Practical Introduction

# Sequence Mapping and Assembly

December 8, 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
  - Many methods/pipelines, we cover 1
- 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 for you to follow
  - Walkthrough of how to produce aligned reads
  - Screenshots with explanations
- Raise your hand if you have any questions/problems
  - Someone will come help

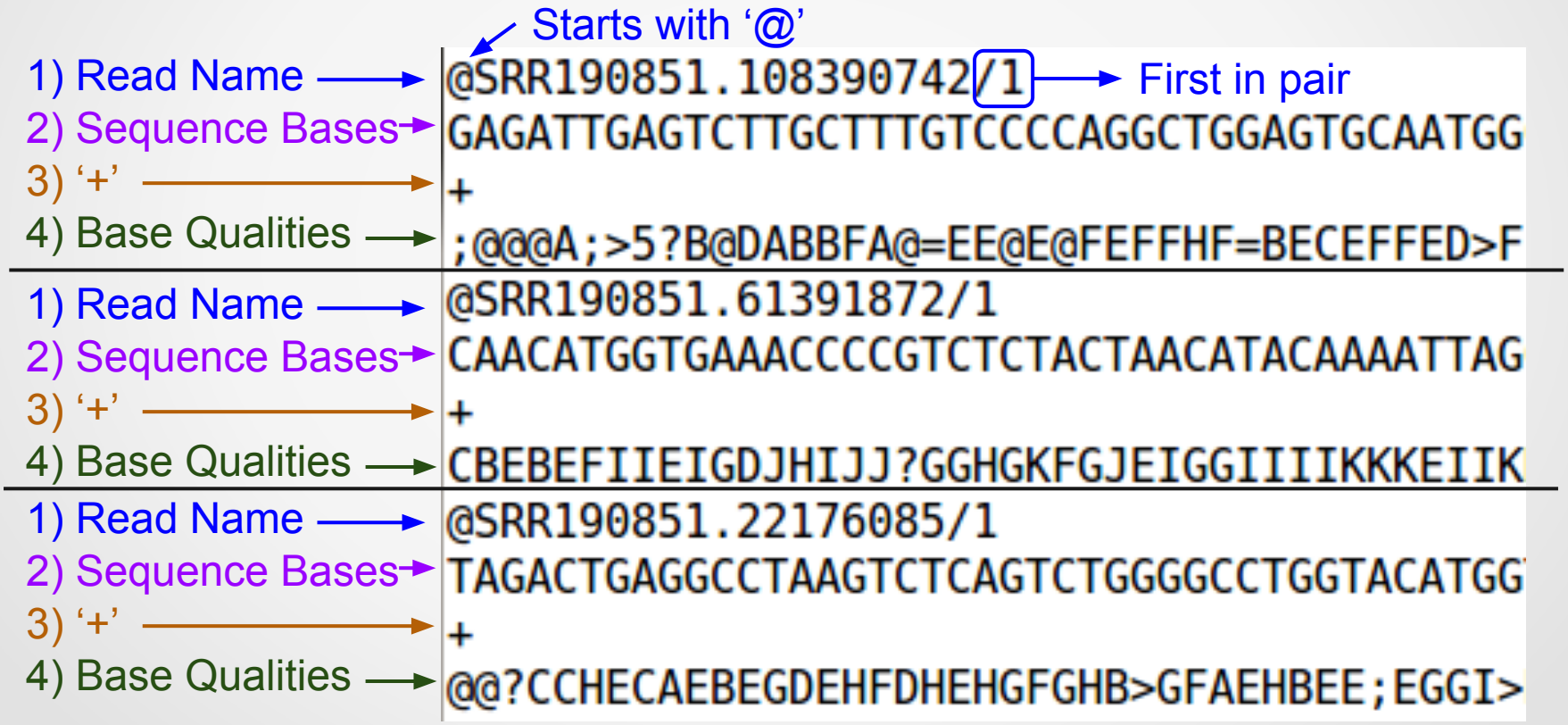
# Raw Sequence Reads (FASTQs)

- Standard file format from sequencing
  - Sequencing done as series of reads
  - Not associated with a chromosome/position
  - Reads can be in pairs
    - Typically separate file for 1st/2nd in pair

