Bioinformatic characterization of non-annotated transcription-units related to Celiac Disease

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Silvia Fernandez Portela

Figure 1. *The symbol of the Barred spike is internationally recognized for safe for consumption by people with celiac disease.*

Introduction

Celiac disease (CD) is a chronic, immune-mediated intolerance to gluten. It is a complex multigenic disease with genetic and non-genetic components. Molecular pathways underlying pathogenesis of celiac disease are poorly understood. Although the 40% of population carries the major risk factor (HLA-DQ2 and DQ8 polymorphisms [1]), only 1% develops the disease.

When a reference genome is available, RNA-seq analysis will normally involve mapping the reads onto the reference genome or transcriptome to infer which transcripts are expressed. Mapping solely to the reference transcriptome of a known species precludes the discovery of new, unannotated transcripts and focuses the analysis on quantification alone.

This project focuses on those transcripts that are detected by the RNAseq and have not been annotated yet.

Identifying novel transcripts using the short reads provided by Illumina technology is one of the most challenging tasks in RNA-seq [2]. Several methods, such as Cufflinks [3], incorporate existing annotations by adding them to the possible list of isoforms and novel transcripts.

Hypothesis

RNAseq serves to detect differential expression of transcripts across the whole genome. So, it could also be possible to identify unannotated significantly expressed genomic regions.

These regions would give new useful information to understand when and why they are expressed in CD.

Objective

Detect, identify, characterize and validate novel transcripts involve in CD patients.

Methods

Table 1. Patients and biopsies for RNAseq and for qPCR analyses. *Active CD: children at diagnosis (on a gluten-containing diet, with CD-associated antibodies, atrophy of intestinal villi and crypt hyperplasia). Treated CD: same patients in remission after being treated with GFD for >2 years (asymptomatic, antibody negative and normalized intestinal epithelium). Control: tissue samples from non-celiac individuals not suffering from inflammation at the time of endoscopy, used as controls.*

	Active CD	Treated CD	Control
RNA_Seq	4		4
Validation	16	16	15

RNAseq

For RNAseq analysis, first *Sickle* [4] was used to remove low quality reads. Then the *Tuxedo* protocol [5] was followed. Briefly, sequenced reads were mapped against human reference genome (hg38) using *Tophat* [6] and providing GENCODE 24 [7] as the reference transcriptome. *Cufflinks* [3] was used to find new transcripts.

Bioinformatic characterization

The bioinformatic characterization was perfomed using public databases (*Figure 2*).

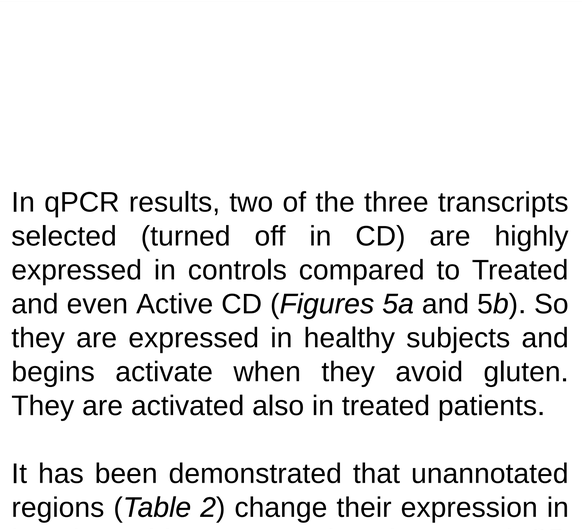

Figure 2. Bioinformatic characterization workflow*. Result is significant when Query Cover* and *Identity >90%*

qPCR and statistical analysis

qPCR and Statistical analysis followed the next workflow (*Figure 3*).

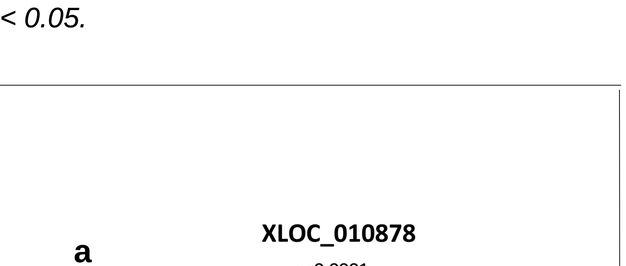

Figure 3. qPCR experimental workflow. *Differences were assumed significant when p < 0.05.*

Results and Discussion

From the RNAseq experiment 276 unannotated transcription units are extracted and referred to four different groups: 135 differentially expressed, 33 which are completely off in celiac patients; 13 completely off in healthy patients; 95 relation changes, genes that are correlated with each other.

Bioinformatics classification was very successful, with only 5% of unannotated regions that were not similar to anything (*Figure 4*). Nevertheless, bioinformatic analysis deserves a further study because it may happen that unannotated transcript is really annotated; this can happen due to a misreading of the bioinformatic analysis after RNAseq.

