Sequence Mapping and Assembly

Sequence Analysis Workshop
June 16, 2014

Mary Kate Wing
University of Michigan
Center for Statistical Genetics

Goals of This Session

- Learn basics of sequence data file formats
  - FASTQ & BAM
- Raw sequence reads -> aligned sequences
  - Get ready for variant calling
- Evaluate quality of sequence data
- Visualize sequence data to examine reads aligned to particular genomic positions

Session Design

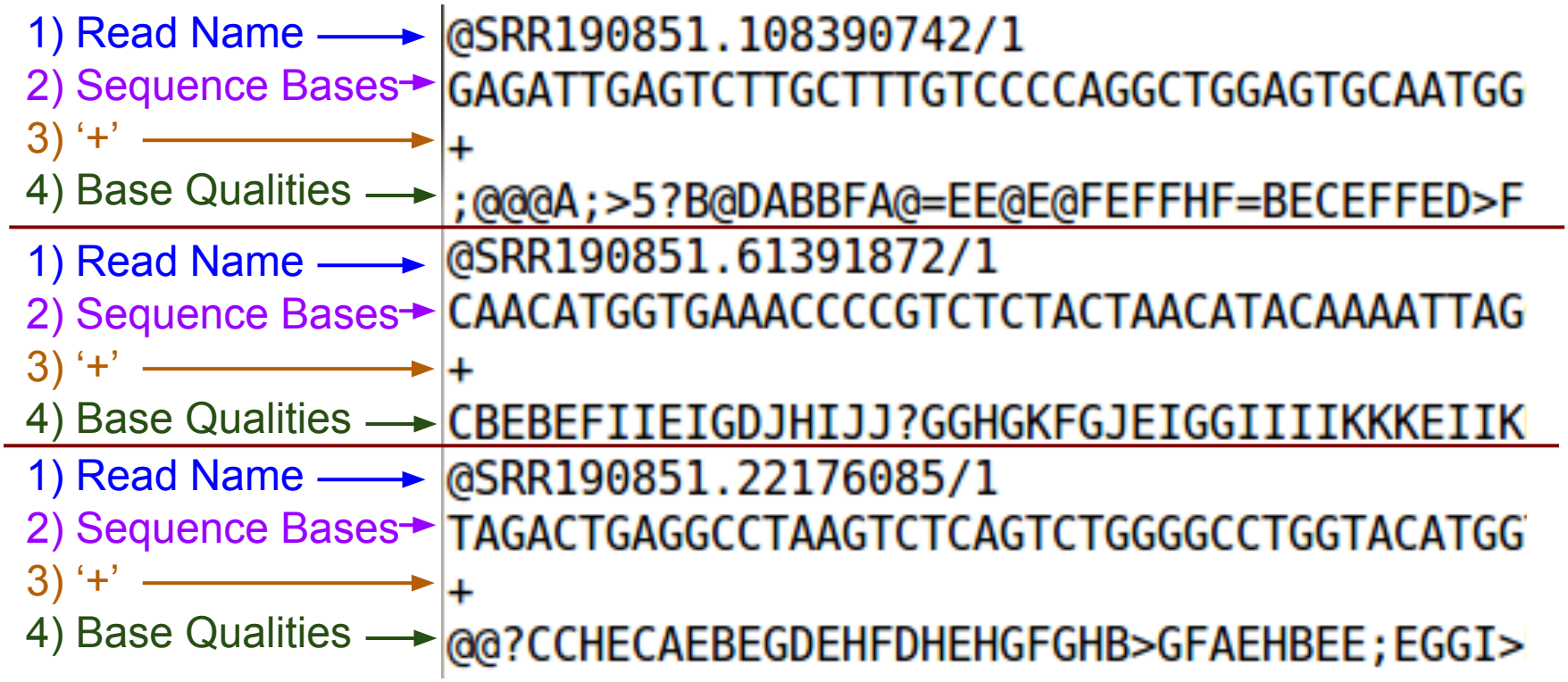
- A few intro slides
  - Introduces you to how to do each of the goals
- Instructions to follow at your own pace
  - Walkthrough of how to produce aligned reads
  - Screenshots with explanations
- Raise your hand if you have any questions/problems
  - Someone will come help

Raw Sequence Reads (FASTQs)

- Standard file format from sequencing
  - Sequencing done as series of reads
  - Not associated with a chromosome/position
- http://en.wikipedia.org/wiki/FASTQ_format

Raw Sequence Reads (FASTQs)

