima H, Patrzyc HB, Budzinski EE, Freund HG, Dawidzik JB, Rodabaugh KJ, and Box HC. 2009. A study of pyrimidine base damage in relation to oxidative stress and cancer. *Br. J. Cancer* **101**: 452–456.

Ingman M, and Gyllensten U. 2006. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res.* **34**: D749–751.

Kim H-S, Patel K, Muldoon-Jacobs K, Bisht KS, Aykin-Burns N, Pennington JD, van der Meer R, Nguyen P, Savage J, Owens KM, et al. 2010. SIRT3 is a mitochondria-localized tumor suppressor required for maintenance of mitochondrial integrity and metabolism during stress. *Cancer Cell* **17**: 41–52.

Kulawiec M, Owens KM, and Singh KK. 2009. Cancer cell mitochondria confer apoptosis resistance and promote metastasis. *Cancer Biol. Ther.* **8**: 1378–1385.

Larman TC, Depalma SR, Hadjipanayis AG, Protopopov A, Zhang Jianhua, Gabriel SB, Chin L, Seidman CE, Kucherlapati R, and Seidman JG. 2012. Spectrum of somatic mitochondrial mutations in five cancers. *Proc. Natl. Acad. Sci. U.S.A.* http://www.ncbi.nlm.nih.gov/pubmed/22891333 (Accessed August 23, 2012).

Larsen NB, Rasmussen M, and Rasmussen LJ. 2005. Nuclear and mitochondrial DNA repair: similar pathways? *Mitochondrion* **5**: 89–108.

Li H, Ruan J, and Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**: 1851–1858.

Lu J, Sharma LK, and Bai Y. 2009. Implications of mitochondrial DNA mutations and mitochondrial dysfunction in tumorigenesis. *Cell Res*. **19**: 802–815.

Maitra A, Cohen Y, Gillespie SED, Mambo Elizabeth, Fukushima N, Hoque MO, Shah N, Goggins M, Califano J, Sidransky David, et al. 2004. The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. *Genome Res*. **14**: 812–819.

Mizutani S, Miyato Y, Shidara Y, Asoh S, Tokunaga A, Tajiri T, and Ohta S. 2009. Mutations in the mitochondrial genome confer resistance of cancer cells to anticancer drugs. *Cancer Sci*. **100**: 1680–1687.

Nouspikel T. 2009. DNA repair in mammalian cells: Nucleotide excision repair: variations on versatility. *Cell. Mol. Life Sci*. **66**: 994–1009.

Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi Pierre, and Wallace Douglas C. 2007. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res*. **35**: D823–828.

Shendure J, and Ji H. 2008. Next-generation DNA sequencing. *Nat. Biotechnol*. **26**: 1135–1145.

Shidara Y, Yamagata K, Kanamori T, Nakano K, Kwong JQ, Manfredi G, Oda H, and Ohta S. 2005. Positive contribution of pathogenic mutations in the mitochondrial genome to the promotion of cancer by prevention from apoptosis. *Cancer Res*. **65**: 1655–1663.

Tseng L-M, Yin P-H, Yang C-W, Tsai Y-F, Hsu C-Y, Chi C-W, and Lee H-C. 2011. Somatic mutations of the mitochondrial genome in human breast cancers. *Genes Chromosomes Cancer* **50**: 800–811.

Wang G, Shimada E, Zhang Jin, Hong JS, Smith GM, Teitell MA, and Koehler CM. 2012. Correcting human mitochondrial mutations with targeted RNA import. *Proc. Natl. Acad. Sci. U.S.A*. **109**: 4840–4845.

