

# Direct transcriptional consequences of somatic mutation in breast cancer

## Authors:

Adam Shlien (1,16), Keiran Raine (1), Fabio Fuligni (15), Roland Arnold (15), Serena Nik-Zainal (1), Serge Dronov (1), Lira Mamanova (1), Andrej Rosic (15), Yeong Seok (1), Susana L Cooke (1), Manasa Ramakrishna (1), Elli Pappaemanuil (1), Helen R Davies (1), Patrick S Tarpey (1), Peter Van Loo (1,2), David C Wedge (1), David Jones (1), Sancha Martin (1), John Marshall (1), Elizabeth Anderson (1), Claire Hardy (1), ICGC Breast Cancer Working Group, Oslo Breast Cancer Research Consortium, Violetta Barbashina (3), Samuel AJR Aparicio (4), Torill Sauer (5), Oystein Garred (5), Anne Vincent-Salomon (6), Odette Mariani (6); Sandrine Boyault (7); Aquila Fatima (8); Anita Langerød (9,10); Åke Borg (11); Gilles Thomas (7); Andrea L Richardson (8); Anne-Lise Børresen-Dale (9,10); Kornelia Polyak (12), Michael R Stratton (1) and Peter J Campbell (1,12,14).

## Institutions:

- (1) Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK
- (2) Department of Human Genetics, VIB and University of Leuven, Leuven, Belgium
- (3) Breakthrough Breast Cancer, The Institute of Cancer Research, London, UK
- (4) British Columbia Cancer Agency, Vancouver, Canada
- (5) Department of Pathology, Ullevål University Hospital, Oslo, Norway
- (6) Institut Marie Curie, Paris, France
- (7) Synergie Lyon Cancer, Centre Léon Bérard, Lyon, France
- (8) Dana-Farber Cancer Institute, Boston, USA
- (9) Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway
- (10) K.G. Jebsen Center for Breast Cancer Research, Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
- (11) Department of Oncology, Lund University, Lund, Sweden
- (12) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, USA
- (13) Department of Haematology, Addenbrooke's Hospital, Cambridge, UK
- (14) Department of Haematology, University of Cambridge, Cambridge, UK
- (15) Department of Genetics and Genome Biology, The Hospital For Sick Children, Toronto, Canada
- (16) Current address: Department of Genetics and Genome Biology, The Hospital For Sick Children, Toronto, Canada

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17

**Address for correspondence:**

Dr Peter J Campbell,  
Cancer Genome Project,  
Wellcome Trust Sanger Institute,  
Hinxton,  
Cambridgeshire CB10 1SA,  
United Kingdom.  
Tel: +44 (0) 1223 834244  
e-mail: [pc8@sanger.ac.uk](mailto:pc8@sanger.ac.uk)

1   **ABSTRACT**

2   The disordered transcriptomes of cancer encompass direct effects of somatic  
3   mutation on transcription; co-ordinated secondary alterations in transcriptional  
4   pathways; and increased transcriptional noise. To catalogue the rules governing  
5   how somatic mutation exerts direct transcriptional effects, we developed an  
6   exhaustive bioinformatics pipeline for analyzing RNA-sequencing data, which we  
7   integrated with whole genome sequencing from 23 breast cancers. Using X-  
8   inactivation analyses, we find cancer cells are more transcriptionally active than  
9   intermixed stromal cells. This is especially true in ER-negative tumours. Overall,  
10   59% of 6980 exonic substitutions were expressed. Compared to other classes,  
11   nonsense mutations showed lower expression levels than expected with patterns  
12   characteristic of nonsense-mediated decay. 14% of 4234 genomic rearrangements  
13   caused transcriptional abnormalities, including exon skips, exon reusage, fusion  
14   transcripts and premature poly-adenylation. We found productive, stable  
15   transcription from sense-to-antisense gene fusions and gene-to-intergenic  
16   rearrangements, suggesting that these mutation classes may drive more  
17   transcriptional disruption than previously suspected. Systematic integration of  
18   transcriptome with genome data therefore reveals the rules by which  
19   transcriptional machinery interprets somatic mutation.

20

21

22

23

## 1    **HIGHLIGHTS**

- 2        • Cancer cells are more transcriptionally active than nearby stromal cells. This  
3           difference in activity appears to be even greater in ER-negative cancers
- 4        • Intron mutations only infrequently affect splicing, even at essential splice  
5           sites
- 6        • Sense-to-antisense & gene-to-intergenic rearrangements have distinctive  
7           RNA effects
- 8        • Exhaustive pipeline for identifying aberrant transcripts from RNA-  
9           sequencing data

10

11

12

# 1 INTRODUCTION

2 Somatic mutation underpins the development of cancer, and most solid tumours  
3 have thousands to tens of thousands of point mutations, coupled with tens to  
4 hundreds of genomic rearrangements and copy number changes (Garraway and  
5 Lander, 2013; Stratton et al., 2009). Small numbers of these, known as ‘driver  
6 mutations’, dysregulate the fundamental cellular processes involved in normal  
7 tissue homeostasis, and confer a selective advantage to the clone. A critical point is  
8 that Darwinian selection acts on *phenotype* and so, for a somatic mutation to drive  
9 cancer, it must manifest a phenotypic effect. Transcription is the primary conduit by  
10 which changes in the genomic code are translated into cellular phenotype, with the  
11 corollary that it is a necessary criterion of driver mutations that they directly induce  
12 a change in transcript structure. Altered transcript structure can take many forms,  
13 including the creation of fusion genes by genomic rearrangement, interference with  
14 RNA splicing at mutated splice sites, alteration of the codon sequence for missense  
15 substitutions and over- or under-expression of genes through copy number  
16 alterations or mutation in regulatory regions.

17  
18 Beyond the primary and direct effects of somatic mutation on transcript structure,  
19 there may be a series of downstream, secondary alterations in the transcriptome  
20 occurring as a consequence of the primary effect. Most studies of the transcriptome  
21 in cancer, including those from large-scale efforts such as TCGA (Kandoth et al.,  
22 2013; The Cancer Genome Atlas Research Network, 2012b), have evaluated these  
23 second-order effects, concentrating predominantly on the magnitude of gene

1 expression using microarray technology (Curtis et al., 2012; Perou et al., 2000;  
2 Sorlie et al., 2001) or RNA-sequencing (Shah et al., 2009; The Cancer Genome Atlas  
3 Research Network, 2012a). They have revealed large-scale disturbances of  
4 transcriptional regulation in most cancers, with expression profiles for many  
5 hundreds of genes differing from profiles of normal cellular counterparts. Within a  
6 tumour type, similarities in transcriptional profiles across individuals allow the  
7 disease to be sub-classified into several groups, many of which have biological,  
8 therapeutic and prognostic significance. In some cases, these changes can be  
9 correlated with underlying driver mutations, such as *ERBB2* amplification in breast  
10 cancer (Sorlie et al., 2001) or specific fusion genes in acute myeloid leukaemia (Valk  
11 et al., 2004). While these studies have concentrated on mRNA profiles, similar  
12 observations are beginning to emerge from studies of microRNA transcription  
13 (Dvinge et al., 2013), long non-coding RNA levels and even expression of  
14 pseudogenes (Kalyana-Sundaram et al., 2012).

15  
16 While it is a necessary criterion for a driver mutation to directly induce modification  
17 of transcript structure, it is not sufficient. Many mutations that do not confer  
18 selective advantage, so-called passenger mutations, will also generate phenotypic  
19 consequences, but consequences of no benefit to the cell. Initial studies correlating  
20 RNA-sequencing data with genomic change in cancer have reported some of these  
21 direct effects, especially for coding point mutations or canonical fusion transcripts  
22 (Shah et al., 2009) but there has been little systematic effort to describe, measure

1 and quantify first-order transcriptional consequences across all classes of somatic  
2 mutation found in well-annotated cancer genomes.

3  
4 Here, we report a comprehensive analysis of the primary transcriptional alterations  
5 induced by somatic mutation in a set of 23 breast cancers. We find that the genomic  
6 variants carried by the cancer cells can have subtle or profound effects on the  
7 transcriptome, many of which could not easily be predicted from the genome, many  
8 of which amalgamate several *in cis* mutations and many of which are stably  
9 expressed at high levels.

## 11 **RESULTS**

### 12 **Whole genome and RNA-sequencing from 23 breast cancer samples**

13 To understand the inter-relationships between somatic mutation and the  
14 transcriptome, we matched RNA-sequencing data to whole genome sequencing data  
15 in 23 breast cancer samples. Of these, 14 were primary breast cancers and 9 were  
16 matched breast cancer cell lines. For the genomes, tumour samples were sequenced  
17 to ~40x coverage and matched normal samples to ~30x coverage, with somatically  
18 acquired substitutions, indels, genomic rearrangements and copy number changes  
19 called by a suite of in-house algorithms. The whole genome sequencing for the 14  
20 primary breast cancer samples has been previously described (Nik-Zainal et al.,  
21 2012a; Nik-Zainal et al., 2012b), although improvements in our bioinformatics  
22 algorithm allowed us to up-date the list of genomic rearrangements (supplementary  
23 table 1). The high-coverage genome sequencing data for 8 breast cancer cell lines is

1 reported for the first time here (somatic mutations in supplementary table 2); for  
2 the other line (HCC2157), we used exome and low-coverage whole genome data  
3 reported previously (Nik-Zainal et al., 2012b; Stephens et al., 2009).

