Search for an aetiological virus candidate in Chronic Lymphocytic Leukaemia by extensive transcriptome analysis

Methods

Bioinformatics analysis of 454 data

An in-house script was used to account for the quality drop at 3' ends of pyrosequencing reads. Further cleaning of the data was performed with SeqClean (http://compbio.dfci.harvard.edu/tgi/software/), trimming reads with 454 adapters, polyA tails, low complexity or ambiguous stretches, and also those with ribosomal, mitochondrial or contaminant signatures. A minimum admitted length of 50 nts was applied for the final sequences.

Digital subtraction strategy. The digital subtraction strategy was conservative in order to avoid false discovery of new viral sequences in the CLL sample (Fig 1A, Fig S1). The objective of the pipeline was to discard from the CLL reads, those with a human transcriptome or genome origin (*ie* to discard all human sequences independently if they belong or not to known human genes or transcripts). In addition, by subtraction of the Normal sample reads, we reinforced this issue, by confirming that each candidate xenobiotic sequence was exclusively observed in the CLL sample. While comparison between both samples was done with BLAST (Altschul *et al*, 1997), the subtractions of the human transcriptome and genome reads were carried out by BLAT (Kent, 2002) search on the references downloaded from Ensembl version GRCh37.53 (Flicek *et al*, 2011; http://www.ensembl.org/). All similarity searches were done with an E-value cutoff of $1x10^{-10}$.

Search of viral reads. For those unmapped reads, BLAST similarity searches were conducted in the non-redundant, viral genome and viral protein databases (http://www.ncbi.nlm.nih.gov/ gb170 and RefSeq-34 releases; Pruitt *et al*, 2009).

Performance. To assess whether our pipeline was able to account for novel viral sequences present in CLL but not in the Normal sample, we took an HIV genome (NC_001802.1) and simulated sequences with 10%, 20%, 30% and 40% of divergence from the original one. The starting genome and the simulated ones were cut achieving a similar length distribution as the

454 reads and 100 pseudo-reads per divergence level were added to the CLL reads. The pipeline was run again and the fate of the HIV simulated reads followed (Fig S1). Sequences with 10 to 30% divergence were all identified as HIV sequences whereas those with 40% divergence as viral ones. All were included among the non-Normal and non-human origin reads.

Expression analysis. To have a glimpse on the transcriptome depth obtained, reads were mapped with GMAP (Wu & Watanabe, 2005) to the human reference genome, using GRCh37.55 as the transcriptomic reference (version: gmap_2007_09_28, options: -m -A). In-house scripts were used to summarize GMAP output and assign read count values to each gene model.

Bioinformatics analysis of Illumina data

Illumina purity filter reads were screened for quality scores and existence of low complexity sections. Python routines were used with the following requirements: presence of no more than two Ns in the first 70 nts; non presence of homopolymers longer than 15 nts and a minimum quality score of 15 in the first 60 nts. For the filtered reads, the quality control indicator was searched and trimmed if present to fix the 3' ends. While this procedure results in reads of variable length (60 to 115 nts), most of the final sequences kept their original size (115 nts).

Digital subtraction strategy. To discard all sequences of presumed human origin, reads were sequentially mapped to well-annotated human rRNA, hg19 genome assembly and transcriptomics databases (Fig 1B). Reference genome and gene models were downloaded from http://genome.ucsc.edu/ (Kent *et al*, 2002; Fujita *et al*, 2011) and Ensembl version GRCh37.60. All mappings were accomplished with Burrows-Wheeler Alignment tool (BWA; Li & Durbin, 2009), a fast light-weighted aligner of short reads which provides gapped alignments and allows a tight control of its behaviour. To increase sensitivity when assigning reads to human origin, we used very permissive options (seed length: 32; maximum mismatches in the seed: 3; maximum number of gaps opened: 2; maximum total number of differences: 6).

Search of viral reads. For remaining unaligned reads, successive nucleotide based BLAST similarity searches were applied to detect candidate viral sequences in the CLL libraries (Fig

1B). First, reads were searched in the viral database. Second, the candidate viral reads (*ie* only those showing a match against a viral genome) were searched once more in human references using BLASTN instead of BWA this time (we expected to have among this candidates some reads of truly human origin but, as a consequence of either polymorphisms or sequencing errors, they were not able to be mapped by BWA). A discriminant function was used to classify the reads as human or viral ones based on both BLAST results (see below). Last, remaining candidate viral reads where searched in the non-redundant database and the final BLAST output carefully inspected. Viral genome, human genome *plus* transcriptome and nucleotide non-redundant (nt) databases were downloaded in March 2011 from NCBI.

