A bioinformatic strategy for large-scale identification and annotation of chromosomal aberrations in tumors

A.Guffanti[*1], L.Luzi[*1], S.Confalonieri*[1], M.Trubia[2], S.Volorio[1], Stéphane Graziani[3],
P.G. Pelicci[2], & P.P. Di Fiore[1]
* These authors contributed equally to this work
[1] IFOM – FIRC Institute of Molecular Oncology, Via Adamello, 16 – 20139 Milan, Italy
[2] IEO – European Institute of Oncology, via G.Ripamonti,45 – 20141 Milano, Italy
[3] ISoft - Chemin de Moulon 91190 Gif-sur-Yvette France

Abstract

We describe here the rationale, implementation and results of a bioinformatic strategy for large-scale identification and annotation of chromosomal translocations in tumours, based on sequence and annotation comparison between human transcriptome and EST partial cDNA sequences derived from tissues or cell lines. We also illustrate how the sequencing and subsequent careful bioinformatic analysis of a number of identified candidate translocation cDNAs revealed the complexity of distinguishing recombination from true translocation events. Finally, we suggest some EST filtering and cleaning strategy for pursuing EST-based "in silico" translocation identification projects.

Rationale

Cytogenetic analysis of tumour cells has revealed that recurring chromosomal abnormalities such as translocations, deletions and inversions are present in many tumours. In leukemias, lymphomas, and sarcomas, these specific chromosomal aberrations are frequently associated with specific morphologic subtypes. Typically, these changes are reciprocal translocations 1 . Similar genetic abnormalities are seen in solid tumours, e.g. the 11;22 translocation in Ewing sarcomas and the inversion of proximal 10q in papillary thyroid carcinomas 2. We may envisage that a noticeable quantity of genetic rearrangements in tumour tissues or tumour cell lines (including solid tumours) has never been addressed on a systematic manner apart from the large collection of clinical cases and cytogenetic evidences represented by the Mitelman database (http://cgap.nci.nih.gov/Chromosomes/Mitelman).

Starting from this working hypothesis, we began a bioinformatic strategy for large-scale identification and annotation of chromosomal translocations, based on sequence and annotation comparison between human transcriptome and EST partial cDNA sequences from tissues or cell lines. We did not aim at answering any biological question such as addressing the causes of this genomic instability, or even the causative relation between cancer and genomic instability, but instead to provide a reliable collection of purely bioinformatic evidences of possible translocations, the genes involved, the tissues and the best possible annotation of the involved genes.

After the completion of the Human Genome Sequences, genome-wide investigations on the genome features that may be associated with recurrent chromosomal aberrations in tumors have appeared 3. Together with the annotation of the Human Genome Sequence, a group at the Sanger Centre performed an effort to detect chimaeric transcripts from oncogenic fusions genes generated by chromosomal translocations, the ends of which mapped to different locations. Fusions were detected to the same degree in both normal and neoplastic tissues, indicating a significant level of false positives but also the complexities of a genomic approach, dogged by low-frequency repeats and multiple, high-fidelity copies of some genes, and the limited amount of variable quality of DNA sequences from cancer cells that was available at the time of publication 4.

We planned to reduce the possible impact of these errors using the following strategies:

- obtaining the carefully selected IMAGE clones and performing single-pass sequencing of the DNA clones in order to reduce sensibly the amount of false positives due to misassembly, mistracking and wrong curation.
- filtering the search results through a step of manual curation performed by biologists with expertise in Oncology and Cancer Genetics in order to shrink drastically the amount of biologically implausible results.
- adding another search strategy using the ORF-centered EST sequences from the ORESTES project.

