Perfect Arthropod Genes constructed with Gigabases of RNA

Don Gilbert, Biology Dept., Indiana University, Bloomington, IN 47405, gilbertd@indiana.edu

Gene construction, not prediction

The decade of gene prediction is over, gene construction from transcript sequence now surpasses predictions for biological validity. To paraphrase others: “...over half the gene predictions were imperfect, with missing exons, false exons, wrong intron ends, fused and fragmented genes”. Gene assembly from RNA has similar problems. This means using all the best data and tools, plus evidence quality tests, to build accurate genes.

Too much data or not enough?

Transcript assemblies can be more accurate than predictions, but effortful to resolve conflicts. RNA data quality sets limits, and nucleotide struggles at both ends of the data range. Sensible data reduction is a major gene construction task, where 10–9 RNA reads are assembled to 10% of competing transcripts, and those filtered with multiple criteria for the closest approach to 10^4.5 biological genes.

Suggested RNA methods:

- 1 Billion short reads, not 50 Million, may be enough.
- Mate paired with staggered inserts (200 – 600 bp), and strand specific is helpful.
- Long (454) > Short (illumina) together works well, both in insert paired ends.

Too much and too little RNA at a locus results in gene join and fragment exons (consensus of assembly tools), as in this data example and evidence summary for Daphnia magna RNA assemblies.

Genes without genomes?

Yes. E.g., Locust gene set is assembled without a genome. Orthology gene family score is higher for locusts than insects with genome-map genes (for Velvet assembly, lower for Trinity).

Alternates, paralogs and bad guesses can be resolved with a genome. Gene copy number variation is hard to resolve without genomic dna. Contaminants don’t map to a genome. E.g. perfect mouse genes I found in two sets of arthropod RNA-seq data. A best gene assembly uses gene structure signals from a geneome. RNA assembly without and without a genome is the best approach to defining perfect genes. Unless your genome is finished, there will be holes that RNA de-novo assembly will fill in.

Is that a honeybee gene in your wasp genome?

Species expressed genes often differ from those mapped from other species.

- Exon changes are common between even closely related species (Honeybee to Jewel wasp). Protein exons may map faithfully yet not be expressed, or rarely, in the related species.
- Big mistakes can be transferred by protein mapping when species are close enough for synteny to hold mistakes. Conversely a majority vote of mapped proteins from several species, including more distant ones, allows phylogeny/evolution to erase such computed mistakes and reveal the biological gene structures.

EvidentialGeneRecipe

http://arthropods.eugenes.org/EvidentialGene/

Gene construction software and methods continue to improve, but are imperfect. A current best strategy employed with EvidentialGene uses several gene modeling and assembly methods, extracting the best of their many results. This is consistent with recent results of others, pertaining to transcriptome assembly (3-4). Rough edges need smoothing: predictor models and transcript assemblies each have qualities the other lacks, for coding sequences and sequence signals, gene holes and mash-ups. Multiple lines of gene evidence can score the quality of competing gene constructions to select, if not perfect, gene set.

EvidentialGeneResults

EvidentialGenes are not perfect yet. But this approach appears to be working. Gene sets for Acyrthosiphon pisum aphid and Nasonia jewel wasp, built with EvidentialGene incorporating RNA assembly and gene prediction directed by evidence, compare superior on several evidence metrics to those of NCS Refseq, built with same available evidence.

Y B Purrfet?

How many gene comparison studies have significant artifacts of quality? In a recent review of gene orthology, "gene annotation emerged as the largest single influence affecting up to 30% of the discrepancies among orthology assessments" [1]. Gene function, derived from orthology or by experiment, is sensitive to imperfections. Differential expression measures are muddled on imperfect genes. Many biology studies that use genome-wide constructed genes hinge on the gene quality.

References: