



Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Interspaced transcription chimeras: Neglected pathological mechanism infiltrating gene accession queries?

Martti Tolvanen^a, Pauli J. Ojala^{b,*}, Petri Törönen^c, Heidi Anderson^b, Jukka Partanen^b, Hannu Turpeinen^b

^aInstitute of Medical Technology, University of Tampere, Finland

^bFinnish Red Cross Blood Service, Research and Development, Kivihaantie 7, 00310 Helsinki, Finland

^cInstitute of Biotechnology, University of Helsinki, Finland

ARTICLE INFO

Article history:

Received 18 August 2008

Available online xxxxx

Keywords:

Clinical transcriptomics

Systems medicine

Intergenic splicing

Chimerism

Transcriptome meta-analysis

ABSTRACT

Over half of the DNA of mammalian genomes is transcribed, and one of the emerging enigmas in the field of RNA research is intergenic splicing or transcription induced chimerism. We argue that fused low-copy-number transcripts constitute neglected pathological mechanism akin to copy number variation, due to loss of stoichiometric subunit ratios in protein complexes. An obstacle for transcriptomics meta-analysis of published microarrays is the traditional nomenclature of merged transcript neighbors under same accession codes. Tandem transcripts cover 4–20% of genomes but are only loosely overlapping in population. They were most enriched in systems medicine annotations concerning neurology, thalassemia and genital disorders in the GeneGo Inc. MetaCore-MetaDrug™ knowledgebase, evaluated with external randomizations here. Clinical transcriptomics is good news since new disease etiologies offer new remedies. We identified homeotic HOX-transfactors centered around BMI-1, the Grb2 adaptor network, the kallikrein system, and thalassemia RNA surveillance as vulnerable hotspot chimeras. As a cure, RNA interference would require verification of chimerism from symptomatic tissue contra healthy control tissue from the same patient.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

In the endeavor towards developing standard gene nomenclature, the upheaval of the “RNA-world” literature confounds unambiguous data retrieval. Recently, data regarding intergenic splicing or transcription induced chimerism (TIC) in the human genome has been published [1–4]. Pioneering studies on the analogous controversial fused transcripts were reported in the mouse [5] and human, from chromosomes 21 and 22 [6], which revealed that over half of the DNA of mammalian genome is transcribed, although less than 4% is translated. There are two types of RNA splicing: cis- and trans-splicing. Molten gene pairs analyzed in this study result from cis-splicing of RNA molecule (Fig. 1). Also trans-splicing has recently been demonstrated to occur regularly in normal human cells [7]. Trans-splicing involves two separate RNA molecules even from genes locating on different chromosomes, whereas cis-splicing is a result of a fusion of two neighboring genes as a single, united messenger RNA molecule.

One challenge has affected genome assemblies until recently and is still present in secondary databases: dozens of transcript extension genes have traditionally been named in what could be considered illicit union in a gene matrimony, in e.g. the Ensembl Gene IDs. We question whether the putative existence and occa-

sional usage of common or *de novo* “molten” exons really justifies assigning the longest transcript ID as the actual gene. Separation of united genes would increase the information content of the genome contig annotation and prevent researchers from futile data-mining. At the moment, use of transcript-driven accessions in the query, for instance, in the Uniprot database [8], the most comprehensive catalog of information on proteins, may generate incorrect results. This presents a systematic caveat in interpreting e.g. microarray meta-analysis results.

Ensembl is the authoritative software system that allows for automatic analysis and annotation of contemporary genome sequence assemblies and visualization of the annotated genomes [9]. Until recently, some entries have resulted in the production of confounding family members due to intersplicing artifacts. Valuable sets of gene clusters with inconsistent topology have been collected along the Ensembl project of various genome assemblies. A new strategy has addressed this dilemma, such that transcript predictions are only clustered into genes when their coding sequences overlap on the genome, in contrast to only overlapping on the transcript level (personal communication, Ensembl project). The artificial, yet official in terms of nomenclature, merged genes of the past, however, shed light on the putative pathology and etiology of clinical transcriptomics.

The number of genes exhibiting occasional intergenic splicing into a single, “tandem” RNA sequence with the potential of encoding a chimerical protein sequence has been estimated to account

* Corresponding author. Fax: +358 9 5801310.

E-mail address: pauli.ojala@bts.redcross.fi (P.J. Ojala).

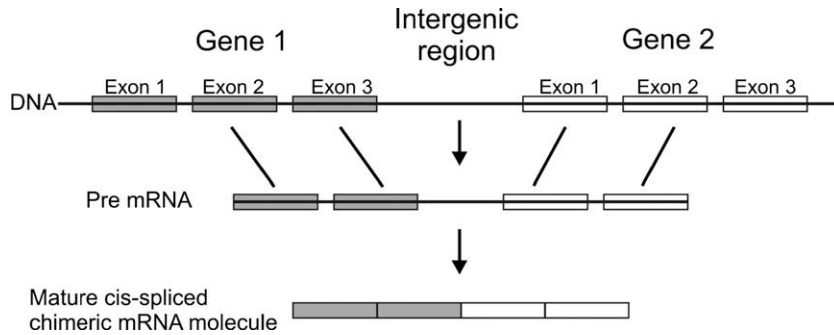


Fig. 1. Schematic representation of cis-splicing leading to chimeric gene fusions. Intergenic splicing combines exons from the upstream and downstream genes to form a chimeric gene fusion. Occasionally, the intergenic region is preserved in the mRNA transcript, too.

for 4–5% of the human genome. This original approximation is according to the “Encyclopedia of DNA elements” pilot project, whose aim was to identify all functional elements in the human genome, based on 1% of the human genome [10]. Although the detailed gene cohorts reported [1–4] have not always been experimentally verified, it is not surprising that these are only loosely overlapping and rather non-redundant.

