Novel domain combinations in proteins encoded by chimeric transcripts

Milana Frenkel-Morgenstern and Alfonso Valencia

Structural Biology and BioComputing Program, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain

ABSTRACT

Motivation: Chimeric RNA transcripts are generated by different mechanisms including pre-mRNA trans-splicing, chromosomal translocations and/or gene fusions. It was shown recently that at least some of chimeric transcripts can be translated into functional chimeric proteins.

Results: To gain a better understanding of the design principles underlying chimeric proteins, we have analyzed 7,424 chimeric RNAs from humans. We focused on the specific domains present in these transcripts, comparing their permutations with those of known human proteins. Our method uses genomic alignments of the chimeras, identification of the gene–gene junction sites and prediction of the protein domains. We found that chimeras contain complete protein domains significantly more often than in random data sets. Specifically, we show that eight different types of domains are over-represented among all chimeras as well as in those chimeras confirmed by RNA-seq experiments. Moreover, we discovered that some chimeras potentially encode proteins with novel and unique domain combinations. Given the observed prevalence of entire protein domains in chimeras, we predict that certain putative domain combinations. Given the observed prevalence of entire protein domains in chimeras, we predict that certain putative domain combinations.

Availability: our scripts are available upon request.

Contact: avalencia@cnio.es

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The splicing of mRNAs is an essential aspect of eukaryotic gene expression. This process can occur in cis, within a single pre-mRNA, or in trans, between two different pre-mRNAs; in the latter case, chimeric transcripts are generated (Gingeras, 2009). Chimeric transcripts can also be produced by other mechanisms, including gene fusion (Li et al., 2008, 2009), recombination and other mechanisms (Gallei et al., 2004; Gmyl et al., 2003). The defining property of a chimeric RNA transcript is a fusion of transcripts from at least two discrete genes.

The pioneering observations evidencing the joining of sequences from separate genes were made in trypanosomes (Sutton and Boothroyd, 1986), though similar events have been shown to occur in nematodes (Krause and Hirsh, 1987) and higher eukaryotes (Li et al., 2009). More recently, chimeric transcripts were demonstrated also in flies (McManus et al., 2010a, b). In the case of the fly mod gene (mdg4), the independently transcribed pre-mRNAs are formed a chimeric transcript with exons from the sense and antisense strands of the mod genomic region. Other genes in fruit flies incorporate both sense and antisense exons, such as lola genes (McManus et al., 2010a, b). In the mosquito, chimeric transcripts were found to involve internal exons from genes at different chromosomal loci (Robertson et al., 2007).

In mammals, chimeric transcripts are frequently associated with cancer (Lackner and Bähler, 2008; Maher et al., 2009a, b). Indeed, abnormal MYC mRNA transcripts, a ubiquitous feature of tumor cells, typically involve one or more exons from another gene spliced to exon 2 of MYC (Chen et al., 2005). Additionally, fusion genes that arise from chromosomal rearrangements have been found in an expressed sequence tag (EST) library (Hahn et al., 2004) and in a more recent study, chimeric proteins were identified in different human cell lines (Frenkel-Morgenstern et al., 2012).

The interesting empirical evidence of human chimeras made up of two genes (Breen and Ashcroft, 1997) was supported in a subsequent study of the human cholesterol acyltransferase-1 (ACAT-1) hybrid mRNA (Li et al., 1999), in which mapped the 5′ Untranslated Region (5′ UTR) of this gene was mapped to Chromosome 7 and the remaining sequence to Chromosome 1 (Li et al., 1999). Paired-end RNA sequencing of human cancer cell lines led to the discovery of tissue specific converging, diverging and overlapping mRNA chimeras (Djebali, 2010; Maher et al., 2009a, b). Finally, high-throughput sequencing of fruit fly mRNAs identified chimeric transcripts with complex genomic architecture (McManus et al., 2010a, b).

