

Ann E. Loraine, Gregg A. Helt, Melissa Cline and Michael A. Siani-Rose
Affymetrix, Inc. 6550 Vallejo Street, Emeryville, CA 94608 USA

Abstract

Understanding the functional significance of alternative splicing and other mechanisms that generate RNA transcript diversity is an important challenge facing modern-day molecular biology. Using homology-based, protein sequence analysis methods, it should be possible to investigate how transcript diversity impacts protein structure and function.

To test this, a data mining technique ("DiffHit") was developed to identify cases where transcript diversity changes protein function. This method identifies protein isoforms produced by the same gene that contain different numbers or types of conserved amino acid motifs. Using this, we found that out of a test set of over 1,300 alternatively spliced genes with solved genomic structure, over 30% exhibited a differential profile of conserved InterPro and/or Blocks protein motifs across distinct isoforms. These results suggest that motif databases such as BLOCKS and InterPro are useful tools for investigating how alternative transcript structure affects gene function.

Gene and Transcript Classification

Gene assignments.

17,811 high-quality transcript alignments were grouped into genes and transcript groups by comparing their genomic alignments.

Regions of continuous alignment were used to infer exons, while gaps in the cDNA partner of the alignment were used to delimit introns.

Transcripts sharing at least 50 bp of in-frame, continuous coding sequence were grouped in genes. Thus, all transcripts from the same gene group encode variant isoforms of the same protein.

Splice group assignments.

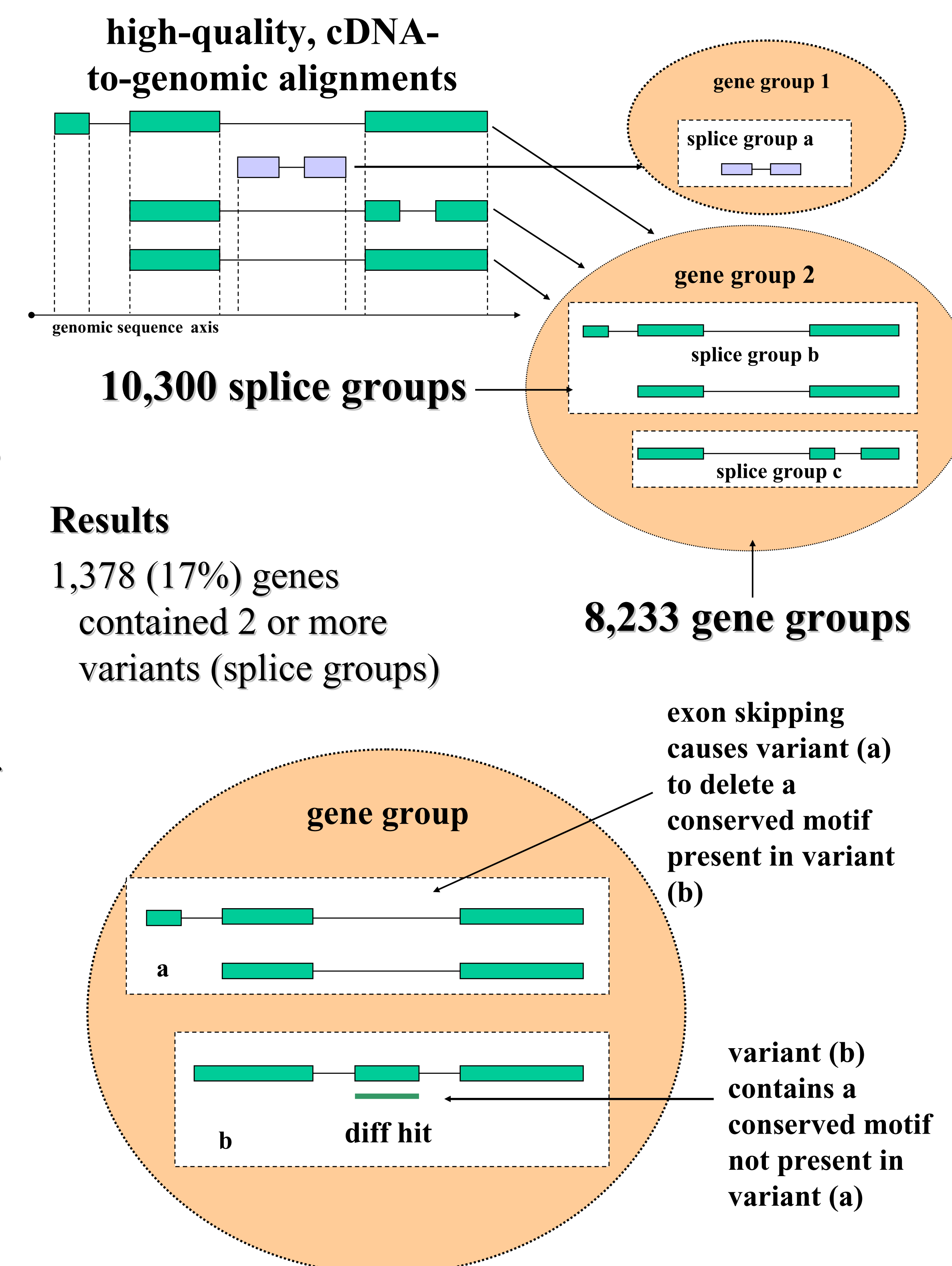
Transcripts with compatible alignments were assigned to the same splice group using the following simple rule: if an inferred intron from one transcript alignment overlapped an inferred exon from another, then the two transcripts were assigned to different splice groups.

