Solexa NlaIII Tag Tables

What are tag tables?

Next generation sequencing produces short sequence tags for identification of transcripts. The Solexa gene expression protocol (GEX) generates 17-mers, which retain their original strandedness. In order to get gene expression data from GEX tags, we need some way of identifying which tags came from what transcript.

Tag tables are the heart of the Solexa tag interpretation process. These comprise a set of fastas (for mapping against) and a set of database tables (for tag annotations). The tables represent all possible 17-mer sequence tags which a given genome can generate, given a restriction enzyme and a set of known/predicted genes and exons. The enzyme will determine the restriction site, which in turn determines the tags produced (17bp downstream, + revcomp of 17bp upstream if site is a palindrome). All unique tag sequences are classified, annotated, and converted into these tables. Aligning experimental tags to the tag tables vastly simplifies their annotation & interpretation.

Why do we need them?

Mapping GEX tags to a genome requires a set of tag tables, if the mapping is to be useful. Although the odds of generating a given 17-bp sequence, randomly speaking, are less than 1 in 17 billion (far larger than most known genomes), genomes are far from random sequences, and the search space for tags must be restricted.

Mapping 17-mer tags to the whole genome fails to produce useful measures of gene expression. Because of the unrestricted searchspace, tags may map to additional positions that could not have generated the tag, enzymatically speaking (i.e. there is no adjacent restriction site). This creates high percentages of repeat reads. Furthermore, basecalling errors or SNPs in the true tag sequence may produce 17-mers which map uniquely to other positions, thus ignoring the original site.

To correct this issue one must realize that all tags were generated with an anchoring enzyme, which can cleave RNA fragments only at specific sites. These may be CATG (if using NlaIII) or GATC (if using DpnII). Therefore each 17-mer must lie immediately 3' to a restriction site. Rather than add "CATG" (or "GATC") to each tag and then search them against the genome, it is faster and easier to annotate if one creates a database of all such 17-mer sequences adjacent to a restriction site. Since both NlaIII and DpnII cut at palindromes, each site is usually capable of generating 2 tags, one running 3' (genomic) on the (+) strand, and one running 5' (genomic) on the (-) strand. Exceptions are when the site lies < 17 bp from the end of the chromosome (or transcript), and cannot generate a 17-mer.

The difference in mapping outcomes is vast. Based on 3 vertebrate GEX runs (human, chicken, mouse, 7 lanes each), mapping against the open genome results in a near-random distribution of tags. Proportions of tags falling into exonic vs. intronic vs. intergenic space is

roughly the same as the genomic proportions of these spaces, as though the tags had been randomly scattered across the chromosomes. Repeat & unique mapping ratios are roughly 75% and 25%, respectively, but repeat hits are useless, so the usable data is down to 25% of the entire run. Furthermore, only 2/3 of these remaining tags actually associate with restriction sites, so 1/3 are biologically false!

Constraining tag mappings to the NlaIII tables results in more 1- and 2-mismatch mappings, as well as missing tags, but the unique mapping percent soars to over 80%, and repeats drop to less than 20%. These numbers are far more useful for gene expression analysis.

What tables are available?

Illumina provides two tables, for human and mouse genomes. The human is based on hs18 (NCBI36) and annotated by UCSC transcript mappings (04/07) and Unigene 201 (03/07). Mouse is older, based on mm8 (NCBI36) and annotated with UCSC mappings (11/06) and Unigene mm159 (also 11/06).

However, according to our correspondence with Illumina, no new tables or table updates are planned. Therefore we built our own pipeline for developing NlaIII tables, although we do not follow Illumina's classification rubric. Currently, the pipeline requires Ensembl or "Ensembl-ized" core genome databases, but flatfile-based genome data will be supported soon. We are using JGI's Aspergillus niger v1.0 to add .gff-based feature definition support to the pipeline. As of 03/18/08, we have NlaIII tables for the following organisms (with Ensembl versions):

- Human: 46, 49
- Mouse: 47, 48
- Chicken: 49

How are the tables constructed?

Illumina:

The pipeline, if one exists, appears to be proprietary. Rubrics for table construction and tag classification are found with the tables:

- Human README.txt
- Mouse README.txt

Illumina also uses a detailed tag classification schema, which informs the table creation process: Tag Classes PPT

Tags have 6 categories; read the READMEs for exact definitions, otherwise they can be loosely defined as follows:

1. Canonical: 3'-most exonic tags from refseq or other known mRNAs, + high-copy ESTs. May include portions of the poly-A tail.
2. Noncanonical: All other exonic tags.
3. Mitochondrial: Any mitochondrion-associated sequence, whether from chrM or genomic mitochondrial genes. **Takes precedence over canonical, noncanonical, and**

ribosomal.