4  
5 RNA sequencing was performed on the 23 breast cancer samples together with 8  
6 organoids freshly isolated from uncultured normal breast milk ducts (Choudhury et  
7 al., 2013). We developed a suite of algorithms to exhaustively characterise the  
8 cancer transcriptome: in so doing we aimed to wring maximum detail on the  
9 structure of cancer transcripts from RNA-sequencing data. Previous work has  
10 examined gene and mutation expression alone or has focused exclusively on one  
11 facet of transcript structure (such as fusion genes or alternative splicing) without  
12 allowing for the discovery of multiple or complex events or the involvement of the  
13 antisense strand. We implemented a seed-and-extend mapping algorithm to find  
14 reads that span different regions of the genome, and then developed a discordant  
15 pair analysis algorithm, drawing these results together with a set of methods to  
16 arrange the results into biologically meaningful categories (described in detail in  
17 Supplementary Methods and Supplementary Figure 1).

18  
19 The primary advantage of our software pipeline, which we call RNA Architect, is the  
20 comprehensive detection of transcriptional alterations, including events missed by  
21 other methods. These would include compound events present in *cis*, such as fusion  
22 transcripts involving alternative splice forms and exon skips with cryptic splice  
23 sites; internal exon shuffling (reusage); post-transcriptional modifications such as



1 early polyadenylation sites; non canonical transcript junctions, for example fusions  
2 between the sense and anti-sense of different genes or those involving lowly-  
3 expressed transcripts that are not present in reference databases. While there  
4 exists a number of methods for aligning RNA-Seq and detecting fusions (Asmann et  
5 al., 2011; Chen et al., 2012; Kim and Salzberg, 2011; McPherson et al., 2012;  
6 Swanson et al., 2013; Torres-Garcia et al., 2014), there have been few efforts to  
7 simultaneously characterize the cancer transcriptome for multiple types of  
8 alterations.

9

#### 10 **Transcription derived from cancer cells and stromal cells**

11 Tumours are comprised of a complex admixture of clonal cancer cells and polyclonal  
12 stromal cells. In breast cancer, the proportion of cells deriving from the malignant  
13 clone is typically 30-70%, while the remaining cells encompass endothelial cells,  
14 supporting connective tissue, inflammatory cells, lymphocytes and normal breast  
15 epithelium. RNA samples extracted from primary breast cancers therefore represent  
16 an amalgam of gene expression signatures derived from multiple cell lineages,  
17 compounding interpretation.

18

19 In females, a randomly selected X chromosome is inactivated in each cell of the  
20 inner cell mass of the early blastocyst, and this choice is transmitted through every  
21 subsequent cell division. Since cancer cells are derived from a single ancestral cell,  
22 all have the same X chromosome inactivated (Fialkow et al., 1981), whereas the  
23 polyclonal stromal tissue has a broadly equivalent fraction of cells with maternal or

1 paternal X chromosomes inactivated. As a result, genes undergoing X inactivation  
2 with heterozygous germline SNPs will be monoallelically expressed in the cancer  
3 cells and biallelically expressed in stromal cells (supplementary figure S1A).

4  
5 We identified heterozygous germline SNPs in expressed regions of the X  
6 chromosome from the genomic sequencing data across the 14 primary breast  
7 cancers, excluding regions that were not diploid in the cancer. From the RNA-  
8 sequencing data, we extracted the number of reads expressing each allele. In  
9 PD4120a, for example, 385 heterozygous SNPs on the X chromosome were  
10 expressed. The observed reference/variant ratio in the RNA-sequencing data at each  
11 position ranged from transcripts whose expression was exclusively monoallelic  
12 through transcripts with skewed ratios to genes that had an approximately equal  
13 expression of both alleles (figure 1A). Respectively, these three scenarios represent  
14 genes expressed exclusively in cancer cells, genes expressed in both cancer and  
15 stromal cells and genes expressed exclusively in stromal cells.

16  
17 We developed a statistical algorithm based on a Bayesian hierarchical Dirichlet  
18 process to model the fraction of transcripts along the X chromosome derived from  
19 cancer cells (supplementary methods). For each heterozygous SNP, the algorithm  
20 estimates what fraction of reads covering that base derived from cancer cells,  
21 allowing for the uncertainty of whether the reference or variant allele is inactivated  
22 in the tumor (figure 1B, supplementary figure S1B). When amalgamated across

1 SNPs from the whole X chromosome, we can estimate the relative contribution of  
2 stromal and cancer cells to transcription as a general distribution (figure 1C).

3

4 Across the 14 patients in which primary breast cancer samples were sequenced, we  
5 find a considerable portion of transcripts that are exclusively expressed in cancer  
6 cells (figure 1D, supplementary figure S1C). Strikingly, many tumors had a set of  
7 transcripts that were 80-90% derived from cancer cells and 10-20% from stromal  
8 cells, whereas there were only small numbers of genes expressed predominantly  
9 from stromal cells. We can also integrate all the data for a given patient to estimate  
10 the overall fraction of transcripts derived from tumor cells, and compare this to the  
11 overall fraction of cancer cells in the sample estimated from the genomic DNA  
12 (figure 1E). This indicates that cancer cells contribute a higher fraction of  
13 transcripts in the RNA sample than expected for their cellular proportion, indicating  
14 that they are more transcriptionally active than the stromal cells. Thus, even though  
15 cancer cells comprise, on average, 30-70% of all cells in a breast tumor, they  
16 contribute 70-90% of all RNA molecules.

17

18 Strikingly, it appeared that the magnitude of the difference between transcriptional  
19 output of cancer cells and stromal cells was greater in ER-negative tumours than  
20 ER-positive tumours (figure 1E). These findings were validated in an independent  
21 set of primary breast cancers (661 ER-positive and 176 ER-negative) using a larger  
22 set of variants (38,337 somatic substitutions; supplementary figure S7A and S7B).  
23 Further, it appears that the number of mutations expressed in a breast tumour, a

1 measure of its transcriptional output, is significantly associated with the amount of  
2 estrogen receptor it expresses ( $-0.2433$ ,  $p < 0.0001$ ). That is, tumours with high levels  
3 of ER express fewer mutations than cancers with low ER. We formally modelled this  
4 relationship and determined that for every 1% decrease in ESR1 expression, 15  
5 more mutations are expressed in breast cancer (supplementary figure S7C and  
6 S7D).

7

### 8 **Effects of point mutations on structure of the transcriptome**

9 We identified all somatically acquired base substitutions in the 23 breast cancers  
10 that were in expressed regions, and compared the fraction of sequencing reads  
11 reporting the mutant allele in the transcriptome to that expected from the genome  
12 (supplementary figure S3A-B). As anticipated, there was a strong overall correlation  
13 between the genomic and transcriptomic variant allele fraction ( $r^2 = 0.59$ ;  $p < 0.0001$ ).  
14 Overall, 6980 substitutions were found in exons, of which 4751 were expressed to a  
15 sufficient degree that five or more sequencing reads covered the base. Of the 6980  
16 variants identified in exonic regions of the 23 samples, 4152 (59%) had discernible  
17 expression in the corresponding transcriptome.

18

19 There were some differences in the transcription levels of base substitutions  
20 according to the predicted consequence on the protein (figure 2A). We find that  
21 silent, missense and UTR mutations have the same strong correlation between  
22 variant allele fractions in the genome and transcriptome, whereas nonsense  
23 mutations have a weaker relationship. Indeed, nonsense mutations had a

1 significantly lower expression than predicted from the genome compared to other  
2 classes of mutation ( $p < 0.0001$ ).

3

4 Several reasons could explain the lower expression of nonsense mutations.  
5 Nonsense-mediated decay could selectively target transcripts with nonsense  
6 mutations for degradation. Nonsense-mediated decay depends upon the cell  
7 distinguishing a premature termination codon from a proper termination codon.  
8 Generally, stop signals in the last exon are considered proper, whereas those  
9 appearing more than 50-55bp upstream of the last exon-exon junction, and  
10 therefore upstream of the exon-junction complex, are more likely to be targeted for  
11 nonsense-mediated decay (Nagy and Maquat, 1998). We did find evidence for  
12 nonsense-mediated decay, since the decreased allele fraction in transcriptome  
13 relative to genome was significantly more pronounced for nonsense mutations if  
14 they were more than 50bp upstream of the last exon-exon splice junction ( $p = 0.003$ ;  
15 figure 2B).

16

17 Another possible explanation for the low expression of nonsense mutations is that  
18 they are tolerated only in genes not expressed in the cancer cells – those occurring  
19 in important genes would be subject to negative selection. To explore this  
20 possibility, we compared the expression levels from the organoids of normal breast  
21 epithelium for genes mutated in the cancer samples. We found no clear-cut  
22 differences across the mutation categories for whether the mutated genes were  
23 expressed in normal breast epithelial cells (supplementary figure S3C), suggesting

1 that this reason does not explain the lower expression levels of nonsense mutations.  
2 Therefore, it appears as if only nonsense mediated decay explains the lower  
3 expression of these mutations.

4  
5 Point mutations can directly affect RNA splicing, leading to retention of introns  
6 (especially for splice donor site mutations), exon skipping (splice acceptor site  
7 variants) or enhancement of alternative splice sites (other exonic or intronic  
8 variants). We assessed the frequency of alternative splicing events related to  
9 somatic base substitutions, where the splice isoform was not present in the normal  
10 breast organoids (figure 2C). We found no excess of abnormal splice isoforms  
11 associated with mutations in exons near splice sites. We found that mutations  
12 affecting the essential splice sites at +1, +2 and -1 into the intron were the most  
13 strongly associated with altered splicing in the given sample ( $p=0.002$ ,  $p=0.0001$   
14 and  $p=0.0005$  respectively compared to intronic mutations more than 100bp from  
15 the nearest exon). Nonetheless, despite this enrichment, the actual fraction of such  
16 mutations at essential splice sites that generated detectable abnormal splice  
17 isoforms was <25%, suggesting that most such variants do not affect splicing or the  
18 transcripts that result are rapidly degraded. Further into the introns, there were  
19 some positions at which mutations caused significantly more splicing abnormalities  
20 than expected (-49,  $p=0.04$ ; +23,  $p=0.02$ ; +46,  $p=0.01$ ; +60,  $p=0.003$ ). Strikingly,  
21 several of these isolated positions coincide with sites of reduced germline  
22 polymorphism. For example, the regions from +21 to +26 and from +45 to +50 both  
23 show strongly significant reductions in genetic variation in the germline (Lomelin et

al., 2010), suggesting that functional motifs regulating splicing may reside in these sites.

