Data Analysis with CASAVA v1.8 and the MiSeq Reporter

Eric Smith, PhD
Bioinformatics Scientist

September 15th, 2011
CASAVA v1.8: Integrated Secondary Analysis Package

- Bcl conversion
- De-multiplex
- Alignment
- Variant Calling
- RNA counting
- Visualization (Genome Studio)
- Tertiary Analysis

CASAVA v1.8 was released in June 2011.
Bcl Conversion with CASAVA v1.8

- CASAVA (not OLB) will perform Bcl conversion and de-multiplexing in a single step.
- To support **standard file formats**, Fastq will replace Qseq as the output format.
- The Fastq quality score encoding will use the **standard** offset value of 33 rather than the previous Illumina-specific offset value of 64.
# Project/Sample-based Sample Sheet

The Sample Sheet is provided by the user for each sequencing run.

<table>
<thead>
<tr>
<th>FCID</th>
<th>Lane</th>
<th>Sample</th>
<th>SampleRef</th>
<th>Index</th>
<th>Description</th>
<th>Control</th>
<th>Recipe</th>
<th>Operator</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC62DBU</td>
<td>1</td>
<td>NA12156_Index_1</td>
<td>Human</td>
<td>ATCACG</td>
<td>1:250method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>1</td>
<td>NA11992_Index_2</td>
<td>Human</td>
<td>CGATGT</td>
<td>1:250method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>1</td>
<td>NA11882_Index_4</td>
<td>Human</td>
<td>TGACCA</td>
<td>1:250method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>1</td>
<td>NA11881_Index_3</td>
<td>Human</td>
<td>TTAGGC</td>
<td>1:250method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>1</td>
<td>lane1</td>
<td>unknown</td>
<td>Undetermined</td>
<td>unknown barcode</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>Undetermined_indices</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>2</td>
<td>NA12156_Index_1</td>
<td>Human</td>
<td>ATCACG</td>
<td>1:500method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>2</td>
<td>NA11992_Index_2</td>
<td>Human</td>
<td>CGATGT</td>
<td>1:500method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>2</td>
<td>NA11882_Index_4</td>
<td>Human</td>
<td>TGACCA</td>
<td>1:500method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>2</td>
<td>NA11881_Index_3</td>
<td>Human</td>
<td>TTAGGC</td>
<td>1:500method1</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>2</td>
<td>lane2</td>
<td>unknown</td>
<td>Undetermined</td>
<td>unknown barcode</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>Undetermined_indices</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>3</td>
<td>NA12156_Index_1</td>
<td>Human</td>
<td>ATCACG</td>
<td>equal Volume</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>3</td>
<td>NA11992_Index_2</td>
<td>Human</td>
<td>CGATGT</td>
<td>equal Volume</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>3</td>
<td>NA11882_Index_4</td>
<td>Human</td>
<td>TGACCA</td>
<td>equal Volume</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>3</td>
<td>NA11881_Index_3</td>
<td>Human</td>
<td>TTAGGC</td>
<td>equal Volume</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>testProject1</td>
</tr>
<tr>
<td>FC62DBU</td>
<td>3</td>
<td>lane3</td>
<td>unknown</td>
<td>Undetermined</td>
<td>unknown barcode</td>
<td>N</td>
<td>test.xml</td>
<td>AT</td>
<td>Undetermined_indices</td>
</tr>
</tbody>
</table>
‘Project-based’ Run Folder
The Sample Sheet provides names of Project and Samples
‘Project-based’ Run Folder
The Sample Sheet provides names of Project and Samples

Run folder
  - Unaligned
  - Aligned
  - Build

- Bcl conversion and De-multiplexing
- Alignment
- Variant calling
‘Project-based’ Run Folder
The Sample Sheet provides names of Project and Samples

- **Run folder**
  - Unaligned
  - Aligned
  - Build

- **Project DNA**
  - Sample Rat
    - Rat_ATGACG_L002_R1.001.fastq.gz
    - Rat_ATGACG_L002_R1.002.fastq.gz
  - Undetermined indexes

- **Export and Summary files**
  - Build
  - Project DNA
  - Sample Rat

- **BAM and variant files**
Reference sequences no longer need to be ‘squashed’ – fasta format is supported.

The GERALD summary file will be modified in accordance with the new directory structure

The runtime of Alignment Resolver has been significantly reduced
Repeat resolution with Eland v2e
- First pass: align using a single seed
- Take reads that did not get matched or hit a repetitive sequence
- Second – fourth passes: Align using overlapping seeds
- Report seeds that hit a non-repetitive sequence
- **Increases sensitivity across repeat regions**
Sometimes only one read in a paired-end fragment (orphaned read) aligns to the genome, this tends to happen around repeat regions:

CASAVA 1.8 uses a sensitive alignment algorithm around the region where the missing read is expected:

This helps recover reads in highly repetitive regions.
Alignment Accuracy with ELANDv2e

Comparative assessment

- 25 million simulated read pairs, introduced error rate <2%, introduced SNPs and short indels

<table>
<thead>
<tr>
<th></th>
<th>bowtie</th>
<th>bwa</th>
<th>ELANDv2e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly mapped</td>
<td>88.1%</td>
<td>94.1%</td>
<td>94.1%</td>
</tr>
<tr>
<td>Incorrectly mapped</td>
<td>3.0%</td>
<td>5.0%</td>
<td>0.09%</td>
</tr>
</tbody>
</table>

Diagram: Correctly mapped pairs, incorrectly mapped pairs, unmapped pairs.
CASAVA v1.8 Variant Calling

- Uses a new probabilistic Bayesian method (like MAQ and GATK)
- Single-sample variant caller
- Calculates genotype probabilities under two prior distributions - genomic and polymorphic
- The output format are text-based genotype calls and the **standard file format** BAM.
CASAVA 1.8: Indel Detection with Local Realignment

1. Cluster Shadow and Partially Aligned Reads

2. Cluster Anomalous Reads

3. Merge Cluster from Same Event
CASAVA 1.8: Indel Detection with Local Realignment

4. De novo Assemble Clusters into Contigs

5. Align Contigs to Reference

6. Send Contig and Read Alignments to Genotyper for genotype calling
Sequenced human Yoruba trio: NA18507, NA18506, NA18508
- 95 Gb raw data each, 2x100 bp, 200G Chemistry

Violations of Mendelian inheritance indicate miscalled variants

Results for chromosome 20

<table>
<thead>
<tr>
<th></th>
<th>CASAVA 1.7</th>
<th>CASAVA 1.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sites (‘N’ excluded)</td>
<td>59,505,520</td>
<td>59,505,520</td>
</tr>
<tr>
<td>Mapped in all 3 of trio</td>
<td>58,188,449 (97.8%)</td>
<td>58,988,964 (99.1%)</td>
</tr>
<tr>
<td>Called in all 3 of trio</td>
<td>56,881,257 (95.6%)</td>
<td>57,696,195 (97.0%)</td>
</tr>
<tr>
<td>Mendelian conflicts</td>
<td>5,072</td>
<td>162</td>
</tr>
<tr>
<td>Conflict rate</td>
<td>0.0087% of called sites</td>
<td>0.00028% of called sites</td>
</tr>
</tbody>
</table>
Analysis Visual Controller (AVC)
Analysis Visual Controller (AVC) – a GUI for CASAVA

- AVC is a Windows GUI to run CASAVA (support for v1.8 coming soon).
- It’s designed to allow biologists to run their own analyses without learning Linux commands.
- AVC supports multi-cores Linux machines as well as SGE clusters, and offers archive and deletion functions to manage your data (and disk space!). No installation is required on the Linux side.
- AVC can be downloaded on iCom for free and is supported by Tech Support.

AVC version 1.7.1
RNA Workflows
# RNA Workflows: TopHat / Cufflinks or CASAVA to fit your application

<table>
<thead>
<tr>
<th>Feature</th>
<th>TopHat + Cufflinks</th>
<th>CASAVA v1.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting of exons, genes, splice junctions</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Differential gene/isoform expression</td>
<td>Yes</td>
<td>Possible</td>
</tr>
<tr>
<td>Novel isoform discovery</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Paired-End Reads</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Analysis without gene/exon annotations</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>SNP detection</td>
<td>No (Yes with pileup)</td>
<td>Yes</td>
</tr>
<tr>
<td>Short indel detection</td>
<td>No (Yes with pileup)</td>
<td>Yes</td>
</tr>
<tr>
<td>Technical Support</td>
<td>No (documentation provided by Illumina)</td>
<td>Yes</td>
</tr>
</tbody>
</table>
iGenomes – Reference Files for CASAVA and TopHat/Cufflinks

- Includes commonly analyzed reference sequences and annotation files.
- Files downloaded from NCBI, UCSC, and Ensembl.
- 25 species supported; available on iCom and ftp.
- Each iGenome includes reference sequences, Bowtie index (for Tophat/Cufflinks) and annotation files (for RNA workflows).
Align small RNA sequences using ELAND to:
  - known small RNA sequences (eg miRBase)
  - genome sequence

Built from previous release of ‘Flicker’
Adaptor sequences are trimmed
Output will be in SAM format

Downstream applications include:
  - Differential expression
  - De novo discovery
Use Software Tools That Meet Your Project’s Needs

- Illumina software tools are available for you to use.
- Many 3rd party tools are also very useful – use what suits your needs best!
- All Illumina software has full technical support at your disposal.
MiSeq Reporter
The resequencing workflow will have the option to only generate FASTQ files for users that want to use their own tools.
MiSeq Specifications

- Up to ~5 million clusters per run
- Single and paired-end runs, with support for indexing
- Data analysis takes <2 hours
- Seven reference genomes delivered; users can add new ones
- Data volume for 2 x 150bp run (yields ~1.5 Gbp):

  - .bcl files: 1.6GB
  - .fastq files: 1.5GB
  - .bam files: 1GB
  - Variant output files: 0.1GB

For more info see [http://www.illumina.com/help/miseq_reporter/default.htm](http://www.illumina.com/help/miseq_reporter/default.htm)
MiSeq Reporter: General Summary Report
MiSeq Reporter -- Resequencing

- Resequencing Workflow
  - Output files: BAM, VCF
  - Algorithms: based on CASAVA v1.8

Workflows

- Resequencing
- Amplicon
- Library QC
- De novo Assembly
- SmallRNA
- Metagenomics

Amplicon and Library QC workflows will have similar output
MiSeq Reporter – *De novo Assembly*

**Workflows**
- Resequencing
- Amplicon
- Library QC
- *De novo* Assembly
- SmallRNA
- Metagenomics

**De novo Assembly Workflow**
Output files: Fasta of contigs; optional dotplot
Algorithms: based on Velvet

![MiSeq Analysis Software](image)

- Syntenic dot-plot
- Contig metrics
MiSeq Reporter – Small RNA

**Workflows**

- Resequencing
- Amplicon
- Library QC
- *De novo* Assembly
- SmallRNA
- Metagenomics

**Small RNA Workflow**

Output files: Count text files of small RNAs by type

Algorithms: based on Flicker & CASAVA v1.8

Distribution of reads by type of small RNA

Histogram of most abundant small RNAs

Trimmed Read-Length Distribution (in Summary Report)
MiSeq Reporter -- Metagenomics

Workflows

- Resequencing
- Amplicon
- Library QC
- De novo Assembly
- SmallRNA
- Metagenomics

Distribution of reads by alignment to 16S rRNA database

Metagenomics Workflow
Output files: Annotation counts at different taxonomic levels
Algorithms: alignment to 16S rRNA
Questions?
DNA Alignment with CASAVA
If in a paired-end fragment, each read aligns to several locations in the genome:

The reads are expected to have a particular orientation:

And occur within a given fragment length:
Variant Calling with CASAVA
Other Bioinformatics Tools

- Support for **Tophat/Cufflinks** for RNA-Seq (see Illumina’s HowTo document)
- The Analysis Visual Controller (**AVC**) is a GUI for running CASAVA
- The **iGenomes** are a collection of references sequence and annotation files for commonly analyzed species. Includes files needed for CASAVA and Tophat/Cufflinks. Downloadable on iCom under ‘Downloads’.
- **Flicker** is small-RNA tool that works with CASAVA – coming soon.
- **GenomeStudio** is a tool that helps you visualize CASAVA output – now taking input in BAM format.
CASAVA Methods – RNA counting

Raw read counts are normalized by:

- **Gene/exon size** – to allow comparisons between genes/exons
- **Read coverage** – to allow comparisons between experiments to produce RPKM values

