TSS to probeset data and methods.

Martin Taylor, mst@ebi.ac.uk
May 2009.

TSS to probeset mapping

Files: humanTssProbesetMapping, mouseTssProbesetMapping

These files provide TSS to probeset mapping informatin. All mappings are either supported by **(a)** transcript annotation (a transcript that overlaps the probeset also has its annotated 5' end within 500 nucleotides of the TSS reference position (refPos)), or **(b)** they are supported by TSS defined by method (a) in the other species (mouse/human) who's reference position has been projected through genomic multiple sequence alignments into the target genome. The distance threshold for this orthology rescue strategy is +/- 50 nucleotides, this threshold is applied such that the distance between TSS reference position and projected reference position is <=50 nucleotides when measured in **both** mouse and human genomic coordinates. The 50 nucleotide threshold was chosen to match the CAGE tag clustering threshold used to define TSS reference positions.

Coordinates are given are zero based (standard for UCSC, add one to map into Ensembl)

ensID	Ensembl ID, for gene targeted by probeset. Being used as probeset ID.
ensAnID	Identifier of sequence feature used to link TSS to probeset, e.g. Ensembl gene, RefSeq transcript.
tssID	Unique identifier for TSS (h95_* and mMac_* are human and mouse macrophage derived respectively).
refPos	TSS reference position (modal tag position calculated summing tags per million over libraries used).
chrom	TSS maps to this chromosome (alternate haplotypes, mitochondrial and *_random fragments excluded)
strand	Chromosome strand
closest_transcript	Identifier for transcript with annotated 5' end closest to TSS. This is marked as rescue where TSS orthology projection between species has been used to **rescue** a TSS in the absence of same-species annotation linking TSS to probeset.
closest_dist	Distance between TSS reference position and annotated 5' end of transcript
transcripts	List of all transcripts that link TSS to probeset using the criteria defined in the methods section.
tag_num	Cage tags per million in macrophage libraries used.

TSS orthology projection

The tables described above give a pretty stringent mapping of TSS to microarray probesets but does not in any way guarantee that a human TSS and a mouse TSS defined for orthologous genes are in any way homologous themselves. For many downstream analyses this is fine and there need not be an assumption of one-to-one orthology for TSS. But for some analyses this is exactly what we need an this is what the orthologyProjection table provides, a carefully filtered set of one-to-one orthology relationships between mouse and human TSS. This table was constructed using the same method for TSS rescue as defined above but includes both rescued and directly assigned TSS. A by-

product of this assignment allows us to quantify “TSS-turnover” between species, e.g. where a TSS site in the mouse is used, but transcription in the human for the orthologous gene is driven by a non-orthologous TSS.

humanTssID	Unique identifier for human TSS
mouseTssID	Unique identifier for mouse TSS
humanDist	Distance (nt) between orthologous TSS measured in human
mouseDist	Distance (nt) between orthologous TSS measured in mouse
humanEnsID	Ensembl identifier for human gene/probeset that humanTssID is mapped to.
mouseEnsID	Ensembl identifer for mouse gene/probeset that mouseTssID is mapped to.
humanDirectProbesetMap	0 or 1. 1=Human TSS directly mapped to probeset by annotation. 0= Human TSS was rescue mapped using info from mouse.
mouseDirectProbesetMap	0 or 1. 1=Mouse TSS directly mapped to probeset by annotation. 0= Mouse TSS was rescue mapped using info from human.