#### **Direct effects of genomic rearrangements on transcriptome structure**

Genomic rearrangements contribute to cancer development through several mechanisms, including changing the copy number of a gene or genes, altering the regulatory apparatus of a gene and reorganizing the exon sequence within a gene or between two genes. To evaluate effects of genomic rearrangements on transcriptome structure, we classified somatically acquired structural variants across two variables: type of rearrangement (deletion; tandem duplication; inverted; interchromosomal) and whether genes were involved at either side of the breakpoint (gene-to-gene, same or opposite orientation; gene-to-intergenic; within gene, same or different introns; local genomic complexity, where more than one rearrangement affected one or other gene). For each rearrangement, we identified any aberrant transcript arising from the genes involved, excluding any splice form seen in the normal breast organoids or the Ensembl database.

Even in cancer samples without rearrangements affecting a given gene, we often find evidence for previously undocumented transcripts, such as novel splice forms, read-through transcripts and non-canonical splice acceptor or donor sites. It is therefore difficult to argue categorically for a given rearrangement that an abnormal transcript arises as a direct consequence of the genomic change. Instead, since we are more interested in the overall patterns of abnormal transcription caused by

1 somatic mutation, we study the excess of aberrant transcription associated with the  
2 different categories of genomic rearrangement. The normalized expression level of  
3 aberrant transcripts was ranked for the sample in which the rearrangement was  
4 found, relative to aberrant transcripts in the other 22 cancer samples (figure 3A). If  
5 a given rearrangement had no effect on transcription, the ranking would be  
6 effectively uniformly distributed across ranks 1-23, whereas those rearrangements  
7 that cause significant alterations to transcriptome structure would garner the  
8 highest rank.

9

10 There is a clear excess of genomic rearrangements with the maximum ranking for  
11 aberrant transcription. Using maximum likelihood methods, we estimate that this  
12 excess represents 11.6% (95% confidence interval, 10.4-12.8%) of 4234 genic  
13 rearrangements identified in these samples (supplementary methods) – that is,  
14 11.6% of somatically acquired genomic rearrangements affecting genes are  
15 associated with evidence of aberrant transcription beyond the background rate in  
16 breast cancer. This varied by the pattern of genes at the breakpoint, with  
17 particularly high rates observed for intragenic rearrangements leading to  
18 alterations of exon order but minimal evidence for aberrant transcription arising  
19 from rearrangements confined to a single intron within a gene.

20

21 We observed a number of different patterns of aberrant transcription (figure 3B).  
22 These included fusion transcripts between two genes, alternative splicing events,  
23 exon reuse and premature polyadenylation. To some extent, the alterations in



1 transcript structure could be predicted by the underlying genomic rearrangements,  
2 such as exon skips caused by intragenic deletions, but in many cases the  
3 abnormalities were rather surprising. In the following sections, we review the  
4 transcriptional consequences associated with each of the major classes of genomic  
5 rearrangement.

6

### 7 **Within-gene rearrangements**

8 Across the 23 breast cancer samples studied, we identified 631 intrachromosomal  
9 rearrangements confined to a single intron of a gene, mostly deletions and tandem  
10 duplications (358 and 192, respectively). Of these, we believe that very few have  
11 discernible consequences on transcriptome structure, since there is no apparent  
12 excess of such rearrangements generating the highest rank for transcriptional  
13 aberration across samples (figure 3A). For the 38 rearrangements with highest rank,  
14 the commonest effect on transcript structure was to skip an exon (69%; figure 3B).

15

16 As expected, rearrangements that went across different introns of the same gene  
17 had considerably more effect on transcript structure than those confined to one  
18 intron (figure 3A). In general, the transcriptome reflected the rearranged gene  
19 structure in an entirely predictable way, with deletions causing exon skips and  
20 tandem duplications causing exon reusage (figure 3B). Of 341 genomic  
21 rearrangements involving different introns of the same gene, 84 had the highest  
22 rank for transcriptional abnormality. Of these 84, 23 (27%) caused multiple  
23 disruptions in the transcriptome at the same gene, mostly alternative splice

1 isoforms. Particularly common were exons skips of not just deleted exons but  
2 neighbouring exons as well.

3

#### 4 **Gene-to-gene rearrangements in the same transcriptional orientation**

5 We identified 205 somatic rearrangements that juxtaposed one or more exons of  
6 two protein-coding genes in the same transcriptional orientation – these would be  
7 predicted to generate fusion genes. Overall, 70 (34%) of these were expressed  
8 (figure 4A). As seen with the within-gene rearrangements, the transcriptome  
9 structure was generally as predicted from the genomic rearrangement, although  
10 more than one splice isoform was present in 20 of the 70 rearrangements,  
11 increasing the range of transcripts observed (figure 4B). The only recurrent fusion  
12 transcript that we observed in this cohort was between *NCOA7* and *TRMT11*,  
13 adjacent genes on chromosome 6, caused by tandem duplications in HCC1954 and  
14 PD4005a (supplementary figure S4).

15

16 We examined the protein reading frame of transcripts arising from gene fusions and  
17 within-gene rearrangements that spanned more than one intron (figure 4B). In  
18 133/501 (27%) such events, the resulting exon structure would be predicted to  
19 generate an in-frame gene from transcript isoforms reported in Ensembl – we found  
20 RNA-sequencing reads supporting these in-frame transcripts in 35/133 (26%). Of  
21 the 368 rearrangements predicted to be out-of-frame or involve the non-coding  
22 UTR, we found evidence for in-frame transcripts in 25 (7%). In many cases, the in-  
23 frame transcript was more heavily expressed than the canonical, out-of-frame

1 transcript, suggesting that nonsense-mediated decay may be acting on the latter  
2 (figure 4C). Overall, then, these data indicate that 60/501 (12%) of genomic  
3 rearrangements reordering exons of one or two genes in the same orientation have  
4 the potential to generate transcripts encoding in-frame proteins. Many of these are  
5 expressed at appreciable levels, mostly driven by the upstream regulatory  
6 apparatus.

7

### 8 **Gene-to-gene fusions in opposite transcriptional orientation**

9 We would expect half of the genomic rearrangements linking two genes to join them  
10 in opposing orientation, which would be split equally between gene pairs pointing  
11 inwardly at one another and gene pairs pointing away from one another. In the  
12 former, the 5' regulatory apparatus and transcriptional start site of both genes  
13 would be retained, and could start transcripts that would extend into the partner  
14 gene on the antisense strand. We identified 171 somatic rearrangements generating  
15 gene-to-gene fusions in opposite orientation, of which 114 were pointing inwards  
16 (5'-to-5' orientation). While there was not much evidence of aberrant transcription  
17 arising from 3'-to-3' fusions, we found an unexpectedly high frequency of stable  
18 transcription at gene pairs pointing inwardly (figure 5A).

19

20 In total, 50 (44%) of all 5'-to-5' rearrangements generated transcripts that fused the  
21 sense portion of one gene with novel exons on the antisense strand of the partner  
22 gene. Mostly, the novel transcribed sequence from the antisense strand of the distal  
23 gene mapped to intronic regions, although a few fusions did generate transcripts

1 that partially or fully overlapped with exons (figure 5B). This is to be expected since  
2 splice sites are directional, so the GT...AG structure of an intron is not recapitulated  
3 on the antisense strand. However, where one might have expected the reads derived  
4 from the antisense strand to be rather scattered, the antisense exons were in fact  
5 surprisingly fixed (figure 5C). That is, the antisense component of the fusion  
6 transcript tended to reuse the same latent splice acceptor and donor sites on the  
7 antisense strand. These were almost always associated with consensus GT-AG splice  
8 signals. For a small number of examples, multiple antisense exons were recurrently  
9 included in the transcript. None of these novel antisense exons was seen in the  
10 absence of the given 5'-to-5' rearrangement, suggesting that it is the genomic  
11 rearrangement that unmask the latent transcriptional potential of these regions.

12  
13 It is unclear what functional potential these sense-to-antisense gene fusions might  
14 have. Notably, we find an example involving the estrogen receptor, *ESR1*, which  
15 generates transcripts linking the sixth exon into a multiply spliced antisense  
16 transcript of *SYNE1* (figure 5C). Fusion transcripts involving the same intron of *ESR1*  
17 are recurrent in breast cancer, and there is evidence they have important functional  
18 consequences, largely conferred by the C-terminally truncated estrogen receptor (Li  
19 et al., 2013). There is also a rearrangement that fuses the first 20 exons of the  
20 transcriptional co-activator, *CREBBP*, to the antisense strand of *CLUAP1*. *CREBBP* is a  
21 well-known cancer gene that can be targeted by inactivating point mutations  
22 (Pasqualucci et al., 2011) or, in leukemias, involved in canonical fusion genes  
23 (Camos et al., 2006).

1

## 2 **Gene-to-intergenic rearrangements**

3 We identified 473 genomic rearrangements that joined the 5' portion of a gene to an  
4 intergenic region, and 461 rearrangements linking 3' ends of genes to intergenic  
5 space (figure 6A). As seen with the 3'-to-3' gene-to-gene fusions, in the absence of  
6 promoters, only one 3' gene-to-intergenic rearrangement led to a detectable RNA  
7 transcript. In contrast, 16 (3.4%) of 5' gene-to-intergenic rearrangements led to  
8 stable expression of abnormal transcripts related to the rearrangement.

