The Picard Pipeline

Tim Fennell
MPG Next Generation Sequencing Workshop
February 4th 2010
What’s Picard?

• The sequencing platform’s pipeline for processing and delivering illumina data
  – A run level pipeline
  – A sample aggregation pipeline
• A set of tools that anyone can use
• A sourceforge project
• A character in Star Trek: The Next Generation
High Level Pipeline

- Extract data from Illumina pipeline
- Align a subset of data with MAQ
- Calibrate Qualities (non-GATK)
- Align all PF reads
- Mark Duplicate Reads
- Recalibrate Qualities (GATK)
- SNP Calling, Metrics, Etc.
- Triage

- Secondary base calling happens during first step
- Adapter trimming/marking also happens in first step and is used during alignment
- Indexed runs are demultiplexed at first step; files created per index
- Likely moving to BWA in the near future
- Not yet using GATK unified genotyper; move is in the works
### Where to find data, tools, code

<table>
<thead>
<tr>
<th>What</th>
<th>Where</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline Outputs</td>
<td>/seq/picard/{flowcell}/...</td>
</tr>
<tr>
<td>Aggregation Outputs</td>
<td>/seq/picard_aggregation/{project}/{sample}/...</td>
</tr>
<tr>
<td>Picard Binaries</td>
<td>/seq/software/picard/current/bin</td>
</tr>
<tr>
<td>Metrics Documentation</td>
<td><a href="http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html">http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html</a></td>
</tr>
<tr>
<td>Source Code</td>
<td><a href="https://svn.broadinstitute.org/picard/trunk/">https://svn.broadinstitute.org/picard/trunk/</a></td>
</tr>
<tr>
<td></td>
<td><a href="https://picard.svn.sourceforge.net/svnroot/picard/trunk">https://picard.svn.sourceforge.net/svnroot/picard/trunk</a></td>
</tr>
</tbody>
</table>
Metrics Definitions

- Generated from comments in code
- Automatically updated and released as part of the Picard release process
### Standard Pipeline Outputs (e.g. WGS)

<table>
<thead>
<tr>
<th>Output</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAM Files (obviously!)</td>
<td>Aligned, duplicate marked BAM files w/all reads; sorted, indexed etc.</td>
</tr>
<tr>
<td>Internal Control Metrics</td>
<td>Error rates by IC sequence etc.</td>
</tr>
<tr>
<td>Quality Calibration Data</td>
<td>Calibration table used to calibrated qualities</td>
</tr>
<tr>
<td>Alignment Summary Metrics</td>
<td>Lots of high level alignment metrics</td>
</tr>
<tr>
<td>GC Bias Metrics</td>
<td>GC Bias metrics and plots</td>
</tr>
<tr>
<td>Quality by Cycle Plot</td>
<td>Plot of mean quality score by machine cycle</td>
</tr>
<tr>
<td>Quality Distribution Plot</td>
<td>Plot of distribution of quality scores in lane/file</td>
</tr>
<tr>
<td>Duplication Metrics</td>
<td>% Duplication, Estimated Library Size etc.</td>
</tr>
<tr>
<td>Insert Size Metrics &amp; Histogram Plot</td>
<td>Insert size information (PE only)</td>
</tr>
<tr>
<td>“Low pass” Concordance Metrics (where data available)</td>
<td>Concordance of sequencing information to known genotypes if available</td>
</tr>
</tbody>
</table>
# Outputs for Other Analysis Types

<table>
<thead>
<tr>
<th>Output</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid Selection Metrics</td>
<td>Lots of metrics to assess capture experiments</td>
</tr>
<tr>
<td>SNP Fingerprinting Metrics</td>
<td>Genotype matches at 24 SNP loci</td>
</tr>
<tr>
<td>Jumping Library Metrics</td>
<td>Metrics to assess jumping (long mate pair) library performance</td>
</tr>
<tr>
<td>QC SNP/Genotype Calls</td>
<td>Genotype calls made using the a very simplistic Bayesian SNP caller</td>
</tr>
<tr>
<td>dbSNP “Concordance” Metrics</td>
<td>Break down of how many called SNPs are in dbSNP vs. not</td>
</tr>
<tr>
<td>SNP Concordance Metrics</td>
<td>Concordance of SNP calls to known genotypes for the sample if available</td>
</tr>
</tbody>
</table>
Internal Control Metrics

• In lane control that is independent of genomic library
• Error rates by cycle based on four internal control sequences
• Sequences are “grown” in bacteria with adapters; no PCR in sample prep
• Matching algorithm is very simple and robust
• Outputs plots, summary metrics and detailed by IC by cycle error metrics

java -jar $PICARD/CollectInternalControlMetrics.jar I=428C0AAXX.4.unmapped.bam \ SUMMARY_METRICS=428C0AAXX.4.internal_control_summary_metrics \ PER_CYCLE_METRICS=428C0AAXX.4.internal_control_per_cycle_metrics \ CHART=428C0AAXX.4.internal_control_error_rates.pdf
Alignment Summary Metrics

