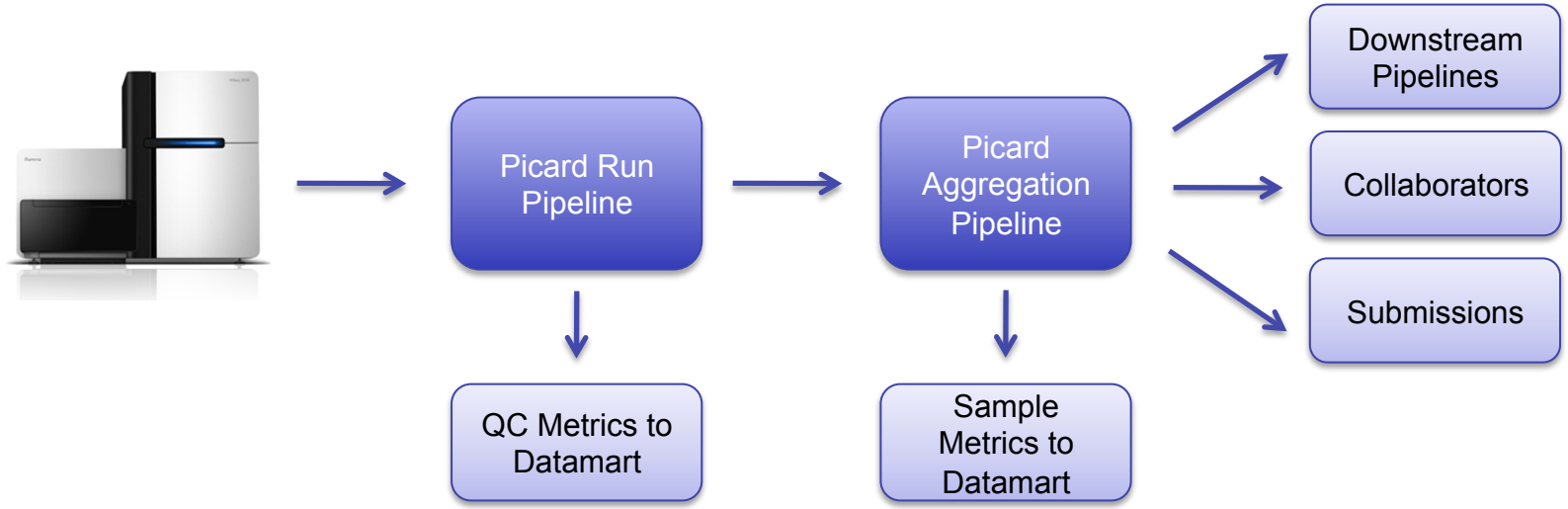
The Picard Pipelines

Tim Fennell
Sequencing Pipeline Informatics
Broad Institute

Pipeline Context



- Production pipelines feed into computational pipelines of other groups at Broad including:
  - Cancer mutation and rearrangement detection
  - Medical Genetics variant calling
  - Automated microbial assembly
- Datamarts allow both rapid reporting of QC metrics and correlation of metrics to LIMS information