Although, a large variety of chimeras have been described at the RNA level, there are still only a few examples in humans, where it has been shown empirically by mass spectrometry experiments that the chimeric RNAs produce corresponding chimeric proteins (Frenkel-Morgenstern et al., 2012). Some examples of chimeric proteins found in tumors, are those that result from recurrent chromosomal translocations, which can be found in different patients with the same tumor type (Mitelman et al., 2005). Although several studies have addressed the potential influence of local sequence features on such recurrence (Aplan, 2006; Miranda et al., 2006; Ortiz de Mendivil et al., 2009), the design principles of chimeric proteins still remain unclear, for examples the preferential incorporation of a full protein domain from the parental protein.

To better understand the design principles underlying chimeric proteins, here, we present an analysis of a data set of 7,424 human chimeric transcripts (Kim et al., 2006, 2010; Li et al., 2009). Recently, the expression of 175 chimeras in this...
We used a data set combined from the study of Ortiz de Mendibil et al., 2012. We focused on protein domains present in the chimeric transcripts and compared their permutations to those contained in protein coding transcripts generally found in humans. We found that chimeras contain complete protein domains significantly more often than would be expected if gene or exon splicing were assumed to be a random process. Specifically, we show that AT-hook, GTP_EFTU, MHC, SH2, SH3, TyrKc, EF-h domains and WD40 repeats are over-represented among chimeras. Moreover, we discovered that some chimeras potentially encode proteins with novel and unique combinations of domains. Given the observed prevalence of entire protein domains in chimeras, we predict that certain putative chimeras with missing activation domains may actively compete with their parental proteins, thereby exerting dominant negative effects. Interestingly, fusion proteins produced by chromosomal translocations in cancers also incorporate the tyrosine kinase (TyrKc), Runt, AT-hook (DNA binding), NUP repeats (nuclear signals) and coiled coil domains. Therefore, we propose that production of chimeric transcripts may produce active competition with original proteins and produce dominant negative phenotypes in cancer. Finally, it must be noted that chimeras can drive diversification of a given transcriptome offering the opportunity to acquire novel functional proteins.

2 METHODS

2.1 Data sets

The data for chimeric transcripts was taken from ChimerDB (Kim et al., 2006, 2010) and the study of Li et al. (2009). In addition, we used the data set of the chimeric junction sites covered by the RNA-seq reads from a recent study (Frenkel-Morgenstern et al., 2012). Only chimeric transcripts involving two genes were included in our chimera data set, and chimeras with a gap of one or more nucleotides in the junction site were excluded from the analysis. In addition, we separated one chimera data set from known fusion proteins catalogued in dCRID (Kong et al., 2011), an inventory of fusion proteins produced by chromosomal translocations in cancers and other diseases. The GENCODE (Harrow et al., 2006) and ENSEMBL (Flicek et al., 2011) databases were used, and to create the ‘all proteins’ data set. To avoid bias in the analysis due to over-representation of identical or similar sequences, we compared the protein sequences using BLAST (Altschul et al., 1997). When sequences were found by BLAST to be similar with an E-value $\leq 10^{-3}$, only one representative was kept in the data set. In addition, short protein coding regions, <100 amino acids, were excluded from the analysis. Domains were defined as in Pfam (Bateman et al., 2004; Finn et al., 2008; Gould et al., 2010) and SMART (Mulder et al., 2005). Finally, for each protein or chimera, we created a list of all possible domain-domain combinations. The basic units in our analysis are these domain pairs.

2.2 Fusion proteins in cancer

We used a data set combined from the study of Ortiz de Mendibil et al., (2009) and the dCRID database (Kong et al., 2011), which incorporates fusion sequences of gene-mapped translocation breakpoints in cancer (Mitelman et al., 2001). Sequences of fusion transcripts shorter than 100 amino acids were excluded from the analysis and thus, we analyzed a total of 323 gene fusions in which both partner genes contributed an annotated protein domain to the chimeric fusion protein generated by the translocation. Most these fusions (>60%) were reported in hematological tumors.

2.3 Random data sets and the null model

A random chimera data set was generated by randomly mixing and matching known protein coding sequences to create ‘artificial’ chimeric transcripts, and a set of 10,000 of these random chimeras was used in the analysis. To test if the chimeras incorporate full protein domains, the null hypothesis was that frequencies of full domains in the ‘real’ and randomly generated chimeras are similar when these sets are compared. We used a null hypothesis to test if certain domains are enriched in the chimeras and if they produce unique combinations of domains, i.e., whether the frequency of the domains is equivalent in the chimeras and in all human proteins when all the chimeras are compared with all proteins.

