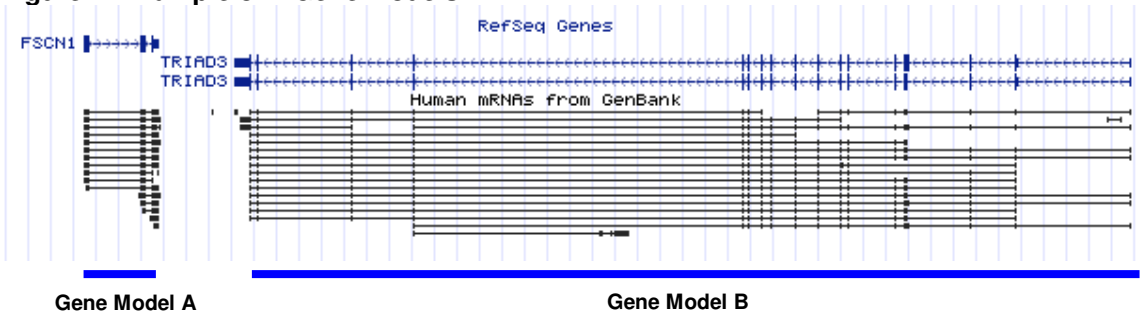
TECHNICAL NOTE

SwitchGear methods for gene model construction and transcription start site prediction

Gene Models

Generally speaking, a SwitchGear gene model is made up of one or more cDNAs that align to the same region of the human genome. Specifically, SwitchGear gene models are defined as clusters of cDNA alignments that have overlapping exons on the same strand (see Fig. 1). The SwitchGear gene model algorithm has processed over 250,000 human cDNA alignments from Genbank and Refseq databases to construct a genome-wide set of ~37,000 gene models. More than 22,000 of these gene models comprise 2 or more cDNAs, with the remaining ~17,000 gene models represented by a single cDNA.

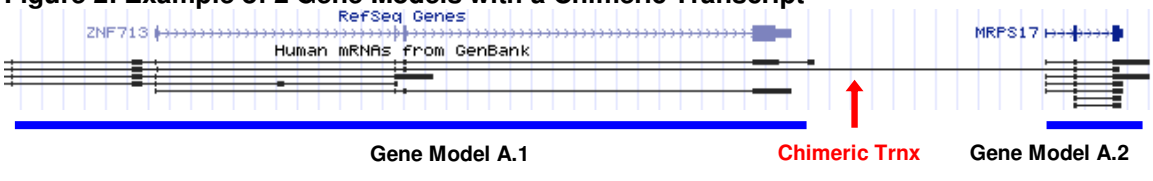
Figure 1: Example of 2 Gene Models



SwitchGear gene models employ a simple naming system for easy referencing, identifying each model by its chromosome number, strand, and unique identifier. For example, SwitchGear Gene Model ID "CHR7_P0362" indicates a cDNA cluster (0362) aligning to the plus strand (P) of chromosome 7 (CHR7). Existing gene annotation is mapped to the SwitchGear gene models through the NCBI annotation associated with Refseq accession numbers (accession numbers begin with "NM_"). See http://www.ncbi.nlm.nih.gov/RefSeq/ for more information.

In some cases, a chimeric transcript or alignment artifact may incorrectly combine two clusters based on their shared exon structure. The SwitchGear gene model algorithm identifies these aberrant cases and seeks to avoid combining two distinct gene models (see Fig. 2). Independent gene models that are broken out from likely chimeric transcripts/alignment artifacts are indicated in the model's unique identifier code by the addition of a "dot number" suffix on the end such as "CHR7_P0362.1" or "CHR7_P0362.2".

Figure 2: Example of 2 Gene Models with a Chimeric Transcript



Transcription Start Sites

Using the genome-wide set of gene models, the SwitchGear Transcription Start Site (TSS) prediction algorithm identifies the most likely sites of transcription initiation for each model. The algorithm employs a scoring metric to assign a confidence level to each TSS prediction based largely on existing transcript evidence. In addition to the ~250,000 human cDNAs listed in Genbank, more than 5 million additional 5' human cDNA sequence tags have been generated using a combination of approaches. While these short sequence reads do not reveal gene structure and, therefore, are not incorporated into the gene model construction, they provide a significant amount of experimental evidence for identifying transcript start sites. SwitchGear utilized a pool of 5' sequence tags gathered from the DBTSS (http://dbtss.hgc.jp/), the FANTOM project (http://fantom31p.gsc.riken.jp/), and the GIS-PET project (http://www.gis.a-star.edu.sg/), assigning each tag to the gene model with the cDNA closest to the 5' end of the tag.

For each gene model, the algorithm counts the number of TSSs (defined as the 5' end of a cDNA) within 200 bp of one another. The TSS score is based on the total number of TSSs identified within this window, with each TSS weighted according to several discriminating features: cDNA library source (RefSeq or full-length cDNA library), relative location within the gene model, and exon structure of the transcript (unique first exon or unspliced transcript). For each cluster of TSSs, the mean, standard deviation, and overall score is calculated and reported for the individual TSSs included in the gene model. Furthermore, the TSSs are ranked to identify the TSS representing the most likely transcription initiation site for a gene model. Rankings are indicated in the model's unique identifier code by the addition of a "dash R number" suffix (i.e. CHR7_P0362_R1 or CHR7_P0362_R2).

In the rare cases where the 5' most TSS for a gene model is derived exclusively from 5' tags, the model's unique identifier code is augmented with a "dash P number" suffix (i.e. CHR7_P0362_P4056). The additional code specifies the base pair distance the tag clusters lie from the 5' end of the gene model. In these cases, the tags represent possible 5' extensions of the gene model to which their linked, but no conclusive evidence exists for their connection to the gene model other than their relative proximity upstream of the 5' end of the gene model.

The distibution of genome-wide TSS scores is shown in Fig. 3.

Figure 3: Distribution of TSS scores in the human genome