

The Picard Pipeline

Tim Fennell

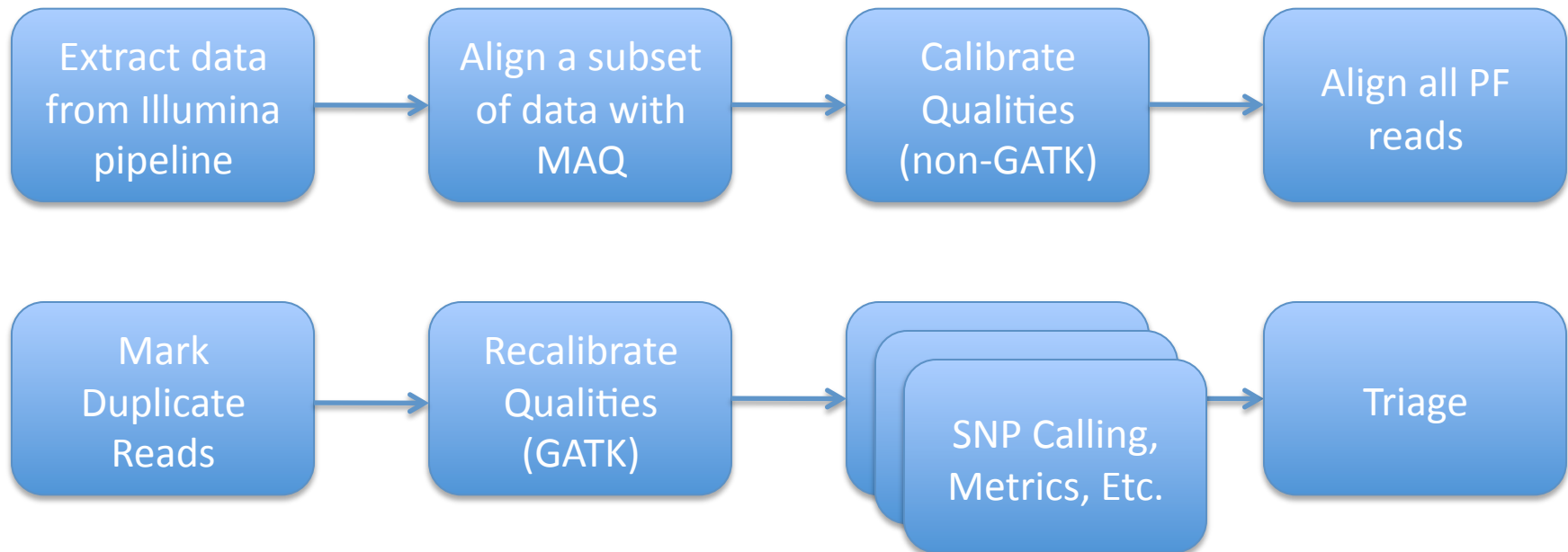
MPG Next Generation Sequencing Workshop

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



- 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

What	Where
Pipeline Outputs	/seq/picard/{flowcell}/...
Aggregation Outputs	/seq/picard_aggregation/{project}/{sample}/...
Picard Binaries	/seq/software/picard/current/bin
Metrics Documentation	http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html
Source Code	https://svn.broadinstitute.org/picard/trunk/
	https://picard.svn.sourceforge.net/svnroot/picard/trunk