- 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

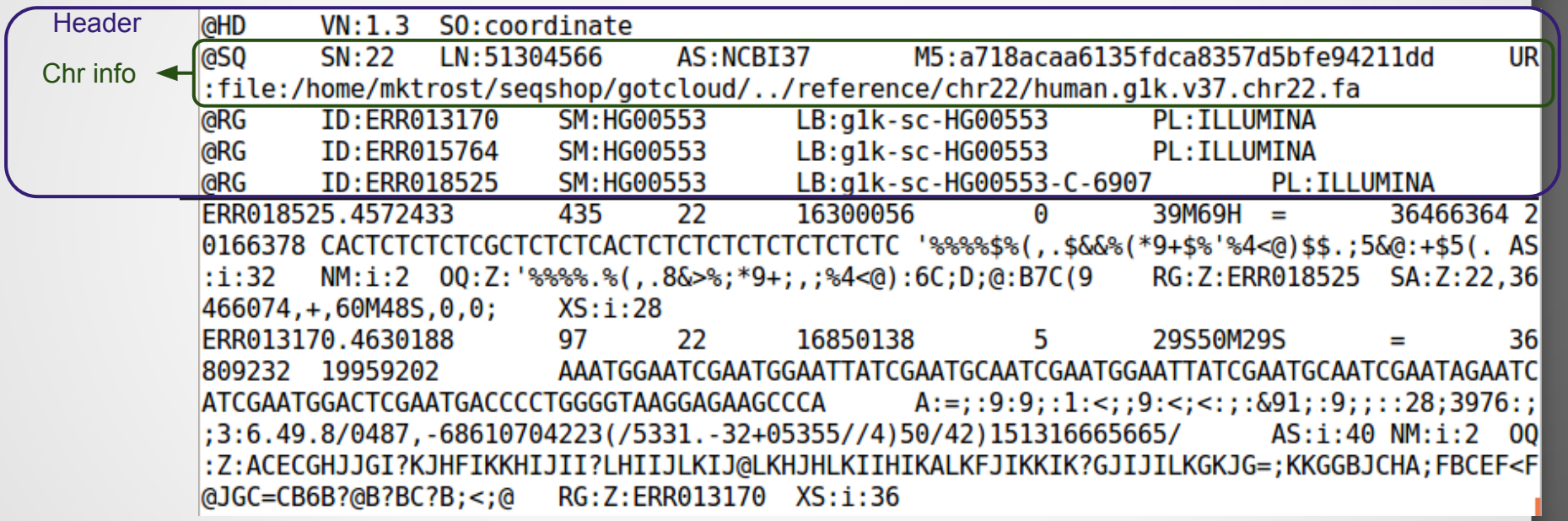
- Maps read to Chromosome & Position
  - Spec: http://samtools.github.io/hts-specs/SAMv1.pdf
  - More Info: http://genome.sph.umich.edu/wiki/SAM
- Header lines
  - Each line starts with '@'
- Records
  - One for each sequence read/FASTQ record
  - FASTQ info PLUS Chr/Pos