2.4 Enrichment of the protein domains

To verify the functionality of protein domains, the Eukaryotic Linear Motifs resource (ELM; Gould et al., 2010) was used. To recognize potential membrane proteins participating in a given chimera, we used TargetP 1.0 (Emanuelsson et al., 2007) and ELM (Gould et al., 2010). The 20-top domain frequencies in chimeras were considered for the over-representation analysis of specific domains versus the domain frequencies in all the human proteins. The standard error was calculated using the distributions of the domain frequencies in chimeras and all proteins as follows:

$$\text{STD}_{\text{ERR}} = \text{STD EV}/\sqrt{N} \quad (1)$$

where $N$ equals 20 (top domains) and STDEV was calculated as a standard deviation of the absolute differences between the observed and expected frequencies of 20-top domains.

An expected frequency for every domain was calculated using an assumption that a domain can be owned from any of two parental proteins and it calculated by the formula:

$$P_{\text{domain}} = 2 \times P_{\text{ALLPRTs}(\text{domain})} - (1 - P_{\text{ALLPRTs}(\text{domain})})$$

(2)

where $P_{\text{ALLPRTs}(\text{domain})}$ is the observed frequency of the domain in a data set of all human proteins. The domains were considered as over-represented in the chimeras if their frequencies were at least two standard errors bigger than the corresponding frequencies in all human proteins. The 20 most common (20-top) domains were chosen for the analysis because all the other domains had a frequency of <1% (Supplementary Material). These 20-top corresponded to 69% of the domains identified in all proteins, and 61% of those in all the chimeras.

2.5 Annotation of chimeras

The sequence similarities between the chimeric RNA transcripts and human-genomic regions were identified using in-house software and the UCSC BLAT search (Kent, 2002) to annotate the genes participating in each chimera. Using information on known transcripts from ENSEMBL (Flicek et al., 2011) and an in-house Perl program, the aligned exons, introns and untranslated regions in the chimeras were recognized. NCBI BLAST (Altschul et al., 1997) was applied to delineate the parental protein domains corresponding to the genomic regions contained within each chimeric mRNA. All the domain annotation results were manually inspected. Finally, WU BLAST (Lopez et al., 2003) was employed to define more precisely short or ‘strange’ genomic regions, as WU BLAST has proven most efficient when transcript composition is unknown (Elizabeth Cha and Rouchka, 2005).

2.6 Analysis of chimeric proteins translated in frame

To identify possible in frame translations of a given chimera, we translated all the chimeric transcripts in six frames. We ran UCSC BLAT search (Kent, 2002) and NCBI BLAST (Altschul et al., 1997) and used ENSEMBL (Flicek et al., 2011) to identify a frame enabled correct translation of exons from both proteins participating in chimeras. Thus, only the exons of two parental
We surveyed all the functional domains in proteins encoded by human genes, appearing in the sequence of ESTs, were used for the domain analysis of the chimeras (Fig. 1). In our chimera data set, 1,643 chimeras had a correct frame for two parental proteins that ensured a directed translation without a stop codon at the junction site of chimeras. We analyzed this collection separately and found an over-representation of protein domains, as in the whole chimera data set.

### 2.7 Comparison of domain pairs

We carried out a paired Wilcoxon test for each collection of protein sequences (Fay and Proschan, 2010), comparing the differences in the permutations of protein domains within the domain pairs of all the chimeras with that of all the human proteins, and that of all the chimeras against the random chimeras. The average length of a specific domain was computed considering all the proteins predicted to have this domain.