Metrics Definitions

Picard Metrics Definitions

http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html

Picard Metrics Definitions

Table Of Contents

- [AlignmentSummaryMetrics](#): High level metrics about the alignment of reads within a SAM file, produced by the CollectAlignmentSummaryMetrics program and usually stored in a file with the extension ".alignment_summary_metrics".
- [DbSnpMatchMetrics](#): Metrics about how genotypes called by the pipeline match up to dbSNP, created by the CollectDbSnpMatches program and usually stored in a file with the extension ".dbsnp_matches".
- [DuplicationMetrics](#): Metrics that are calculated during the process of marking duplicates within a stream of SAMRecords.
- [ExtractIlluminaBarcodes.BarcodeMetric](#): Metrics produced by the ExtractIlluminaBarcodes program that is used to parse data in the Bustard directory and determine to which barcode each read should be assigned.
- [GcBiasDetailMetrics](#): Class that holds detailed metrics about reads that fall within windows of a certain GC bin on the reference genome.
- [GcBiasSummaryMetrics](#): High level metrics that capture how biased the coverage in a certain lane is.
- [GenotypeConcordanceMetrics](#): Statistics about how well a given set of input genotypes matches a set of well known reference genotypes.
- [HsMetrics](#): The set of metrics captured that are specific to a hybrid selection analysis.
- [InsertSizeMetrics](#): Metrics about the insert size distribution of a paired-end library, created by the CollectInsertSizeMetrics program and usually written to a file with the extension ".insert_size_metrics".
- [InternalControlCycleMetrics](#): Metrics about observations of an internal control sequence in an individual cycle.
- [InternalControlSummaryMetrics](#): Summary metrics about internal controls within a lane.
- [LowPassConcordanceMetrics](#): Concordance statistics for a set of low coverage reads against well known well known genotypes for the same sample for the purpose of ensuring that the sample being sequenced is the sample we think it is.
- [SamFileValidator.ValidationMetrics](#):

AlignmentSummaryMetrics

High level metrics about the alignment of reads within a SAM file, produced by the CollectAlignmentSummaryMetrics program and usually stored in a file with the extension ".alignment_summary_metrics".

Column Definitions

CATEGORY: One of either UNPAIRED (for a fragment run), FIRST_OF_PAIR when metrics are for only the first read in a paired run, SECOND_OF_PAIR when the metrics are for only the second read in a paired run or PAIR when the metrics are aggregated for both first and second reads in a pair.

TOTAL_READS: The total number of reads including all PF and non-PF reads. When CATEGORY equals PAIR this value will be 2x the number of clusters.

PF_READS: The number of PF reads where PF is defined as passing Illumina's filter.

PCT_PF_READS: The percentage of reads that are PF (PF_READS / TOTAL_READS)

PF_NOISE_READS: The number of PF reads that are marked as noise reads. A noise read is one which is composed entirely of A bases and/or N bases. These reads are marked as they are usually artifactual and are of no use in downstream analysis.

PF_READS_ALIGNED: The number of PF reads that were aligned to the reference sequence. This includes reads that aligned with low quality (i.e. their

- Generated from comments in code
- Automatically updated and released as part of the Picard release process

Standard Pipeline Outputs (e.g. WGS)