4. Ribosomal: Tags from rRNA loci.
5. "Just Genome": Tags which do not map to an exon (i.e. intronic & intergenic tags.)
6. Repeats: Any tag found in over 100 locations genomewide.

Please note, these tables assume a complete digestion model, in that tags with internal CATGs are excluded.

In-House:

Our pipeline uses one 30kb Perl script and is run once per new database, a 2-5 hour process depending on network traffic, server load, and genome size. This creates several reports, a set of fastas for Eland (or anything else), and two mysql tables. Annotation of Eland results can then be done by query, which is much more efficient than the batch grep used with the Illumina tables.

Our rubric differs from Illumina's somewhat. We have 7 categories:

1. Canonical: Exonic tags from protein-coding transcripts, including those which span, or are generated from, known splice junctions. May include portions of the poly-A tail. (Ensembl's "blessed" gene predictions are included with known genes.)
2. Noncanonical: All other genic tags, including intronic tags and exon-intron / exon-intergenic overlaps. May also include portions of the poly-A tail, if the sequence also contains intronic bases (yes, some exons are that short!).
3. Mitochondrial: Tags from any mitochondrion-associated transcript, whether from chrM or genomic mitochondrial genes.
4. Ribosomal: Tags from an rRNA or tRNA locus.
5. ncRNA: Tags from any other noncoding RNA loci.
6. Intergenic: Tags which map entirely to intergenic space.
7. Repeats: Any tag found in 2+ locations, **unless** all locations are from the same gene (this is not a rare event). If from same gene, canonical assignments will take precedence. Tags occurring in the overlap of two or more genes are treated somewhat differently: if the tag is contained within one exon, it is attributed to that gene; if it is contained by more than one exon, or by none (including exon overhangs) then it is assigned to repeats, as there is no clear evidence for which gene(s) originated the tag.

Table precedence is: (i.e. each tag will be assigned to the lowest possible table number)

1. Repeats
2. Mitochondrial
3. Ribosomal
4. ncRNA
5. Canonical
6. Noncanonical
7. Intergenic

An incomplete digestion model is supported, as tags with internal CATGs are allowed. Tag classes are not currently taken into consideration, but a hexadecimal tag classification model is in the works. Illumina's tag classes seem designed for detailed transcriptome investigations,

like testing for alternative splicing or alternative polyadenylation. The utility of a given classification scheme will depend on what kinds of biological and technical considerations exist.

Upgrades to the tables are already planned. The current process does not take EST data into account, nor can it detect novel tags which arise from alternative splicing/polyadenylation or backwards transcription (the latter, apparently, is a real issue). Later, these may be added to the "noncanonical" table, or the tables may get restructured somewhat, with a first-pass set and a second-pass set (only for tags which didn't match the first time). Trans-splicing models will never be supported due to the resulting table sizes.

What if some experimental tags are not found in the tables?

If a tag is not found in the tables, it may be the result of unforeseen biology or technical error. Ideally, we want the tables to cover as many biologically genuine tags as possible while keeping the potential for repeats at a minimum, but there will always be leftovers. Unless the unmapped percentage is high, then confidence about unmapped tag quality should be correspondingly low. However, we can set up a BLAT or MEGABLAST for any unmapped tags, if this is important.

Biological explanations include:

1. Tag contains multiple SNPs (causing 3+ mismatches)
2. Tag contains indels (Eland doesn't support gapped alignments)
3. Tag is a product of trans-splicing or other very unusual transcriptional events (not supported by the table pipeline)
4. Tag does not actually come from the target organism (reporter construct, contamination, pathogens, commensals, whatever)

Procedural errors include:

1. Sample contamination (see #4 above)
2. 'Noise' or PCR artifacts in the library preparation (molecular bio can be fuzzy)
3. Quirks in the tag generation process that made the tag too short (i.e., part of the "tag" comes from the 3' sequencing adaptor)
4. Multiple basecalling errors (it happens)

ResearchWiki: ArielPaulson/TagTables (last edited 2008-05-07 16:29:00 byArielPaulson)