### 3 RESULTS

We surveyed all the functional domains in proteins encoded by chimeric transcripts translated and found that 27% of chimeras tend to contain complete functional domains of their parental proteins, significantly higher than in the corresponding random data set of chimeras (Wilcoxon test, \( P < 10^{-21} \)). We noted that chimeras contain various well-characterized complete protein domains (Supplementary Material), representing most protein domains (69%) of those present in the GENECODE 3C (human) database (Table 1). Remarkably, we found that the chimeras in our data set were significantly enriched in AT_hook, GTP_EFTU, MHC, SH2, SH3, TyrKc, EF-hand domains and WD40 (see ‘Results’ section below). These findings complement our earlier discovery of enrichment in transmembrane domains among chimeras (Frenkel-Morgenstern et al., 2012). Moreover, we found that 30.2% of chimeras confirmed by RNA-seq reads in the study of (Frenkel-Morgenstern et al., 2012) encode complete protein domains. Most of the chimeras are weakly expressed transcripts and they are very tissue-specific (Frenkel-Morgenstern et al., 2012). Given the prevalence of entire protein domains in chimeras, we suggest that certain chimeras may actively compete with the wild-type parental proteins, especially transcription factors.

#### 3.1 Domains from membrane proteins and receptors in chimeras

The domains of membrane proteins and receptors, namely, MHC and receptor tyrosine kinases-catalytic domain (TyrKc) are enriched in chimeras \((P < 10^{-5})\). Moreover, we noted that short coiled-coil domains \((<50 \text{aa})\) are also significantly enriched in human chimeras, likely reflecting the enrichment of parental membrane proteins (Barr and Short, 2003; De Matteis and Morrow, 2000; Short et al., 2005). These findings extend our previous observations that chimeras incorporate transmembrane domains and signal peptides to change a cellular localization of parental proteins (Frenkel-Morgenstern et al., 2012). Notably, 15% of chimeras confirmed by RNA-seq reads (Frenkel-Morgenstern et al., 2012) indeed incorporate coiled-coil and MHC domains of the parental membrane proteins (Table 1).

#### 3.2 EF-hand domains in chimeras

The ‘EF-hand’ (EFh) domain name is the Calcium (Ca)-binding motif discovered in the structure of

---

**Table 1.** The enriched protein domains in chimeras

<table>
<thead>
<tr>
<th>Chimeric data set</th>
<th>All human genes(^a)</th>
<th>All chimeras(^b)</th>
<th>All chimeras confirmed by RNA-seq reads(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total genes</td>
<td>22 304</td>
<td>7424</td>
<td>175</td>
</tr>
<tr>
<td>Identified domains(^d)</td>
<td>18 045</td>
<td>1318</td>
<td>55</td>
</tr>
<tr>
<td>ANK</td>
<td>1.9%</td>
<td>2%</td>
<td>0%</td>
</tr>
<tr>
<td>AT_hook</td>
<td>0.2%</td>
<td>1.6%</td>
<td>0%</td>
</tr>
<tr>
<td>Coiled Coil</td>
<td>18.1%</td>
<td>18.6%</td>
<td>10%</td>
</tr>
<tr>
<td>EF-hand</td>
<td>1.1%</td>
<td>3.9%</td>
<td>2%</td>
</tr>
<tr>
<td>EGF-like</td>
<td>5.5%</td>
<td>3.2%</td>
<td>2%</td>
</tr>
<tr>
<td>GTP_EFTU</td>
<td>0.5%</td>
<td>1.9%</td>
<td>4%</td>
</tr>
<tr>
<td>HOX</td>
<td>1.5%</td>
<td>1.2%</td>
<td>0%</td>
</tr>
<tr>
<td>Ig-like</td>
<td>7.2%</td>
<td>4.4%</td>
<td>2%</td>
</tr>
<tr>
<td>LRR</td>
<td>7%</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>MHC</td>
<td>0.3%</td>
<td>1.8%</td>
<td>4%</td>
</tr>
<tr>
<td>PDZ</td>
<td>0.8%</td>
<td>1.9%</td>
<td>0%</td>
</tr>
<tr>
<td>Phakinase</td>
<td>4%</td>
<td>2.2%</td>
<td>0%</td>
</tr>
<tr>
<td>RING</td>
<td>2.4%</td>
<td>1.9%</td>
<td>0%</td>
</tr>
<tr>
<td>RRM</td>
<td>1.6%</td>
<td>2.2%</td>
<td>6%</td>
</tr>
<tr>
<td>Runt</td>
<td>0%</td>
<td>1.2%</td>
<td>0%</td>
</tr>
<tr>
<td>SH2</td>
<td>0.8%</td>
<td>2.5%</td>
<td>2%</td>
</tr>
<tr>
<td>SH3</td>
<td>2%</td>
<td>3.5%</td>
<td>2%</td>
</tr>
<tr>
<td>TyrKc</td>
<td>0.7%</td>
<td>2.5%</td>
<td>0%</td>
</tr>
<tr>
<td>WD40</td>
<td>2.3%</td>
<td>5.5%</td>
<td>8%</td>
</tr>
<tr>
<td>ZnF</td>
<td>8%</td>
<td>8%</td>
<td>6%</td>
</tr>
<tr>
<td>(P)-value(^e)</td>
<td>&lt;10E(^{-5})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)All human genes from GENCODE (Harrow et al., 2006).