4 lines per read



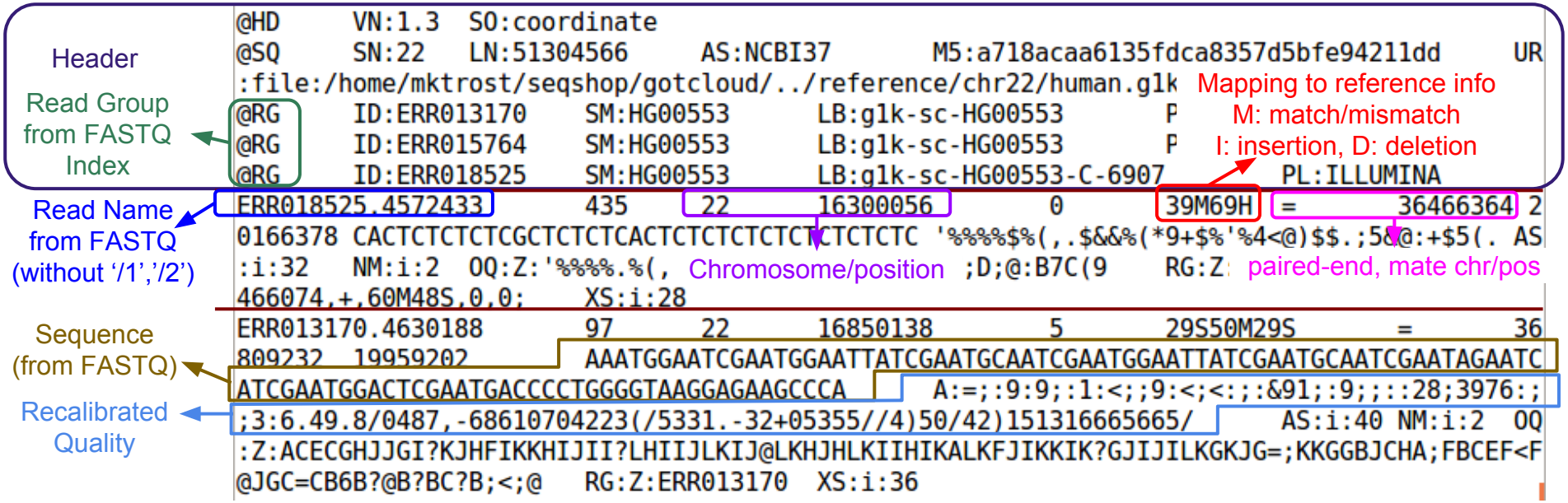
Raw Sequence Reads (FASTQs)

- Base Qualities
  - ASCII quality code for each base
    - 33 + phred scale = 33 + $-10\log_{10}$ e
      - e is estimate probability of an incorrect base
    - Lower qualities: special characters/digits
      - ! (Q=0), “ (Q=1), # (Q=3), + (Q-10), / (Q=14)
      - 0 (Q=15), 5 (Q=20), 9 (Q=24)
    - Higher qualities (>Q30): alphabetic characters
      - : (Q=25), ? (Q=30), @ (Q=31)
      - A (Q=32), B (Q=33), G (Q=38)
  - Will be recalibrated in alignment pipeline
    - By sequencing run/fastq pair
    - Become more accurate

Sequence Alignment/Map Format: SAM/BAM

- Maps read to Chromosome & Position
  - Spec: http://samtools.github.io/hts-specs/SAMv1.pdf
  - More Info: http://genome.sph.umich.edu/wiki/SAM
- Header lines
  - Each line starts with '@'
- Records
  - One for each sequence read/FASTQ record
  - FASTQ info PLUS Chr/Pos