9

10 The predominant transcripts that resulted from these 5' gene-to-intergenic  
11 rearrangements were fusions linking the 5' portion of the broken gene to exon 2 of  
12 the first intact, sense gene downstream of the breakpoint (figure 6B). Occasionally,  
13 splicing into novel intergenic exons or into exon 1 of the downstream gene was  
14 observed, but compared to splicing into exon 2, these transcripts were infrequent  
15 and represented minor RNA species. In general, the first exon of a gene commences  
16 with the transcription start site, and therefore does not contain a splice acceptor  
17 site, explaining why 5' gene-to-intergenic rearrangements fuse into exon 2. Since the  
18 first exon of many genes carries the ATG that initiates translation, many of these  
19 gene-to-intergenic rearrangements could translate into *bona fide* fusion proteins.  
20 Indeed, we identified three fusion transcripts caused by gene-to-intergenic  
21 rearrangements that were potentially in-frame (figure 6B). The length of the novel  
22 intron created by these transcribed gene-to-intergenic fusions was typically in the  
23 50-100kb range, but could be as high as 250kb (figure 6C).

1

## 2 **Regions of local complexity**

3 We defined a region of local complexity as any gene footprint that contained two or  
4 more genomic rearrangements. Typically, these represented sites of extensive  
5 genomic amplification, such as around *ERBB2* or *CCND1*, or they were regions of  
6 chromothripsis, a mutational process generating tens to hundreds of localized  
7 genomic rearrangements in a one-off catastrophic event (Stephens et al., 2011).  
8 Given the complexity of the genomic changes in many of these regions, a surprising  
9 number of rearrangements led to measurable transcriptional consequences (figure  
10 7A). Indeed, when compared with genes hit by simple rearrangements, the fractions  
11 of rearrangements from regions of local genomic complexity giving aberrant  
12 transcripts were broadly similar. This suggests that the regulatory apparatus  
13 enabling transcription initiation remains at least partially intact in many of these  
14 heavily rearranged regions, and that the genomic structure supports production of  
15 stable transcripts.

16

17 We find that the transcripts that arise in these regions often represent an  
18 integration across multiple rearrangements (figure 7B, supplementary figure S5). In  
19 PD4107a, for example, we find a fusion transcript that links *QKI* to the antisense  
20 strand of *TRPS1* (blue arc, figure 7B), which is in fact driven by two *in cis* genomic  
21 rearrangements linking *QKI* to *ANKRD11* and then *ANKRD11* to *TRPS1*. Due to the  
22 massive number of rearrangements sometimes found in these regions of local  
23 complexity, there can be a considerable degree of aberrant transcription. In

1 PD4103a, for example, among the hundreds of clustered rearrangements localized  
2 to a small number of genomic regions, we find 12 different fusion transcripts, as  
3 well as 7 alternatively spliced isoforms driven by within-gene rearrangements  
4 (supplementary figure S5).

5

## 6 **DISCUSSION**

7 The disturbed transcriptional landscape of cancer cells results from three main  
8 forces: (1) direct, primary consequences of somatic mutation; (2) co-ordinated,  
9 secondary gene expression changes resulting from altered cellular signaling,  
10 transcriptional regulation and chromatin landscape; and (3) general loss of  
11 transcriptional fidelity, manifesting as shorter 3' UTRs (Mayr and Bartel, 2009),  
12 retained introns, trans-splicing (Li et al., 2008) and so on. Here, we have  
13 concentrated on dissecting the immediate impact that the repertoire of somatic  
14 mutations has on the transcriptome in breast cancer, exploring the rules that govern  
15 how the transcriptional machinery interprets somatic mutation. In some ways, this  
16 is the most straightforward analysis of a cancer transcriptome to perform – the  
17 causation chain is short and, in theory, predictable.

18

19 One striking conclusion of the analysis is that transcription, once started, will  
20 attempt to complete. We found an unexpectedly high number of transcripts  
21 resulting from structural variants that sow the 5' seeds of a gene, namely upstream  
22 enhancers, promoter and first few exons, into seemingly infertile ground, such as  
23 intergenic space or the antisense strand of another gene. Indeed, in our data, the

1 fraction of such events generating productive transcription was not dissimilar to  
2 that observed for rearrangements predicted to cause canonical gene fusions. In the  
3 case of gene-to-intergenic rearrangements, the transcriptional machinery can scan  
4 many tens of kilobases in search of a splice acceptor site, often contributed by the  
5 second exon of a downstream intact gene in the same orientation. For sense-to-  
6 antisense fusions, the sense transcript will often splice into novel exons within the  
7 gene footprint of the antisense gene. In one example, this generated a truncated  
8 version of the estrogen receptor gene, *ESR1*. Recently, it has been reported that  
9 fusion transcripts arising from breaks in the same intron of *ESR1* are recurrent in  
10 breast cancer and can confer resistance to endocrine therapy (Li et al., 2013).

11  
12 It is a necessary condition of a somatic mutation to be oncogenic that it induces  
13 some transcriptional consequence, but it is far from sufficient. We find that 59% of  
14 exonic point mutations are expressed and 11.6% of genomic rearrangements hitting  
15 a gene footprint generate aberrant transcripts. These transcripts are poly-  
16 adenylated, stable and have the potential to generate protein products. In the case of  
17 cancer, even those that generate proteins will be mostly inconsequential to cell  
18 biology, although there will be some that are oncogenic. In the case of species  
19 evolution, however, such a high proportion of genomic rearrangements generating  
20 stable fusion transcripts, novel exons and splicing isoforms could readily provide a  
21 substrate for further genomic evolution over many generations.

## 22 23 **ACKNOWLEDGEMENTS**



1 This work was supported by the Wellcome Trust (grant reference 077012/Z/05/Z).  
2 PJC is personally funded through a Wellcome Trust Senior Clinical Research  
3 Fellowship (grant reference WT088340MA). AS is supported by a HL Holmes Award  
4 from the National Research Council Canada and an EMBO Fellowship. SD, SM and AL  
5 are funded through the BASIS project which is a European research project funded  
6 by the European Community's Seventh Framework Programme (FP7/2010-2014)  
7 under the grant agreement number 242006. AL and ALBD are supported South-  
8 Eastern Norway Regional Health Authority (2011079), Norwegian Cancer Society  
9 (0332) and The Norwegian Radium Hospital Research Foundation. PVL is a  
10 postdoctoral researcher of the Research Foundation - Flanders (FWO). Genome  
11 sequence data have been deposited at the European Genome-Phenome Archive  
12 (<http://www.ebi.ac.uk/ega/>, hosted by the EBI). SNP6 and expression array data  
13 has been deposited with ArrayExpress Archive (EBI). We also would like to  
14 acknowledge the Core Sequencing Facility and IT groups of the Wellcome Trust  
15 Sanger Institute, support for samples, sample banking and processing from the  
16 Breakthrough Breast Cancer Unit and members of the ICGC Breast Cancer Working  
17 Group as well as the Pathology Review subgroup of the Breast Cancer Working  
18 Group.

## 21 **FIGURE LEGENDS**

22 **Figure 1. Separating expression of X-linked genes into stromal and tumor**  
23 **compartments.**

(A) Fraction of RNA-sequencing reads reporting reference allele of heterozygous germline SNPs on the X chromosome in one of the patients (PD4120a). The depth of colour reflects the level of expression.

(B) Fraction of transcripts derived from tumor cells for each heterozygous germline SNP shown in (A), estimated with a Bayesian Dirichlet process.

(C) Estimated distribution and 95% posterior intervals for relative gene expression in cancer versus stromal cells for PD4120a. The y axis reports the estimated density of genes; the x axis reports the fraction of transcripts for each gene deriving from cancer cells. Thus, the transcripts for most genes in PD4120a are 80-100% derived from cancer cells and 0-20% from stromal cells, with only a small peak of genes predominantly expressed from stromal cells.

(D) Distributions for several selected primary cancers, as for (C).

(E) Overall fraction of transcripts derived from cancer cells (y axis) compared to the estimated proportion on tumor cells in the sample (x axis; estimated from genomic DNA using copy number profiles).

**Figure 2. The effect of somatic point mutations on expression and aberrant splicing.**

(A) Comparison of the variant allele fractions in the transcriptome to the genome, for all classes of point mutation. The squared correlation coefficient between the genome and transcriptome are in brackets. Only expressed coding changes are shown ( $\geq 5$ x coverage).

1 (B) Variant allele fractions in the transcriptome relative to the genome. Nonsense  
2 mutations >50 base pairs from the terminal 3' exon-intron junction are the only  
3 variants to show a significant difference.

4 (C) Positional effect of mutations on aberrant splicing.  
5

6 **Figure 3. The transcriptional consequences of structural rearrangement.**

7 All rearrangement types and their position with respect to genes are shown as a  
8 matrix in both panels. Transcriptional disruptions caused by each rearrangement  
9 type are shown within the matrix.

10 (A) Number of rearrangements causing aberrant transcription. Normalised  
11 aberrant transcription levels were contrasted between the sample that  
12 contained the rearrangement and all others. Plotted is the aberrant transcription  
13 ranking of the rearranged sample relative to all others, for the same genes (red  
14 bars). The pie charts show the fraction of all rearrangements of that type that  
15 are excess in the final rank compared to the number expected under a uniform  
16 distribution.

17 (B) Types of aberrant transcriptional events caused by rearrangements.  
18

19 **Figure 4. Rearrangements between and within genes.**

20 (A) Fusions caused by rearranged genes in the same orientation.

21 (B) Proportion of rearrangements predicted to lead to an in-frame event contrasted  
22 to the proportion actually expressing in-frame transcripts (top).  
23 Characteristics of expressed fusions (bottom).