\(^b\)All ESTs and mRNAs from the ChimerDB collection (Kim et al., 2010), 200 transcripts of human data set (Li et al., 2009).

\(^c\)All chimeric transcripts confirmed by RNA-seq from all three aforementioned data sets (Frenkel-Morgenstern et al., submitted for publication).

\(^d\)The additional 3701 transmembrane domains and 2339 signal peptides were identified for the same data set of all chimeras (Frenkel-Morgenstern et al., 2012). For all chimeras confirmed by RNA-seq reads, the additional overrepresented domains are: ACTIN (4%), ATP synth_A (13%) and Ribosomal (11%).

\(^e\)10% of proteins having multiple EF-hand domains incorporate them from both of the parental proteins.
Among other functions, HMG proteins are also involved in the transcription regulation of genes containing, or in close proximity to, AT hooks, which are type of DNA-binding motifs with a preference for A/T rich regions and found in the high mobility group (HMG) proteins (Andrade et al., 2000). In particular, 9% of chimeras confirmed by RNA-seq reads at the chimeric junction site incorporate WD40 repeats (Table 1). However, it is notable that the number of repeats in chimeras is on average 1.3, as opposed to 6.1 in the parental proteins (Table 1). This reduction in the number of repeats raises the possibility that these domains are only partially folded and may be non-functional, and thus, they may contribute to dominant negative effects (see below) if they are fused into transcription factors.

### 3.3 WD40 repeats in chimeras

WD40 repeats of about 40 amino acid residues are found in a wide variety of eukaryotic proteins (Andrade et al., 2000). WD40 proteins play roles as adaptor and regulatory modules in various processes including signal transduction, pre-mRNA processing, transcriptional activation, cytoskeleton assembly, and cell cycle regulation (Andrade et al., 2000). WD40 repeats form β-propeller structures, which act as a platform for stable association between proteins. Additionally, the WD40 repeat propeller structure is an adaptable module that can form a multimer or bind DNA, it can antagonize the function of the wild-type protein (Maki et al., 2008). Given the prevalence in our data set of entire DNA-binding domains in the absence of transcriptional activation domains, we suspect that such chimeras compete with the parental proteins in chimeras as dimers/multimers have the potential to exert dominant negative effects (Fig. 2). In support of our hypothesis, a chimera from the two parental proteins (Table 2). Moreover, the same order and amount of exons are incorporated in the chimera as found in both parental proteins. Expression of the novel chimeras was confirmed at the RNA level by RNA-seq reads, with 4.1 chimeric reads on average in different human tissues (Frenkel-Morgenstern et al., 2012). These observations support the premise that there are specific design principles underlying the formation of chimeras.

### 3.4 DNA and GTP binding domains in chimeras

AT hooks are type of DNA-binding motifs with a preference for A/T rich regions and found in the high mobility group (HMG) proteins (Reeves and Beckerbauer, 2001) and others (Singh et al., 2006). Among other functions, HMG proteins are also involved in the transcription regulation of genes containing, or in close proximity to, AT-rich regions. The AT hooks domains are enriched in chimeras (Wilcoxon test, P < 10E−5) and incorporate with other functional domains, which may influence the activity of parental proteins especially transcription factors incorporated in chimeras.