Sequence Alignment/Map Format: SAM/BAM Records



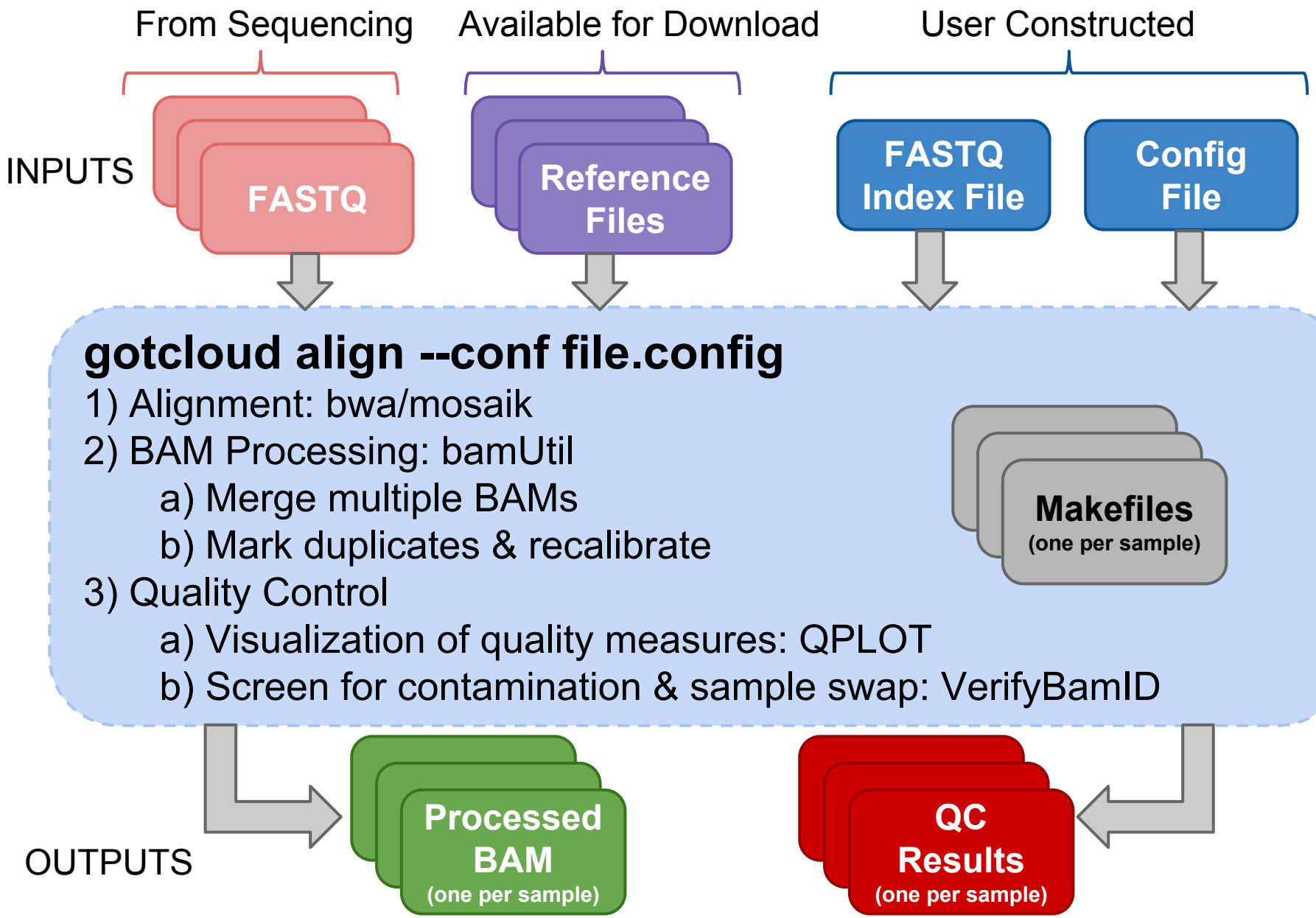
Viewing SAM/BAM Files

- Samtools
  - http://samtools.sourceforge.net/
  - view
    - read group, library, MAPQ >, region
  - tview
    - text alignment viewer - visualize reads by position
- BamUtil
  - http://genome.sph.umich.edu/wiki/BamUtil
  - Lot's of SAM/BAM tools

Genomes on the Cloud (GotCloud): Alignment Pipeline

- All-in-one sequence analysis pipeline
  - You don't need to know the details of individual components
  - Automates steps for you
- Robust parallelization
  - Automatically partitions multi-sample jobs
  - Takes advantage of clusters
    - Supports MOSIX, slurm, SGE, pbs (flux)
    - Can setup a cluster on Amazon
  - via GNU make
    - Reliable and fault-tolerant
    - Restart where it stopped upon unexpected crash

GotCloud Alignment Pipeline Overview



User Constructed Input: FASTQ Index File

- GotCloud needs to know about each FASTQ
  - Where to find it
  - Sample name
    - Each sample can have multiple FASTQs
    - 1 FASTQ only has a single sample

- Format
  - Tab delimited
  - Header line
  - One line per single-end
  - One line per paired-end