Protein Classification - BLOCKS & InterPro

Proteins from each gene were searched against the BLOCKS and InterPro amino acid profile databases to detect conserved, functional motifs in each protein.

Diff Hits

Conserved motifs across splice groups were compared. When different splice groups from the same gene contained different profiles of conserved motifs, these were counted as "diff hits." Together, BLOCKS and InterPro detected "diff hits" in 475 genes (35% of 1,378). These were genes where alternative transcript structure coincides with changes in conserved regions in the encoded proteins.



Method	genes recognized (total possible 8,223)	diff hits (total possible 1,378)
InterPro	6,017 (66%)	366 (27%)
BLOCKS	5,389 (73%)	340 (25%)

Protein Analysis Results

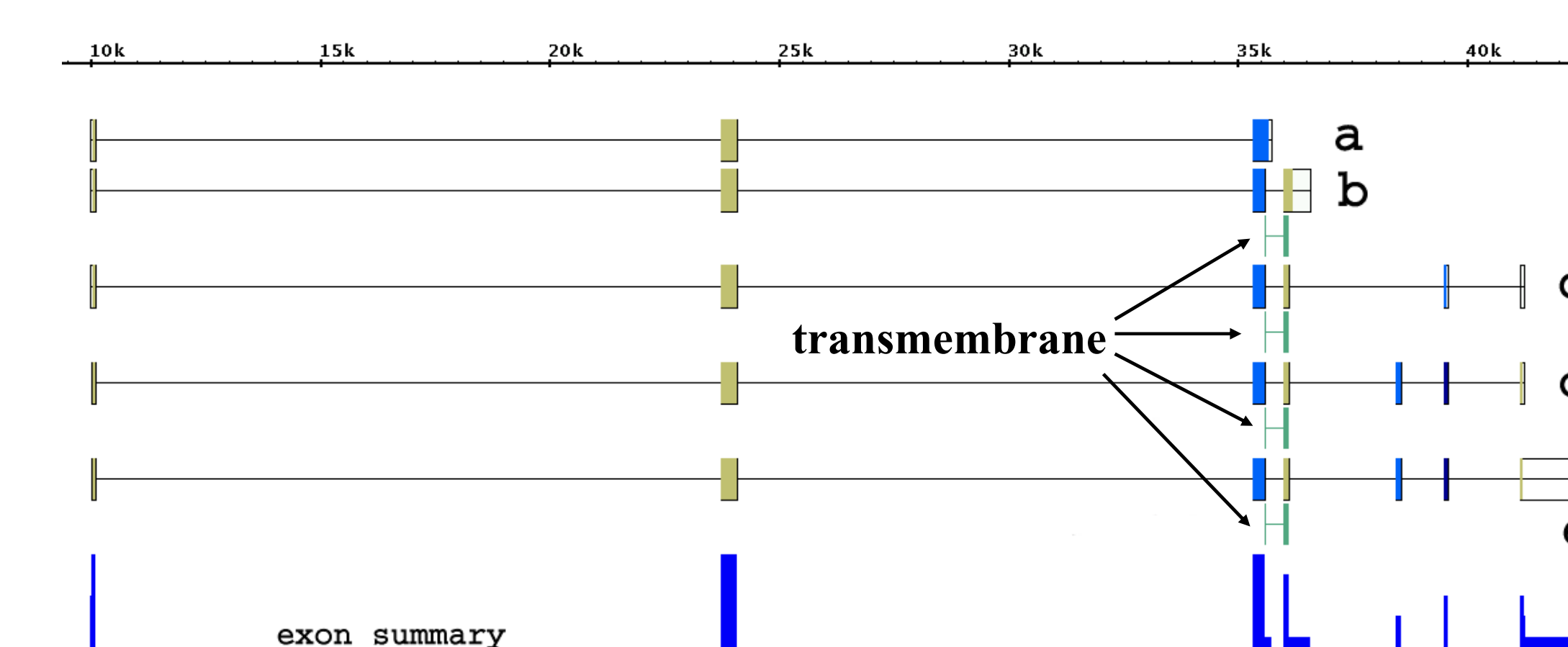
Table 1. InterPro homologies ranked by number of genes matched. (a) Top ten InterPro motifs affected by alternative transcript structure. (b) Ranks for these same motifs among alternatively spliced genes. (c) Ranks for these motifs among all genes, including genes with just one variant.

In general, the frequency of motif types detected as "diff hits" among alternatively spliced genes resembles the frequency of motif types found for all genes. The same pattern is true for the BLOCKS results, not shown here.

Diff Hits examples

A Java program ("ProtAnnot") was developed to display protein sequence annotations alongside gene structures. ProtAnnot colors coding region exons to indicate relative frame of translation, while the amino acid motifs these exons encode are shown below each transcript in green. When motif elements are split across an intron, glyphs representing these are linked by a thin, horizontal line.

CD84 antigen (immunoglobulin-associated beta) . An immune cell surface protein.