### 3.5 Novel combinations of protein domains in chimeras

We compared the permutations of domain pairs in chimeras versus those in all proteins. We found that most domain pair combinations are represented in chimeras but also noted novel domain pairs, including HLH (Helix–Loop–Helix) and Pfam:GTP_EFTU (GTP-binding domain) as well as Pfam:Hydrolase_3 domain and Pfam:Polyprenyl_synth, coiled coil domain and ZnF_C2C2 and others (Table 2). Notably, the domains participating in a novel domain combination in a given chimera were found in multiple ESTs from the two parental proteins (Table 2). Moreover, the same order and amount of exons are incorporated in the chimera as found in both parental proteins. Expression of the novel chimeras was confirmed at the RNA level by RNA-seq reads, with 4.1 chimeric reads on average in different human tissues (Frenkel-Morgenstern et al., 2012). These observations support the premise that there are specific design principles underlying the formation of chimeras.

### 3.6 Dominant negative effects

Mutated transcription factors and mutated proteins that function as dimers/multimers have the potential to exert dominant negative effects competing with the parental proteins (Fig. 2). Namely, if a mutated protein has lost a particular activity but is still able to form a multimer or bind DNA, it can antagonize the function of the wild-type protein (Maki et al., 2008). Given the prevalence in our chimera data set of entire DNA-binding domains in the absence of transcriptional activation domains, we suspect that such chimeras compete with the parental transcription factors, thus exerting dominant negative effects (Fig. 2). In support of our hypothesis, a chimera that comprises a ligand-dependent transcriptional factor, the Peroxisome proliferator-activated receptor gamma-2 (PPARγ2) and the amino-terminal domain of the nuclear receptor co-repressor (CoR) has already been shown to exert a dominant negative effect (Suzuki et al., 2010).

Based on our hypothesis, we predict that a chimera from our data set, comprising the E2-α transcription factor, TCF3 (Immunoglobulin enhancer-binding factor E12/E47) and the Ribosomal protein, RPS19 (EST = BC009346.1) exerts dominant negative effects. Although the full sequence of RPS19 is present, encoding a known transcriptional repressor that is located in
Novel domain combinations in proteins

Table 2. The novel domain combinations found in chimeras

<table>
<thead>
<tr>
<th>Domain 1</th>
<th>Domain 2</th>
<th>Chimeric ESTs</th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Potential function</th>
</tr>
</thead>
<tbody>
<tr>
<td>VHS (membrane targeting/cargo recognition role in vesicular trafficking)</td>
<td>BRIX (ribosomal RNA processing)</td>
<td>AW977393</td>
<td>GGA2 (ADP-ribosylation factor binding protein 2)</td>
<td>GNG5 (RNA processing factor 1)</td>
<td>Golgi-trafficking</td>
</tr>
<tr>
<td>HLH (DNA-binding)</td>
<td>Pfam:GTP-EFTU (GTP binding domain)</td>
<td>HE514178</td>
<td>TCF3 (transcription factor E2-alpha)</td>
<td>EEF1A1 (Eukaryotic Translation Elongation Factor 1 Alpha)</td>
<td>Dominant negative</td>
</tr>
<tr>
<td>Pfam:Protein kinase (kinase catalytic domain)</td>
<td>Pfam:ubiquitin (ubiquitin, location or trafficking of the protein)</td>
<td>BF349450</td>
<td>UBD (ubiquitin D)</td>
<td>CSNK1D (casein kinase 1, delta isofrom 1)</td>
<td>Change localization or trafficking</td>
</tr>
<tr>
<td>Pfam:Hydrolase_3</td>
<td>Pfam:Polyprolyl synth</td>
<td>IG491331</td>
<td>PMM2 (phosphomannomutase 2)</td>
<td>FDPS (farnesyl diposphate synthase isofrom a)</td>
<td>Dominant negative</td>
</tr>
<tr>
<td>Znf_C2C2</td>
<td>Coiled-coil (DNA binding) domain</td>
<td>BG164187</td>
<td>Prickle-like 2 (Drosophila homolog)</td>
<td>TCEA3 (transcription elongation factor A (SII), 3)</td>
<td>Dominant negative</td>
</tr>
<tr>
<td>PHD-zinc finger</td>
<td>Coiled-coil (DNA binding) domain</td>
<td>AB032253</td>
<td>BAZ1B (briondomain adjacent to zinc finger domain, 1B), transcription factor</td>
<td>GRID1 (Glutamate receptor delta-1 subunit)</td>
<td>Dominant negative</td>
</tr>
<tr>
<td>PHD-zinc finger</td>
<td>Coiled-coil (DNA binding) domain</td>
<td>DA092156</td>
<td>PHF14 (PHD finger protein 14 isoform 1)</td>
<td>DCM4466</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Every domain combination is confirmed by the putative chimeric ESTs from our data set. Potential function of the resulting chimera is proposed.