User Constructed Input: FASTQ Index File

Header Row

MERGE_NAME	FASTQ1	FASTQ2	RGID	SAMPLE	LIBRARY	PLATFORM
HG00551	HG00551.SRR190851.fastq	.	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00551	HG00551.SRR190851_1.fastq	HG00551.SRR190851_2.fastq	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00553	HG00553.ERR013170.fastq	.	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR013170_1.fastq	HG00553.ERR013170_2.fastq	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764.fastq	.	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764_1.fastq	HG00553.ERR015764_2.fastq	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR018525.fastq	.	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00553	HG00553.ERR018525_1.fastq	HG00553.ERR018525_2.fastq	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00640	HG00640.ERR013174.fastq	.	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR013174_1.fastq	HG00640.ERR013174_2.fastq	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768.fastq	.	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768_1.fastq	HG00640.ERR015768_2.fastq	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR018527.fastq	.	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00640	HG00640.ERR018527_1.fastq	HG00640.ERR018527_2.fastq	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00641	HG00641.SRR069531.fastq	.	SRR069531	HG00641	Solexa-41496	ILLUMINA
HG00641	HG00641.SRR069531_1.fastq	HG00641.SRR069531_2.fastq	SRR069531	HG00641	Solexa-41496	ILLUMINA

User Constructed Input: FASTQ Index File

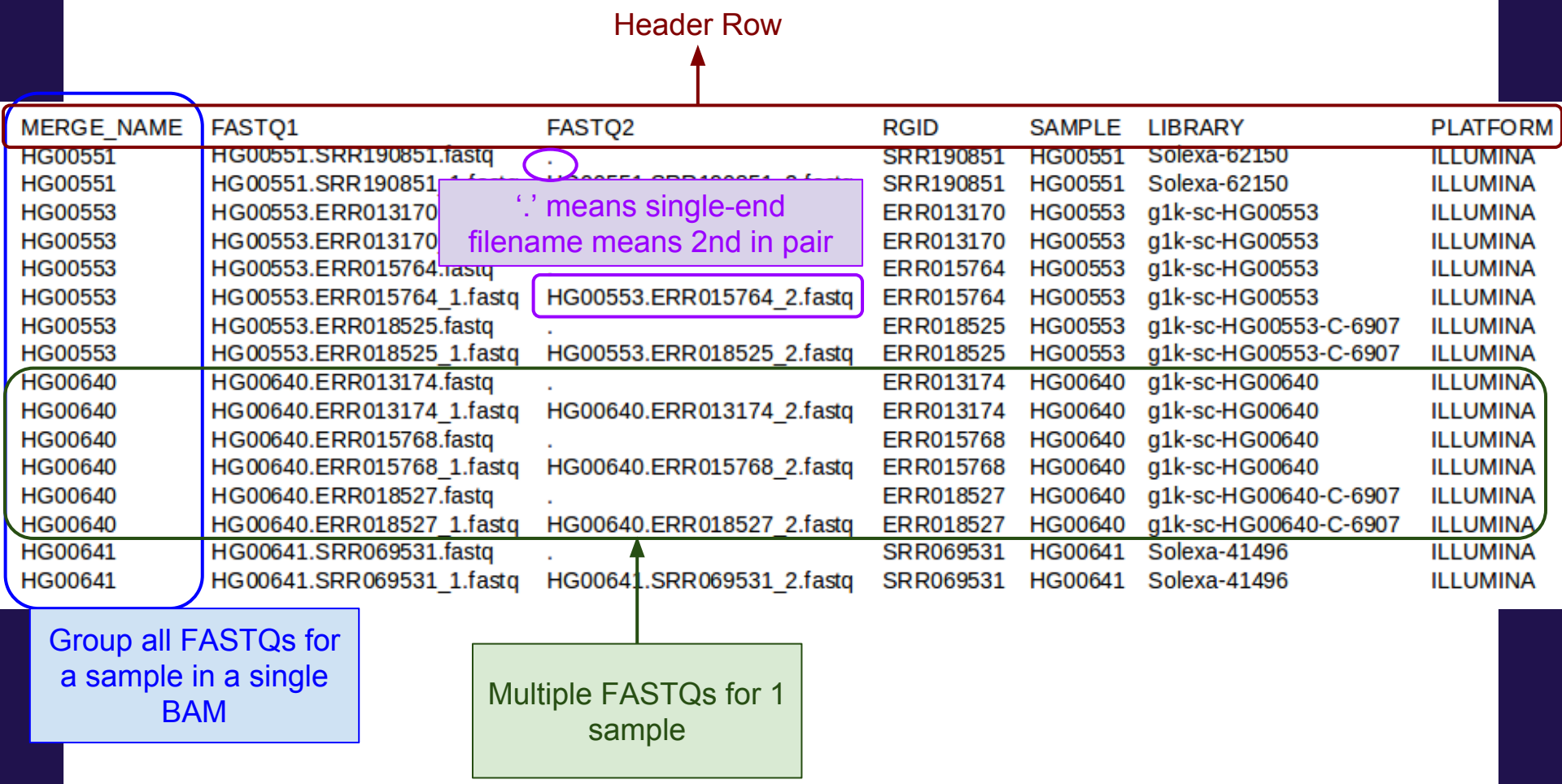