- Lots of metrics about reads, alignments etc.
- Metrics are provided per read and by read-pair
- Outputs in Picard’s “MetricsFile” format that is very easy to parse
- Snapshot on right is of one lane dumped into excel and transposed

```
java -jar $PICARD/CollectAlignmentSummaryMetrics.jar I=428C0AAXX.4.aligned.duplicates_marked.bam \
O=428C0AAXX.4.alignment_summary_metrics \ 
R=/seq/references/Homo_sapiens_assembly18/v0/Homo_sapiens_assembly18.fasta
```
GC Bias Metrics

- Shows:
  - Distribution of GC in genome
  - Bias of aligned sequence by GC
  - Mean quality score by GC

- Looks at every overlapping 100bp window in the genome

- Calculated %GC of the window and how many reads start at the first base of that window

```
java -jar $PICARD/CollectGcBiasMetrics.jar I=429G5AAXX.3.aligned.bam \
O=429G5AAXX.3.gc_bias.detail_metrics CHART=429G5AAXX.3.gc_bias.pdf \
R=/seq/references/Homo_sapiens_assembly18/v0/Homo_sapiens_assembly18.fasta
```
Quality By Cycle Plot

- Simple plot of mean quality by cycle across all PF reads (including unaligned reads)
- Data for plot is also written out to a file

```
java -jar $PICARD/QualityScoreDistribution.jar \
I=429G5AAXX.3.aligned.duplicates_marked.bam \
O=429G5AAXX.3.quality_distribution_metrics CHART=429G5AAXX.3.quality_distribution.pdf
```
Quality Distribution Plot

- Simple histogram plot of quality scores
- Data for plot is also written out to a file

```
java -jar $PICARD/MeanQualityByCycle.jar \
  I=429G5AAXX.3.aligned.duplicates_marked.bam \
  O=429G5AAXX.3.quality_by_cycle_metrics CHART=429G5AAXX.3.quality_by_cycle.pdf
```
Duplication Metrics

- Calculated during the duplicate marking process
- Estimated library size – a coverage independent way to assess library complexity
- Histogram data that can be used to generate ROI curve

**UNPAIRED_READS_EXAMINED** | 930829
---|---
**READ_PAIRS_EXAMINED** | 13,715,861
**UNMAPPED_READS** | 8,390,609
**UNPAIRED_READ_DUPLICATES** | 185,281
**READ_PAIR_DUPLICATES** | 68,797
**PERCENT_DUPLICATION** | 1.14%
**ESTIMATED_LIBRARY_SIZE** | 1,362,670,191

<table>
<thead>
<tr>
<th>IN</th>
<th>UNIQUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1.989985</td>
</tr>
<tr>
<td>3</td>
<td>2.970055</td>
</tr>
<tr>
<td>4</td>
<td>3.940311</td>
</tr>
<tr>
<td>5</td>
<td>4.900849</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Library Type</th>
<th>Target Library Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Genome Shotgun</td>
<td>Billions (1-5bn is normal)</td>
</tr>
<tr>
<td>Whole Exome Hybrid Selection</td>
<td>Hundreds of Millions</td>
</tr>
<tr>
<td>Small Design Hybrid Selection</td>
<td>Tens of Millions (depends on target size)</td>
</tr>
</tbody>
</table>

```
java -jar $PICARD/MarkDuplicates.jar 
I=429G5AAXX.3.aligned.bam \ 
O=429G5AAXX.3.aligned.duplicates_marked.bam M=429G5AAXX.3.duplicate_metrics
```
Insert Size Histogram

- Simple insert size histogram
- Also summary metrics:
  - Mean
  - Stdev
  - “Width” to capture 10%, 20% etc. of the distribution

```
java -jar $PICARD/CollectInsertSizeMetrics.jar I=429G5AAXX.3.aligned.duplicates_marked.bam \
O=429G5AAXX.3.insert_size_metrics H=429G5AAXX.3.insert_size_histogram.pdf
```
Low Pass Concordance

- Take a large set of well-known genotypes (e.g. 250k, 1m array)
- Segregate known genotypes into hom ref, het, hom non-ref
- For each of the three categories count:
  - Reference bases observed
  - Non-reference bases observed
- By default uses only Q30+ mappings and Q20+ bases
- Robust even in fairly low quality sequence

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>REFERENCE</th>
<th>NON_REFERENCE</th>
<th>PCT_CONCORDANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOMOZYGOUS REF</td>
<td>331,192</td>
<td>428</td>
<td>99.87%</td>
</tr>
<tr>
<td>HETEROZYGOUS</td>
<td>108,866</td>
<td>107,519</td>
<td>99.38%</td>
</tr>
<tr>
<td>HOMOZYGOUS NON REF</td>
<td>261</td>
<td>160,221</td>
<td>99.84%</td>
</tr>
</tbody>
</table>
Hybrid Selection Metrics

- Lots of metrics to assess the efficiency of hybrid capture experiments
- Kris will be talking in more detail about this in a few minutes
Future Directions

• Integrate GATK Unified Genotyper in single-sample mode (and variant evaluation)
• Move all large-genome workflows to BWA
• Implementing a Methylation pipeline
• Integrate the GATK Indel cleaner (probably)
• Support for HG19/GRCh37
Questions?