The cytoplasm and nucleus, and that can form homodimers and heterodimers (Fig. 3A), only the DNA-binding domain of the nuclear transcription factor TCF3 is incorporated in the chimera, the activation domain of TCF3 is missing (Fig. 3A). Therefore, the resulting chimera can potentially bind TCF3’s target DNA but it cannot activate transcription and accordingly it could compete with wild-type TCF3 to bind DNA (Fig. 3B). Overall, around 14% of the chimeras in our data set have the potential to exert dominant negative effects. We also noted changes in the permutations of domain pairs among chimeras, particularly when considering domains found in enzymes that function as dimers. This finding suggests that such chimeric enzymes could exert dominant negative effects, but when considering enzymes, it is important to note that the final phenotype would be strongly dependent on the expression level of the chimeric protein. For example, in the human data set (Li et al., 2009), a translocation between the myeloid/lymphoid or mixed-lineage leukemia (MLL) protein and guanine monophosphate synthetase (GMPS) produces a chimera with potential to exert dominant negative effects. Typically, MLL positively regulates the expression of target genes, including multiple HOX genes (Li et al., 2005), but some MLL gene fusions have been shown to exert dominant effects (Li et al., 2005). GMPS is an enzyme that functions as a homodimer, and although the functional domain is missing in the chimera, the dimerization domain is preserved along with the Zinc finger and PHD domains (Fig. 3C). Therefore, the resulting chimeric enzyme can potentially dimerize with the wild-type functional GMPS enzyme but interfere with its function due to the non-functional part of MLL (Fig. 3D), and thus exert a dominant negative effect. Similarly, the AML1/EVI-1 chimera generated by translocation has been shown to exert dominant negative effects (Mitani, 2004) and the MLL/AF4 fusion product has been suggested to function as a dominant negative protein (Domer et al., 1993).

We focused our analysis on chimeras that incorporate parts of transcription factors (activators or repressors) as it has been shown that even low expression of mutated transcription factors can interfere with gene activation repression. The abundance of chimeras containing transcription factor among DNA-binding proteins was determined to be (19.5%, P < 6.5E-6). Given the prevalence of entire protein domains in chimeras, we predict that even low expressed chimeras that incorporate the DNA-binding domains of transcription factors but have lost activation domains actively compete with the parental proteins, thus exerting dominant negative effects.
biosynthesis regulated by GMPS with parental GMPS to form dysfunctional heterodimers, impairing the MLL/GMPS protein lacking a functional GMPS domain likely competes homodimer, and mixed-lineage leukemia (MLL). (RPS19). (view of guanine-monophosphate-synthetase (GMPS), which functions as a

Fig. 2. A schematic representation of the competition mechanism between the wild-type transcription factor (TF) and the dominant-negative chimera (DNC). (A) A lowly expressed DNC compete with TF. (B) As a result of the competition, the DNC binds a promoter region but cannot activate a target gene inserting dominant-negative effect.