Header Row

MERGE_NAME	FASTQ1	FASTQ2	RGID	SAMPLE	LIBRARY	PLATFORM
HG00551	HG00551.SRR190851.fastq	.	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00551	HG00551.SRR190851_1.fastq	HG00551.SRR190851_2.fastq	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00553	HG00553.ERR013170.fastq	.	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR013170_1.fastq	HG00553.ERR013170_2.fastq	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764.fastq	.	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764_1.fastq	HG00553.ERR015764_2.fastq	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR018525.fastq	.	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00553	HG00553.ERR018525_1.fastq	HG00553.ERR018525_2.fastq	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00640	HG00640.ERR013174.fastq	.	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR013174_1.fastq	HG00640.ERR013174_2.fastq	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768.fastq	.	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768_1.fastq	HG00640.ERR015768_2.fastq	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR018527.fastq	.	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00640	HG00640.ERR018527_1.fastq	HG00640.ERR018527_2.fastq	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00641	HG00641.SRR069531.fastq	.	SRR069531	HG00641	Solexa-41496	ILLUMINA
HG00641	HG00641.SRR069531_1.fastq	HG00641.SRR069531_2.fastq	SRR069531	HG00641	Solexa-41496	ILLUMINA

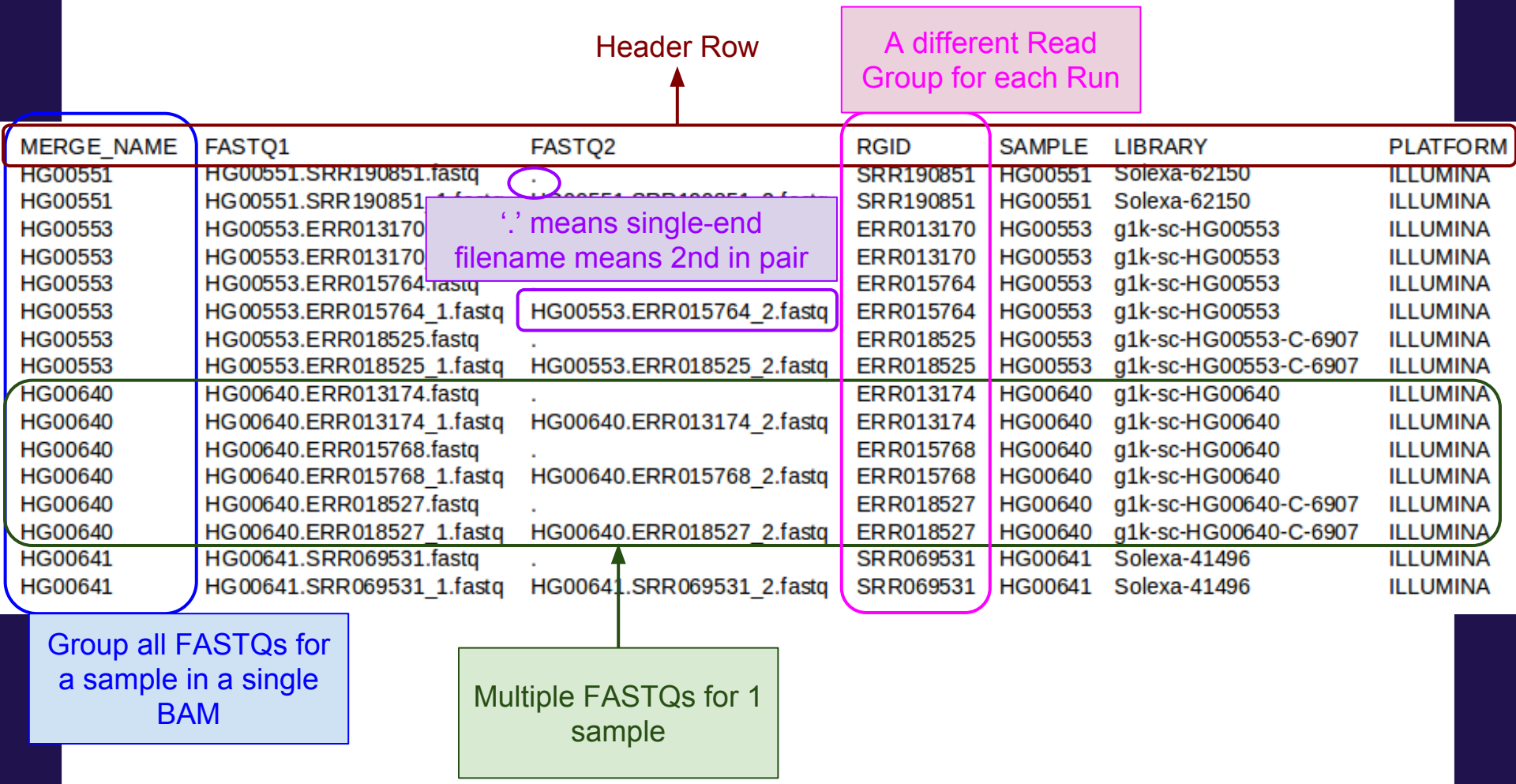
Group all FASTQs for a sample in a single BAM