BLASTN protocols. In relation to the amount of Illumina unmapped sequences to be aligned (about two million reads per sample), nucleotide BLAST searches need much more computational resources and a compromise between sensitivity and cost had to be achieved when choosing the BLASTN protocol. To evaluate this issue, we simulated two viral genomes, HIV type 1 and EBV, with increasing levels of divergence (randomly mutating 10% to 50% of the genome nucleotides) to the chosen starting genomes (NC_001802.1 and NC_007605.1; Fig S2). From each original and mutated genome, random sequences of 115 nts were extracted and the pool aligned against the viral database with different BLASTN protocols: (i) default megablast; (ii) default blastn; (iii) “blastn sensitive” (word_size 9, reward 1, penalty -1); (iv) “blastn w7” (word_size 7, reward 1, penalty -1). Execution times and hit counts were recorded. As expected, both variables grew with improved BLAST protocols, showing maximum sensitivity with the “blastn w7” one (Table SI). However, “blastn sensitive” also performed suitable for much lower run-time requirements so it was the chosen protocol to BLAST the unaligned reads in each step above described (compared to “blastn sensitive”, “blastn w7” showed a 10-fold time increase; Table S1).

Linear discriminant analysis. To compare BLAST outputs and assign reads to a viral or human origin, a linear discriminant analysis (LDA; Fisher, 1936) was implemented. Simulated viral reads were matched against viral genomes and hg19 reference using the previously chosen BLAST protocol. Similarly, *in silico* extracted human genome reads (115 nts) were also aligned to human and viral databases with same parameters. These resulted in combined BLAST output files for 597 true positive viral reads and 597 true positive human reads, used in turn as a training set for the LDA. Accordingly to previous observations, the discriminant function was built on the “alignment length” and “E-value” variables. Specifically, it was based on the observed difference in alignment length and E-value for the “virus *versus* human”

BLAST results comparison. The classifier showed 80% of true positives for the viral class while zero false positives were observed for the human class. The obtained LDA coefficients were later employed to filter out candidate viral reads showing better alignments to human genome *plus* transcriptome. LDA was implemented with the MASS package (http://cran.r-project.org/web/packages/MASS/) in the R environment (Venables & Ripley, 2002; R Development Core Team, 2011).

Expression analysis. To explore whether sample preparation protocol and Illumina sequencing succeed to improve transcript detection, SAM files from BWA alignments to hg19 genome assembly were input into HTSeq (HTSeq: Analysing high-throughput sequencing data with Python; http://www-huber.embl.de/users/anders/HTSeq/). Gene read counts were obtaining using the "intersection-nonempty" option and the same gene models employed in the 454 analysis (GRCh37.55).

Results

Candidate Mason-Pfizer monkey virus reads

Almost all reads showing a viral match mapped anti-sense to the primer binding site (PBS) of the Mason-Pfizer monkey virus (MPMV). This primate betaretrovirus carries at its 5' end two long terminal repeats separated by 63 nucleotides complementary to tRNA-Lys, the usual primer for reverse transcription of the viral genome (Sonigo *et al*, 1986). To study the presence of additional MPMV reads in our transcriptomes, we conducted a BLAST search of all non-ribosomal reads of each sample against the three MPMV genomes currently available (AF0033815.1, M12349.1 and U85506.1). An average of 801.0 ± 317.5 reads per sample were identified as matching anti-sense to the PBS, whereas none other region of the MPMV genome appeared covered. Following these reads throughout our pipeline, for instance 1,244 reads from the CLL250 sample, 41 of them had been mapped by BWA to hg19 assembly at positions annotated as tRNA-Lys. Remaining 1,203 reads were first recovered when searched with BLAST in the viral database and then recovered again as human genome plus transcriptome reads in the next pipeline step. All these reads were classified as "human" by the classifier but 367 (including the 336 candidate MPMV reads) had posteriors lower than 0.9

and thus were finally searched in the non-redundant database, eventually exhibiting the highest similarity to MPMV. Further analysis of these 1,203 reads revealed that they are chimeric-like sequences, where the first stretch of the read derives from a tRNA-Lys precursor while the following region is variable (see GC content decay in Fig S3A), pointing that none single analysed read could be extended beyond the retroviral PBS. Of note, the 336 MPMV reads are enriched with the 3' CCA extension of the mature tRNA (Fig S3B) raising the possibility that additional tRNA processing modifications could have occurred (Findeiβ *et al*, 2011), thus impairing their similarity to human references. Lastly, as they seemed to be tRNA derived sequences, we assumed that this phenomenon could be widespread and searched putative MPVM's PBS reads in other DSN normalized human samples as well as other tRNA derived sequences among our unmapped reads (data not shown). Both approaches gave positive results thus definitively excluding a true viral origin of these reads.