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

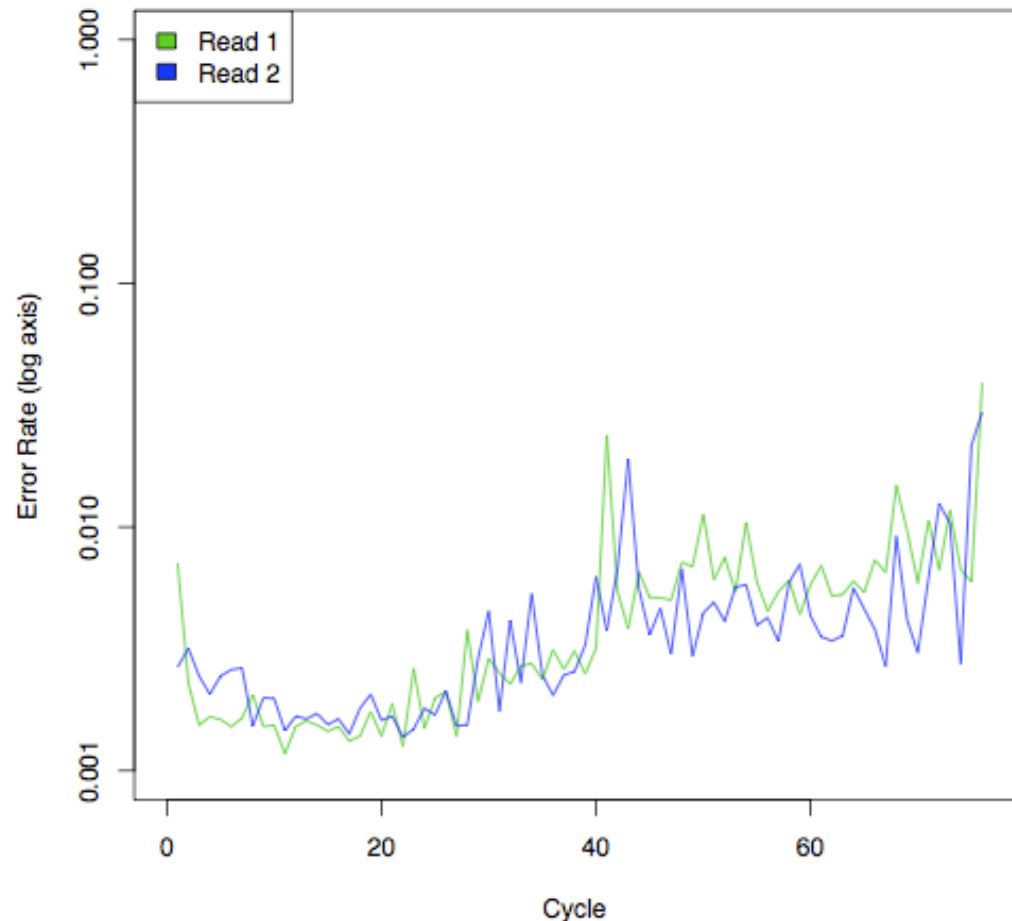
Outputs for Other Analysis Types

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

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

428C0AAXX.4.unmapped.bam Total (n=46848) IC Error Rate by Cycle



```
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

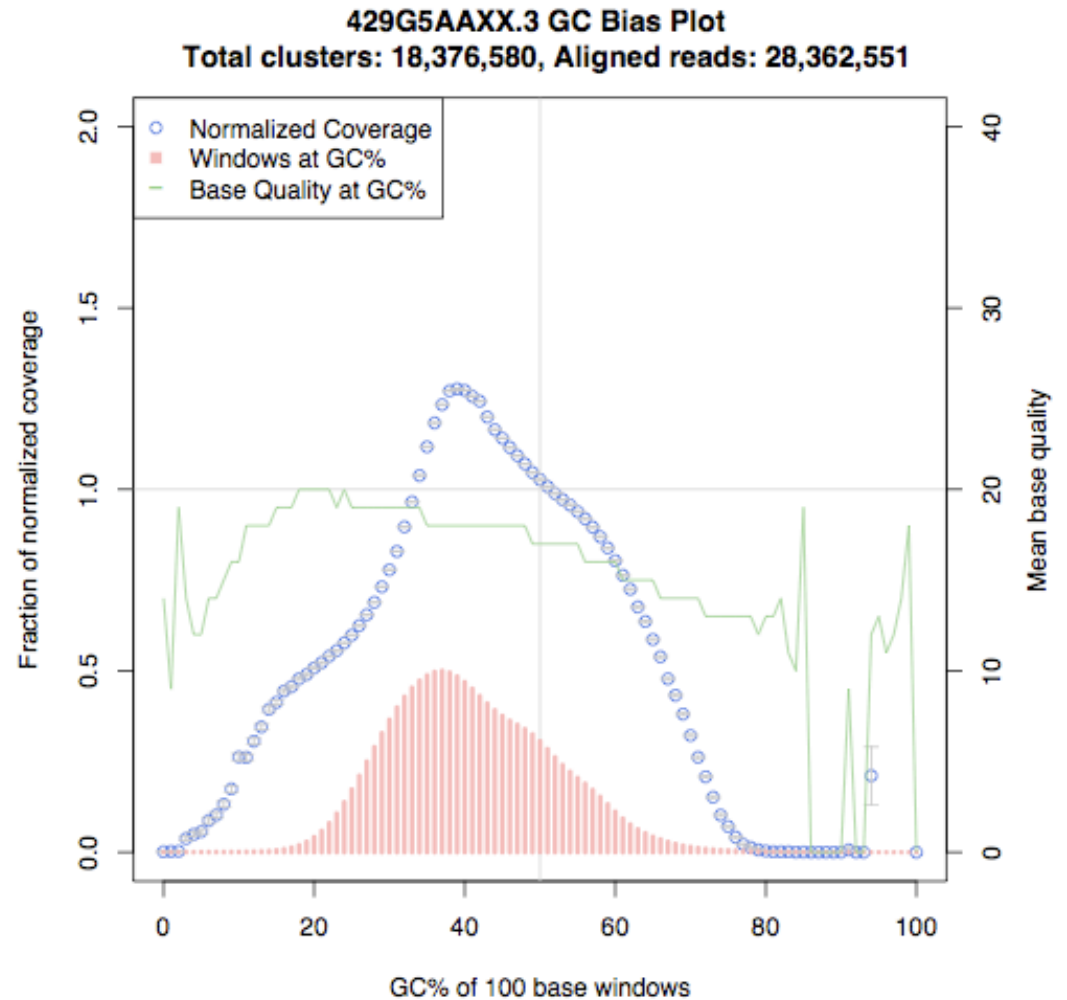
- Lots of metrics about reads, alignments etc.
- Metrics are provided per read and by read-pair
- Outputs in Picard's "MetricsFile" format that is very easy to parse
- Snapshot on right is of one lane dumped into excel and transposed

CATEGORY	FIRST_OF_PAIR	SECOND_OF_PAIR	PAIR
TOTAL_READS	27,191,296	27,191,296	54,382,592
PF_READS	19,840,618	19,840,618	39,681,236
PCT_PF_READS	72.97%	72.97%	72.97%
PF_NOISE_READS	1,278	3,798	5,076
PF_READS_ALIGNED	15,524,456	16,399,698	31,924,154
PCT_PF_READS_ALIGNED	78.25%	82.66%	80.45%
PF_HQ_ALIGNED_READS	14,399,769	15,140,729	29,540,498
PF_HQ_ALIGNED_BASES	1,454,057,669	1,529,035,036	2,983,092,705
PF_HQ_ALIGNED_Q20_BASES	734,931,725	733,556,680	1,468,488,405
PF_HQ_MEDIAN_MISMATCHES	2	1	1
PF_HQ_ERROR_RATE	3.66%	2.76%	3.20%
MEAN_READ_LENGTH	101	101	101
READS_ALIGNED_IN_PAIRS	14,476,222	14,475,681	28,951,903
PCT_READS_ALIGNED_IN_PAIRS	93.25%	88.27%	90.69%
BAD_CYCLES	0	0	0
STRAND_BALANCE	49.93%	50.05%	49.99%
PCT_CHIMERAS	1.09%	1.03%	1.06%
PCT_ADAPTER	0.54%	0.01%	0.28%