Identification of translocation events from IMAGE EST clones (end sequencing)

This data mining project is based on the identification of putative rearranged chimeric transcript in IMAGE EST clone inserts by association of 5' and 3' respective EST ends from the public domain to the human transcriptome extracted from genomic resources and detection of reliable clone mismatches, diagnostics of possible translocations; The outline of the procedure is summarized in Fig. 1

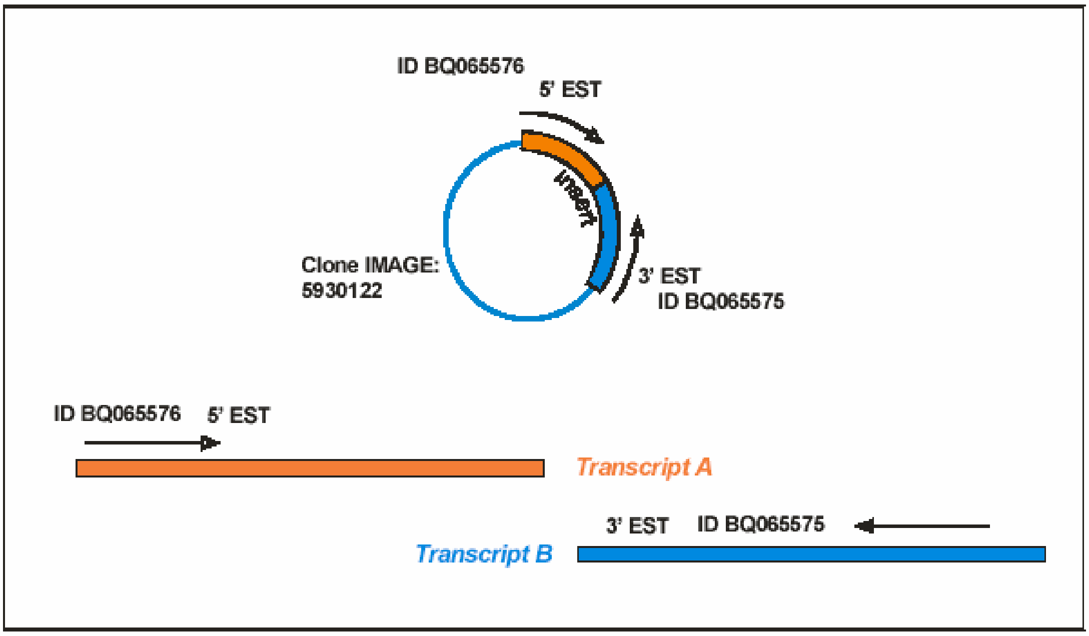

Fig.1 – schema of the IMAGE EST-based translocation identification project. A single cDNA IMAGE clone is sequenced on both ends and the EST sequences should contain the same clone information. In a clone derived from a putative translocation, the two EST sequences will be associated with different cDNA transcripts.

By sequencing EST clones after a number of steps of selection we eliminated the mismatches due to annotation errors. By careful annotation both of EST sequences and transcripts we restricted the list of possible candidates to a more realistic set, with two possible approaches (genetic and functional). In general terms, we intended to include in the final database the greatest possible number of sequences available.

We used both the SRS (Sequence Retrieval System) sequence database indexing and retrieval system, mirrored at IFOM (http://bio.ifom-firc.it/srs6/) and a series of perl scripts or the commercial Data Transformation tool, Amadea from ISoft (www.isoft.fr) and its bio-informatics dedicated module used in the HKIS project (www.hkis-project.com) to retrieve all the IMAGE EST clones with two end sequences from the EMBL database and check for discordances between the annotation and the sequence identifiers (r1/s1 and x1/y1 identifiers). We started from 3,089,202 human IMAGE EST sequences contained in EMBL rel. 71 and ended up with 2,054,606 EST with full annotation and tissue information. Of these, a total of 583,140 total clones had at least two sequences, but 362,189 with missing or incomplete tissue information.
In order to identify the putative translocations, we selected two human transcriptome sources: cDNAs from the EnsEMBL project (http://www.ensembl.org) and the RefSeq project (http://www.ncbi.nlm.nih.gov/RefSeq/). For EnsEMBL we selected "known" and "novel" transcripts.
The pipeline for identification of translocations based on the IMAGE EST approach consists in a series of BlastN searches of the complete query cDNA dataset, one transcript at a time, versus the IMAGE EST dataset, and in the subsequent parsing of the results. We retrieved, from each Blast run, the single best-scoring HSP (based on Blast scores) for the first 3.000 matches with the target EST dataset. The Blast searches are filtered with the dust low-complexity DNA filter from NCBI.
We used an E-value threshold of 1e-100, requiring only one HSP (High Scoring Segment Pair) for each EST vs. cDNA match. The parsing of the search results followed this schema:
1) For each Blast alignment, record the EST clone identifier, the EST Accession Number and the query transcript identifier. The main identifier becomes the IMAGE clone identifier, and all the other identifiers (query cDNA and EST Accession Number) will be linked to that.

