MAKER: An easy to use genome annotation pipeline

Carson Holt
Yandell Lab
Department of Human Genetics
University of Utah
Introduction to Genome Annotation

- What annotations are
- Importance of genome annotations
- Effect of next generation sequencing technologies on the annotation process
What Are Annotations?

- Annotations are descriptions of features of the genome
  - Structural: exons, introns, UTRs, splice forms etc.
  - Functional: metabolism, hydrolase, expressed in the mitochondria, etc.
- Annotations should include evidence trail
  - Assists in quality control of genome annotations
- Examples of evidence supporting a structural annotation:
  - *Ab initio* gene predictions
  - ESTs
  - Protein homology
Why should I care about genome annotations?

Background

SUCCESS
Why should I care about genome annotations?

Incorrect annotations poison every experiment that uses them!!
Advances in Technology Promise to Make Whole Genome Sequencing “Routine” for Even Small Labs

Pacific Biosciences Preparing the 15-Minute Genome by 2013

BY KEVIN DAVIES
Feb. 12, 2008 | Marco Island, FL — Midway through this year’s “Advances in Genome Biology and Technology” conference, Pacific Biosciences sponsored a beachfront fireworks display to promote its name and celebrate its emergence from years in stealth mode. Perhaps the 600 or so attendees were intended to imagine the exploding multi-colored fireworks as a metaphor for the captured fluorescence at the heart of the company’s novel DNA sequencing technology.

But it turns out that Pacific Biosciences didn’t really need to burn money on pyrotechnics after all. The closing talk, by company founder and Chief Technology Officer Stephen Turner, was all the delegate could talk about.

“How cool was that?!” parroted Washington University’s Elaine Mardis, following Turner’s talk.

In the Light
Pacific was founded in 2004, but the technology dates back to Turner’s days as a grad student and post-doc at Cornell University. The SMRT (single molecule real time) system monitors the real-time procession of a DNA template as it interacts with a single DNA polymerase enzyme. Using four fluorescently tagged nucleotides, the system images each nucleotide as it is bound by the enzyme. The polymerase is tethered to the bottom of a zero mode waveguide (ZMW) — a sub-microscopic, 20-nanometer well that the company claims is “the world’s smallest detection volume.” All this happens at a speed of about 10 bases/second (in nature, the polymerase moves 50-70 times faster).

Using the ZMW concept that Turner and his former Cornell colleagues, physicists Harold Craighead and Walt Webb, published in Science in January 2006, the SMRT system was able to produce fragment sequences directly from native DNA — without the need for any enrichment or purification step. The result was a fixed number of long reads — approximately 300 bases or 1000 with sequencing error rates of one per 10,000 reads.

“SMRT is the technology for the future,” Turner said. “It’s the universal language for the next generation of sequencing.”

The technology’s potential applications are vast. The ability to sequence the human genome for $1000, for example, could enable the discovery of new genetic markers, and help researchers to better understand genetic diversity among individuals.

But the biggest impact of SMRT could be its ability to sequence the whole human genome in a matter of hours. Currently, the process can take weeks or months.

“With SMRT, we can sequence the genome in a single pass,” Turner said. “Our goal is to sequence the genome in 2013, in just 15 minutes.”

For the moment, however, the technology is still in its infancy. Turner said that the company is currently working on improving the accuracy of its reads, and is also working to reduce the cost of sequencing.

“We are aiming to make SMRT affordable for everyone,” Turner said. “We are working to make it as simple to use as a cell phone.”

Although the technology is still a work in progress, Turner said that the company is already receiving interest from major players in the biotechnology industry.

“We are seeing a lot of interest in SMRT,” Turner said. “We have already received investment from some of the biggest names in the industry.”

For Turner, the ultimate goal is to bring SMRT to the masses. “We want to make SMRT affordable for everyone,” he said. “We want to make it as simple to use as a cell phone.”

