Chromosome translocations in cancer: computational evidence for the random generation of double-strand breaks

Francisco J. Novo and José L. Vizmanos. Department of Genetics. University of Navarra. c/Irunlarrea s/n. 31080 PAMPLONA. SPAIN.

Corresponding author: F.J. Novo (fnovo@unav.es)

Abstract: We show that introns harboring translocation breakpoints in tumors are significantly longer than non-translocated introns of the same genes, but are not significantly enriched in sequence elements potentially involved in chromosomal rearrangements. Our findings provide evidence that double-strand breaks, the type of DNA damage that leads to translocations in tumors, are created at random in the genome, and that sequence elements do not play a general role in the localization of the breaks.

Translocation breakpoints in cancer

Chromosome aberrations are a common feature of cancer initiation and progression. Several types of aberrations are found in tumors, reciprocal translocations being one of the best characterized types of genetic alteration found in known cancer genes [1]. It is widely accepted that chromosome rearrangements in cancer are mediated by the repair of DNA double-strand breaks (DSB) through non-homologous end joining (NHEJ), which seems to be the major repair pathway in human somatic cells [2-4].

Analysis of some translocations, especially in hematological malignancies, revealed that the breakpoints are some times distributed in a non-random fashion, either densely clustered in particular regions of some genes, or more diffusely clustered along one or a few specific introns in other genes [5]. This has led to suggestions that the presence of local chromatin features or specific sequence motifs can determine genomic domains that are particularly prone to sustain a DSB in the vicinity. In line with this view, various sequence elements have been reported to be associated with chromosomal rearrangements in tumors (reviewed in [6]). However, since clustering of breakpoints is not observed in most translocations identified to date, there is considerable debate as to whether the presence of DNA domains with a high risk of sustaining a DSB is important only in a few specific translocations, or it is a general feature common to all translocations in human tumors. The lack of a significant number of breakpoints cloned at the genomic level has delayed progress in this area, because the precise localization of the breakpoints and the sequence surrounding them is known only in a limited number of cases.

In order to address this issue we have performed computational studies that enabled us to precisely locate a large number of cancer translocation breakpoints on the human genome reference sequence, and to analyze the genes involved with a view to unveil any potential factors accounting for the localization of breakpoints in human tumors.

Features of the genes involved in reciprocal translocations in tumors

We collected 268 genes whose involvement in reciprocal translocations in various types of tumors has been reported in the literature, and extracted their sequence and annotations from the Ensembl database [7] using perl scripts and the Application Programming Interface (API) provided by the Ensembl project [8]. We compared various sequence features between these genes and a group of 9,406 genes that are not involved in chromosomal rearrangements in tumors (see supplementary Methods for details on data collection and analysis). Table 1 shows clear differences in gene structure between both groups of genes. In particular, translocated genes are significantly longer than non-translocated genes. This increase in total gene size was due to a marked increase in the size of the longest intron, because this was the variable showing the strongest correlation with gene size (Spearman correlation coefficient = 0.925, $p=1.0x10^{-6}$). It might be argued that the difference in intron size is due to sequence composition bias, since GC-rich regions of the genome have been shown to contain genes with significantly shorter introns [9]. Thus, the longer intron size of translocated genes could be accounted for by the preferential localization of these genes in GC-poor regions. However, we did not observe a significantly lower GC content in translocated genes ($p=0.028$, Mann-Whitney test), excluding the possibility that the difference in intron size is simply due to compositional bias.

Since translocated genes could be grouped according to the type of tumor in which they are rearranged, we also compared all these variables in the groups of genes translocated in hematological, mesenchymal and epithelial tumors (see supplementary Methods). No significant differences were found for any of the variables, suggesting that genes rearranged in these three

types of tumors have a similar gene structure in terms of gene size, intron size and frequency of repeated elements.

We also asked whether the group of genes involved in reciprocal translocations in tumors represents a particular group of molecular functions. To answer this question, we analyzed the Gene Ontologies of our gene sets using FatigGO [10] and found that the ontologies comprising "DNA binding" and "transcription factor binding" are significantly over-represented in the genes translocated in tumors (see supplementary Figure S1). Almost half of translocated genes contain a DNA binding annotation, whereas only 18% of non-translocated genes contain that ontology ($p<10^{-5}$ adjusted for the False Discovery Rate, FDR [11]). Likewise, the proportion of genes with a transcription factor binding ontology is clearly higher in this group of genes (FDR-adjusted $p=0.00008$).

Translocation breakpoints are located in significantly large introns

We also determined the localization of the translocation breakpoints on the human genome reference sequence. To do this, we first searched the literature and Genbank® for all translocation junction sequences available for our group of 268 translocated genes (see supplementary Methods). We could obtain translocation sequence data for 249 genes (93%), either as fusion mRNAs or as genomic breakpoint junction sequences. Using these sequences as queries, we performed BLAST searches against the complete genomic sequence of those same 249 genes, and this enabled us to identify the specific introns involved in each translocation. Following this procedure, we were able to identify 334 introns containing a total of 962 different translocation events. As control group, we collected all the introns that belong to the same genes but are not involved in translocation events (2,754 introns).