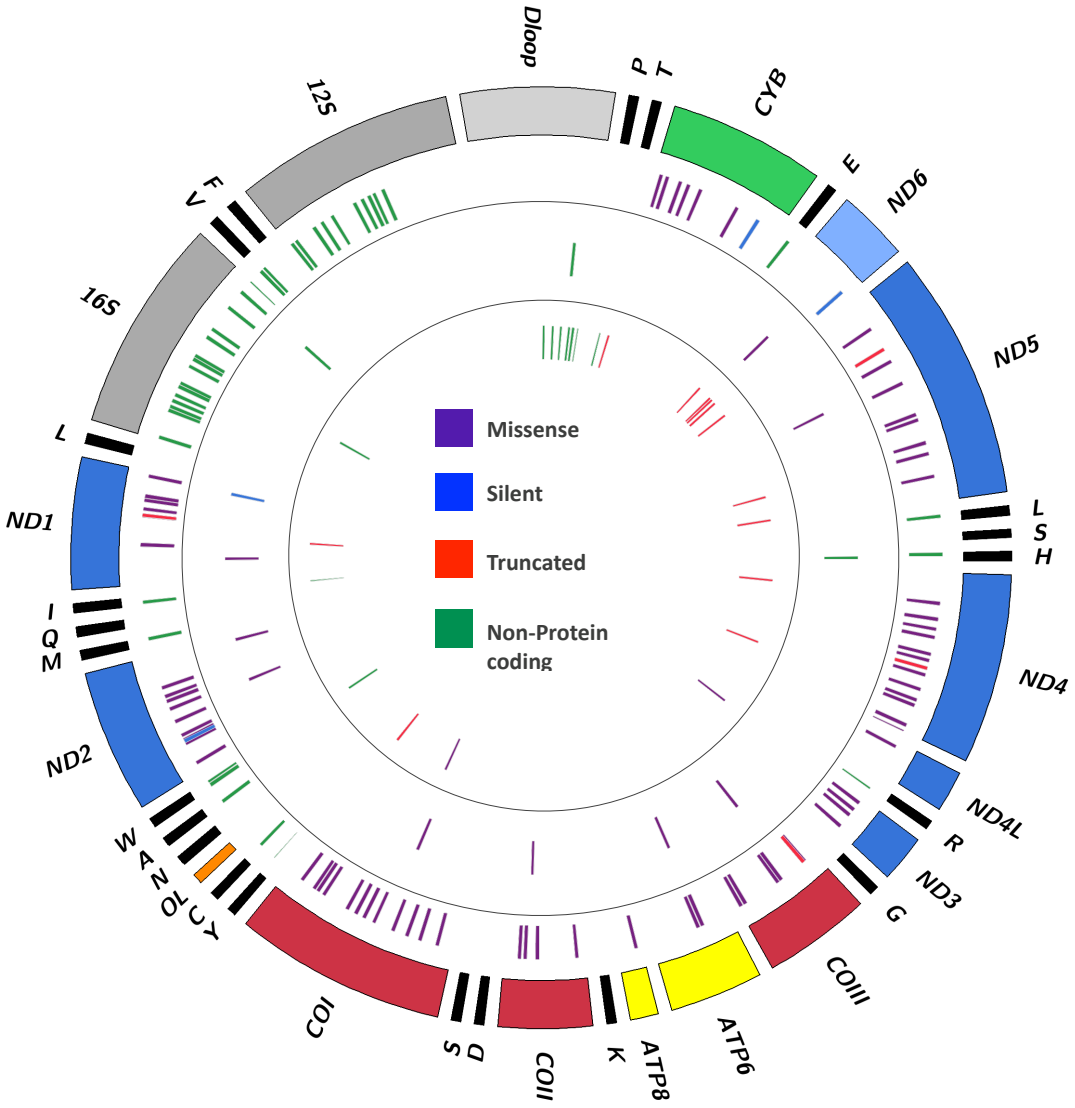

Dloop
P
T
CYB
E
ND6
ND5
L
S
H
ND4
ND4L
R
ND3
G
COIII
ATP6
ATP8
K
COII
D
S
COI
C
Y
O
L
N
A
W
ND2
M
Q
I
ND1
L
16S
V
F
12S
Missense
Silent
Truncated
Non-Protein coding