```
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

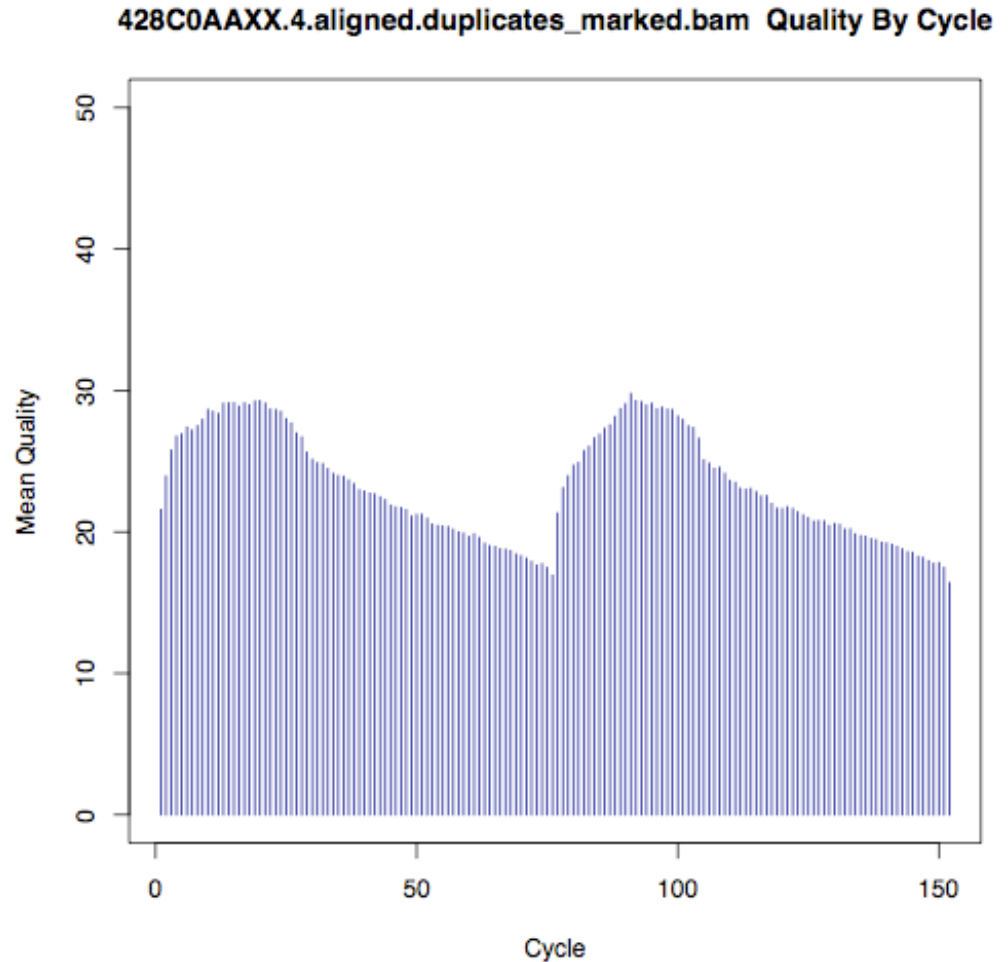
- Shows:
 - Distribution of GC in genome
 - Bias of aligned sequence by GC
 - Mean quality score by GC
- Looks at every overlapping 100bp window in the genome
- Calculated %GC of the window and how many reads start at the first base of that window



```
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

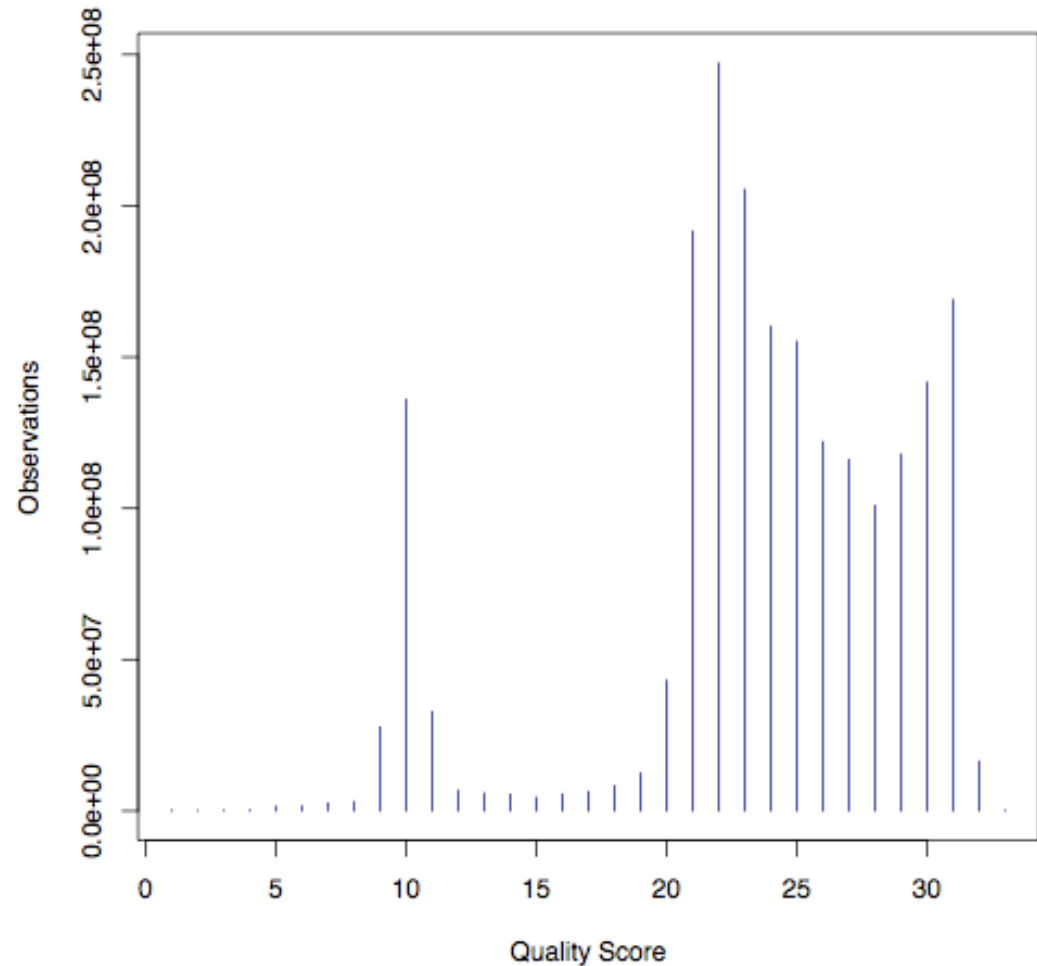