1 (C) Many fusions are expressed in multiple isoforms.

2

3 **Figure 5. Antisense expression caused by rearranged genes in opposite**  
4 **orientation.**

5 (A) Stacked bar plot showing the number of expressed transcripts per sample  
6 resulting from gene fusions in opposite orientation.

7 (B) The diversity of chimeric transcripts produced by gene-to-gene rearrangements.

8 The expression level of each transcript is plotted on the y axis. Tail to tail gene  
9 pairs (green) are rarely expressed whereas, surprisingly, sense-to-sense and  
10 sense-to-antisense fusions show similar levels of expression (blue and red,  
11 respectively). Transcripts are placed on the x axis according to the type of read  
12 joining the two genes. Genes adjoined by exonic reads are plotted to the right on  
13 the x axis, and genes brought together only by exon-to-intron reads are on the  
14 left.

15 (C) Examples of productive, stable antisense fusion transcripts. Plotted on the y axis  
16 are the read depths supporting the fusion. Hatched lines indicate rearrangement  
17 breakpoints. In most cases, we observe a single donor gene, which expresses  
18 sequence from its sense strand (yellow), and a single acceptor gene that  
19 expresses sequence from its antisense strand (red). Rarely both promoters are  
20 used, leading to reciprocal sense-antisense fusions (both genes express sense  
21 and antisense sequence). The fusions *SZT2-SLC6A9* and *SLC6A9-SZT2* are  
22 examples of a reciprocal pair. In general antisense transcripts display features

of traditional exons: they are stably expressed, around 200bp, and are frequently spliced at GT-AG splice sites (asterisks).

**Figure 6. Non-canonical fusions caused by gene to intergenic breakpoints.**

(A) Percentage of gene to intergenic rearrangements causing fusions.

(B) Length of the intron created.

(C) Genes involved in non-canonical fusions. We observed 18 fusions where a broken gene (donor) splices to another gene (acceptor) that is itself unbroken and often distant. These fusions can be highly expressed (width of line) and cause in-frame transcripts (red line).

**Figure 7. Regions of local complexity give rise to unique transcriptional consequences.** A region of local complexity is any gene footprint that contains two or more genomic rearrangements. Local complexity can occur in regions of chromothripsis and high-level amplification.

(A) Proportion of simple and complex rearrangements that lead to an expressed transcript, grouped by sample.

(B) Regions of local complexity and their transcriptional consequences. Two samples' regions of complexity are shown as pairs of Circos plots. The genomic events one would predict to be expressed are highlighted (blue arcs). Often the tumours do not express these events, or they amalgamate multiple *cis* rearrangements and express a transcript that combines genes only indirectly linked to another.

1

## 2 REFERENCES

- 3 Asmann, Y.W., Hossain, A., Necela, B.M., Middha, S., Kalari, K.R., Sun, Z., Chai, H.S.,  
4 Williamson, D.W., Radisky, D., Schroth, G.P., *et al.* (2011). A novel bioinformatics  
5 pipeline for identification and characterization of fusion transcripts in breast cancer and  
6 normal cell lines. *Nucleic acids research* 39, e100.
- 7 Camos, M., Esteve, J., Jares, P., Colomer, D., Rozman, M., Villamor, N., Costa, D.,  
8 Carrio, A., Nomdedeu, J., Montserrat, E., *et al.* (2006). Gene expression profiling of  
9 acute myeloid leukemia with translocation t(8;16)(p11;p13) and MYST3-CREBBP  
10 rearrangement reveals a distinctive signature with a specific pattern of HOX gene  
11 expression. *Cancer Res* 66, 6947-6954.
- 12 Chen, K., Wallis, J.W., Kandath, C., Kalicki-Veizer, J.M., Mungall, K.L., Mungall, A.J.,  
13 Jones, S.J., Marra, M.A., Ley, T.J., Mardis, E.R., *et al.* (2012). BreakFusion: targeted  
14 assembly-based identification of gene fusions in whole transcriptome paired-end  
15 sequencing data. *Bioinformatics* 28, 1923-1924.
- 16 Choudhury, S., Almendro, V., Merino, V.F., Wu, Z., Maruyama, R., Su, Y., Martins,  
17 F.C., Fackler, M.J., Bessarabova, M., Kowalczyk, A., *et al.* (2013). Molecular profiling  
18 of human mammary gland links breast cancer risk to a p27(+) cell population with  
19 progenitor characteristics. *Cell Stem Cell* 13, 117-130.
- 20 Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D.,  
21 Lynch, A.G., Samarajiwa, S., Yuan, Y., *et al.* (2012). The genomic and transcriptomic  
22 architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346-352.
- 23 Dvinge, H., Git, A., Graf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y.,  
24 Hirst, M., Armisen, J., Miska, E.A., *et al.* (2013). The shaping and functional  
25 consequences of the microRNA landscape in breast cancer. *Nature* 497, 378-382.
- 26 Fialkow, P.J., Faguet, G.B., Jacobson, R.J., Vaidya, K., and Murphy, S. (1981). Evidence  
27 that essential thrombocythemia is a clonal disorder with origin in a multipotent stem cell.  
28 *Blood* 58, 916-919.
- 29 Garraway, L.A., and Lander, E.S. (2013). Lessons from the cancer genome. *Cell* 153, 17-  
30 37.
- 31 Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.M., Cao,  
32 X., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J., *et al.* (2012). Expressed  
33 pseudogenes in the transcriptional landscape of human cancers. *Cell* 149, 1622-1634.
- 34 Kandath, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson,  
35 A.G., Pashtan, I., Shen, R., Benz, C.C., *et al.* (2013). Integrated genomic characterization  
36 of endometrial carcinoma. *Nature* 497, 67-73.
- 37 Kim, D., and Salzberg, S.L. (2011). TopHat-Fusion: an algorithm for discovery of novel  
38 fusion transcripts. *Genome biology* 12, R72.
- 39 Li, H., Wang, J., Mor, G., and Sklar, J. (2008). A neoplastic gene fusion mimics trans-  
40 splicing of RNAs in normal human cells. *Science* 321, 1357-1361.
- 41 Li, S., Shen, D., Shao, J., Crowder, R., Liu, W., Prat, A., He, X., Liu, S., Hoog, J., Lu, C.,  
42 *et al.* (2013). Endocrine-Therapy-Resistant ESR1 Variants Revealed by Genomic  
43 Characterization of Breast-Cancer-Derived Xenografts. *Cell Rep* 4, 1116-1130.

1 Lomelin, D., Jorgenson, E., and Risch, N. (2010). Human genetic variation recognizes  
 2 functional elements in noncoding sequence. *Genome Res* 20, 311-319.  
 3 Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative  
 4 cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138, 673-684.  
 5 McPherson, A., Wu, C., Wyatt, A.W., Shah, S., Collins, C., and Sahinalp, S.C. (2012).  
 6 nFuse: discovery of complex genomic rearrangements in cancer using high-throughput  
 7 sequencing. *Genome research* 22, 2250-2261.  
 8 Nagy, E., and Maquat, L.E. (1998). A rule for termination-codon position within intron-  
 9 containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* 23, 198-  
 10 199.  
 11 Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K.,  
 12 Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., *et al.* (2012a). Mutational processes  
 13 molding the genomes of 21 breast cancers. *Cell* 149, 979-993.  
 14 Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau,  
 15 K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., *et al.* (2012b). The life history  
 16 of 21 breast cancers. *Cell* 149, 994-1007.  
 17 Pasqualucci, L., Dominguez-Sola, D., Chiarenza, A., Fabbri, G., Grunn, A., Trifonov, V.,  
 18 Kasper, L.H., Lerach, S., Tang, H., Ma, J., *et al.* (2011). Inactivating mutations of  
 19 acetyltransferase genes in B-cell lymphoma. *Nature* 471, 189-195.  
 20 Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack,  
 21 J.R., Ross, D.T., Johnsen, H., Akslen, L.A., *et al.* (2000). Molecular portraits of human  
 22 breast tumours. *Nature* 406, 747-752.  
 23 Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A.,  
 24 Gelmon, K., Guliany, R., Senz, J., *et al.* (2009). Mutational evolution in a lobular breast  
 25 tumour profiled at single nucleotide resolution. *Nature* 461, 809-813.  
 26 Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen,  
 27 M.B., van de Rijn, M., Jeffrey, S.S., *et al.* (2001). Gene expression patterns of breast  
 28 carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U*  
 29 *S A* 98, 10869-10874.  
 30 Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance,  
 31 E.D., Lau, K.W., Beare, D., Stebbings, L.A., *et al.* (2011). Massive genomic  
 32 rearrangement acquired in a single catastrophic event during cancer development. *Cell*  
 33 144, 27-40.  
 34 Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T.,  
 35 Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J., *et al.* (2009). Complex landscapes of  
 36 somatic rearrangement in human breast cancer genomes. *Nature* 462, 1005-1010.  
 37 Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. *Nature* 458,  
 38 719-724.  
 39 Swanson, L., Robertson, G., Mungall, K.L., Butterfield, Y.S., Chiu, R., Corbett, R.D.,  
 40 Docking, T.R., Hogge, D., Jackman, S.D., Moore, R.A., *et al.* (2013). Barnacle: detecting  
 41 and characterizing tandem duplications and fusions in transcriptome assemblies. *BMC*  
 42 *genomics* 14, 550.  
 43 The Cancer Genome Atlas Research Network (2012a). Comprehensive genomic  
 44 characterization of squamous cell lung cancers. *Nature* 489, 519-525.  
 45 The Cancer Genome Atlas Research Network (2012b). Comprehensive molecular  
 46 portraits of human breast tumours. *Nature* 490, 61-70.