# SAM/BAM Records

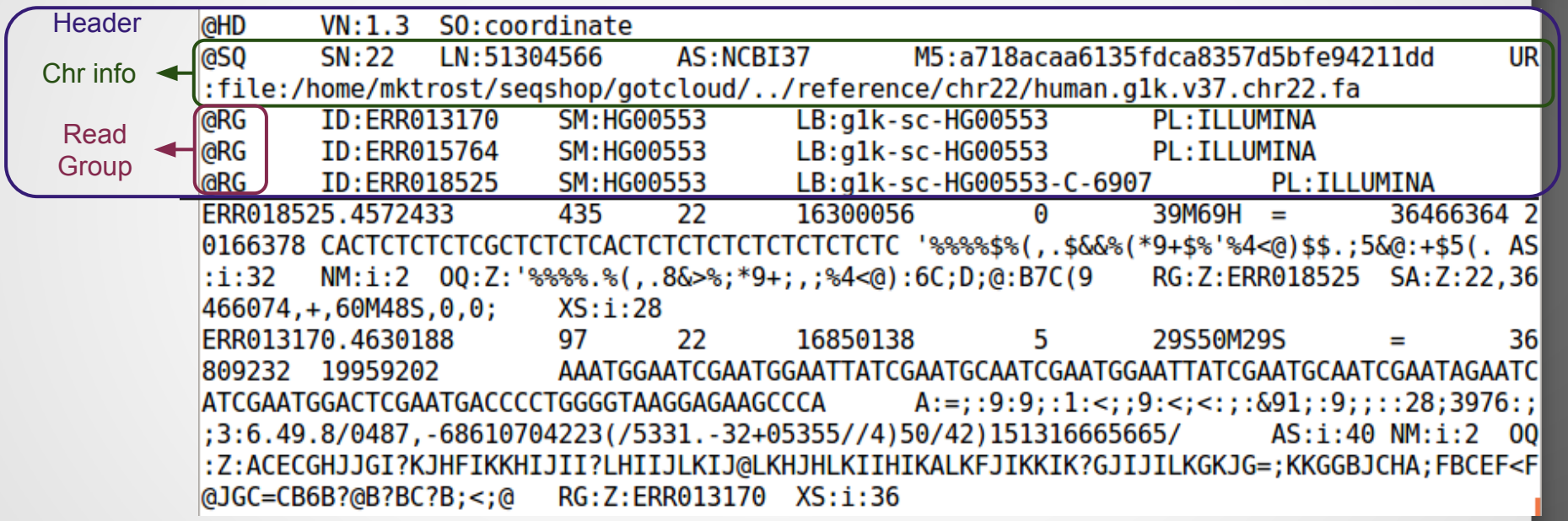
Header

```
@HD     VN:1.3  SO:coordinate
@SQ     SN:22   LN:51304566     AS:NCBI37       M5:a718acaa6135fdca8357d5bfe94211dd     UR
:file:/home/mktrost/seqshop/gotcloud/../reference/chr22/human.g1k.v37.chr22.fa
@RG     ID:ERR013170    SM:HG00553      LB:g1k-sc-HG00553       PL:ILLUMINA
@RG     ID:ERR015764    SM:HG00553      LB:g1k-sc-HG00553       PL:ILLUMINA
@RG     ID:ERR018525    SM:HG00553      LB:g1k-sc-HG00553-C-6907        PL:ILLUMINA
ERR018525.4572433       435     22      16300056        0       39M69H  =       36466364 2
0166378 CACTCTCTCTCGCTCTCTCACTCTCTCTCTCTCTCTC '%%%%$%(,.$&&%(*9+$%'%4<@)$$.;5&@:+$5(. AS
:i:32   NM:i:2  OQ:Z:'%%%%.%(,.8&>%;*9+;,;%4<@):6C;D;@:B7C(9    RG:Z:ERR018525  SA:Z:22,36
466074,+,60M48S,0,0;    XS:i:28
ERR013170.4630188       97      22      16850138        5       29S50M29S       =       36
809232  19959202        AAATGGAATCGAATGGAATTATCGAATGCAATCGAATGGAATTATCGAATGCAATCGAATAGAATC
ATCGAATGGACTCGAATGACCCCTGGGGTAAGGAGAAGCCCA      A:=;:9:9;:1:<;;9:<;<:;:&91;:9;;::28;3976:;
;3:6.49.8/0487,-68610704223(/5331.-32+05355//4)50/42)151316665665/      AS:i:40 NM:i:2  OQ
:Z:ACECGHJJGI?KJHFIKKHIJII?LHIIJLKIJ@LKHJHLKIIHIKALKFJIKKIK?GJIJILKGKJG=;KKGGBJCHA;FBCEF<F
@JGC=CB6B?@B?BC?B;<;@   RG:Z:ERR013170  XS:i:36
```

