Whole Eukaryotic Genomes, \textit{de novo}, in the CLC Genomics Workbench

Jeffrey Boore
CEO, Genome Project Solutions, Inc.
Adjunct Professor, UC Berkeley
Guest Researcher, Berkeley National Lab / DOE Joint Genome Institute

Outline

- A little about us at Genome Project Solutions
- Roles of the CLC Genomics Workbench
- Examples from our genome projects
- PHRINGE and whole genome presentations

What is Genome Project Solutions?

Non-profit independent research group

- Conduct genomics research, especially regarding the evolution of genes, gene families, and genomes
- Facilitate researchers’ interface with high-throughput genomics techniques
- Develop and implement software and web tools for analyzing and presenting genomes
- Educate scientists and the public about genome science and evolutionary genomics

Import in MANY formats

Import as
Std import ...
Tracks ...
Roche 454 ...
Illumina ...
Solid ...
Fasta ...
Sanger ...
Ion Torrent ...
BAM/SAM ...
Seqs in table ...

http://GenomeProjectSolutions.com
Viewing and Manipulation

- Many different, customizable views
- Sequences can be sorted by several criteria
- Sequences can be sorted into other sequence lists

Sequence Quality Evaluation
(excellent substitute for FASTQC)

Trimming and Reports

<table>
<thead>
<tr>
<th>Trim</th>
<th>Input reads</th>
<th>No Trim</th>
<th>Trimmed</th>
<th>Nothing left or Decoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trim on quality</td>
<td>44,294,024</td>
<td>29,138,805</td>
<td>5,112,422</td>
<td>41,829</td>
</tr>
<tr>
<td>Ambiguity trim</td>
<td>44,052,229</td>
<td>44,023,592</td>
<td>21,089</td>
<td>7,822</td>
</tr>
<tr>
<td>Filter on length</td>
<td>44,294,027</td>
<td>44,195,159</td>
<td>0</td>
<td>49,272</td>
</tr>
</tbody>
</table>

Genome Assembly

- Uses a de Bruijn graph method
- Has heuristics that make it very fast
  (400 MB genome with 50x Illumina coverage in about a day on a 6 quad-core Mac)
- Modest RAM requirements
  (using 96 GB RAM)
- Often adequate, but lacks many features found for optimization in alternatives such as the Celera Assembler or Trinity

Read Mapping
Example-Correcting 454 errors

Homopolymer corrections ~1:2000 nts

Annotating with GFFs / GTFs

- Can BLAST for exons and do similar manual mapping, but does not have any sophisticated gene modeling such as Maker or Glean
- Good tools for annotating with GFFs produced by other gene modelers and for viewing in browser format and for exporting the resulting annotations in many forms (e.g., export all tRNAs as fasta, etc.)

Annotating with GFFs / GTFs
MANY Export Features

- Downloadable into CLC Genomics Workbench
- Simple tool to create BLAST databases
- Similar interface to that above for BLAST searches to internal or public databases

http://GenomeProjectSolutions.com

Outline

- A little about us at Genome Project Solutions
- Roles of the CLC Genomics Workbench
- Examples from our genome projects
- PHRINGE and whole genome presentations

http://GenomeProjectSolutions.com
Outline

- A little about us at Genome Project Solutions
- Roles of the CLC Genomics Workbench
- Examples from our genome projects
- PHRINGE and whole genome presentations

PHRINGE

- Create thousands of clusters of homologous genes from many genomes
- Uses a graph-based method
- Based on calculations of evolutionary distance from full-length alignments
- Incorporates information from the phylogeny of the organisms

http://GenomeProjectSolutions.com

PHRINGE

- Create phylogenetic trees for each gene cluster
- Compare gene and species trees to assign orthology and paralogy

Each gene has a "Details Page"

Gene duplication in metaphytes

Each cluster has a "Compare Page"

"Synteny Viewer"
Acknowledgements

Developers at CLC Bio and organizers of this session

Paramvir Dehal and Bob Stiles - Database work on PHRINGE system

MANY collaborators on whole genome projects

http://GenomeProjectSolutions.com