1 Torres-Garcia, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger,  
2 M.F., Weinstein, J.N., Getz, G., and Verhaak, R.G. (2014). PRADA: pipeline for RNA  
3 sequencing data analysis. *Bioinformatics* 30, 2224-2226.  
4 Valk, P.J., Verhaak, R.G., Beijen, M.A., Erpelinck, C.A., Barjesteh van Waalwijk van  
5 Doorn-Khosrovani, S., Boer, J.M., Beverloo, H.B., Moorhouse, M.J., van der Spek, P.J.,  
6 Lowenberg, B., *et al.* (2004). Prognostically useful gene-expression profiles in acute  
7 myeloid leukemia. *N Engl J Med* 350, 1617-1628.  
8



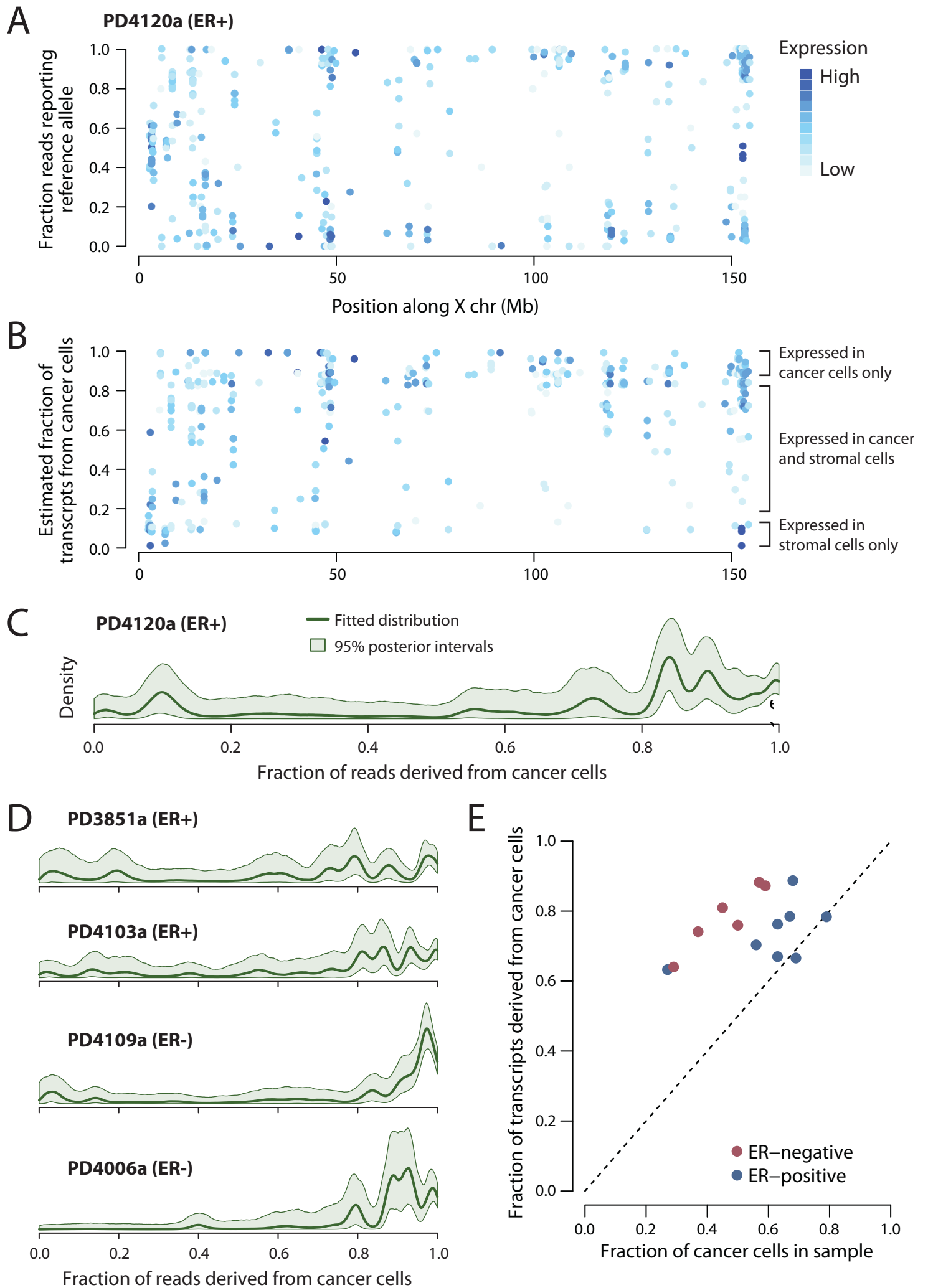


Figure 1

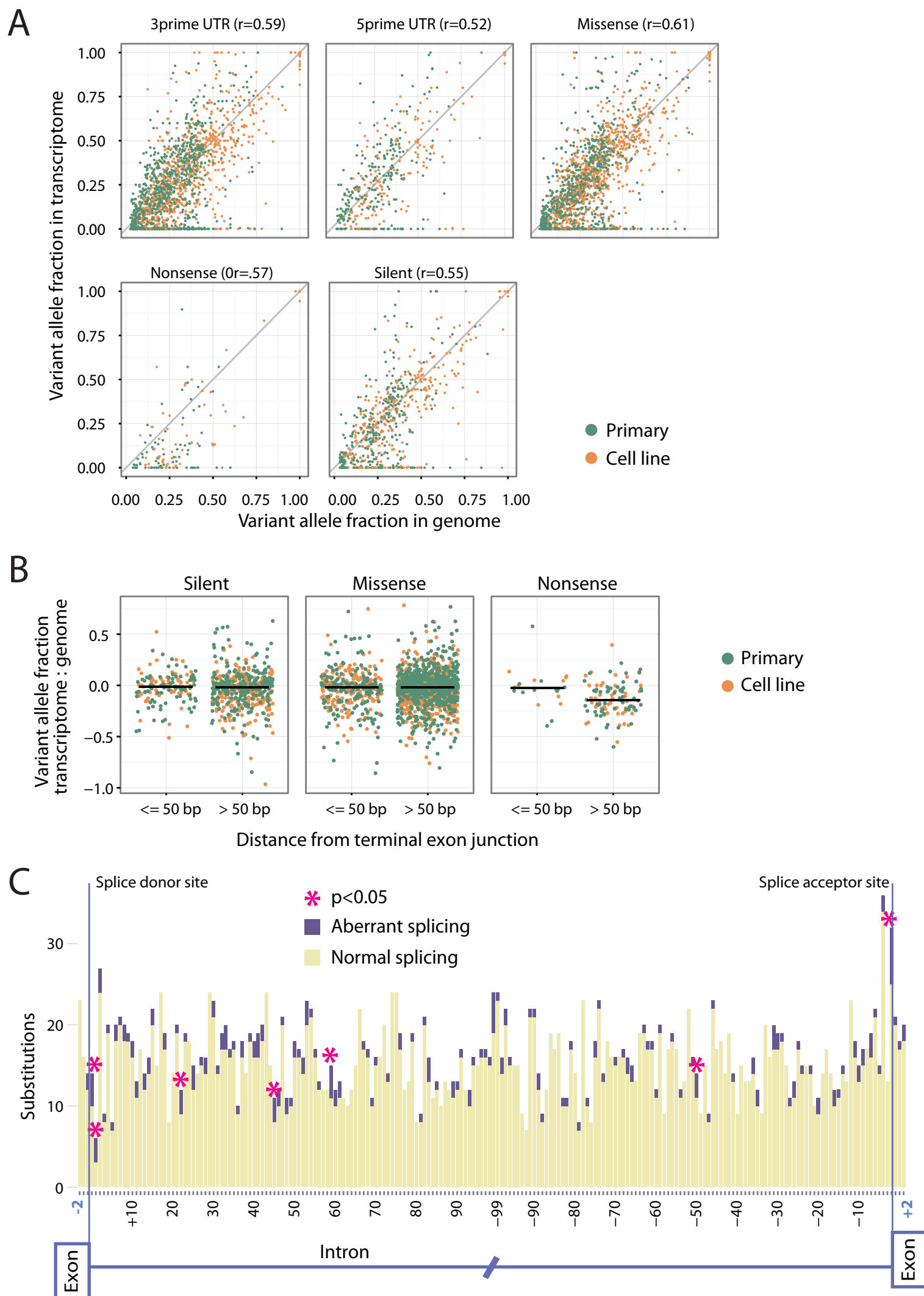
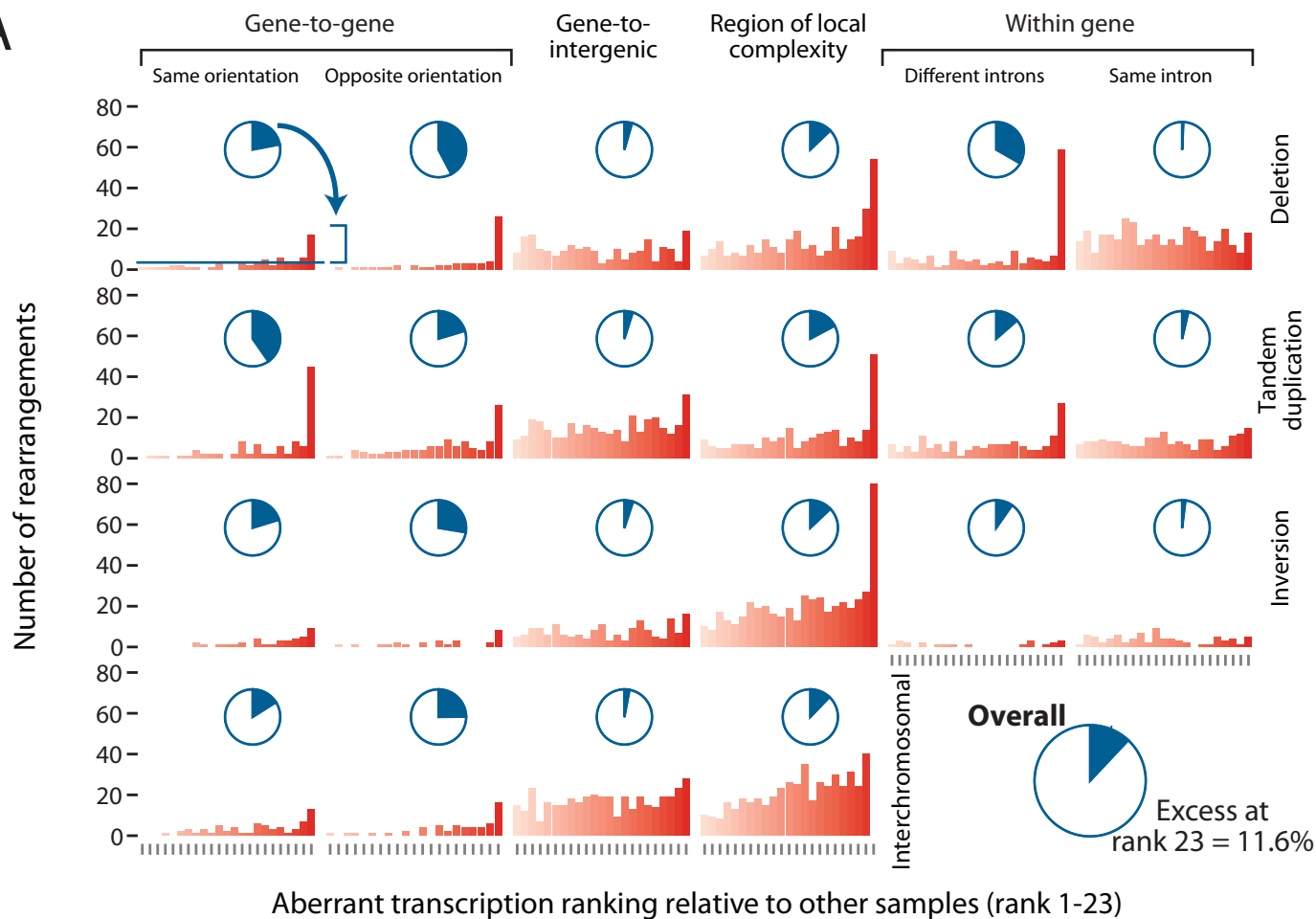


