Proposed project: RNA-seq data mining to uncover the role of Immunoglobulin germ-line transcription.

Abstract

Antibody production is an essential feature of adaptive immune responses. After subsequent encounters with the same pathogen, antigen-specific memory B cells and high affinity antibody-secreting cells are generated. This process is genetically defined by class switch recombination (CSR) and somatic hypermutation (SHM), which represents the molecular basis of immunological memory (Xu *et al*., 2012).

In particular, CSR is preceded by and depends on non-coding transcription ("Sterile" or 'germ-line' transcripts) up-stream of the coding region of the antibody constant region (Chr14. q32.33, spanning ~390 Kb) (Xu *et al*., 2012). Germ-line transcripts (GLT) are about 300 bp and have been studied in mice. In humans, however, little information is known regarding its structure and expression patterns in healthy immune responses and diseases.

While performing 454 Rep-seq from human peripheral blood (Georgiou *et al*. 2014, Cortina-ceballos *et al*., 2015), we came to notice that a small percentage of the sequences (0.5-4% aprox) were mapping in an unannotated region upstream from the coding exons of IgG1 and IgG3, suggesting that they could represent GLT's. Although biological replicates were few, we observed some trends in post-vaccinated individuals and in patients with rheumatoid arthritis. We have now developed a qPCR assay to quantitate GLT.

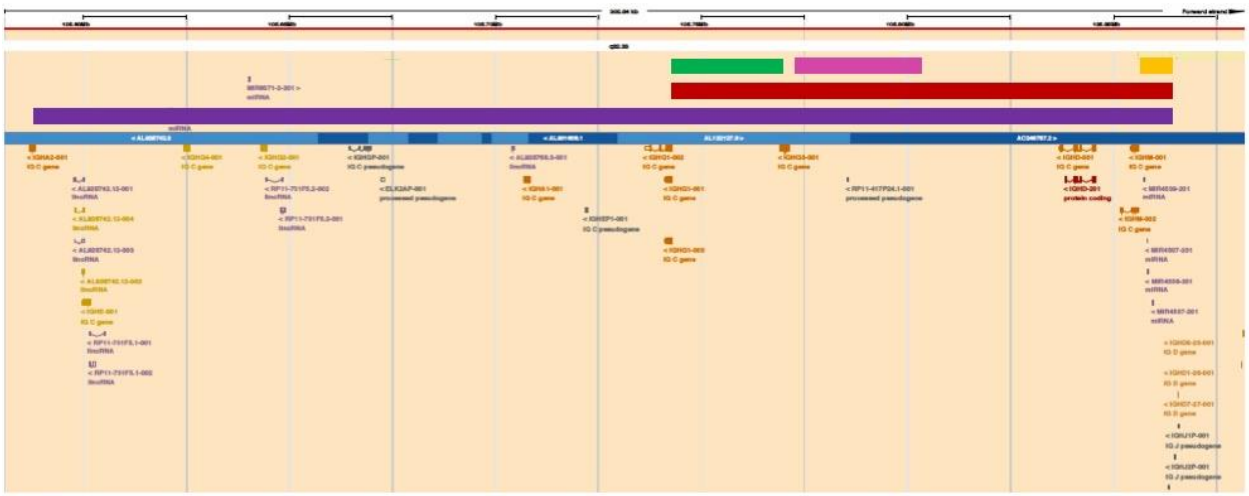
Objective: To take advantage of the wealth of data present in the SRA to mine the full genomic ranges of GLT transcription and to identify differentially expressed regions in the IGH locus.

Region of interest: The whole locus of interest spans 390 Kb and includes annotated protein coding genes as well as some non-coding RNA's and pseudogenes (Figure. *purple* track).

http://www.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000270474;r=14:105580968-105881806;t=ENST00000604817

Agnostic analysis of the whole locus would be nice since we could compare GL transcription vs IGHC transcription (protein coding). However, this may be a technically more complex task. Additionally, since coding transcription may be significantly higher than non-coding, this could represent an additional statistical problem.

Alternatively, we could restrict the region of interest only to the IGHG1-IGHM range (Figure, red track), or simply restricting the search to the intergenic regions of the corresponding IGHC gen where we have evidence of GL transcription (Figure. *yellow*, *pink* and *green* tracks).



🟪 Complete locus: 390 Kb
🟥 IgG1-IGHJ locus: 128 Kb
🟩 IgG1-IgG3 intergenic locus: 28 Kb
🟪 IgG3-IgM intergenic locus: 30 Kb
🟨 IgM-IGHJ intergenic locus: 7 Kb

Proposed methodology: Please take into consideration that we lack experience in R and big data analysis. However as far as we can understand from Collado-Torres, *et al*. 2016 manuscripts, we would like to propose the following approach:

1) If feasible, we would have to define which of the SRA projects have significant germ line transcription in the proposed locus (Figure). In principle this would provide valuable information regarding the conditions and type of experiment (mRNA, poly-A -, miRNA) in which GLT are not expressed.
   We assume that this could be done from bigWig file, but we don´t know if is possible to automate. If we understand correctly, this could also be done by calculating a coverage matrix for all SRA samples and then filtering out all the regions that don´t map in the proposed locus (Figure 1). Is this correct? Which approach would be easier?

2) Then, it would be great to identify differential expression of GLT and to identify splicing variants. In conjunction with available metadata in the SRA, this could provide us invaluable information of the biological role of GLT. This is particularly of interest, since this GLT may serve as biomarkers of vaccination response, certain types of B cell lymphomas and leukemias.
3) In case the proposed approach is too complicated, we could focus our attention to those RNA-seq experiments directly related to human immune responses (lymphoid organs, and peripheral blood, vaccination, infection, leukemia/lymphomas and autoimmune diseases.

References:

- Xu Z, Zan H, Pone EJ, Mai T, Casali P. 2012. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. Nat Rev Immunol. Jun 25;12(7):517-31

- Georgiou, George and Ippolito, Gregory C and Beausang, John and Busse, Christian E and Wardemann, Hedda and Quake, Stephen R. 2014. The promise and challenge of high-throughput sequencing of the antibodyrepertoire**.** Nat Biotech. 32(2):158-168

- Cortina-Ceballos B, Godoy-Lozano EE, Téllez-Sosa J, Ovilla-Muñoz M, Sámano-Sánchez H, Aguilar-Salgado A, Gómez-Barreto RE, Valdovinos-Torres H, López-Martínez I, Aparicio-Antonio R, Rodríguez MH, Martínez-Barnetche J. 2015. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. Genome Med. Nov 25;7:124.

- Leonardo Collado-Torres, Abhinav Nellore, Alyssa C. Frazee, Christopher Wilks, Michael I. Love, Ben Langmead, Rafael A. Irizarry, Jeffrey T. Leek, Andrew E. Jaffe. 2016. Flexible expressed region analysis for RNA-seq with derfinder. *Nucl Acids Res*; 45 (2)

- Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, Jeffrey Leek. 2016. recount: A large-scale resource of analysis-ready RNA-seq expression data. doi: https://doi.org/10.1101/068478