Figure 1

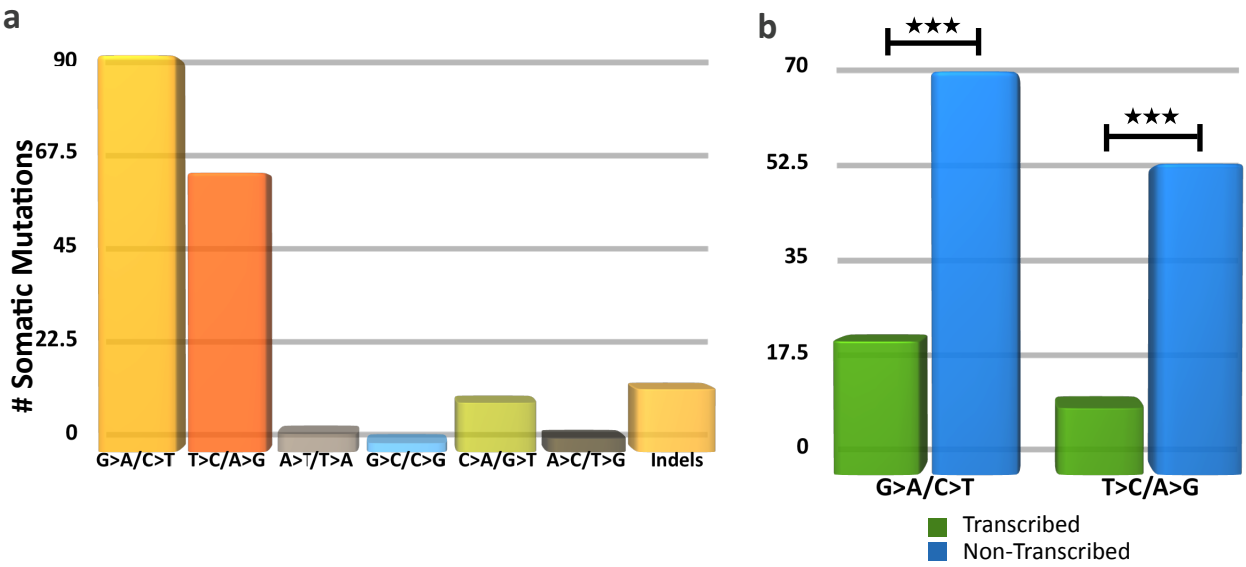

a
Somatic Mutations
90
67.5
45
22.5
0
G>A/C>T
T>C/A>G
A>T/T>A
G>C/C>G
C>A/G>T
A>C/T>G
Indels
b
70
52.5
35
17.5
0
★★★
★★★
G>A/C>T
T>C/A>G
Transcribed
Non-Transcribed

Figure 2

ACKNOWLEDGMENTS

Primero quisiera agradecer a mi director de tesis y mentor, el Dr. Ignacio Varela, quien me dio esta gran oportunidad. Nacho, a lo largo de los años me has brindado una orientación paciente y reflexiva, con el equilibrio adecuado entre ofrecer consejos y sugerir ideas que me permitió formarme y crecer como científico. Tu continuo entusiasmo por la ciencia me ha servido de inspiración.

Al equipo Confetti, Laura, Carlos, Antonio, Laurita "peque", Bea, quienes me ayudaron tanto y de los cuales aprendi tanto, pero en especial a Lau, por tu constante apoyo, por nuestras eternas discusiones en el café, científicas y no tan científicas. Por nuestros "miércoles de cine", por compartir conmigo tu amor por la ciencia, por enseñarme a ser mas pragmática, mejor compañera, por aceptarme con mis rarezas y por reírte conmigo y de mi.

To Paola and her Cancer Epigenetics group in The Francis Crick Institute. I have great gratitude for all. The time I shared with you guys was definitely a defining moment for me and marked a turning point in my life. I want to give special thanks to Paola for taking the time for teaching me so many things and especially for making me want to be a better scientist. To Cristina for all the help that you provide me and for making me feel like home. To Eleanor, thanks to you I have my own "Swan Epiphany". To Josep, Tom, Tristan, and Matt, for making me feel part of the team.

A mis primeros compis de laboratorio. A Magda y a Iago. Iago, siempre recordaré tu frase "Estoy a punto de colapsar". Magda, thanks for make me feel "normal", for sharing the way I see the world, because from the minute I first met you, you offered me your friendship and honesty and selfless advice and you became a true friend.

A todas las personas que pasaron por el laboratorio. A Silvia, tantas cosas que aprendí de ti, eres un ejemplo para mi y te admiro por tu constancia y optimismo. Ana, Raquel, Lety, Alba e Isabel, de quienes aprendí tantas cosas.

Al grupo de Javier Leon y Dolores Delgado, por hacerme sentir bienvenida desde el primer momento. A sus chicas, Judit, Esther, (mis cheerleaders), Rosa y Lorena. A Lucia, gracias por las risas y las anécdotas compartidas.

Al grupo de Piero Crespo, por toda la ayuda que me brindaron. En especial a Berta, por sacar tiempo para ayudarme siempre que lo necesitaba.

A mis evalúadores Dolo y JuanMa, gracias por el apoyo, guía y concejo que he recibido durante los últimos 4 años.

A las fuentes de financiación y los biobancos, ya que sin ellos esta tesis no seria posible.

Por ultimo, a mi familia, a quienes les debo todo lo que tengo y lo que soy. Sin su apoyo incondicional nada de esto hubiera sido posible.