Alignment files

These are standard Phylip sequential format. First line is two numbers, first is the number of sequences and the second is the total length of the alignment including gaps. Subsequent lines have a genus_species name followed by white space and then the complete aligned sequence for that species as a single line. The alignments are true multiple sequence alignments (not stacked BlastZ as has often been used) and alignment blocks are threaded relative to a reference sequence. The reference sequence is always the first sequence in the file and will be Homo_sapiens for human TSS and Mus_musculus for mouse TSS. These alignments contain 10,000 nucleotides of reference sequence and whatever that aligns with (see methods below)

Methods

{Note: There is some redundency between parts of this section and some of the slightly more verbose text above. This section is intended to provide the basis for a methods (supplemental?) section in the manuscript.}

TSS identification

CAGE tag sequences were generated from human HMDM, BMM cells prepared as described for microarray analysis. CAGE libraries were prepared from cells prior to LPS stimulation and the at the same time-points post stimulation as for the microarray data. CAGE tags from each library (representing a single cell type and time-point) were mapped to the reference human (NCBI36) and mouse (NCBIM37) genomic assemblies using a rescue mapping strategy (Faulkner et al. 2008).

TSS to microarray probeset mapping

Mouse and human microarray probes were mapped to the corresponding reference genome assemblies (human: NCBI36, mouse: NCBIM37) using Blat. We then identified every Ensembl or RefSeq transcript annotated (Ensembl version 53) to overlap the mapped probes. If there was a TSS reference position (as defined above) within 500 nucleotides of the annotated transcript 5' end it was considered to map to the probeset through transcript annotation. In the case that there were TSS

within the 1,000 nucleotide window (-500,+500), the TSS who's reference position was closest to the annotated transcripts 5' end was taken to be the TSS mapped to the probeset. For TSS-to-probeset mapping we also required that the TSS reference position be upstream of the 5' most microarray probe. This procedure led to at least one TSS being defined for 89% (2238/2505) of genes on the targeted human microarray and 92% (2297/2505) for the mouse (see Supplementary Notes for the distribution of number TSS identified per gene and the distribution of distances between transcript annotation and TSS reference position).

Through this fairly conservative approach of TSS-to-probeset mapping, a difference in annotation rather than a difference in biology, could lead to a TSS being mapped to a probeset in one species but not so for the orthologous TSS in the other species. To overcome this and to define conserved orthologous TSS, we projected the annotation of human probeset-mapped-TSS onto mouse and vice versa through genomic alignments (see below). Where there was a probset mapped TSS reference position within 50 nucleotides of a TSS position projected from the aligned species, they were considered an orthologous pair.

If no orthologous pair could be defined using probeset-mapped TSS then if any non-probest mapped TSS were within 50 nucleotides of the projected mapped TSS it was rescued and mapped to the TSS as guided by the projected mapped TSS. There were 35 human TSS rescued in this manner by mouse annotation and 83 mouse TSS rescued by human annotation {This needs to be explained better – perhaps a diagram as a supplemental figure}.

Genomic alignments and substitution rate estimates

The EPO9 whole genome alignment dataset from Ensmebl was used for coordinate transformations between species and to extract orthologous sequence alignments from the reference genomes of human (hg18, NCBI36), chimpanzee (panTro2, CHIMP2.1), orang utan (PPYG2), macaque (MMUL_1), mouse (mm9, NCBIM37) rat (RGSC3.4), dog (BROADD2), horse (EquCab2) and cow (Btau_4.0). Ensembl API (version 53) and custom perl code (available on request, MST) was used to perform alignment extractions, resolve overlapping alignment blocks, filter paralogous aligned sequence and for the transformation of coordinates between genomes. All alignments used are available as supplementary material (http://www.#).

Substitution rates were estimated using BASEML from PAML (version 4; Yang 2007) with the HKY model of nucleotide substitution (Hasegawa et al, 1985), categorical gamma (n=6) and assumed the phylogenetic topology: ((((human, chimpanzee) orang) macaque (mouse, rat)) ((cow, horse), dog))) or consistent subsets of it. Comparison of relative evolutionary rate was implemented as a likelihood ratio test where tree scaling parameters were either constrained to be identical between compared sequence categories (Mgene=0) or independent (Mgene=4). The test statistic was two times the difference in log-likelihood between constrained and unconstrained models compared to a a chi-squared distribution with one degree of freedom. False discovery rates were estimated using the R package Qvalue (version 1.1; Story and Tishirani 2003).