And with advances in technology like SMRT, that may not be far off.
Advances in annotation technology have not kept pace with genome sequencing, and annotation is rapidly becoming a major bottleneck affecting modern genomics research.

- As of October 2009, 222 eukaryotic genomes were fully sequenced yet unpublished.
- Currently there are over ~900 eukaryotic genome projects underway, assuming 10,000 genes per genome, that’s 9,000,000 new annotations.
- There is a limit to how much data can be managed, maintained, and updated by a single organization.
- Small research groups affected disproportionately by difficulties related to genome annotation.

• GOLD: Genomes OnLine Database. 2009.
• MAKER is an easy-to-use annotation pipeline designed to help smaller research groups convert the mountain of genomic data provided by next generation sequencing technologies into a usable resource.
MAKER Overview

• What does MAKER do?
• What sets MAKER apart from other tools (ab initio gene predictors, etc.)?
The easy-to-use annotation pipeline.

<table>
<thead>
<tr>
<th>User Requirements:</th>
<th>Can be run by a single individual with little bioinformatics experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Requirements:</td>
<td>Can run on laptop or desktop computers running Linux or Mac OS X</td>
</tr>
<tr>
<td>Program Output:</td>
<td>Output is compatible with popular GMOD annotation tools like Apollo and GBrowse</td>
</tr>
<tr>
<td>Availability:</td>
<td>Free open source application (for academic use)</td>
</tr>
</tbody>
</table>

**MAKER** identifies repeats, aligns ESTs and proteins to a genome, produces *ab-initio* gene predictions, automatically synthesizes these data into gene annotations, and produces evidence-based quality values for downstream annotation management.

Other Features
MPI Support

- Message Passing Interface (MPI) is a communication protocol for computer clusters which essentially allows multiple computers to act like a single powerful machine.
MPI Maker

MPI Throughput

Megabases/Hour

Nodes (CPUs)
What sets MAKER apart from other tool (i.e. ab initio gene predictors)?

Gene-prediction ≠ gene annotation
Model versus Emerging genomes

Model genomes:

- Classic experimental systems
- Much prior knowledge about genome
- Large community
- Big $$

Examples: *D. melanogaster, C. elegans*, human, etc
Model *versus* Emerging genomes

**Emerging genomes:**

- New experimental systems
  - Genome will be the central resource for work in these systems
- Little prior knowledge about genome
  - Usually no genetics
- Small communities
- Less $ 

Examples: flatworms, oomycetes, the cone snail, etc.
Comparison of gene models produced by state-of-the-art algorithms against a REFERENCE genome.

Table 1. MAKER’s performance on the C. elegans genome

<table>
<thead>
<tr>
<th>Performance category</th>
<th>Ab Initio</th>
<th>Evidence based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Snap</td>
<td>Augustus</td>
</tr>
<tr>
<td>Genomic overlap (gene)</td>
<td>SP</td>
<td>92.48%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>95.44%</td>
</tr>
<tr>
<td>Exon overlap</td>
<td>SP</td>
<td>18.88%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>87.63%</td>
</tr>
<tr>
<td>Exact transcript</td>
<td>SP</td>
<td>3.92%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>12.22%</td>
</tr>
<tr>
<td>Full exact transcript</td>
<td>SP</td>
<td>0.41%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>1.22%</td>
</tr>
<tr>
<td>Exact UTR5</td>
<td>SP</td>
<td>1.38%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>5.80%</td>
</tr>
<tr>
<td>Exact UTR3</td>
<td>SP</td>
<td>6.40%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>31.36%</td>
</tr>
<tr>
<td>Exact all exons</td>
<td>SP</td>
<td>19.02%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>93.48%</td>
</tr>
<tr>
<td>Start stop</td>
<td>SP</td>
<td>7.05%</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>35.95%</td>
</tr>
</tbody>
</table>