```
java -jar $PICARD/QualityScoreDistribution.jar \  
I=429G5AAXX.3.aligned.duplicates_marked.bam \  
O=429G5AAXX.3.quality_distribution_metrics CHART=429G5AAXX.3.quality_distribution.pdf
```

Quality Distribution Plot

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

428C0AAXX.4.aligned.duplicates_marked.bam Quality Score Distribution



```
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

IN	UNIQUE
1	1
2	1.989985
3	2.970055
4	3.940311
5	4.900849
...	...

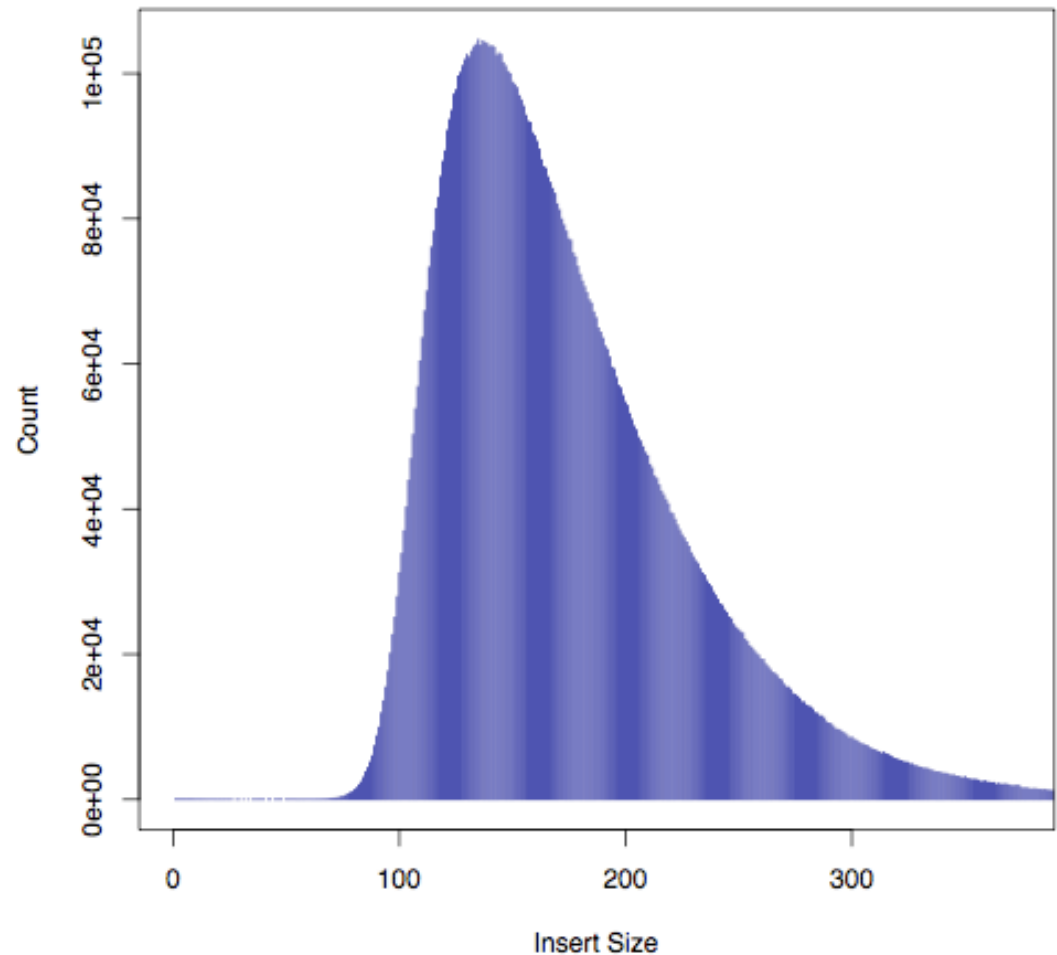
Library Type	Target Library Size
Whole Genome Shotgun	Billions (1-5bn is normal)
Whole Exome Hybrid Selection	Hundreds of Millions
Small Design Hybrid Selection	Tens of Millions (depends on target size)

```
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

428C0AAXX.4.aligned.duplicates_marked.bam 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

CATEGORY	REFERENCE	NON_REFERENCE	PCT_CONCORDANCE
HOMOZYGOUS REF	331,192	428	99.87%
HETEROZYGOUS	108,866	107,519	99.38%
HOMOZYGOUS NON REF	261	160,221	99.84%

- 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