Whether the fused proteins with chimerical domains reflect a dysregulation of the transcription termination by RNA polymerase or serve a genuine teleonomic purpose remains an open question. Is this phenomenon a functional one, or does it constitute a novel pathological mechanism?

2. Methods

2.1. Gene symbols

We examined interspersed chimeras from 60 genes constituting 30 gene pairs from the 421 individual RefSeq transcript IDs verified from tissues by Akiva et al. [1] as intergenic splice isoforms, and 106 genes constituting 53 gene pairs from the 352 RefSeq IDs listed by Parra et al. [2] as tandem pairs with ESTs linking two transcripts. Since only 11 of the “molten” gene pairs were in both major reports, the total number of the controversial genes was 145. Twenty gene pairs of the 30 transcript fusion pairs forced under a single Ensembl gene ID from Akiva et al. and 30 gene pairs of the 53 fusion pairs from Parra et al. exhibited both unconventional topology and multiple symbols. The reference information for the 421+352 human genome gene fusions was derived using Clone Gene ID-converter [11] and SRS 7 [12].

2.2. Randomizations of the enrichment and depletion analyses

Significance of the scoring and prioritization of networks and pathways according to the relevance to the chimeric gene sets was evaluated based on the size of the intersection between input data set and set of genes/proteins corresponding to the network module in question. The setup can be considered as a sampling without replacement and the probability to randomly obtain intersection of certain size between the input set, following hypergeometric distribution. This means a null-hypothesis, whereby the genes are regarded as independent with each others in terms of annotation. Noteworthy, the manually curated MetaCore-MetaDrug™ platform takes into account also nodes representing literature entities without vis-a-vis gene correspondence, such as protein complexes. The probability of a subset of size n to include r marked ones provided that n and R are unrelated follows the hypergeometric distribution

$$P(r, n, R, N) = \frac{C_R^r C_{N-R}^{n-r}}{C_N^n} = \frac{C_n^r C_{N-n}^{R-r}}{C_N^R} \\ = \frac{R!(N-R)!}{N!} \frac{n!(N-n)!}{r!(R-r)!(n-r)!(N-R-n+r)!}$$

The mean of this distribution is equal to:

$$\mu = \sum_{r=0}^n rP(r, n, R, N) = \frac{nR}{N} = nq$$

where q defines the proportion of the network or pathway in the whole genome/dataset. The dispersion/variance of this distribution is described as:

$$\sigma^2 = \sum_{r=0}^n r^2 p(r, n, R, N) - \mu^2 = \frac{nR(N-n)(N-R)}{N^2(N-1)} \\ = nq(1-q) \left(1 - \frac{n-1}{N-1}\right)$$

These equations are invariant in terms of exchange of n for R which means that the “subset” and “marked” are equivalent and symmetrical sets. In the cases of

$$r > n, r > R \text{ or } r < R + n - N, P(r, n, R, N) = 0$$

the z-score is used for comparison and prioritization of the networks in the MetaCore™

$$z\text{-score} = \frac{r - \frac{nR}{N}}{\sqrt{n \left(\frac{R}{N}\right) \left(1 - \frac{R}{N}\right) \left(1 - \frac{n-1}{N-1}\right)}} = \frac{r - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation of the hypergeometric distribution. z-Score represents a measure of relative deviation of r from its expected mean value. For the evaluation of statistical significance, the null-hypothesis which states that the subsets R and n are independent and, therefore, the size of their intersection follows the hypergeometric distribution, is considered. The alternative hypothesis states that there is positive correlation between the subsets. Based on these assumptions, p -value is calculated as the probability that intersection of two randomly selected subsets of N would have the size of n or larger:

$$pVal(r, n, R, N) = \sum_{i=\max(r, R+n-N)}^{\min(n, R)} P(i, n, R, N) = \frac{R!n!(N-R)!(N-n)!}{N!} \\ \times \sum_{i=\max(r, R+n-N)}^{\min(n, R)} \frac{1}{i!(R-i)!(n-i)!(N-R-n+i)!}$$

In short, the randomization is carried out beforehand in the GeneGo MetaCore™ platform and taken into account in the output p -values.

In the present investigation, we have also critically evaluated the enrichment method of the GeneGo MetaCore™ as end-users who do not have the full access to its MetaBase (i.e. local downloading of the databases found in the knowledgebase). More specifically, we performed an independent evaluation of the aforementioned null-hypothesis assumption, by carrying out randomization of corresponding input accession codes against the public Gene Ontology (GO) classes that are also found in the GeneGo MetaCore platform aside its proprietary systems medicine and toxicology annotation databases. In this evaluation, we compare randomization-derived empirical p -values with the p -values reported by GeneGo platform against the public GO. In the cases of the putatively flawed nomenclature possibly infiltrating the Ensembl platform with identical gene accessions for both the up and downstream fusion genes, these gene pairs were considered as one input item when calculating the size of the input gene sets.