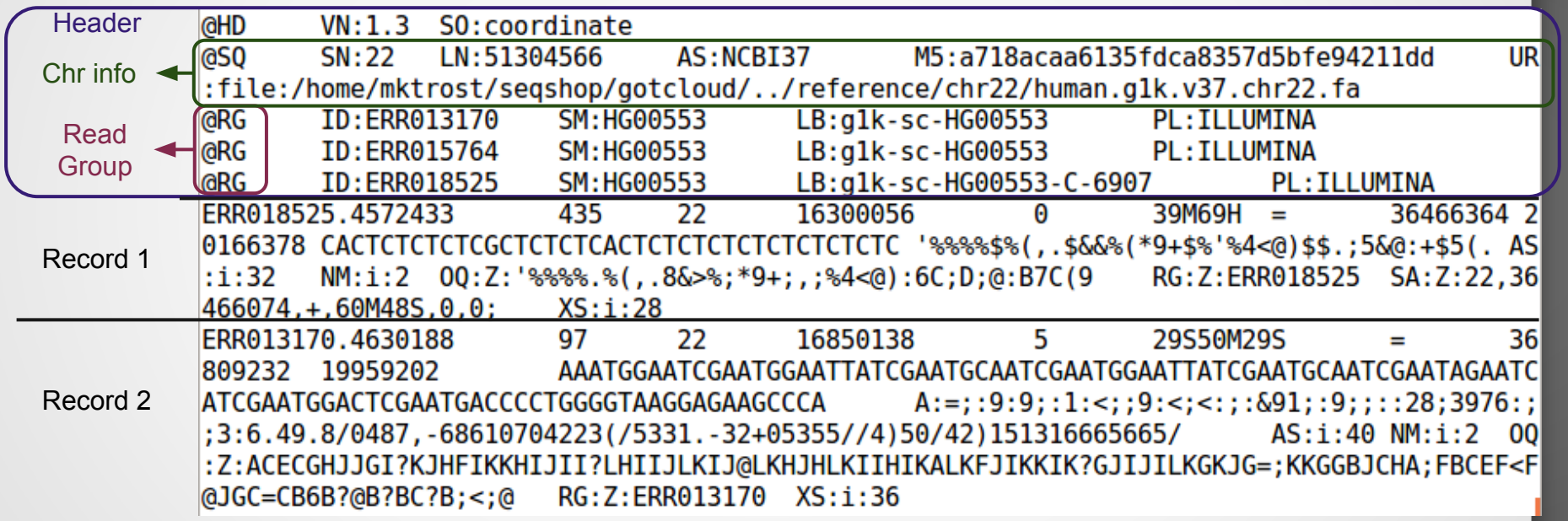
# SAM/BAM Records



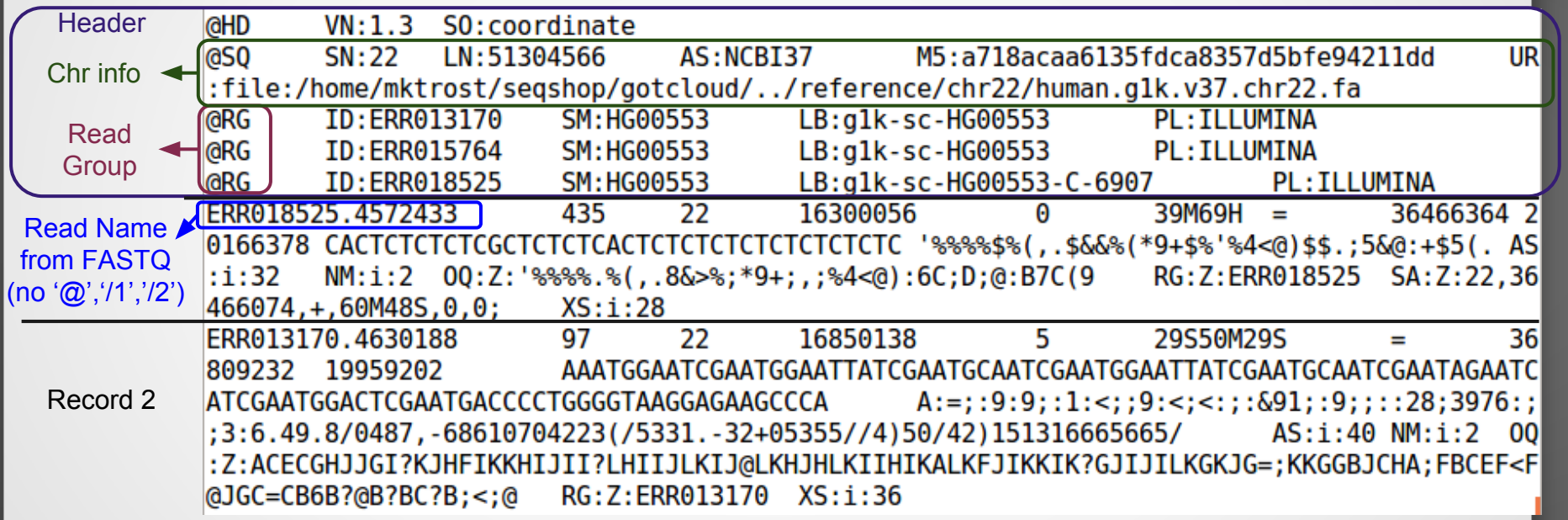
# SAM/BAM Records