Multiple FASTQs for 1 sample

User Constructed Input: FASTQ Index File

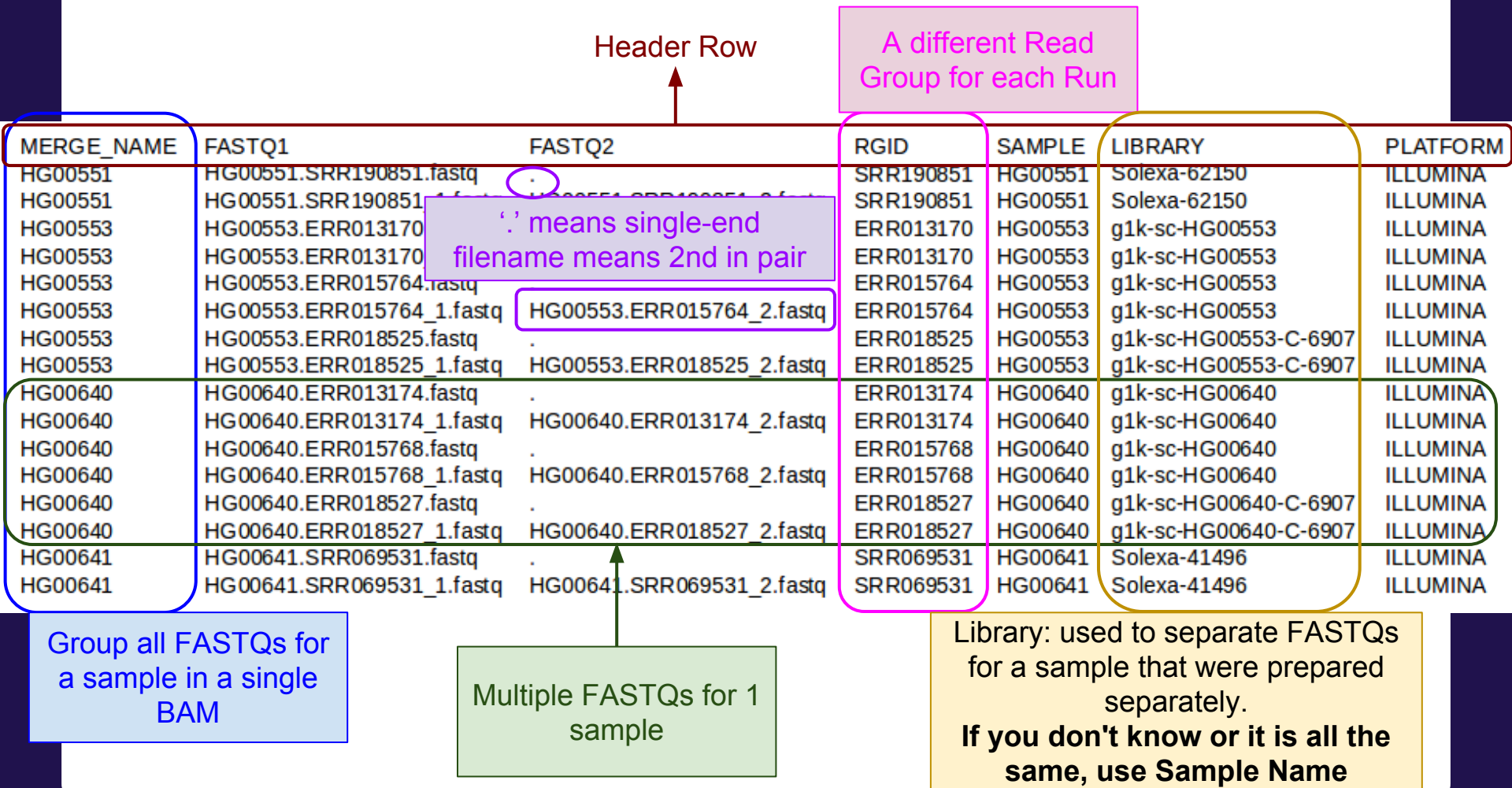


User Constructed Input: FASTQ Index File



MERGE_NAME	FASTQ1	FASTQ2	RGID	SAMPLE	LIBRARY	PLATFORM
HG00551	HG00551.SRR190851.fastq	.	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00551	HG00551.SRR190851_1.fastq	HG00551.SRR190851_2.fastq	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00553	HG00553.ERR013170.fastq	.	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR013170_1.fastq	HG00553.ERR013170_2.fastq	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764.fastq	.	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764_1.fastq	HG00553.ERR015764_2.fastq	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR018525.fastq	.	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00553	HG00553.ERR018525_1.fastq	HG00553.ERR018525_2.fastq	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00640	HG00640.ERR013174.fastq	.	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR013174_1.fastq	HG00640.ERR013174_2.fastq	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768.fastq	.	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768_1.fastq	HG00640.ERR015768_2.fastq	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR018527.fastq	.	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00640	HG00640.ERR018527_1.fastq	HG00640.ERR018527_2.fastq	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00641	HG00641.SRR069531.fastq	.	SRR069531	HG00641	Solexa-41496	ILLUMINA
HG00641	HG00641.SRR069531_1.fastq	HG00641.SRR069531_2.fastq	SRR069531	HG00641	Solexa-41496	ILLUMINA

User Constructed Input: FASTQ Index File



MERGE_NAME	FASTQ1	FASTQ2	RGID	SAMPLE	LIBRARY	PLATFORM
HG00551	HG00551.SRR190851.fastq	.	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00551	HG00551.SRR190851_1.fastq	HG00551.SRR190851_2.fastq	SRR190851	HG00551	Solexa-62150	ILLUMINA