References

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Findeiβ S., Langenberger D., Stadler P.F. & Hoffmann S. (2011) Traces of post-transcriptional RNA modifications in deep sequencing data. *Biological Chemistry*, **392**, 305-313.

Fisher R. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179-188.

Flicek P., Amode M.R., Barrell D., Beal K., Brent S., Chen Y., Clapham P., Coates G., Fairley S., Fitzgerald S., Gordon L., Hendrix M., Hourlier T., Johnson N., Kähäri A., Keefe D., Keenan S., Kinsella R., Kokocinski F., Kulesha E., Larsson P., Longden I., McLaren W., Overduin B., Pritchard B., Riat H.S., Rios D., Ritchie G.R., Ruffier M., Schuster M., Sobral D., Spudich G., Tang Y.A., Trevanion S., Vandrovcova J., Vilella A.J., White S., Wilder S.P., Zadissa A., Zamora J., Aken B.L., Birney E., Cunningham F., Dunham I., Durbin R., Fernández-Suarez X.M., Herrero J., Hubbard T.J., Parker A., Proctor G., Vogel J. & Searle S.M. (2011) Ensembl 2011. *Nucleic Acids Research*, **39**(Database issue), D800-D806.

Fujita P.A., Rhead B., Zweig A.S., Hinrichs A.S., Karolchik D., Cline M.S., Goldman M., Barber

G.P., Clawson H., Coelho A., Diekhans M., Dreszer T.R., Giardine B.M., Harte R.A., Hillman-Jackson J., Hsu F., Kirkup V., Kuhn R.M., Learned K., Li C.H., Meyer L.R., Pohl A., Raney B.J., Rosenbloom K.R., Smith K.E., Haussler D. & Kent W.J.. (2011) The UCSC genome browser database: update 2011. *Nucleic Acids Research*, **39**(Database issue), D876-D882.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M. & Haussler D. (2002) The human genome browser at UCSC. *Genome Research*, **12**, 996-1006.

Kent W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Research*, **12**, 656-664.

Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754-1760.

Pruitt K.D., Tatusova T., Klimke W. & Maglott D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, **37**(Database issue), D32-D36.

R Development Core Team. (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Sonigo P., Barker C., Hunter E. & Waint-Hobson S. (1986) Nucleotide sequence of a Mason-Pfizer monkey virus: an immunosuppresive D-type retrovirus. *Cell*, **45**, 375-385.

Venables W.N. & Ripley B.D. (2002) Modern applied statistics with S. Fourth edition. Springer, New York.

Wu T.D. & Watanabe C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859-1875.

Figure legends

Fig S1. Computational subtraction approach developed to identify non-human sequences in the CLL transcriptome obtained with the 454 technology. The illustration summarizes the number of reads remaining at each step of the pipeline. The fate of 500 simulated reads from HIV is also followed.

Fig S2. Simulated reads from HIV and EBV genomes were pooled and searched using different BLASTN protocols in the viral genome database. To simulate divergent genomes,

10% to 50% of the original genomic nucleotides were randomly changed and 100 reads of 115 nts length were extracted for each divergence level (three replicates per level were performed). The number of reads aligning to the starting genome decreased with the divergence level but increased through more exhaustive BLASTN protocols. The inset shows the modified parameters in each protocol.

Fig S3. Compositional analysis of 1,203 reads that aligned anti-sense to the primer binding site PBS of the MPMV genome (sample CLL250). (A) Change in the G+C content of all 1,203 reads along position. While there is a clear signature in the first 63 nts (coincident with the PBS length), the G+C content turns to 50% in the second part of the reads, pointing to a random composition. This suggest the presence of chimeric reads with a common first region (the one that aligns anti-sense to the PBS) and a variable second part. (B) The frequency of the CCA triplet is followed along the reads. The black line corresponds to the 336 reads recovered as MPMV-like reads in the final BLASTN search. The grey shadow corresponds to the 867 reads assigned to “human” by the LDA. The expected CCA end added during tRNAs maturation (arrow) is more frequent in the MPMV-like reads, suggesting the presence of additional changes in these reads. It is known that modified bases in the template end up as sequencing errors (although certain base pairing rules prevail), thus impairing read similarity to the reference sequence when aligned.

Fig S4. Median values of Illumina raw read counts for genes with 0-5 read counts when previously analysed by 454. Medians are shown for the four samples analysed. Neither normalization nor standardization was performed among samples. Horizontal read lines repeat the 454 observed value to ease comparison.