2)

a) If the clone identifier has been entered for the first time, record it together with the EST Accession Number and the corresponding query cDNA sequence identifier.

b) If the clone has been already found during the parsing procedure, compare the current query cDNA identifier with the list of the already recorded identifiers:

- if they are identical, it is a multiple match with the same cDNA or with a very similar gene family member. This uncertainty will be solved in a successive step of manual analysis.
- if the identifier of the current cDNA query is different from all the others associated with the current EST clone, it could point to a putative translocation and it will be recorded in a different list.
- if an IMAGE clone finds on the two ends the same cDNA query and also two different cDNA matches, we will keep both possibilities (i.e. a normal, non translocated cDNA and a putative translocation) only if the value of the alignments are comparable.
- matches of more than two ESTs with a query cDNA sequence are allowed; however, if three EST sequences from the same clone match two different query cDNAs, this clone is recorded as diagnostic of a putative translocation. There is a reduced amount of IMAGE clones, which have more than two sequences in dbEST.

3) The final list of the IMAGE clones that are predicted having two or more different cDNA sequences is completed, through a series of further programs, with a number of annotations referred both to the EST and to the genes involved in the putative translocation.

The most important immediate aim of the manual annotation of the results generated by this sequence comparison pipeline was reducing the number of false positives due to high sequence similarities between different query cDNA sequences. Although using the human transcriptome derived from the human genome instead of directly the genome sequence reduced the impact of this problem, we still may have cases in which two DNA sequences from the same EST clone match different transcripts simply because the two cDNA are almost identical (although identified with different names) or have significant portion of similarity due, for instance, to alternative splicing or to a very highly conserved region.

We established a sequence annotation procedure devoted to the analysis of the selected cDNA query matches and also we correlated the coordinates of the EST matches with the conserved sequence region to rule out this kind of false positives.

Also, by relating the gene identifiers from the two different query dataset (Ensembl and RefSeq) with a common secondary DNA sequence database such as UniGene, we were able to compare the results between the different strategies regardless of annotation errors and point our attention to the putative translocations present in the intersection of the two datasets. Ensembl, in fact, gives a link to RefSeq entries but the opposite is not true. In addition, it may happen that different transcripts belonging to the same UniGene cluster could be alternative splice isoforms. Thus, correlating the two query cDNA datasets with UniGene, we ruled out another sizeable part of false positives.

Results and perspectives

Our bioinformatics procedure identified 188 clones corresponding to translocation or recombination events. We ordered from HGMP Gene Services and sequenced 42 interesting targets, ending with 16 full-length cDNA sequences, corresponding to 10 putative translocation events. We will discuss the detailed sequence analysis of these sequences, highlighting some new findings about the nature of IMAGE EST cloning procedures that make difficult the use of this material for these ‘in silico translocation identification’ investigations. We will also propose some filtering procedures to try and eliminate recombination events background. We are currently applying our bioinformatic pipeline to sequences extracted from EMBL Rel. 77 and addressing a different translocation identification strategy, based on the comparison of ORESTES EST sequences with human transcriptome and detection of incomplete alignments. We are also addressing the problem of incomplete tissue classification by applying the SAMBI eVOC ontologies to our EST dataset.

References

1. Rabbitts, T.H. Chromosomal translocations in human cancer. Nature 372, 143-9 (1994).
2. Vecchio, G. & Santoro, M. Oncogenes and thyroid cancer. Clin Chem Lab Med 38, 113-6 (2000).
3. Kolomietz, E., Meyn, M.S., Pandita, A. & Squire, J.A. The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genes Chromosomes Cancer 35, 97-112 (2002).
4. Futreal, P.A. et al. Cancer and genomics. Nature 409, 850-2 (2001).