We next studied the structural and sequence features of both sets of introns, computing the G+C content, the size of the intron and the presence of various types of repeats and sequence motifs known to be associated with chromosome rearrangements. We found that translocated introns are

significantly longer than non-translocated introns of the same genes (median of 3,733 bp *versus* 1,554 bp, Mann-Whitney rank test $p=1.3x10^{-25}$).

Are DSBs created at random throughout the genome?

We also found that translocated introns are enriched in all the sequence elements studied, with the proportion of introns with at least one repeat or sequence motif being significantly higher in the group of translocated introns (Figure 1). However, this difference could be due to the fact that these introns are also significantly longer, since longer introns will be more likely to contain any sequence element by chance alone. In fact, we tested the presence of random motifs of various sizes and found that a high proportion of them were also significantly over-represented in the group of translocated introns (not shown). Therefore, we adopted a binning strategy in order to correct the effect of intron size. After segmentation of our dataset into ten categories according to the size of the introns, no statistically significant differences were found for intron size between translocated and non-translocated introns within each category. We then analyzed the presence of repeats or sequence motifs separately for each category, finding that none of the sequence elements studied was over-represented in the group of translocated introns (see supplementary Figure S2).

In conclusion, we show that the most distinctive single feature of introns that contain translocation breakpoints is their increased size. This finding is consistent with the hypothesis that DSBs are initially created randomly throughout the genome, as longer introns will be more likely to contain a breakpoint by chance alone. On the contrary, if specific sequence or chromatin features were generally responsible for the localization of a DSB in their vicinity, one would not expect to find such a strong association between intron size and the presence of breakpoints. Moreover, none of the repeated elements or sequence motifs that we have studied was over-represented in translocated introns. This further suggests that such elements do not play a general role in the genesis of translocations in tumors.

Future directions

Several steps are necessary for the generation of reciprocal translocations in cancer. First of all, two breaks must be created in different genes at the same time. The free ends must then come close to each other within the cell nucleus. Finally, some repair pathway must join the broken ends together and the resulting molecule must provide some proliferative advantage to the cell. Taken together, our findings support a model in which: i) chromosome breaks are generated randomly in the genome; and 2) the localization of breakpoints, as a general rule, is the result of functional constraints, whereas the presence of specific sequence elements does not seem to play a general role in this regard. It will be interesting to see whether specific types of translocations are mediated by some of these elements.

Although we have collected all the sequences available in public databases and in the literature, hematological malignancies are better represented than mesenchymal and epithelial tumors. We could not detect significant differences in gene structure between genes rearranged in these three types of tumors (see supplementary Methods), therefore we believe that our findings apply generally to all neoplasms. However, this might change as new translocations are identified in the future.

Finally, the timing and localization of the breaks within the cell nucleus, as well as the functional status of the genes involved, are also crucial factors in the process of creating an oncogenic translocation. In this regard, recent advances in chromosome positioning in normal and cancer cells will help us to gain fundamental insights into the mechanisms involved in the genesis of chromosome translocations in tumors.

Acknowledgements: We thank Iñigo Landa and Beatriz Sánchez for their technical assistance. This work would not have been possible without the effort of the Ensembl team. Supported by grant 6/2003 from the Department of Health of the Government of Navarra, Spain.

References

1. Futreal, P.A. et al. (2004) A census of human cancer genes. Nat. Rev. Cancer 4, 177-183.

2. Khanna, K.K. and Jackson, S.P. (2001) DNA double-strand breaks: signaling, repair and the cancer connection. Nat. Genet. 27, 247-254.

3. Pierce, A.J. et al. (2001) Double-strand breaks and tumorigenesis. Trends Cell. Biol. 11, S52-59.

4. van Gent, D.C, Hoeijmakers, J.H., Kanaar, R. (2001) Chromosomal stability and the DNA double-stranded break connection. Nat. Rev. Genet. 2, 196-206.

5. Greaves, M.F. and Wiemels, J. (2003) Origins of chromosome translocations in childhood leukaemia. Nat. Rev. Cancer 3, 639-649.

6. Kolomietz, E. et al. (2002) The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genes Chromosomes Cancer. 35, 97-112.

7. Hubbard, T. et al. (2002) The Ensembl genome database project. Nucleic Acids Res. 30, 38-41.

8. Curwen, V. et al. (2004) The Ensembl automatic gene annotation system. Genome Res. 14, 942-950.

9. Versteeg, R. et al (2003). The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. 13, 1998-2004.

10. Al-Shahrour, F. et al. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms to groups of genes. Bioinformatics 20, 578-580.

11. Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate under dependency. Ann Stat. 29, 1165-1188.

12. Myers, S. et al. (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310, 321-324.

13. Nowicki, M.O. et al. (2004). BCR/ABL oncogenic kinase promotes unfaithful repair of the reactive oxygen species-dependent DNA double-strand breaks. Blood. 104, 3746-3753.