HG00553	HG00553.ERR013170.fastq	.	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR013170_1.fastq	HG00553.ERR013170_2.fastq	ERR013170	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764.fastq	.	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR015764_1.fastq	HG00553.ERR015764_2.fastq	ERR015764	HG00553	g1k-sc-HG00553	ILLUMINA
HG00553	HG00553.ERR018525.fastq	.	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00553	HG00553.ERR018525_1.fastq	HG00553.ERR018525_2.fastq	ERR018525	HG00553	g1k-sc-HG00553-C-6907	ILLUMINA
HG00640	HG00640.ERR013174.fastq	.	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR013174_1.fastq	HG00640.ERR013174_2.fastq	ERR013174	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768.fastq	.	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR015768_1.fastq	HG00640.ERR015768_2.fastq	ERR015768	HG00640	g1k-sc-HG00640	ILLUMINA
HG00640	HG00640.ERR018527.fastq	.	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00640	HG00640.ERR018527_1.fastq	HG00640.ERR018527_2.fastq	ERR018527	HG00640	g1k-sc-HG00640-C-6907	ILLUMINA
HG00641	HG00641.SRR069531.fastq	.	SRR069531	HG00641	Solexa-41496	ILLUMINA
HG00641	HG00641.SRR069531_1.fastq	HG00641.SRR069531_2.fastq	SRR069531	HG00641	Solexa-41496	ILLUMINA

User Constructed Input: GotCloud Configuration

```
############ ← #'s are comments
# References
REF_DIR = ref22
AS = NCBI37  # Genome assembly identifier
REF = $(REF_DIR)/human.g1k.v37.chr22.fa
DBSNP_VCF =  $(REF_DIR)/dbsnp_135.b37.chr22.vcf.gz
HM3_VCF = $(REF_DIR)/hapmap_3.3.b37.sites.chr22.vcf.gz
INDEL_PREFIX = $(REF_DIR)/1kg.pilot_release.merged.indels.sites.hg19
OMNI_VCF = $(REF_DIR)/1000G_omni2.5.b37.sites.PASS.chr22.vcf.gz

######### General ########
BAM_INDEX =  $(OUT_DIR)/bam.index

######### ALIGNMENT ########
MAP_TYPE = BWA_MEM
INDEX_FILE =align.index

######### Variant Calling ########
CHRS = 22
```

Use $(KEY) to refer to other KEYs
GOTCLOUD_ROOT is built in

Path to chr22 reference files

For align: output - where to write BAM list
For others: input - where to find list of BAMs