Figure 2

A



B

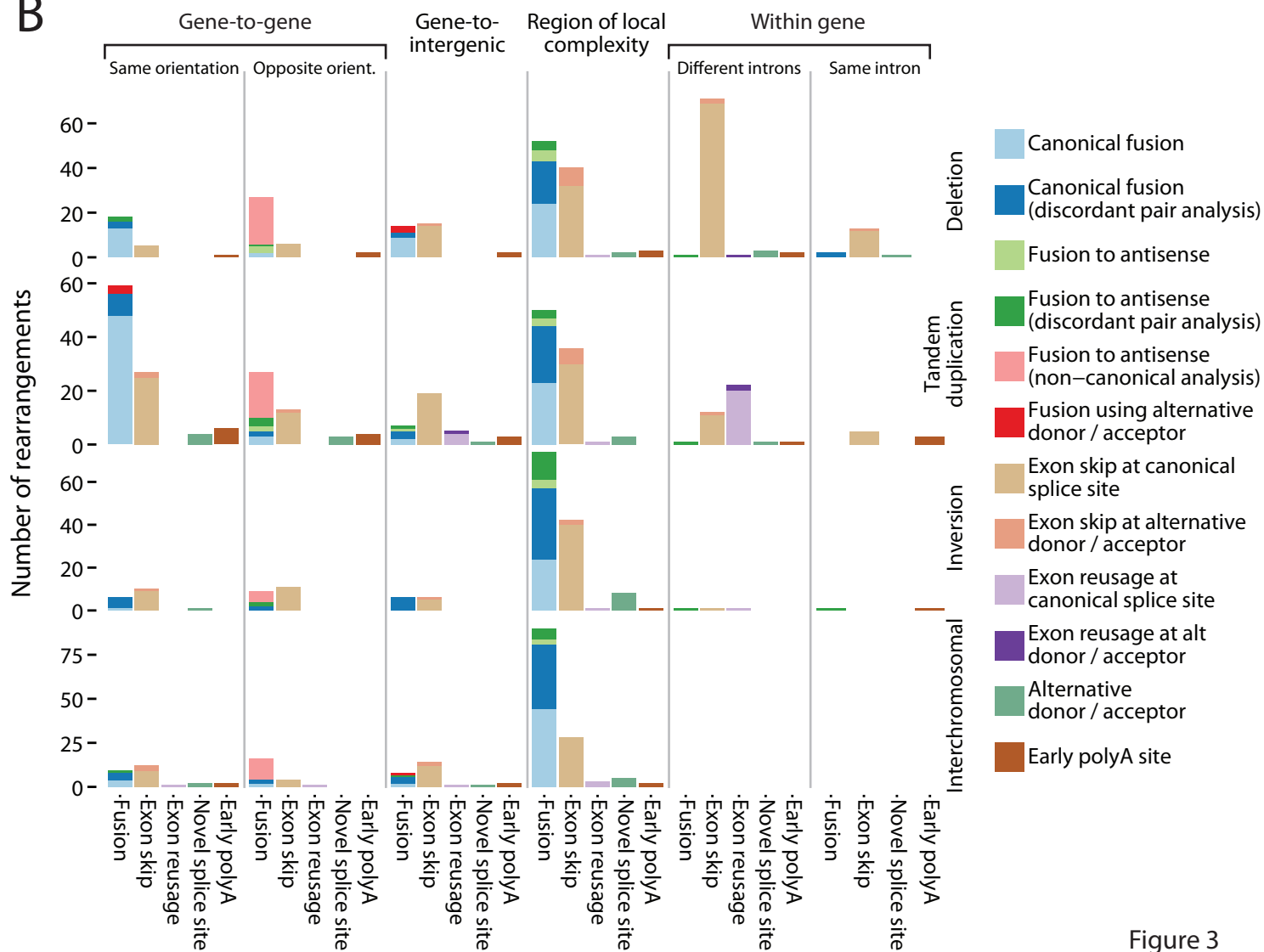
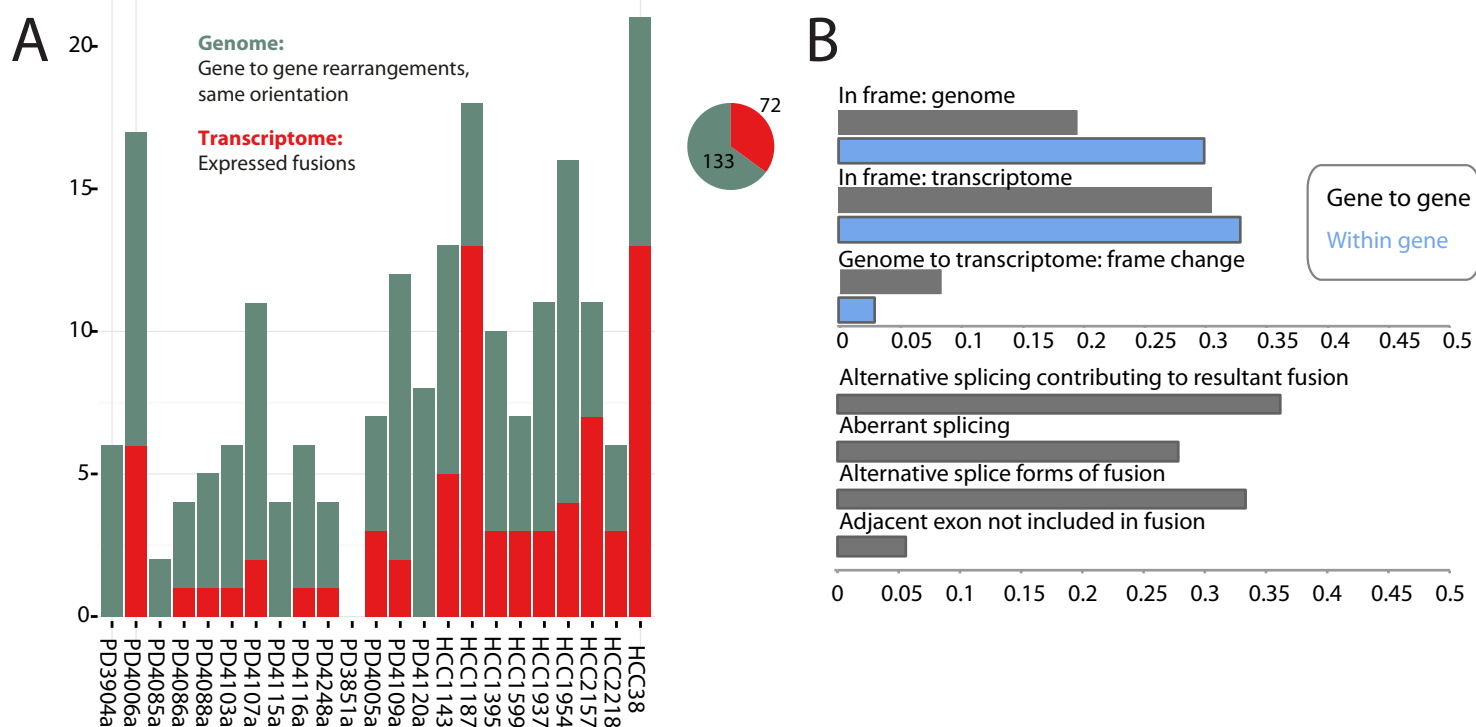
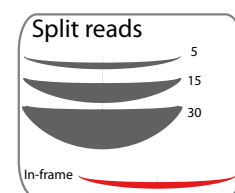
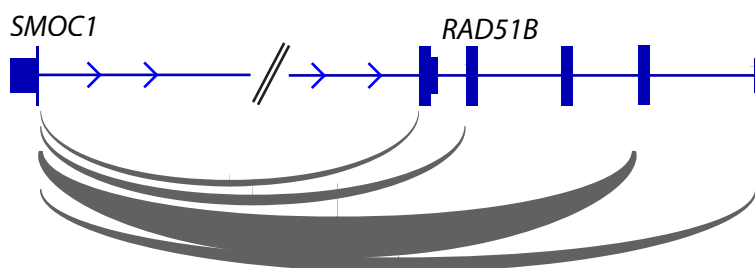