Classes were obtained by using the Ensembl gene IDs and obtaining the classifications from ID-Converter web tool [11]. Obtained GO mappings were used to link genes to reported and parental GO classes. Only classes that had more than three genes were included to the dataset. Analysis of GO classes was done using the one-sided hypergeometric test, designed to monitor both over- and under-representation [13]. We omitted the genes that were not mapped to GO, according to a procedure verified before [14] as for these genes there is no clear information whether they belong or do not belong to the classes. Some classes showed strong signal in empirical p -value analysis, with all the results from the randomization showing weaker signal. This generates the p -value = 0, and as the results were analyzed as 10 based logs, these cases were un-calculatable. Therefore these cases were modified by adding 0.5 to number of better observations. The generated GO data included 6090 classes. This generates a significant Multiple Testing problem, generating seemingly significant results. We performed, therefore, an analysis of 500 random gene lists for each analyzed gene sets with exactly identical size and identical GO data matrix. The randomizations allow us the generation of empirical p -values for each rank in the sorted list of positive results, by using the ordered results from each randomization with the same rank as a reference. A similar idea has been used for example in the analysis of the hierarchical trees with GO classes [15]. The drawback of this method is that as the reference p -values for the lower ranks in the sorted GO class list are naturally weaker, one can actually get better empirical p -values for the lower ranks in the GO class list. In addition, the method does not account for the correlation occurring between the GO classes. Therefore, we added an *over-conservative empirical p -value analysis* to the procedure, where each rank of p -values is compared to the best result from each randomization. This is otherwise similar to the method represented in [15] but with the more stringent criterion that our novel method maintains the same, hardest reference pool for all the ranks. The methods produce the same result for the first class in the ordered GO class list, but in the descending ranking, the over-conservative method starts to penalize the obtained results with an extra cost. Our method reports not only enrichment, but also depletion or deprivation signals, in contrast to the GeneGo MetaCore™. These are relevant, however, for the annotation classes containing many gene members. We consider that these two represented empirical p -values give a reliable estimate on the actual probability on seeing similar or stronger results by random. We are not aware of more stringent methods for the analysis of true significance of the class enrichment or depletion methods.

2.3. Text mining

Quantitative text mining analysis of the published interacting partners of the chimera constituents was derived by Ariadne

Genomics Pathway Studio Enterprise 5.0 based on ResNet 5.0 Knowledgebase and MedScan natural language processing against PubMed abstracts (Ariadne Genomics Inc., Rockville, USA) [16]. The common and unique attributes and property enrichment of the genes undergoing intergenic splicing were also compared against the GeneGo Inc's MetaCore-MetaDrug platform, the putatively widest, human-specific, manually curated systems medicine database available [17]. This systems biology knowledgebase only takes into account causal relations (in contrast to unverified high-throughput screening associations).

2.4. Ensembl discrepancy analysis

In the Ensembl project discrepancy analysis from [Supplementary Table S2](#), the multiple category criterion refers to transcripts of a particular gene associated with more than one gene symbol associated with RefSeq or Swiss-Prot records for the same species. In regard to the topology criterion, a gene has at least three transcripts and one transcript ends before another one starts, which would be the case if there was a longer transcript that would join two genes into one cluster. In the clustering criterion, overlapping of the exon start or end coordinates was detected. In the topology criterion, two transcripts have been clustered into the same Ensembl gene based on genomic overlap, and the script behind provides a more stringent search for overlapping exon boundaries. In the case when there is no overlap, the script indicates sub clusters where it finds matching exon coordinates. Splice variants are assumed to have at least one exon overlap, according to conventional definition, but this appears to not always be observed in empirical biology. The problematic contigs were originally derived from the Ensembl assembly 43, and the script used to classify the genes is described in [Supplementary Table S2](#). At the contemporary most recent assembly release 50, the Ensembl deposit is no longer maintained and thus can not be used as a source of information on the transcript level fusions and other discrepancies.

3. Results

3.1. Enrichment of the transcript fusions in gene ontologies and disease annotations

Akiva et al. [1] reported 213 gene pairs and Parra et al. (2006) [2] 176 gene pairs that exhibited intergenic splicing, whereas the Ensembl project listed 709 genes with multiple symbols, 460 genes with inconsistent cluster topology, 264 problematic genes based on clustering, and 217 genes with discrepant gene names ([Supplementary Table S2](#)). Only 105 individual genes were common to both Akiva et al. and Parra et al. which contained 268 and 198 unique genes, respectively ([Supplementary Fig. S1](#)). Therefore, we analyzed their common and unique attributes and report their property enrichment to the putatively widest, manually curated, human-specific relation database available [17], and compared all 1760 most controversial genes from the Akiva et al. (2006) and Parra et al. (2006) reports, as well as Ensembl deposit against GeneGo's MetaCore-MetaDrug platform. In the GeneGo platform, 349 gene IDs had been annotated to canonical maps, 593 in diseases, 742 in public gene ontology, and 1197 in networks.

In the manually validated GeneGo MetaCore-MetaDrug™ systems medicine platform, the controversial gene sets or their combinations were most significantly associated with the neurological Cockayne syndrome, pregnancy complications, genital disorders, epilepsies, thalassemia, alcohol-induced neurological delirium and neuroectodermal tumours. Notably, four essential genes out of the 28 annotated to thalassemia were included among the transcript fusions: hemoglobin alpha-1, hemoglobin gamma-1,

hemoglobin gamma-2, and glucose-6-phosphate isomerase. The toxicological networks enriched were nucleotide-excision repair in DNA damage, antigen presentation, various circuits of the cell cycle, and transmission of nerve impulses.