**SP**, specificity; **SN**, sensitivity. Genomic overlap is based upon all annotations; other categories are for complete, confirmed genes only. Overlap indicates that prediction overlaps reference annotation on the same strand; exact, coordinates of prediction are identical to reference annotation; full exact transcript, all exons match reference annotation coordinates, as do the start and stop codons. Gramene data are from ensembl.gff; Augustus ab initio results are for augustus_cat1v2.gff; Augustus evidence-based results are from augustus_cat3v1.gff. SNAP and MAKER data are from snap.gff, and makerZ_testset.gff, respectively. All data are from files available at http://www.wormbase.org/wormbase/NCAPS. WormBase release WB160 was used as the reference. Sensitivity and specificity were calculated using EVA (Keller and Brent 2003).

**MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.**

Genome Res 18(1) 188-196
With *enough* training data, *ab-initio* gene predictors can match or even out-perform annotation pipelines*

*ngASP - the nematode genome annotation assessment project*  Avril Coghlan, Tristan J Fiedler, Sheldon J McKay, Paul Flicek, Todd W Harris, Darin Blasiar, The ngASP Consortium and Lincoln D Stein  
Ab initio gene predictors don’t do nearly so well on emerging genomes*

*MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.
Benefits of MAKER

• Provides gene models as well as an evidence trail correlations for quality control and manual curation
• Provides a mechanism to train and retrain *ab initio* gene predictors for even better performance.
• Output can be loaded into a GMOD compatible database for annotation distribution
• Annotations can be automatically updated by new evidence by simply passing existing annotation sets back into the pipeline
What is Happening Inside MAKER

- RepeatMasking
- *Ab Initio* Gene Prediction
- EST and Protein Evidence Alignment
- Polishing Evidence Alignments
- Integrating Evidence to Synthesize Final Annotations
Annotating the Genome – Apollo View

Current evidence

Current Assembly
Identify and Mask Repetitive Elements

Current evidence

Current Assembly
Current evidence

Current Assembly

Identify and Mask Repetitive Elements

- RepeatMasker
  - RepBase
  - Species specific library
- RepeatRunner
  - MAKER internal protein library
Identify and Mask Repetitive Elements

Current evidence

Current Assembly
Generate *Ab Initio* Gene Predictions

Current evidence

*Ab initio* Predictions

Current Assembly
Generate *Ab Initio* Gene Predictions

- MAKER currently supports:
  - SNAP
  - Augustus
  - GeneMark
  - FGENESH

- Remember to supply HMM’s for each
Generate *Ab Initio* Gene Predictions

Current evidence

*Ab initio* Predictions

Current Assembly
Align EST and Protein Evidence

- Identify regions being actively transcribed (i.e. EST data)
- Identify region with homology to a known protein
Polish BLAST Alignments with Exonerate

Current evidence

Ab initio Predictions

Current Assembly

Polished EST

Polished protein
Polish BLAST Alignments with Exonerate

- All base pairs must align in order.
- No HSP overlap is permitted
- Aligns HSPs correctly with respect to splice sites.
Polish BLAST Alignments with Exonerate

Current evidence

Ab initio Predictions

Current Assembly
Pass Gene Finders Evidence-based ‘hints’

Current evidence

Ab initio Predictions

Hint-based SNAP
Hint-based FgenesH

Current Assembly
Identify Gene Model Most Consistent with Evidence*

*Quantitative Measures for the Management and Comparison of Annotated Genomes
Karen Eilbeck, Barry Moore, Carson Holt and Mark Yandell BMC Bioinformatics 2009
Revise it further if necessary; Create New Annotation