<table>
<thead>
<tr>
<th>Normalized Values</th>
<th>Counting Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes and Exons RPKM = 10^9 x Cb / NbL</td>
<td>• RPKM = Reads Per Kilobase of exon model per Million mapped reads</td>
</tr>
<tr>
<td></td>
<td>• Cb = the number of bases that fall on the feature</td>
</tr>
<tr>
<td></td>
<td>• Nb = total number of mapped bases in the experiment</td>
</tr>
<tr>
<td></td>
<td>• L = the length of the feature in base pairs</td>
</tr>
<tr>
<td>Splice Junctions RPKM = 10^9 x Cr / NrL</td>
<td>• Cr = the number of reads that cover the junction point</td>
</tr>
<tr>
<td></td>
<td>• Nr = total number of mapped reads in the experiment</td>
</tr>
<tr>
<td></td>
<td>• L = the length of the feature in base pairs</td>
</tr>
</tbody>
</table>
TruSeq Exome Analysis
TruSeq Exome Script Output

- Run ID: 101022_SN140_0234_A805N8ABXX_2
- Lane: 2
- Read length: 101
- Targeted Regions: TruSeq_exome
  - Pull-down region size (calculated on single probe regions): 340
- Total regions size: 62085295
- Mean coverage ESTIMATE (calculated from read counts): 102
- # of reads (PF + Aligned): 85899275
- # of reads in targeted regions: 62759149
- Enrichment in targeted regions: 73.06
- Enrichment in targeted regions +/-150 bp: 83.88
- # targeted regions: 201071
- # of targeted regions with no reads: 473
- % of targeted regions with no reads: 0.24

General quality metrics

Read counts by GC content

Coverage plots

Control coverage and snp frequencies
Visualization with GenomeStudio
GenomeStudio – Data Visualization, now supports BAM

- Powerful coupling of table-based data with visualization tools

- Modules for DNA, RNA, ChIP-Seq

- License-based product
Data Volume Management
Throughput Continues To Increase…

GB per Week

- 2008
- 2009
- 2010
- 2011

Throughput in GB per Week from 2008 to 2011.
Data Volumes: HiSeq2000 + v3 Chemistry + CASAVA 1.8 (600Gb output per run; Q2, 2011)

<table>
<thead>
<tr>
<th>Data Volume</th>
<th>File Format</th>
<th>Size</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Call / Quality Score Data</td>
<td>Bcl</td>
<td>660 GB</td>
<td></td>
</tr>
<tr>
<td>Read-level Data</td>
<td>compressed FASTQ</td>
<td>660 GB</td>
<td></td>
</tr>
<tr>
<td>Internal alignment output</td>
<td>export.txt</td>
<td>~5 TB</td>
<td>~5TB of temp space is needed per 600Gb HiSeq2000 run</td>
</tr>
<tr>
<td>Alignment Output &amp; Archiving</td>
<td>BAM</td>
<td>660 GB</td>
<td></td>
</tr>
</tbody>
</table>
MiSeq Reporter
Running A MiSeq

- User fills out Sample Sheet at desk
- Excel Add-In verifies that it is valid

- User scans bar-code on flow cell
- Sample Sheet is automatically loaded (if named after bar code)
- Flow cell/Reagents loaded
- Run Begins

- User reviews progress and results via HTTP interface
MiSeq Reporter Workflow Reports

Resequencing – coverage, q-scores, variants (outputs BAM, VCF)

De novo assembly – dotplot and metrics (outputs fasta of contigs)

Small RNA – trimmed reads distribution, reads by RNA type, most abundant small RNAs

Metagenomics – 16S rRNA alignment
MiSeq Reporter: Resequencing details report

- Coverage and error
- Q score
- SNP/indel + annotation
- Table of samples/variants

Scope of view

Zoom in/out

MiSeq Analysis Software

Current Jobs

Completed Jobs

Samples

Variants

<table>
<thead>
<tr>
<th>#</th>
<th>Sample</th>
<th>Genome</th>
<th>Chromosome</th>
<th>ClustersRaw</th>
<th>ClustersPF</th>
<th>Error</th>
<th>NoCall</th>
<th>Coverage</th>
<th>HetSNPs</th>
<th>HomSNPs</th>
<th>Insertions</th>
<th>Deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sample1</td>
<td>PhiX</td>
<td>phiX174</td>
<td>538210</td>
<td>517635</td>
<td>0.08</td>
<td>0.00</td>
<td>2363.00</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
MiSeq Reporter: De Novo details report
Trimmed read lengths
MiSeq Reporter: Small RNA details page

Distribution of small RNA reads by type

Top 10 most abundant small RNA
MiSeq Reporter: 16S metagenomics details report
Future Software Products…
Software Workflow Manager – Under Development

- Run diverse workflows from a single GUI application
- Mix and match software components within a workflow
- Enter custom workflows
- Possible LIMS integration