In terms of functions, the most enriched molecular processes in the systems biology section (in contrast to the systems medicine) of the MetaCore-MetaDrug™ were related to cytoskeletal dynamics, meiotic cell division, transcription, folding and neurohormone signalling. More specifically, the gene set behind Parra et al. (2006) included components of both intermediate filament and tubulin networks affecting cell division. The common and unique genes among the gene sets attributed to intergenic splicing or other inconsistent RNA topologies described above are displayed in [Supplementary Fig. S1](#). The most enriched gene ontologies in GeneGo processes are displayed in [Supplementary Fig. S2](#), disease biomarkers in [Supplementary Fig. S3](#), and GeneGo toxicity classes in [Supplementary Fig. S4](#), taking into account all gene sets except the most divergent Ensembl deposit of disagreeing gene symbols.

Next, we investigated the nature of the fragmentary reports of the transcript fusions and Ensembl depots for controversial gene contigs one by one to gain insight into the actual molecular biology behind the phenomena. In order to compare the enrichment method of the commercial GeneGo MetaCore-MetaDrug™ platform to the analogous results subtracted with randomization, we utilized projection against the public Gene Ontology (GO). Upon a local downloading of the current human genome (Ensembl assembly 50) and the respective 141,631 GO annotations of the 36,396 putative human gene ID's, not only enrichment analysis but also impoverishment (depletion) analysis was carried out. This was done with the same hypergeometric distribution method as inside the GeneGo, but with with extensive external extra randomization round afterwards with artificial input gene sets, to subtract and derive empirical *p*-values against the background noise. Here, the common genes inside the controversial gene sets were not combined with optimal permutations, which is feasible only in the MetaCore™, but the gene sets were individually scrutinized.

As a result, the GO projection from the GeneGo platform brought forth similar enrichment results than our inhouse enrichment with randomizations. [Supplementary Fig. S5](#) shows the GeneGo enrichment to GO Biological Processes, [Supplementary Fig. S6](#) shows the GeneGo enrichment to GO Molecular Functions, and [Supplementary Fig. S7](#) shows the GeneGo enrichment to Cellular Components. In comparison, [Supplementary Table S1](#) shows our inhouse enrichment results and depletion results (marked as “2way” method). Since the number of randomizations was 500, the maximal enrichment in terms of empirical *p*-values could be only <0.001 or E-3 (marked as >3, in the log scale), although the observed *p*-values reached as low levels as E-10.

In the GO Biological Processes, the Ensembl controversies with clustering criterion and multiple symbols were significantly enriched in the classes of nucleosome assembly, chromatin assembly, and DNA packaging; the category of discrepant symbols in antigen presentation; the category of topology criterion in regulation of small GTPases, and the gene set behind Parra et al. (2006) only slightly in protein polymerization. In the GO Molecular Functions, the category of topology criterion was enriched in GTPase regulation; the category of discrepant symbols in MHC II receptor activity and tapasin binding; and the category of clustering criterion in DNA binding. In the GO Cellular Components, the Ensembl controversies with clustering criterion and multiple symbols were enriched in the classes of protein-DNA complex, nucleosome and chromatin; and the the category of topology criterion in M-band.

Our inhouse method for deprivation detected also depletion of metabolic regulation in the gene set of the clustering criterion. The most important enrichments that the MetaCore™ underesti-

mated in comparison to our randomization method were positive regulation of cytokine secretion (clustering criterion), and the enrichment of the genuine intergenic transcript fusions behind Akiva et al. (2006) and Parra et al. (2006) as a merged gene set to mitochondrion and oxidoreductive activity.

The most established, mutual binding interaction network at the protein level for the intergenic splicing transcripts included in both Akiva et al. (2006) [1] and Parra et al. (2006) [2] is displayed in [Fig. 2](#). The most remarkable cluster around the BMI-1 stem cell regulator are inhibitions of transcription, and the most enriched relations include also the blood clotting cluster, which are typically proteolytic cleavages. The GRB2 adaptor network was the most important hub in the physical interactome, connected, for example, to cytoskeletal and histone-regulating proteins.

Next, all five controversial gene sets were studied as a whole, including also the more divergent Ensembl deposit of disagreeing gene symbols and objects unique to only one gene set. The most statistically significant hubs controlling other genes with controversial borders were the transcription factors (causal edges; total number/upstream incoming edges): HNF4-alpha (170/8), E2F1 (13/2), HIF1A (10/0), SMAD4 (8/0), TCF4 (7/0), RUNX2 (6/2), HNF3-beta (5/2), LRH1 (5/0), ESR2 (4/0), CREM regulators, and members of the HOXA family. The most central receptor by far was the proto-oncogene tyrosine-protein kinase receptor RET (5/0), which is relevant to enteric nervous system maturation and neural crest cell migration. The most central secreted peptides were interferon-alpha (3/3), carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1; 3/0) and WNT isoforms. Other protein hubs were histone 2B (10/9), histone 2A (8/8), Cdc42 (9/0), kallikrein 3 (PSA; 8/5), GRB2 (8/0), and BMI-1 (8/0). The cardinal divergence hub among these vulnerable genes was the transcription factor hepatocyte nuclear factor 4-alpha/HNF4A, which regulates blood coagulation and lipid metabolism through transcription from RNA polymerase II promoter. The most enriched processes against the interactome around the HNF4A network were chromatin assembly ($p = 1.901e-05$), nucleosome assembly ($p = 1.948e-05$), and digestion ($p = 1.065e-04$). The main convergent hub was UDP-glucuronosyltransferase UGT, which is significant in heme, bilirubin, retinol and steroid metabolism and response to toxins.