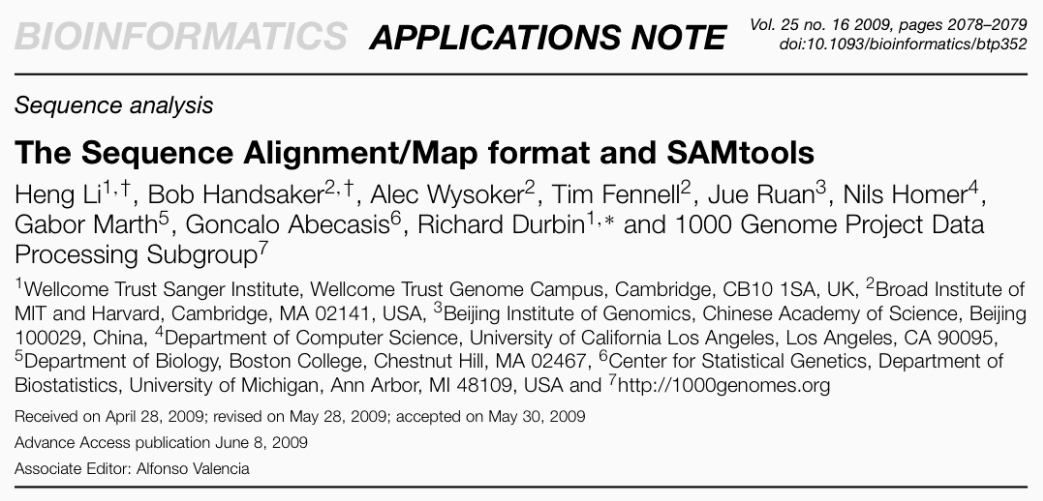
Pipeline requirements

- Produce ***analysis ready*** BAM[1] files
  - "State of the art" data processing
  - Retain all data; flag don't discard
  - Identity and integrity checked
  - Aggregated by "Sample"
- Produce key QC metrics
  - Assess run quality
  - Assess library construction
  - Identify trends
  - Informs downstream analysis
- Produce project tracking metrics
  - "Is my sample done"?
  - Enable project managers
- Efficient use of compute and storage
  - Optimize for throughput; turn around time still important
  - Less compute = cheaper and faster results
  - Storage costs now roughly equivalent to reagent costs!

[1] SAM specification: http://samtools.sourceforge.net

BAM: A few notes about your BAM files

- All primary data is delivered in BAM format, which includes basecalls (the reads), quality scores, alignment data, etc.
- BAM files processed through Picard always contain all reads, including:
  - All unaligned reads (marked as unmapped)
  - All duplicate reads (marked as duplicates)
  - All “non-PF” reads (marked as failing vendor quality)



BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

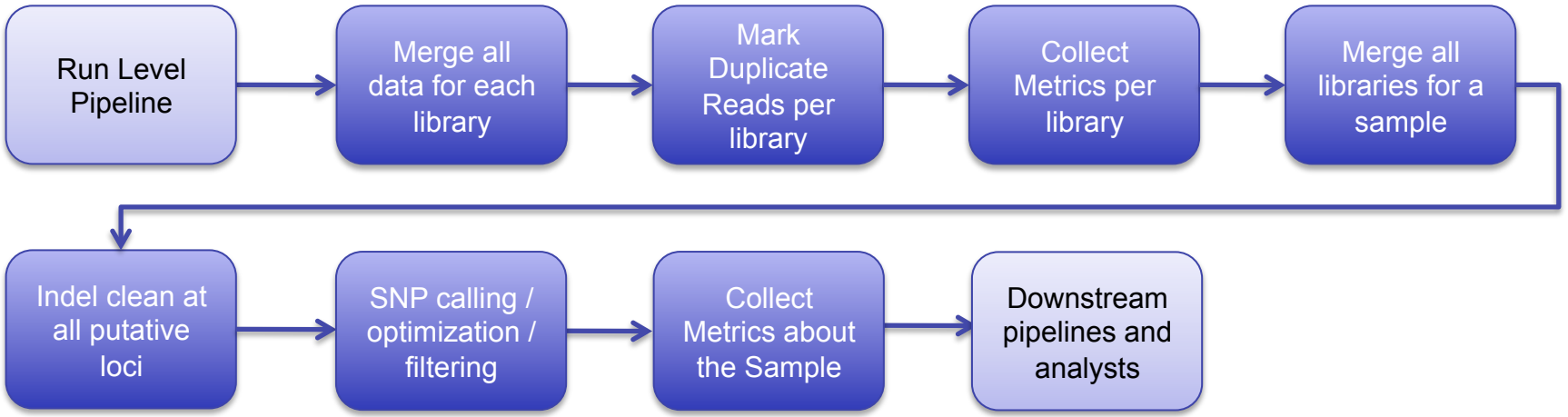
Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

The WGS/Exome Pipeline High Level Overview

- Adapter trimming/marking happens during data extraction from Illumina (information is used during alignment)
- Indexed runs are de-multiplexed during extraction and each index/sample processed independently
- Recalibration only performed for references with dbSNP data available
- Many pipeline variants exist for:
  - Whole methylome and reduced representation bisulfite sequencing
  - Various long mate pair protocols
  - Hybrid selection
  - Pre-assembly QC