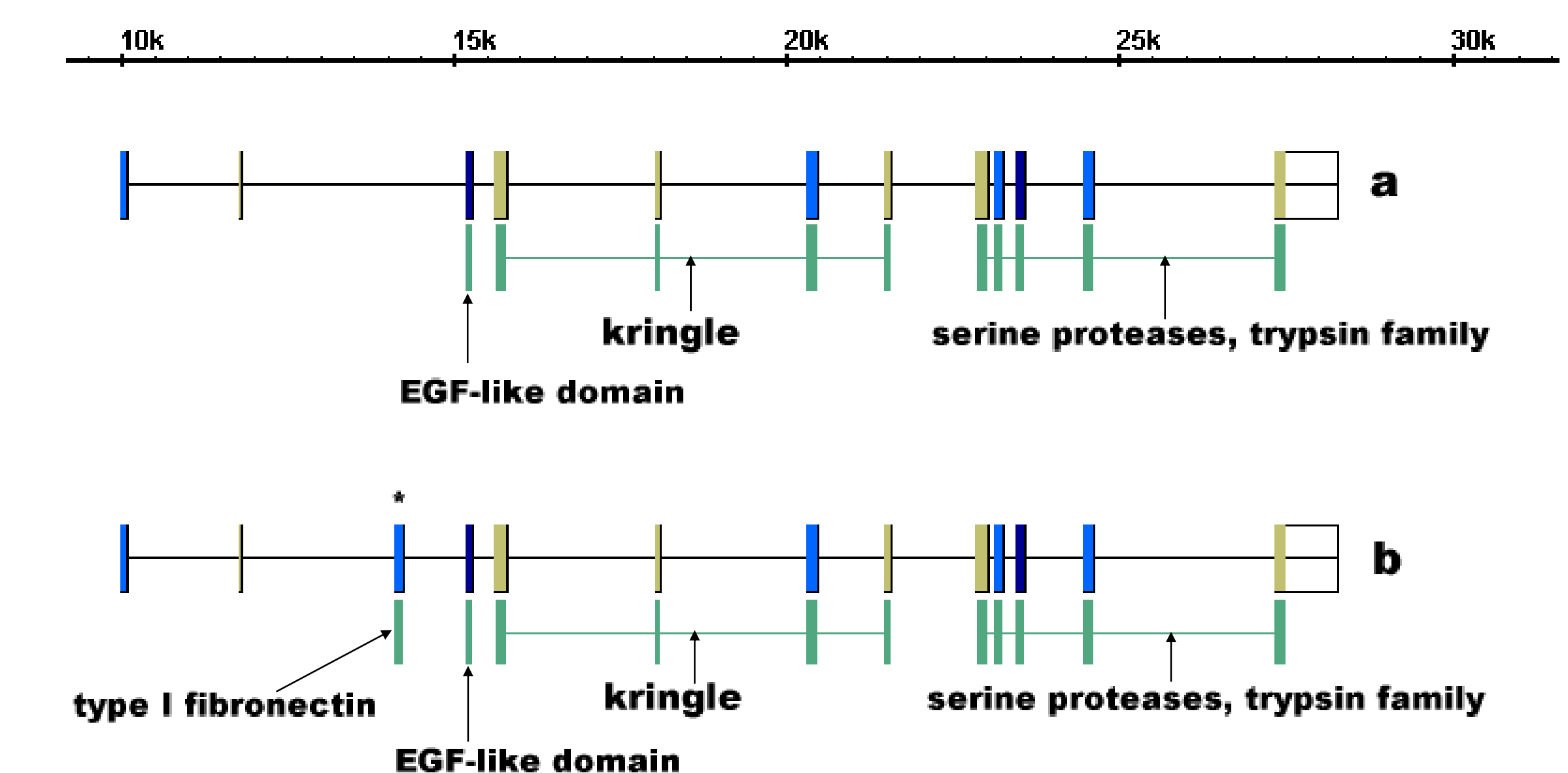


cDNA-to-genomic alignments for five CD84 variants are shown. Proteins encoded by variants (b-e) are recognized by Block IPB00031 (Transmembrane 4 family). One transcript (a) lacks this region because of its longer exon 3, the terminal exon, and may encode a secreted form of the CD84 protein. Alternatively, this form, if it is indeed a true variant and not a sequence database artifact, may localize to the membrane via some other mechanism, such as by association with the membrane-bound form.

Exons 2 and 3 for all 5 transcripts encode an immunoglobulin motif, not shown here.

Description	alt diffs rank (a)	alts rank (b)	all rank (c)
Proline-rich region	1	1	1
Ig_MHC complex domain	2	5	7
G-protein beta WD-40 repeats	3	9	10
Immunoglobulin subtype	4	7	8
Eukaryotic protein kinase	5	3	4
Zinc finger, C2H2 type	6	6	2
Immunoglobulin C-2 type	7	11	25
Serine/Threonine protein kinase	8	2	5
Tyrosine protein kinase	9	4	6
RNA-binding region RNP-1 (RNA recognition motif)	10	13	14

PLAT - tissue type plasminogen activator



The human PLAT locus encodes at least two distinct proteins. Both are forms of a tissue-type plasminogen activator, a secreted serine protease which activates a second protease (plasmin) involved in tissue remodeling and cell migration.

Both variants contain EGF-like, kringle, and serine protease motifs, but one form (b) contains an additional fibronectin motif not present in the other form. The other form (a) is expressed in melanotic melanoma; this suggests that (a) may play a role in the pathology of cancer cells, and that its lack of the fibronectin motif, named for the extracellular matrix protein fibronectin, may be involved.