In their endeavor to predict human genes not yet in the gene catalogs *in silico* using methods that do not require cDNA, Siepel et al. (2007) [4] identified 734 “novel gene fragments” (NGFs) with at least one novel exon, containing altogether 2188 exons. Out of these, 563 were distinct genes, 327 were completely novel clusters, 160 were previously unknown proteins, and hundreds of the rest were analogous significant extensions of transcription induced chimeras. These NGFs seemed to be expressed at low levels and in a tissue-specific manner, but their comparison to the other gene sets is not straightforward, as they were reported as novel accession codes and not in regard to conventional gene symbols and boundaries. In addition, a large proportion of the genomic coordinates of these NGF clusters did not have extensive (or any) overlap with known genes. They were enriched in annotations of motor activity, cell adhesion, connective tissue, and central nervous system development.

3.2. Standardization of the open reading frames

Since standard ORF (open reading frame) is essential to genes by definition, the existence of these untypical and unannotated exons in low abundance (in addition to the fused introns) is an indication towards the pathological nature of transcript fusion. We therefore examined whether any of the peculiar gene categories reported recently in the exhaustive open reading frame con-

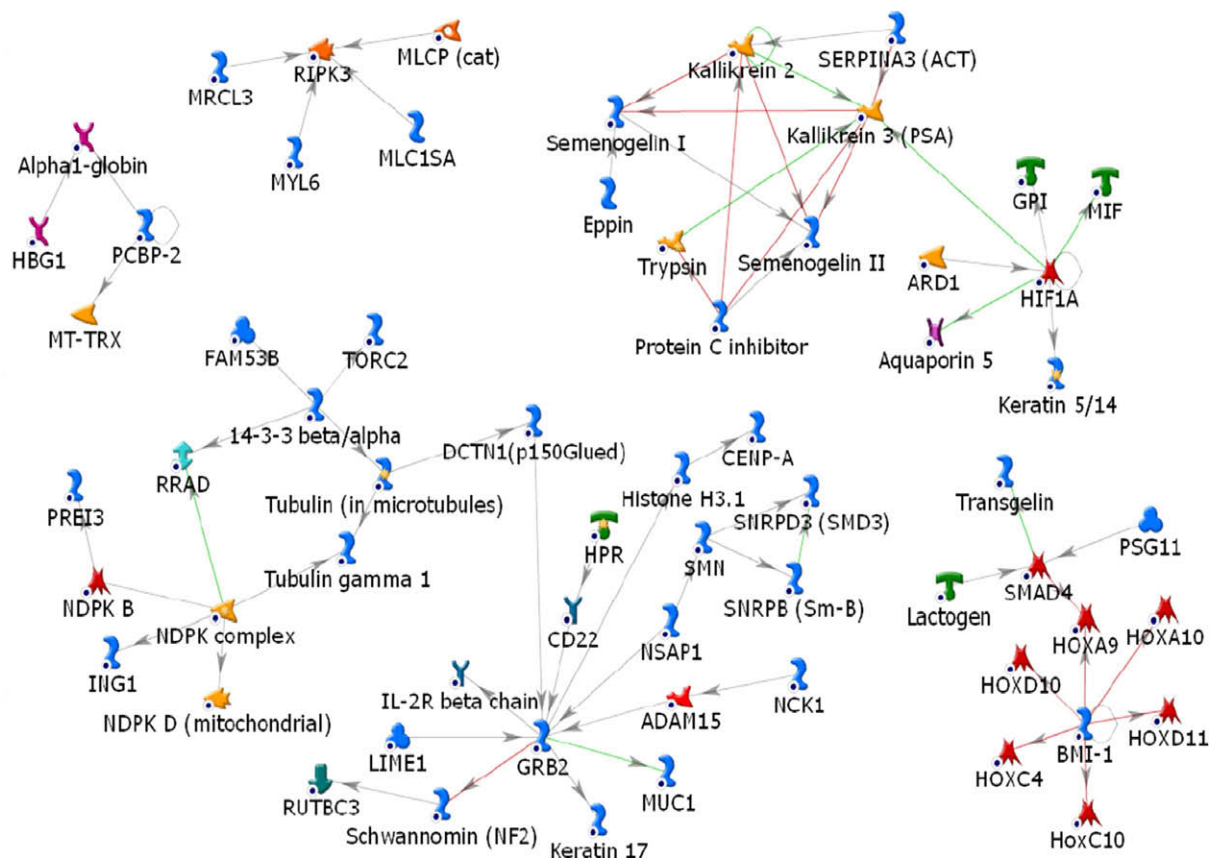


Fig. 2. Mutual interaction network for the putative intergenic splicing transcripts included in the pioneering discoveries by Akiva et al. [1] and Parra et al. [2]. If the observation reveals a novel pathological mechanism, these putative protein complexes are most likely to be affected. In particular, the housekeeping or homeotic HOX-transfactors centered around BMI-1 warrant further attention in developmental biology. Only the direct interactions are displayed. Blue arrows indicate activation and red arrows inhibition. As defined by the GeneGo Inc MetaCore/MetaDrug systems medicine platform, the symbols are as follows: enzyme; kinase; protease; protein; generic binding protein; transfactor; GTPase; G-protein adaptor; receptor; receptor ligand; transporter; GPCR; channel.

servation (RFC) comparative genomics project [18] significantly overlapped with the first fragmentary gene sets of the fused proteins. The RCF database divides human genes into the main categories of ortholog, cross species paralog, human-specific paralog, functional pseudogene, functional transposon, pfam structural family supported, special regions, redundant, pseudogene, transposon, artifact, and 1177 orphan genes (Ensembl 35) that do not share RFC even with primates and display less amino acid sequence identity to other mammals. Assuming that RFC and its loss as such reflect functionality and dysfunction, respectively, Clamp et al. predicted, indeed, that the 1177 human Ensembl genes are not real genes. Out of 22,219 human genes, a total of 14,257 genes displayed 100% RFC with dog (88.5% AA id) and 14,157 genes 100% RFC with mouse (85.2% AA id). In Supplementary Fig. S8, we show that when RFC drops below 100%, the amino acid identity also dramatically drops down when qualitatively ascending from 100% RFC. Alternative splicing, in general, seems to be less conserved than the amino acid sequence identity if a mammal is compared to a representative of other phyla, since splicing is more rare in simpler organisms [19]. In human, only about 6% of human genes are made from a single, linear piece of exon and the maintenance of genetic border regions seems to be important also in general in the light of comparative genomics.