```
java -jar $PICARD/CollectLowPassConcordanceStats.jar \  
I=429G5AAXX.3.aligned.duplicates_marked.bam \  
OUTPUT=429G5AAXX.3.low_pass_concordance_statistics \  
REFERENCE_GENOTYPES=/seq/references/reference_genotypes/hapmap/Homo_sapiens_assembly18/NA19056.geli
```

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

BAIT_SET	tcga_6k_genes
GENOME_SIZE	3,096,521,113
BAIT_TERRITORY	14,506,583
TARGET_TERRITORY	12,187,068
BAIT_DESIGN_EFFICIENCY	84.01%
TOTAL_READS	27,410,716
PF_READS	24,881,914
PF_UNIQUE_READS	24,031,465
PCT_PF_READS	90.77%
PCT_PF_UQ_READS	87.67%
PF_UQ_READS_ALIGNED	21,881,843
PCT_PF_UQ_READS_ALIGNED	91.06%
PF_UQ_BASES_ALIGNED	1,662,932,254
ON_BAIT_BASES	975,625,616
NEAR_BAIT_BASES	306,537,630
OFF_BAIT_BASES	380,769,008
ON_TARGET_BASES	881,474,940
PCT_SELECTED_BASES	77.10%
PCT_OFF_BAIT	22.90%
ON_BAIT_VS_SELECTED	76.09%
MEAN_BAIT_COVERAGE	67.253992
MEAN_TARGET_COVERAGE	75.90869
PCT_USABLE_BASES_ON_BAIT	51.59%
PCT_USABLE_BASES_ON_TARGET	46.61%
FOLD_ENRICHMENT	125.232646
ZERO_CVG_TARGETS_PCT	6.11%
FOLD_80_BASE_PENALTY	6.325724
PCT_TARGET_BASES_2X	90.47%
PCT_TARGET_BASES_10X	78.53%
PCT_TARGET_BASES_20X	69.24%
PCT_TARGET_BASES_30X	61.48%
HS_LIBRARY_SIZE	146,609,384
HS_PENALTY_10X	12.298504
HS_PENALTY_20X	12.626163
HS_PENALTY_30X	12.953821

```
java -jar $PICARD/CalculateHsMetrics.jar \ I=428C0AAXX.4.aligned.duplicates_marked.bam \
METRICS_FILE=428C0AAXX.4.hybrid_selection_metrics \
BAIT_INTERVALS=/seq/references/HybSelOligos/tcga_6k_genes/tcga_6k_genes.baits.interval_list \
TARGET_INTERVALS=/seq/references/HybSelOligos/tcga_6k_genes/tcga_6k_genes.targets.interval_list
```

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?