Hybrid method of using RNA-Seq for genome annotation

Alexandre Souvorov
NCBI
souvorov@ncbi.nlm.nih.gov

Outline of the talk

• Genome-guided gene reconstruction
• Genes are ‘broken’ if parts of them are in an assembly gap
• De novo RNA-Seq assembly can provide missing transcript parts which are used for composite gene models
• Hybrid method annotates the genome and assembles on the fly the missing transcript parts
• Results for Baboon and Rabbit

Genome guided gene reconstruction

1) Align reads, remove artifacts, reduce redundancy
2) Determine exon/intron structures
3) Generate transcripts, select isoforms, determine gene names...

Longest CDS for each alternative

• Better chance to reconstruct the “main” isoform
• Minimal number of transcripts with premature stops
• No chimeric transcripts for overlapping genes
• Alignments of reads are easily mixed with longer alignments (protein/cDNA)
• > 200 annotated genomes

RNAseq and alternative variants

• Short reads don’t show which long variants are actually valid
• Number of variants is extremely high for genes with several variable areas (Drosophila Dscam gene > 38,000 variants)

Genome gaps break genes

• The scale of the problem for unfinished genomes is large - in baboon and rabbit > 1,300 genes have a CDS interrupted by an assembly gap
• Nothing could be done in the framework of the genome guided method
**Composite gene models**

- **Composite** gene models include parts not found in the genome
- Annotated >20 genomes
- Thousands of coding regions could be affected
- Composite genes need relatively short 100-300 bp patches

**Transcript assembly gap filling**

- Partial genes provide partial transcripts assemblies
- We need to find RNA-Seq which bridges the gap on the transcript
- This is a much smaller and easier problem than the genome gap filling

**Connect coding regions only**

- RNA-Seq is contaminated by not fully spliced transcripts
- Detecting splice cites is difficult in the transcriptome space
- We can avoid most not spliced introns if we limit our search to coding regions only - **no in-frame stops**

**De Bruijn subgraph for a broken gene**

- Binary search in sorted kmers+abundances
- The kmer A is the origin of the subgraph
- The maximal extension L limits the size
- The subgraph is small enough to explore all sequences and to find those which connect A and B

**Assembling a coding part of a transcript**

- Both kmers A and B are in the coding regions with known frames
- Valid connections will not have an in-frame stop - any excursion into a not spliced intron or UTRs is cut short
- No repeats
- SNP-type variations are resolved in favor of the higher expression

**Scanning all gaps**

- We can’t always determine which genes are partial and which pairs have to be connected
- We try all coding regions within a reasonable distance from a gap
Baboon

- 1319 genes have a CDS interrupted by an assembly gap
- Average patch size is 140 bp
- Long genes are more frequently affected

Rabbit

- 1347 genes have a CDS interrupted by an assembly gap
- Average patch size is 193 bp

Method limitations

- We need to detect coding regions on both sides of the gap
- We can’t reliably detect very short coding regions

Conclusion

- The hybrid method (genome guided annotation + assembly on the fly) fixes most of the genomic gap related problems with a very high accuracy
Acknowledgements

- Terence Murphy
- Mike DiCuccio
- Slava Chetvernin
- Tatiana Tatusova
- Richa Agarvala