After the bioinformatics analysis, the choice of three regions for validation was based on the number of isometrics and exon (both equal to 1). Furthermore, it is supposed that transcripts that switch on and off are more interesting than others. Thus three transcripts are selected(*Table 2*). One that is turned on in CD (XLOC_022314) and two that are turned off in CD patients (XLOC_010878, XLOC_012919).

In qPCR results, two of the three transcripts selected (turned off in CD) are highly expressed in controls compared to Treated and even Active CD (*Figures 5a* and 5*b*). So they are expressed in healthy subjects and begins activate when they avoid gluten. They are activated also in treated patients.

It has been demonstrated that unannotated regions (*Table 2*) change their expression in healthy subjects, treated patients and CD patients. Therefore these results confirm RNASeq data information.

Consequently, the detection of new transcripts by the RNAseq could open a new line of study for celiac disease research, and can be extrapolated to other areas in Genetics.

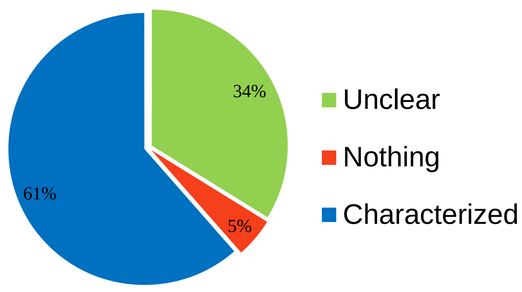

Figure 4*. Bioinformatic characterization results.*

Table 2*. Three chosen transcripts.*

	XLOC_002314	XLOC_012919	XLOC_010878
Localization	chr20:50,278,410-50,279,321	chr16: 5,017,237-5,018,210	chr14:104,082,856-104,083,931
Search	USCS Genome Browser	USCS Genome Browser	Blast
Similar to	LincRNA: LINC01272	mRNA: SEC14L5	Predicted LOC105370691

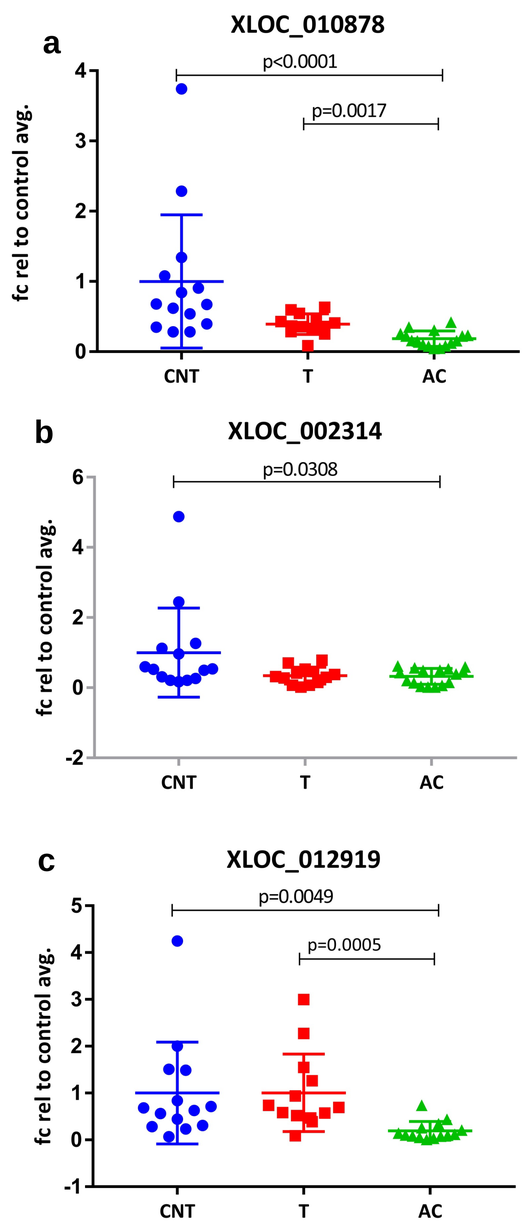

Figure 5*. qPCR results of three transcripts. Fold change to control average vs. Control (CNT) Treated (T) and Active (AC) patients.*

References & Acknowledgements

This work is part of projects ISCIII-PI13/1201 GVSAN- 2011111034 to JRB, and are approved by the Cruces University Hospital and Basque Clinical Trials and Ethics Committees (Codes CEIC- E09/10 and PI2013072) Biopsies of distal duodenum were obtained by endoscopy after informed consent was obtained from all subjects or their parents. qPCR are made by SGIKER by Fluidigm platform.

I would like to thank Dr. **Jose Ramon Bilbao** for his support and encouragement and Dr. **Koldo Garcia-Etxebarria** for his invaluable technical assistance.

[1] Sollid, L. M., & Lie, B. A. (2005). Celiac disease genetics: current concepts and practical applications. *Clinical Gastroenterology and Hepatology*, *3*(9), 843-851.
[2] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A Survey of Best Practices for RNA-seq Data Analysis.
[3] Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).
[4] Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle. 2011 (2011).
[5]Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–78 (2012).
[6] Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
[7]. Harrow, J. et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 22, 1760–1774 (2012).