ABSTRACT: Computational Analysis of Oncogenic Gene Fusions

Shugay M, Ortiz de Mendibil I, Vizmanos JL, Novo FJ*

*Department of Genetics, University of Navarra, Pamplona, Spain. *E-mail: fnovo@unav.es*

The most common class of somatic mutation found in the human cancer-gene census involves chromosomal translocations fusing two different genes that result in a chimeric transcript[1]. Gene fusions lead to deregulated gene expression or are translated as fusion proteins with oncogenic potential due to the presence of protein domains that normally are located in separate proteins. There is compelling evidence that fusions represent an initial event in oncogenesis[2]. It is estimated that around 20% of all cancers are caused by gene fusions driven by chromosomal translocations[3] and the overall role of fusions in cancer morbidity is estimated as 17%[2].

Currently, the knowledgebase available for fusion proteins in human cancers is growing large, with several hundreds of clinical reports contained in Mitelman database[2]. Many putative fusions were predicted solely from EST and mRNA sequences found in publicly available databases[4]. With the advent of next-generation sequencing technologies then number of putative fusions is growing rapidly as novel fusions are being predicted from RNA-Seq data[5–7].

The first aim of our studies was to analyze available high-throughput datasets with computational and robust statistical methods, in order to identify genomic hallmarks of fusion partner genes (FPGs). Taking advantage of our database of fusion sequences[8] we found that fusion genes are overexpressed due to promoter and 3'UTR substitution and that this trend is more general than believed earlier[9]. Furthermore, expression profiling of 5' FPGs and of interaction partners of 3' FPGs indicates that these features can help to explain tissue specificity of hematological translocations. Analysis of protein domains retained in fusion proteins identified specific functional signatures in gene fusions[9,10]. It is generally accepted that chromosomal proximity in the nucleus can explain the specific pairing of 5' and 3' FPGs and the recurrence of fusions, but our analysis of chromosomal contact capture data (Hi-C) showed that neither of these trends is statistically significant. However, we show that FPGs are preferentially located in early replicating regions and occupy distinct clusters in the

nucleus.

The second aim of our study was to use inferred hallmarks of fusion genes to build a classifier to identify 'driver' fusions from the vast majority of 'passenger' events captured by high-throughput fusion screening. We show that a simple, yet highly accurate, Bayesian classifier could be trained on a specific set of features. We rigorously test the performance of our classifier on thousands of putative and known gene fusion instances and shown that: 1) most predicted driver fusions are composed of FPGs that lie in crucial cancer pathways; 2) the classifier scores (p-values) correlate with clinical frequency of fusions, provide clues on tissue-specificity and distinguish driver chromosomal translocations from their reciprocal non-oncogenic counterparts. Copy number analysis using SNP array data allowed us to test classifier predictions by searching for characteristic breakpoint signatures near FPGs. Such signatures were enriched around fusions classified as driver events. We also show and discuss the advantages of our pipeline over existing algorithms and propose that it could be of great benefit to an increasing part of the scientific community who are now performing RNA-Seq studies on cancer samples, as it allows filtering the list of putative gene fusions and focusing on the most reliable targets for experimental validation.

This work has been funded with the help of the Spanish Ministry of Science and Innovation (SAF 2007-62473), the PIUNA Program of the University of Navarra, the Caja Navarra Foundation through the Program "You choose, you decide" (Project 10.830) and ISCIII-RTICC (RD06/0020/0078). MS is funded by a grant from ADA UNAV.

References

1. Futreal, P. A. et al. *Nat Rev Cancer* **4**:177–83 (2004).
2. Mitelman, F., Johansson, B. & Mertens, F. *Nat Rev Cancer* **7**:233–45 (2007).
3. Nambiar, M., Kari, V. & Raghavan, S. C. *Biochim Biophys Acta* **1786**:139–52 (2008).
4. Kim, P. et al. *Nucl Acids Res* **38**:D81–5 (2010).
5. Sakarya, O. et al. *PLoS Comp Biol* **8**:e1002464 (2012).
6. Edgren, H. et al. *Genome Biol* **12**:R6 (2011).
7. Frenkel-Morgenstern, M. et al. *Genome Res* **22**:1231–42 (2012).
8. Novo, F. J., De Mendibil, I. O. & Vizmanos, J. L. *BMC Genomics* **8**:33 (2007).

9. Shugay, M., Ortiz de Mendibil, I., Vizmanos, J. L. & Novo, F. J. *PLoS Comp Biol* **8**:e1002797 (2012).
10. Ortiz de Mendibil, I., Vizmanos, J. L. & Novo, F. J. *PLoS One* **4**:e4805 (2009).