The Aggregation Pipeline
High Level Overview



- A single BAM file is created per Sample (within the context of a project)
- Aggregations are started after data is processed or re-processed through the run-level pipeline (after a 12 hour “quiet period”)
- Outdated aggregations are kept for 2 weeks after newer aggregations are completed

Extracting Bases + Qualities

- Parses file formats produced by all major versions of Illumina pipelines
- Auto-detects Illumina pipeline version and read configuration
- Detects and marks Illumina adapter sequences
  - Allows us to rescue reads from inserts significantly shorter than the read length that would otherwise not align
  - Ensures base quality calibration isn't aversely affected by reads from inserts slightly shorter than the read length
- Can also:
  - Run multithreaded to reduce runtime
  - Parse intensity information
  - De-multiplex multiplexed runs on the fly
- Generates a well-formed "unmapped" BAM complete with
  - Sample and library metadata
  - Adapter clip points

Generating aligned BAM files

Three stage process to generate well-formed aligned BAM file

- SamToFastq creates fastq files with adapter sequences clipped
- Alignment performed with BWA
  - Uses BWA's quality trimming feature with a low cutoff (Q5)
  - Run multithreaded (usually 4 threads per alignment)
  - Otherwise mostly default options
- MergeBamAlignment to create "aligned" bam
  - Merges SAM output from bwa with unmapped bam
  - Carries forward all unmapped reads
  - Carries forward all tags/metadata from unmapped bam
  - Restores bases hard-clipped before alignment (with soft clip)
  - Produces a very "well-formed" BAM file for downstream processing

Optical and PCR duplicate marking

Problem: PCR amplification causes molecular duplicates, sequencing artifacts cause optical duplicates; significant duplication causes problems for variant discovery etc.

MarkDuplicates outputs new BAM with duplicate reads marked

- Supports downstream analyses
  - Duplicates can be accounted for
  - Improves base quality recalibration
- Yields key metrics to assess library quality
  - Percent duplication
  - Optical duplication rate
  - Estimated library size (from non-optical duplicates)
- Works on *all* paired-end and single-end data simultaneously