Current evidence

Ab initio Predictions

Current Assembly
Compute Support for Each Portion of Gene Model

Table 2. Maker quality index summary

<table>
<thead>
<tr>
<th>Maker quality index summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of the 5’ UTR</td>
</tr>
<tr>
<td>Fraction of splice sites confirmed by an EST alignment</td>
</tr>
<tr>
<td>Fraction of exons that overlap an EST alignment</td>
</tr>
<tr>
<td>Fraction of exons that overlap EST or Protein alignments</td>
</tr>
<tr>
<td>Fraction of splice sites confirmed by a SNAP prediction</td>
</tr>
<tr>
<td>Fraction of exons that overlap a SNAP prediction</td>
</tr>
<tr>
<td>Number of exons in the mRNA</td>
</tr>
<tr>
<td>Length of the 3’ UTR</td>
</tr>
<tr>
<td>Length of the protein sequence produced by the mRNA</td>
</tr>
</tbody>
</table>
Using MAKER
Welcome to the MAKER Web Annotation Service (MWAS)

To get started just click on "New Job" above. You can then submit a sequence for annotation or select from a list of pre-loaded example annotation jobs. Once a job has been added to the queue you can see your job's run status below. For more information on using the MAKER Web Annotation Service, click on "Help" above.

**Refresh Job Status**

Your Jobs (1)

<table>
<thead>
<tr>
<th>JobID</th>
<th>Description</th>
<th>Job Status</th>
<th>Start Time</th>
<th>Finish Time</th>
<th>Log</th>
<th>View Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>D. melanogaster : example contig</td>
<td>waiting in queue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MAKER Web Annotation Service
http://www.yandell-lab.org
De novo Annotation of a Newly Sequenced Genome

• You are involved in a genome project for an emerging model organism.
• You have no pre-existing gene models.
• What you do have:
  – ESTs
  – Proteins from other species available from public databases
Go to Web
GFF3 pass-through: How to use external evidence

- You have an existing annotation set.
- You want to update the evidence and allow the annotation to change to reflect the new evidence.
What if I have mRNA-seq data?
RNA-seq is fundamentally changing the field of genome annotation for both model and emerging genomes.
RNA-seq may soon make gene prediction (mostly) a thing of the past

- Still need to de-convolute reads & evidence (for now)
- Still need to archive and distribute annotations
- Still need to manage genome and its annotations
How to use RNA-seq data in MAKER

• Use BowTie and TopHat to produce, aligns reads into expression “islands” and “junctions”

• Pass data through as EST evidence via GFF3 pass-through.
Go to Web
Another issue: legacy annotations

- Many are no longer maintained by original creators
- In some cases more than one group has annotated the same genome, using very different procedures, even different assemblies
- The communities associated with those genomes are going to want mRNA-seq data
- Many investigators have their own genome-scale data and would like a private set of annotations that reflect these data
- There will be a need to revise, merge, evaluate, and verify legacy annotation sets in light of RNA-seq and other data
Merging and Revising Legacy Annotation Sets

Legacy Annotation Set 1

Legacy Annotation Set 2

Legacy Annotation Set n
Align Evidence and Legacy Annotations to Current Assembly

Current evidence

Legacy Annotations

Current Assembly
Pass Gene Finders Evidence-based ‘hints’

Current evidence

Legacy Annotations

Hint-based SNAP
Hint-based FgenesH

Current Assembly
Identify Gene Model Most Consistent with Evidence*

*Quantitative Measures for the Management and Comparison of Annotated Genomes*
Karen Eilbeck, Barry Moore, Carson Holt, and Mark Yandell BMC Bioinformatics 2009
Go to Web
Working with Chado

- maker2chado [OPTION] <database_name> <gff3file1> <gff3file2> ...
- maker2chado [OPTION] -d <datastore_index> <database_name>

This script takes MAKER produced GFF3 files and dumps them into a CHADO database. You must set the database up first according to CHADO installation instructions. CHADO provides its own methods for loading GFF3, but this script makes it easier for MAKER specific data. You can either provide the datastore index file produced by MAKER to the script or add the GFF3 files as command line arguments.
Working with JBrowse

- `maker2jbrowse [OPTION] <gff3file1> <gff3file2> ...`
- `maker2jbrowse [OPTION] -d <datastore_index>`

This script takes MAKER produced GFF3 files and dumps them into JBrowse for you using pre-configured JSON tracks.