3.3. Inconsistent termination of the polymerase reaction

Akiva et al. [1] observed that there was an intergenic distance bias in fused compared with nonfused genes in favor of short

intergenic regions. In regard to the RNA polymerase, transcript fusion could be likened to a train that fails to stop at a station with too short a platform. Deneud et al. [3] reported what they called RACEfrags, tissue- or cell-line-specific transcribed fragments 5' distal to the annotated 5' terminus, which fused the open reading frames (ORFs) of the adjacent transcripts in 20% of the cases. The authors used 5' rapid amplification of cDNA ends (RACE) that allows detection of low-copy-number transcripts/isoforms to argue that at least at this low abundance level, intergenic splicing is much greater than previously anticipated. However, only 132 genes including many synonym names in the conventional gene conventions were cloned in the study, so the cohort does not contribute much to parallel gene set enrichment analysis.

As for the phylogenetic occurrence, genes of similar function tend not only to be maintained in close proximity but also tend to be present or absent together. In comparative genomics, genuine gene fusions are seen for subunits of protein complexes encoded by separate genes in other organisms, which for example is the foundation for the String database [20]. We emphasize that the transcript fusion phenomenon described above is in nearly all cases in altogether different category and not deduced from, or supported by, these functional tasks typically related to substrate channelling in metabolism seen in comparative genomics. Besides, Deneud et al. [3] reported that the analyzed novel exons were relatively poorly conserved as a whole across species, although some conservation of novel internal exons in mammals was noticed.

3.4. Pathological transcript pairs infiltrating gene accession systems

In the emerging era of transcriptomics meta-analysis, the cluster based approach of the Ensembl nomenclature is indispensable in terms of gene coverage when comparing genome-wide expression correlation lists. Based on the hypothesis of epigenetic histone code working on predestined expression programs, the short mRNA phase is an ideal time window to assess the biological role of a gene, instead of its concentration in the effectors' protein level. Transcript fusion, sharing of the same DNA sequence in a different reading frame, extensive antisense transcription, ligation of two separate mRNA molecules (trans-splicing), thousands of noncoding RNAs, as well as alternative splicing and alternative promoters, however, confound this approach [21].

The HGNC and Swissprot symbols of the transcription induced chimeras misinterpreted as mere splice isoforms of the same gene are fatal, but unfortunately do exist, as summarized in [Supplementary Table S2](#). The usage of overlapping Gene IDs (the ENSG...-convention) in these sixty transcripts (the ENST...-convention) has led to autological classification of the extended "genes" as Ensembl family members (the ENSF...-convention) in many cases. In the protein family category, the claim has been more easily further propagated. The annotations, as indicated by the link to the UniProt database, can refer to either up or downstream transcripts. It is, indeed, difficult to track down how the error has been propagated further in the databases and converter servers.

Reassignments lead to displacements of misleading gene IDs to new and exclusive codes. Since the pioneering chimera reports, all 145 fused human gene accessions or 72 pairs challenged here were still in use in the Ensembl release 43 (February 28th 2007), as shown in [Supplementary Table S2](#). Fortunately, Ensembl regularly associates more than one gene symbols to clusters as a link to indicate that more than one biological gene may have been subsumed in a particular cluster. Also, corresponding GeneView pages, the manual Ensembl Vega project, as well as the Ensembl data-mining tool BioMart contain more specific information for genes for a casual end-user. The Ensembl project responded to finally tackle the problem with a new strategy in release 47 (23rd October 2007) that contained a full re-build of the human gene set. In this rebuild, the clustering strategy was also altered so that transcript predictions are only clustered into genes when their coding sequences overlap in the genome, in contrast to only overlapping on the transcript level. This change in strategy improved the situation, although in the current release 50, at least 82 genes or 41 of the transcript pairs identified in either Akiva et al. [1] or Parra et al. [2] are still merged under same gene ID.

4. Discussion

Although over 60% of the bulk DNA in mammalian genomes appears to be transcribed [5], the mRNA half-life is the least understood stage in the way too simplistic idiom of the central dogma ("DNA makes RNA makes protein"). In a corresponding manner to human SNP polymorphism and copy number variation (CNV), sequencing efforts of ESTs and full-length cDNA are only beginning to be extensive enough to quantitatively comprehend transcriptomics. Do the extended fusion transcripts, then, bear a functional significance, or do they just accidentally appear upon RNA processing? Do they merely contribute to variation, or is transcript fusion a novel pathological mechanism?