Indel Realignment

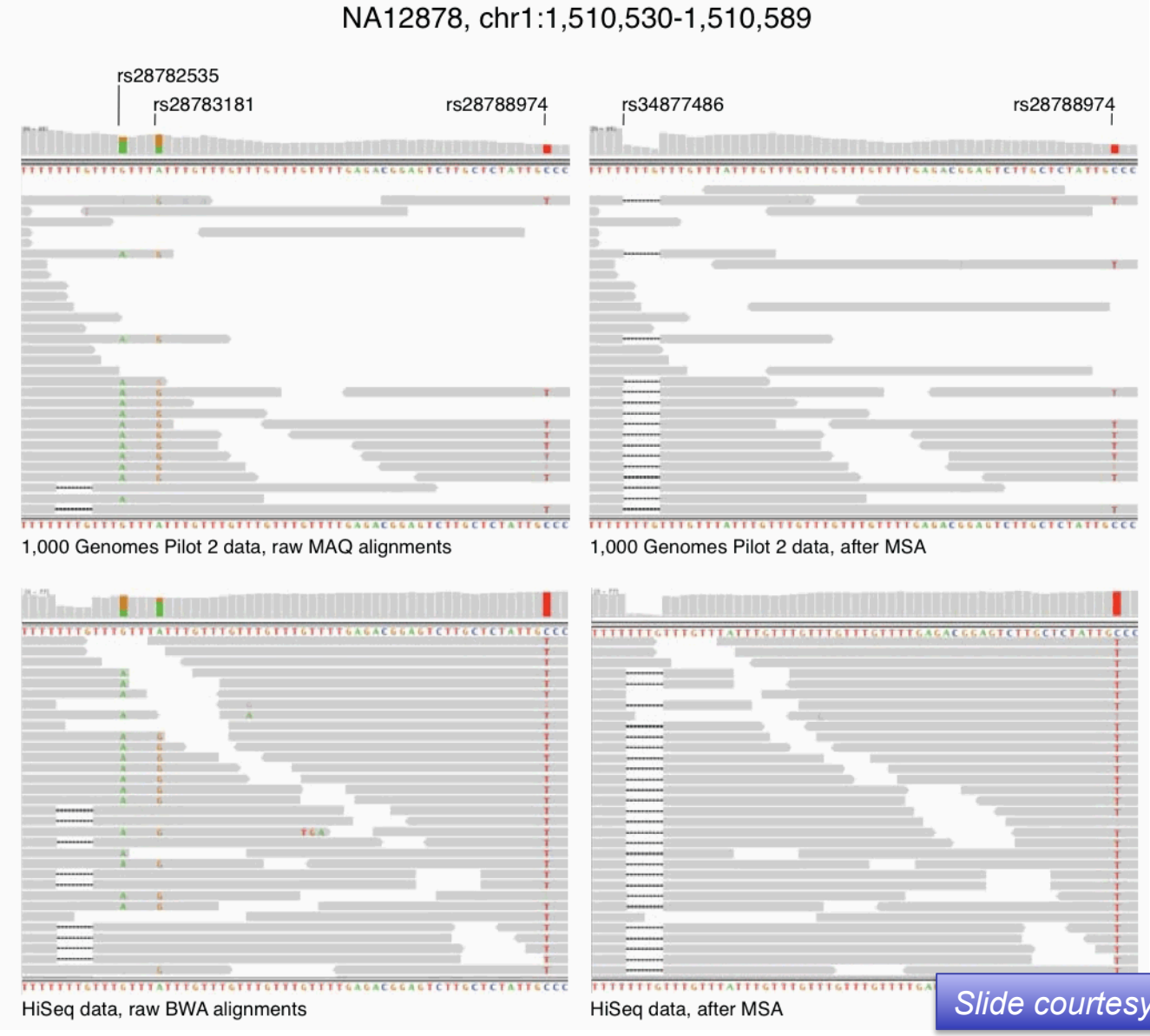
Problem: Short reads are often misaligned around small insertions and deletions

GATK Indel Realigner identifies reads that map across positions of known indels from dbSNP and the 1000 Genomes project and

1. Realigns reads against all known haplotypes
2. Re-writes read alignment information based on best haplotype alignment

- Improves base quality recalibration
- Reduces false-positive SNP calls
- Improves indel genotyping