Figure 3



**C** **SMOC1-RAD51B:**  
Single isoform of donor, multiple isoforms of acceptor



**PLXND1-TMCC1:**  
Multiple isoforms of donor, single isoform of acceptor

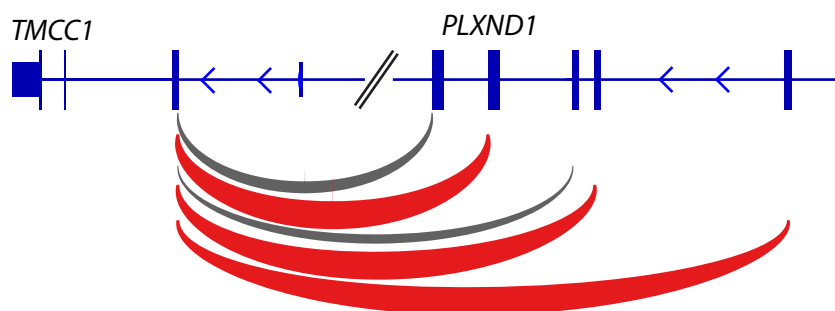


Figure 4

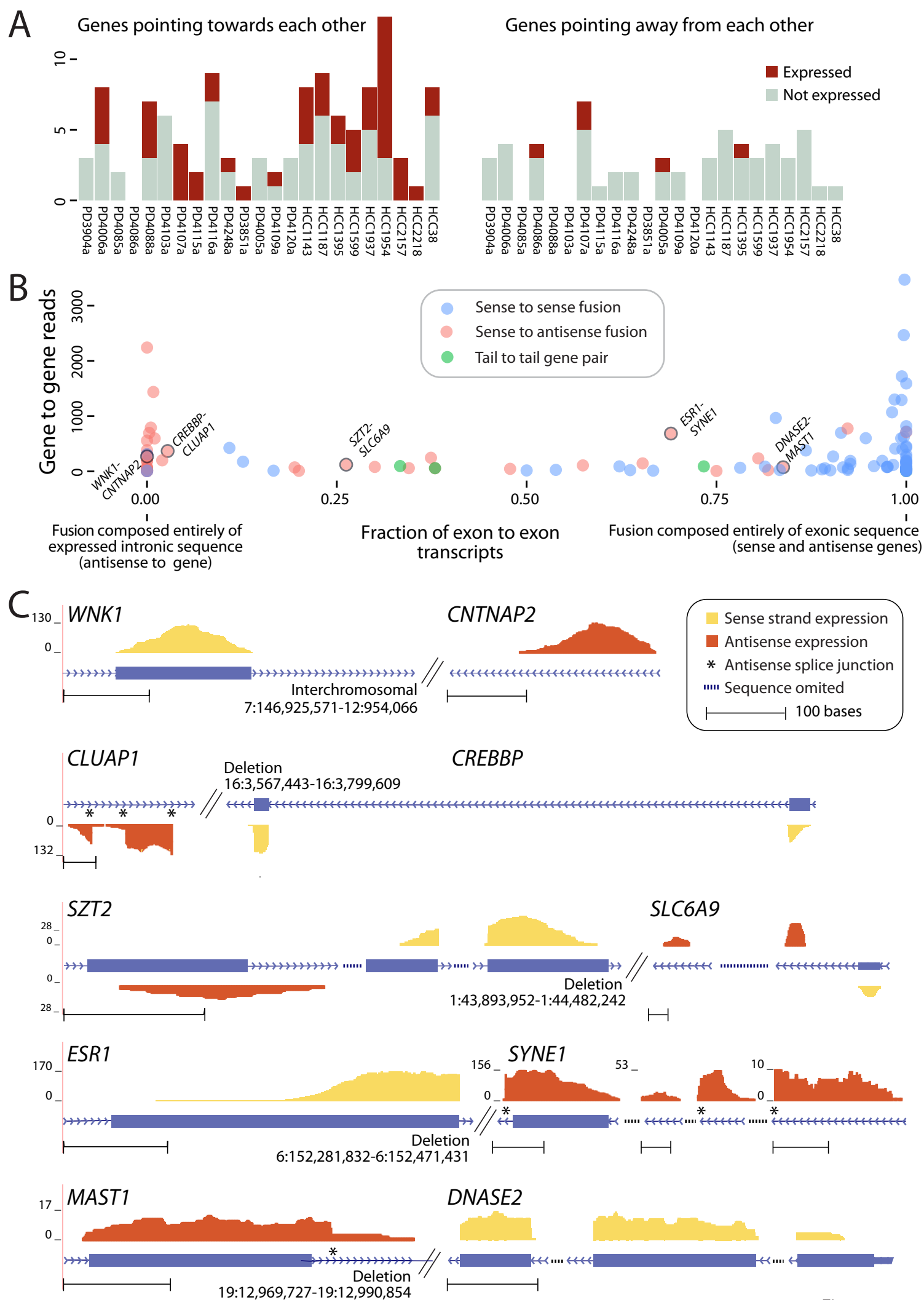


Figure 5

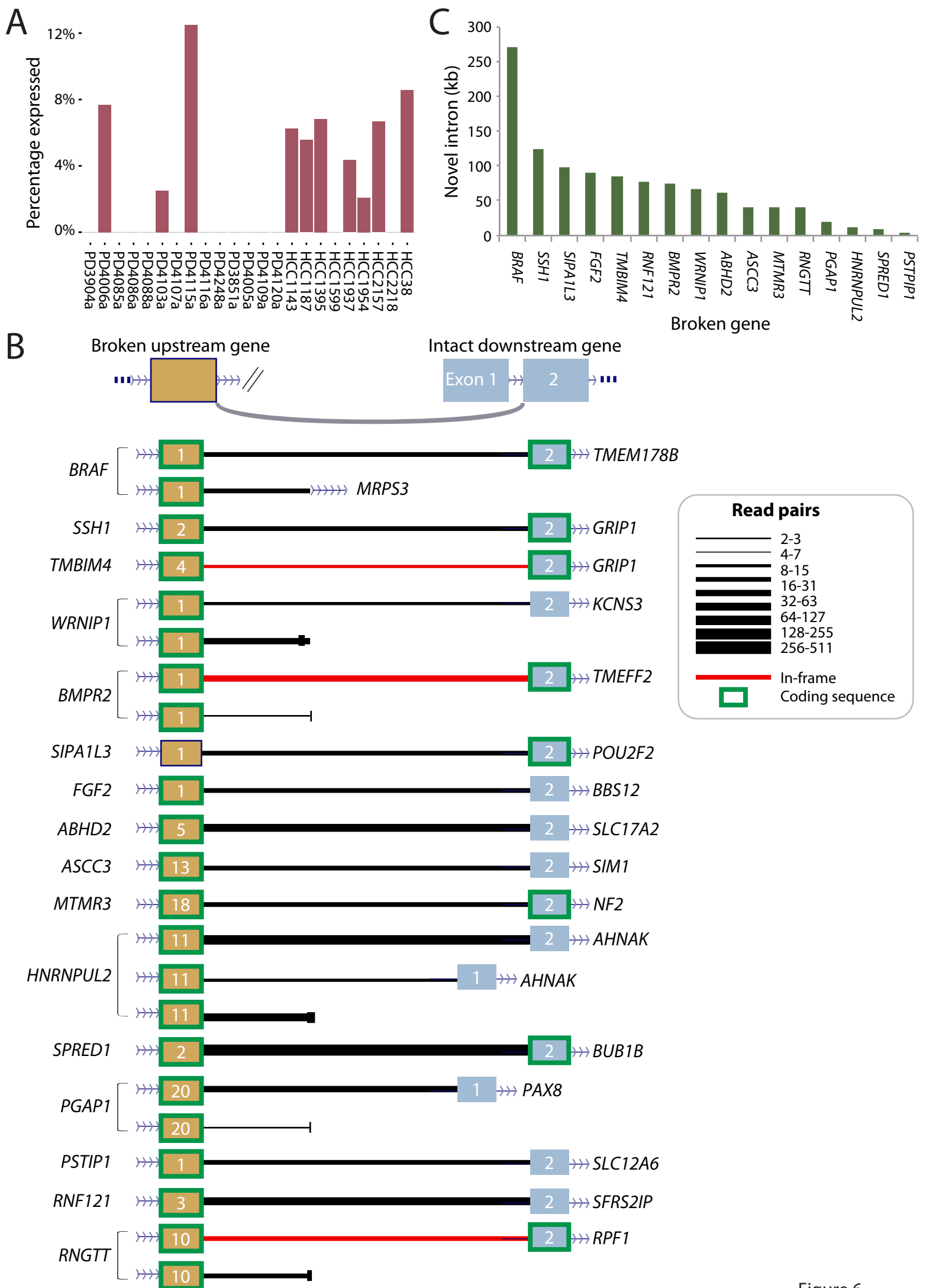


Figure 6