The drop in sequence identity in comparative genomics along with the drop of the conservation of open reading frame (Fig. 3) indicates that the borders of the genes are critically error-prone. Are some RNA polymerase isoforms more prone to miss termination? It is fascinating that there are three codons that stop peptide

synthesis, since some of the transcript fusions are also seen at the protein level. Based on this "splicing wobble", especially at promoter and terminator regions, we suggest that untypical splicing patterns in general, transcript extensions in particular, constitute a novel disease mechanism. The pioneering findings of intergenic splicing have been discussed in the context of a surprising but functional level of gene regulation. The ability of these genes to undergo gene fusion does not seem to be omnipresent across neither the population nor the human body, and seems to be more or less random. Intergenic splicing has been discussed in the context of deriving protein diversity for functional variation that brings selection benefits after reverse transcription and retrotransposition in the genome [1,3], in the same vein as the role of alternative splicing is emphasized in the mammalian genome. In contrast, dysregulation of the transcription termination with pathological outcome has rarely been even proposed. Ruan et al. [22], however, have detected such transcribed retrotransposed loci in cancer genomics. In addition, Kim et al. [23] introduced a server based on a database of 2344 hybrid gene pairs in the human genome, including intergenic splicing as well as trans-splicing and sense/antisense transcription, and emphasized the role of the truncated proteins as candidate risk factors [24].

When the annotations of gene set associated with open chromatin regions were compared to the verified chimeras against the public Gene Ontology (GO) [3], Denoeud et al. did not find any obvious association. Unfortunately, the public annotation databases such as GO do not contain hierarchic or indexed clinical designations for genes. Here, by using the GeneGo Inc's MetaCore/MetaDrug systems medicine platform whose annotation class members are more interwoven in the literature as gene circuits or constituents of the same protein complexes, we show that such enrichments do exist in the disease hierarchy. Extended by the findings of the inconsistent or unconventional gene topologies at the Ensembl genome assembly project and by the various unlikely gene entities reported by the Eric Lander group [18], these gene sets are only loosely overlapping and non-redundant. In the level of annotation, we noticed that the chimeras are, nevertheless, enriched to some biomedical classes. Intergenic splicing seems to be important especially in the etiology of neurological disorders, thalassemia and pregnancy or genital disorders. The statistically most significant processes exposed to the phenomenon of intergenic splicing were related to cytoskeletal dynamics regulating cell division. The statistically most significant indication of teleonomic function in the molecular biology was not derived from the genuine intergenic splicing but from the Ensembl controversy categories of clustering criterion and multiple symbols, both of which were related to chromatin remodelling and histone organization. The gene sets with inconsistent topology were enriched in the classes of GTPase regulation. Discrepant symbols are most often encountered in the disciplines of MHC complex and antigen presentation and processing. Regarding the randomization results as such, the GeneGo method following Hypergeometric distribution performed as a reliable method for gene set enrichment analysis without intensity values or ranking position attached to genes.

Interestingly, the alternative splicing program has been shown to differ particularly in the nervous system tissues compared to other mammalian tissues [25]. Pathways that detect and selectively eliminate defective messenger RNAs are known as surveillance pathways and operate either in a more general way in many cell types or specifically in only certain cell types. Several surveillance pathways have been identified that recognize specific types of mutations in RNA. For instance, one pathway recognizes nonsense mutations that result in an RNA that results in a truncated protein. Duchenne's muscular dystrophy and cystic fibrosis are examples of hereditary diseases that result from nonsense mutations. A recent and notorious finding regards the red blood

cell surveillance problem that causes thalassemia, the most common genetic form of anemia worldwide [26]. Noteworthy, hemoglobin alpha-1, hemoglobin gamma-1, hemoglobin gamma-2, and glucose-6-phosphate isomerase that exhibit transcript fusion are among the vulnerable agents that expose a patient to thalassemia. Final proof for RNA polymerase dysregulation mediated diseases would be to quantitate the RNA aggregates from pathological organs. Denoué et al. also noted the importance of specific organs, especially testis, stomach, kidney and brain, as the stage for the fusion. When the tissue specificity of the pioneering datasets by Akiva et al. [1] and Parra et al. [2] were studied here in the Applied Biosystems based GeneGo platform covering 31 human tissue types, no evident correlation was seen in the tissues.

The Ensembl project has finally reacted and tackled previous challenges with a new strategy so that transcript predictions are only clustered into genes when their coding sequences overlap on the genome, in contrast to just overlapping on the transcript level. This change in strategy has improved the situation considerably, and other databases should react to the dilemma that has led to systematic misinterpretation of HTS data. Likely, the situation is much worse in organisms other than humans. The complete gene catalog even for humans is a moving target since the classic definitions of gene are on trial. Nevertheless, it is critically important that the broad coverage reference system is still developed to better serve the community. This becomes ever more important if transcript chimerism contributes to pathologies, particularly in the versatile splicing program of the nervous system.

In short, instead of teleonomic molecular function, the data indicates that intergenic splicing might stochastically cause pathologies, for which we suggest certain hotspots. We argue that transcript chimeras might constitute an underestimated pathological mechanism akin to the copy number variation due to ribosome crowding and loss of stoichiometric ratios in the subunits of complexes by the law of mass action. The emerging era of clinical transcriptomics would be good news since new disease etiologies offer new remedies. Whether the hotspot networks affected by interspersed chimeras such as the developmental, housekeeping or homeotic HOX-transfactors centered around BMI-1, the adaptor network centered around Grb2, the Hif1A-kallikrein system, or the hemoglobin network identified here could be targeted by interfering strand microRNAs in the future warrants further research. RNA-based therapy is a new avenue for biomedicine, in the advent of drugs based on RNA interference for both coding and noncoding strands. A prerequisite for the advancing RNA interference therapeutic treatment after identification of pathological splicing chimera hotspots is the individual verification of the chimera with corresponding symptoms in the tissue in a patient compared to a healthy control tissue with the absence of symptoms and chimerism from the same individual. The phenomenon is also a new parameter in the organ and stem cell transplantations and *de novo* molten exons are likely to be detected immunologically to cause havoc.