Indel Realignment



Slide courtesy of Eric Banks

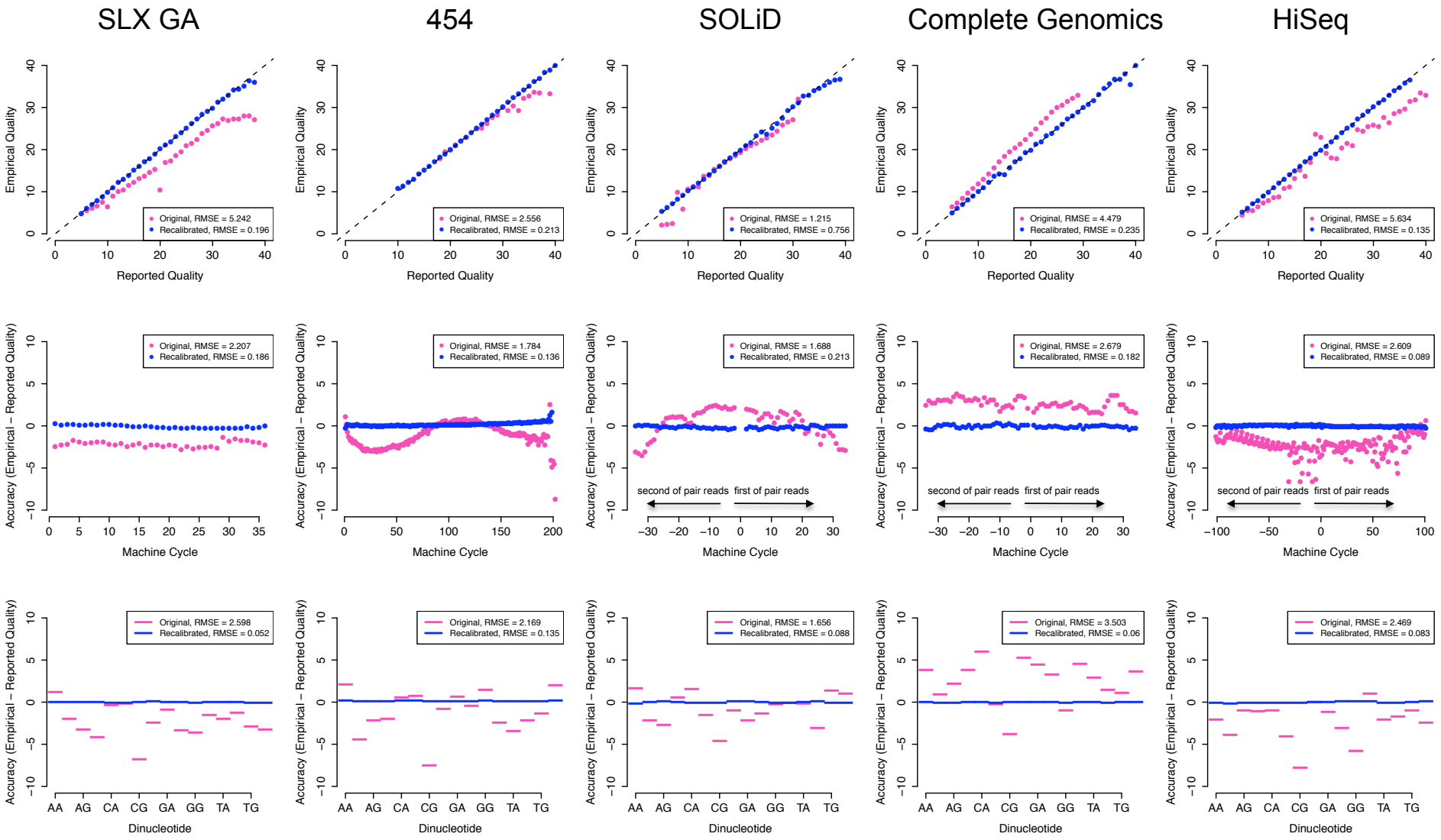
Base Quality Score Recalibration

Problem: Base quality score accuracy is lower than desired

GATK base quality score recalibrator

- Uses multiple covariates to separate high and low quality bases and provides a more information-rich distribution
  - Instrument cycle
  - Assigned quality score
  - Dinucleotide context
- Ensures quality scores are accurate wrt observed error rates
- Retains original quality scores in BAM file alongside calibrated quality scores
- Only performed for references with dbSNP build

Base Quality Score Recalibration



Slide courtesy of Ryan Poplin

Sample Fingerprinting

Problem: In a lab handling thousands of samples a week with rapidly evolving protocols mix ups will happen; need to be able to catch and un-mix informatically

SNP Fingerprinting

- Panels of 24-36 SNPs picked specifically for this purpose
  - High minor allele frequency in all HapMap populations
  - Ideally with many perfect proxies
- Sub $5 assay performed in separate lab from independent aliquot
- Bayesian likelihood model allows us to robustly test
  - Sequence data vs. expected SNP fingerprint
  - Sequence data vs. SNP fingerprint for all samples in lab
  - Sequence data vs. other sequence data

Pipeline manager

- Now on our second pipeline manager, called “Zamboni”
- Key requirements:
  - Robustness, robustness, robustness!
  - Totally independent from our LIMS and environment
  - Workflows defined as graphs; parallelism inferred
  - Scale to tens of thousands of concurrent workflows
  - Small runtime footprint
  - Maximize use of available compute resources
  - Automatic retry for known failure modes
  - Restart of any workflow from where it stopped
  - Workflow versioning; multiple versions live at once
  - Rapid development/modification/testing of pipelines
- In the six weeks it has been live we have run over 50,000 pipelines and processed ~40 terabases of sequence

Where to find tools, data, source code, etc.

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

- And coming soon – BASS
  - Programmatic access to BAM files in BASS available
  - Web page to access BAM files in BASS under construction