14. Konopka, A.K. et al. (1985). Concordance of experimentally mapped or predicted Z-DNA sites with positions of selected alternating purine-pyrimidine tracts. Nucleic Acids Res. 13, 1683-1701.

15. Bacolla, A. et al. (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. Proc. Natl. Acad. Sci. U. S. A. 101, 14162-14167.

16. Raghavan, S.C. et al. (2004) A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex. Nature 428, 88-93.

17. Spitzner, J.R and Muller, M.T. (1988) A consensus sequence for cleavage by vertebrate DNA topoisomerase II. Nucleic Acids Res. 16, 5533-5556.

18. Hesse, J. E. et al. (1989) V(D)J recombination: a functional definition of the joining signals. Genes Dev. 3, 1053–1067.

19. van Drunen, C.M. et al. (1999) A bipartite sequence element associated with matrix/scaffold attachment regions. Nucleic Acids Res. 27, 2924-2930.

20. Kurahashi, H. et al. (2003) The constitutional t(17;22): another translocation mediated by palindromic AT-rich repeats. Am. J. Hum. Genet. 72, 733-738.

Table 1. Gene features in the group of genes involved in reciprocal translocations in cancer and in control genes.

	median (interquartile range)		
	CONTROL GENES n=9,406	**CANCER GENES n=268**	**Mann-Whitney**
GC% [a]	45.2 (12.6)	43.3 (11.1)	n.s.
ALU%	11.1 (15.4)	10.6 (15.0)	n.s.
MIR%	2.2 (2.9)	2.1 (2.4)	n.s.
LINE%	7.9 (12.1)	7.9 (11.0)	n.s.
DNA%	1.8 (3.4)	2.0 (2.8)	n.s.
LTR%	1.4 (4.1)	1.2 (3.1)	n.s.
REPEAT%	39.2 (24.9)	36.5 (23.9)	n.s.
GENESIZE (bp)	25,805 (55,017)	59,934 (100,202)	9.8×10^{-17}
NUMTRANS	1.0 (1.0)	2.0 (2.0)	1.0×10^{-10}
INTSPERTRANS	8.0 (9.0)	9.6 (10.7)	9.6×10^{-5}
INTAVGSIZE (bp)	2,837 (4,645)	4,021 (7,075)	5.4×10^{-9}
MAXINTSIZE (bp)	9,254 (20,675)	22,521 (48,262)	1.3×10^{-16}

[a] GC%: G+C content. ALU%, MIR%, LINE%, DNA%, LTR%: percentage of gene sequence that is comprised by each of these types of repeats, or (REPEAT%) by all types of repeats. NUMTRANS: average number of alternative transcripts per gene. INTSPERTRANS: average number of introns per transcript. INTAVGSIZE: average intron size. MAXINTSIZE: size of the longest intron. For each comparison, the significance value of the Mann-Whitney test is given. Only p-values $<5 \times 10^{-4}$ are considered significant in order to correct for multiple testing (n.s. not significant).

Figure 1. Proportion of introns having at least one sequence motif or repeat.

Legend: For each type of element, we show the proportion of introns with at least one hit either in the group of introns that contain translocation breakpoints (dark gray bars) or in the group of control introns (white bars). Differences in percentages are all significant (Fisher's exact test, p-value shown on the right; only p-values $<5x10^{-4}$ are considered significant in order to correct for multiple testing). REC1 and REC2: CCTCCCT and CCCCACCCC motifs, respectively [12]; GC: G/C-rich tracts of 9 to 30 guanines or cytosines [13]; PUPY: purine/pyrimidine tracts of 5 to 25 alternating purines and pyrimidines [14]; OLIGOPU: tracts of 15 to 25 purines [15,16]; TOPO II: vertebrate topoisomerase II cleavage sites [17]; NONA and HEPTA: nonamer and heptamer recombination signal consensus sequences [18]; S/MAR: scaffold/matrix attachment region sites [19]; PAL: palindromic inverted repeats of repeat length between 30-300 basepairs and arms separated by 0-100 basepairs [20]. LTR, LINE, DNA, MIR and ALU: types of repeated elements (see supplementary Methods for details).

S**upplementary on-line materials:**

Figure S1: Comparison of gene ontologies (level 4 molecular function) between cancer genes (red bars) and control genes (green bars). Gene ontologies were identified for every gene using FatiGO [10] a tool for data mining using Gene Ontology categories that is available through a web interface (http://www.fatigo.org/) and implements Fisher's exact test for the comparison of two groups of genes and corrects for multiple testing using the False Discovery Rate procedure of Benjamini and Yekutieli [11]. For each ontology, the percentage of genes in each group is shown, together with unadjusted and FDR-adjusted p-values.

Figure S2: Introns were divided into 10 categories, depending on their size. Within each category, the proportion of introns with at least one hit for each sequence element is shown.

Table S1: Ensembl and HUGO identifiers for all the genes included in this study. Gene type indicates genes involved in translocations (type=1) or control genes (type=0). Tumor type indicates hematological (type=1), mesenchymal (type =2) or epithelial (type = 3) neoplasms.

Supplementary Methods: Details on data collection and analysis.