5. Conclusion

In this study, we (i) postulate that the fused transcript chimeras constitute a novel and neglected pathological mechanism affecting especially chromatin remodelling and cytoskeleton, instead of a teleonomical function; (ii) identify hotspot networks affected by intergenic splicing comprising homeotic HOX-transfactors centered around BMI-1, Grb2 adaptor network, kallikrein system, and thalassemia RNA surveillance; and (iii) argue against a systematic error in the chimerical gene accession nomenclature. The latter has led transcriptomics meta-analysis of HTS arrays astray in the case of at least 145 human genes, some of which are extremely ac-

tively studied, and much more in other organisms. Finally, we (iv) argue that the emerging era of clinical transcriptomics is of significant benefit, as new disease etiologies also offer new remedies. That is to say, the advancing RNA interference technology may provide a cure after the pathological splicing chimera hotspots have been individually verified in a patient with corresponding symptoms in the tissue with affected chimerism, compared to healthy control tissue with the absence of symptoms and chimerism from the same individual. Our proposal as a solution for the current gene annotation dilemma is (i) Not to use transcript stage as a basis for gene nomenclature at all. (ii) To insert caution remarks for the gene entries which are still named on the basis of intergenic chimeras seen in the transcript stage. We recommend, however, that the genome browsers such as Ensembl continue the maintenance of the depot of the controversial gene contigs, since these exceptional and occasional findings may be, indeed, pathologically informative. At the moment, such an entry has been discontinued.

Acknowledgments

We are grateful to the EBI Ensembl project personnel for quick and cordial technical support and specifically acknowledge Michael Schuster for compiling lists of human genes with controversial architecture. We also acknowledge CSC Center for Science Ltd, administered by the Ministry of Education in Finland, for providing supercomputer facilities for systems biology we feel gratitude towards Mikko Frilander for ex cathedra definitions of splicing subtypes. We want to thank the Academy of Finland for funding.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jbi.2008.11.002](https://doi.org/10.1016/j.jbi.2008.11.002).

References

- [1] Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, et al. Transcription-mediated gene fusion in the human genome. *Genome Res* 2006;16:30–6.
- [2] Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, et al. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* 2006;16:37–44.
- [3] Denoéud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res* 2007;17:746–59.
- [4] Siepel A, Diekhans M, Brejová B, Langton L, Stevens M, Comstock CL, et al. Targeted discovery of novel human exons by comparative genomics. *Genome Res* 2007;17:1763–73.
- [5] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. *Science* 2005;309:1559–63.
- [6] Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002;296:916–9.
- [7] Li H, Wang J, Mor G, Sklar J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* 2008;321:1357–61.
- [8] Uniprot database. Available from: <http://www.ebi.uniprot.org>.
- [9] Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, et al. Ensembl 2007. *Nucleic Acids Res* 2007; 35:D610–617. Ensembl database is Available from: <http://www.ensembl.org>.
- [10] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004; 306: 636–40.
- [11] Alibés A, Yankilevich P, Cañada A, Díaz-Uriarte R. IDconverter and IDlight: Conversion and annotation of gene and protein IDs. *BMC Bioinformatics* 2007; 8:9 link. ID-converter. Available from: <http://idconverter.bioinfo.cnio.es/>.
- [12] Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* 1996; 266:114–28. Sequence Retrieval System. Available from: <http://srs.im.ac.cn/>.
- [13] Toronen P. Selection of informative clusters from hierarchical cluster tree with gene classes. *BMC Bioinformatics* 2004;5:32.
- [14] Kankainen M, Brader G, Törönen P, Palva ET, Holm L. Identifying functional gene sets from hierarchically clustered expression data: map of abiotic stress regulated genes in *Arabidopsis thaliana*. *Nucleic Acids Res* 2006;34:e124.

- [15] Marco A, Marín I. A general strategy to determine the congruence between a hierarchical and a non-hierarchical classification. *BMC Bioinformatics* 2007;8:442.
- [16] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* 2003;19:1699–706.
- [17] Nikolsky Y, Ekins S, Nikolskaya T, Bugrim A. A novel method for generation of signature networks as biomarkers from complex high throughput data. *Toxicol Lett* 2005; 158:20–9. MetaCore-MetaDrug database. Available from: www.genego.com.
- [18] Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci USA* 2007;104:19428–33.
- [19] Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, Jin L, et al. Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56, 419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res* 2006;34:3917–28.
- [20] String database. Available from: http://dag.embl.de/newstring.cgi/show_input_page.pl.
- [21] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; 447:799–816.
- [22] Ruan Y, Ooi HS, Choo SW, Chiu KP, Zhao XD, Srinivasan KG, et al. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* 2007;17:828–38.
- [23] Kim DS, Huh JW, Kim HS. HYBRIDdb: a database of hybrid genes in the human genome. *BMC Genomics* 2007;8:128.
- [24] Hybrid and chimera databases. Available from: <http://www.primate.or.kr/hybriddb/> and <http://genome.ewha.ac.kr/ChimerDB/>.
- [25] Fagnani M, Barash Y, Ip JY, Misquitta C, Pan Q, Saltzman AL, et al. Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol* 2007;8:R108.
- [26] Kong J, Liebhaber SA. A cell type-restricted mRNA surveillance pathway triggered by ribosome extension into the 3' untranslated region. *Nat Struct Mol Biol* 2007;14:670–6.