Use bwa mem instead of just regular BWA

Path to fastq index file

For snpcall & indel -> chr22 only

User Constructed Input: GotCloud Configuration

```
######### THUNDER ########
# Update so it will run faster for the tutorial
#  * 10 rounds instead of 30 (-r 10)
#  * without --compact option
#  Runs faster, but uses more memory, but not a lot for the small example
THUNDER = $(BIN_DIR)/thunderVCF -r 10 --phase --dosage --inputPhased $(THUNDER_STATES)
```

Thunder Settings to speed up LD Refinement Pipeline for the tutorial

```
##############################
## GenomeSTRIP
##############################
GENOMESTRIP_OUT = $(OUT_DIR)/sv
GENOMESTRIP_SVTOOLKIT_DIR = svtoolkit
GENOMESTRIP_MASK_FASTA = $(GENOMESTRIP_SVTOOLKIT_DIR)/ref/human_g1k_v37.chr22.mask.100.fasta
GENOMESTRIP_PLOIDY_MAP = $(GENOMESTRIP_SVTOOLKIT_DIR)/conf/humgen_g1k_v37_ploidy.chr22.map
GENOMESTRIP_PARAM = $(GENOMESTRIP_SVTOOLKIT_DIR)/conf/genstrip_parameters.txt
```

Structural Variation Pipeline Settings

GotCloud Quality Control:
Sample Contamination/Swap (by *VerifyBamID*)

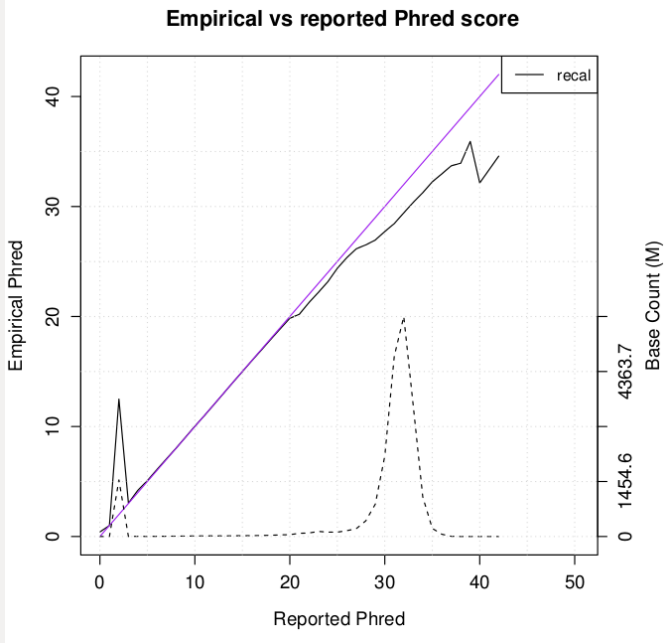
- Genotype-free estimate of contamination
  - 0-1 scale, the lower, the better
  - ‘FREEMIX’ column < 0.03
  - http://genome.sph.umich.edu/wiki/VerifyBamID#A_guideline_to_interpret_output_files
- Estimate of contamination with genotypes
  - 0-1 scale, the lower, the better
  - ‘CHIPMIX’ column
  - We don’t have this in our tutorial

#SEQ_ID	RG	CHIP_ID	#SNPS	#READS	AVG_DP	FREEMIX	FREELK1	FREELK0	FREE_RH	FREE_RA	CHIPMIX
HG00551	ALL	NA	20056	3644	0.18	0.00000	955.44	955.44	NA	NA	NA

GotCloud Quality Control:
Quality Metrics (by *QPLOT*)

- .stats file contains metrics, including
  - mapping rate, coverage, % high quality bases
- .R file that generates a .pdf of plots
  - Empirical vs reported Phred score

```
TotalReads(e6)  0.08
MappingRate(%)  98.93
MapRate_MQpass(%)       98.93
TargetMapping(%)        0.00
ZeroMapQual(%)  0.91
MapQual<10(%)   1.47
PairedReads(%)  98.91
ProperPaired(%) 86.53
MappedBases(e9) 0.01
Q20Bases(e9)    0.01
Q20BasesPct(%)  89.52
MeanDepth       7.43
```



Try it yourself

http://genome.sph.umich.edu/wiki/SeqShop:_Sequence_Mapping_and_Assembly_Practical

- Interested in GotCloud?
  - http://genome.sph.umich.edu/wiki/GotCloud
  - Join the mailing list:
    - http://groups.google.com/group/GotCloud