Fig. 3. Putative chimeric proteins can exert dominant negative effects. (A) Schematic view of transcription factor (TCF3) and ribosomal protein (RPS19). (B) The putative chimeric protein harnesses RPS19 to the DNA-binding domain of TCF3 and likely exerts a dominant negative effect by competing with wild type TCF3 for DNA binding. (C) Schematic view of guanine-monophosphate-synthetase (GMPS), which functions as a homodimer, and mixed-lineage leukemia (MLL). (D) The putative chimeric MLL/GMPS protein lacking a functional GMPS domain likely competes with parental GMPS to form dysfunctional heterodimers, impairing the biosynthesis regulated by GMPS

3.7 Protein domains in cancer fusion proteins

Using the same prediction procedures, we identified the main classes of fusion proteins resulting from translocations: the tyrosine kinase (TyrKc) domains (5.7%), the Runt domain (4.5%), the AT-hook DNA-binding domains (4.8%), the NUP repeats (6%) and coiled-coil domains (18.5%). Interestingly, most of these domains are DNA-binding domains or they incorporate nuclear signals. Our findings indicate that only certain combinations of protein domains are present in fusion proteins, forming non-random domain combinations, and implying that such combinations are subject to potential functional constraints (Ortiz de Mendibil et al., 2009). As mentioned above, the signature of DNA-binding domains and transcription factors can produce dominant negative phenotypes (Fig. 3A and B).

4 DISCUSSION

Here, we show that chimeras are particularly enriched in eight types of protein domains. In conjunction with our earlier study, our data raise the possibility that chimeras result not only in altered cellular localizations due to a sizable enrichment in transmembrane domains (Frenkel-Morgenstern et al., submitted for publication), but also in the acquisition of a number of other new functions.

Using the ELM resource, the position and length of each domain within a given protein was assigned and annotated. Thus, the domain composition and domain order relative to the protein sequence was readily available for each protein, allowing us to identify unique and novel domain combinations in chimeras. Of note, as many chimeras have been evidenced by multiple ESTs (Fig. 1 and Supplementary Material), we propose that some chimeras may be produced by regulated trans-splicing and have functional advantages. Elucidating the possible trans-splicing mechanisms, in humans, remains as a promising field of research.

In this study, we focused on chimeras ‘in-frame’ and chimeras confirmed by RNA sequencing experiments (Frenkel-Morgenstern et al., 2012). For all chimeras confirmed by RNA-seq and ‘in-frame’, we predicted the protein domains they contained and compared the permutations of domain pairs to those found in all proteins or in random data sets. Using this approach, we identified novel domain pairs that are unique to chimeras, which incorporate transcription factors as parental genes. Given the fact that even low expression of mutated transcription factors can interfere with the function of the wild-type transcription factor, we propose that many chimeras exert dominant negative phenotypes. Taking into consideration the protein domains identified in the fusion proteins, we propose that dominant negative effects of fusion proteins may be frequent in cancer. Moreover, chimeras may influence the protein–protein interactions of parental genes in the protein interaction network. Thus, chimeras with domains from the highly connected proteins will produce more consequences than chimeras from less connected proteins. Taking in consideration, the recent findings on the highly expressed genes incorporated in the human chimeras (Frenkel-Morgenstern et al., 2012), and the fact that highly expressed genes tend to have more protein–protein interaction partners (Krylov et al., 2003), we propose that chimeras can influence the protein–protein interaction network of the parental genes.

Having validated the existence of chimeric proteins in eukaryotes, we stress the need to take chimeric proteins into consideration when carrying out experimental studies of protein cellular localization. Moreover, the potential dominant negative effect of chimeras should be taken into account when protein–protein interactions are studied.

Finally, we suggest that future studies investigate if chimeric RNAs represent useful biomarkers for the early diagnosis of different cancers.

5 ACKNOWLEDGEMENTS

The authors thank Dr Miguel Vazquez for help with random data sets, Jose Manuel Rodriguez for the ELM, SignalP and TargetP scripts and
Novel domain combinations in proteins

Dr Dan Michalci for valuable discussions. The authors also thank Prof. Sanghyuk Lee and Prof. Seokmin Shin for making available the chimeric transcrips in the ChimerDB databases.

Funding: Spanish Government grant (CONSOLIDER, CSD2007-00050), B02007-66855, Open PHACTS (IMI-115191-2) and the NHGRI-NIH ENCODE grant (HG00455-04). Miguel Servet (FIS) to M.-F.M.

Conflict of Interest: none declared.

REFERENCES