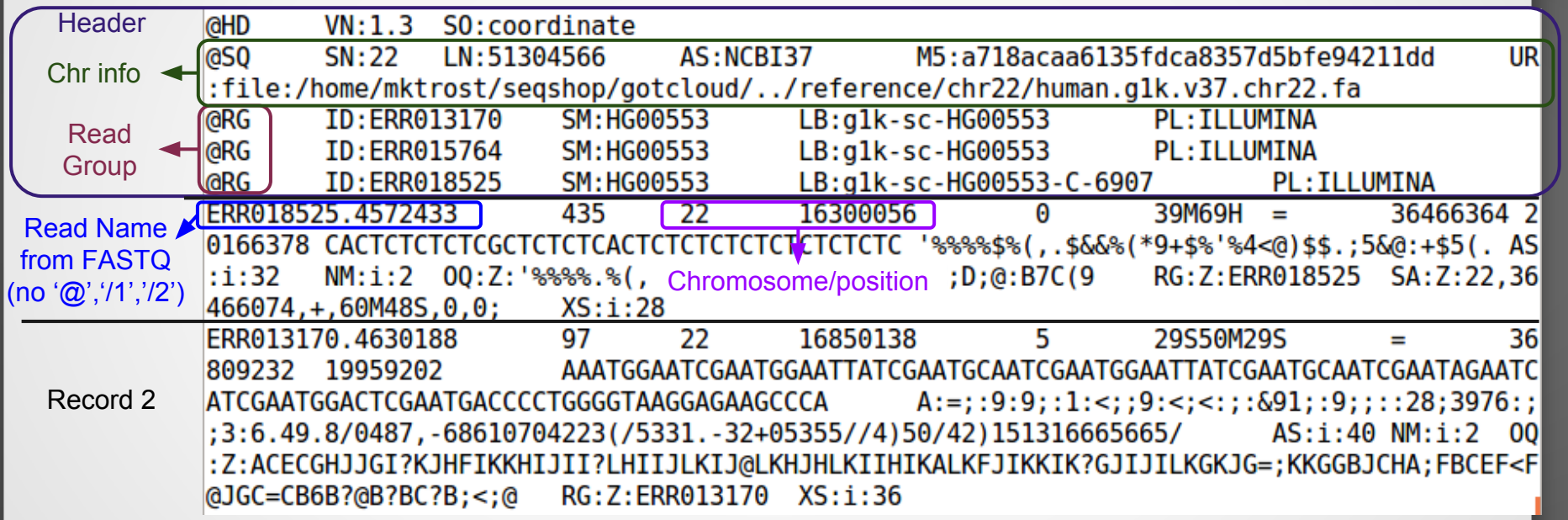
# SAM/BAM Records



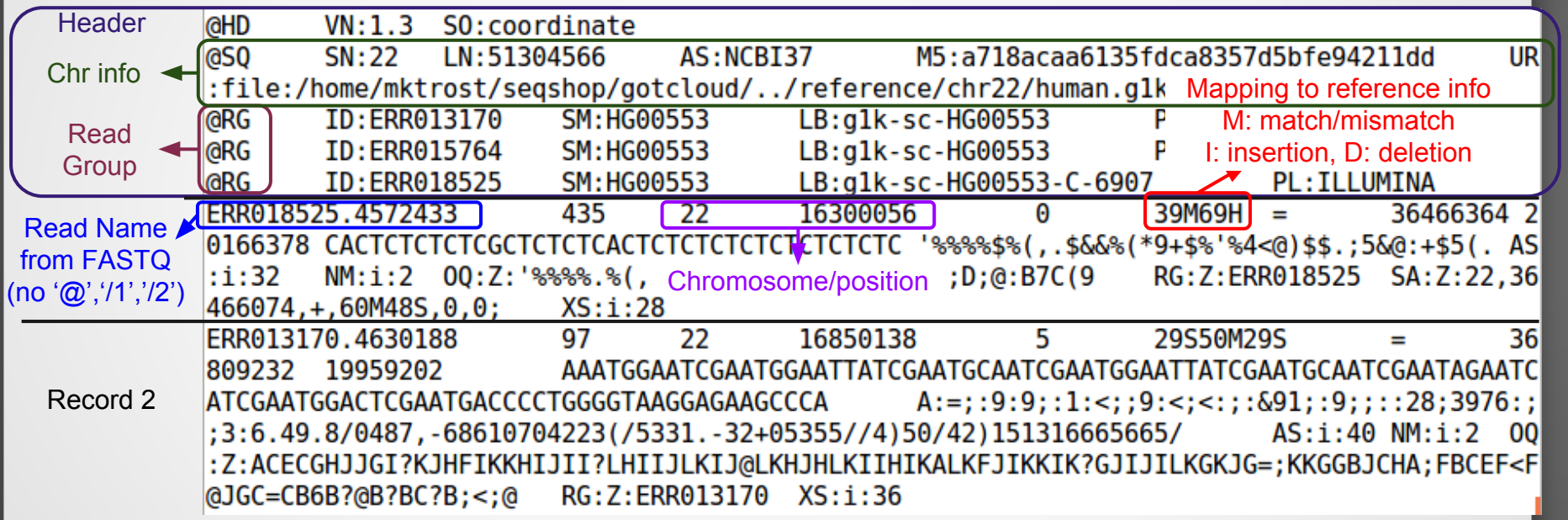
# SAM/BAM Records



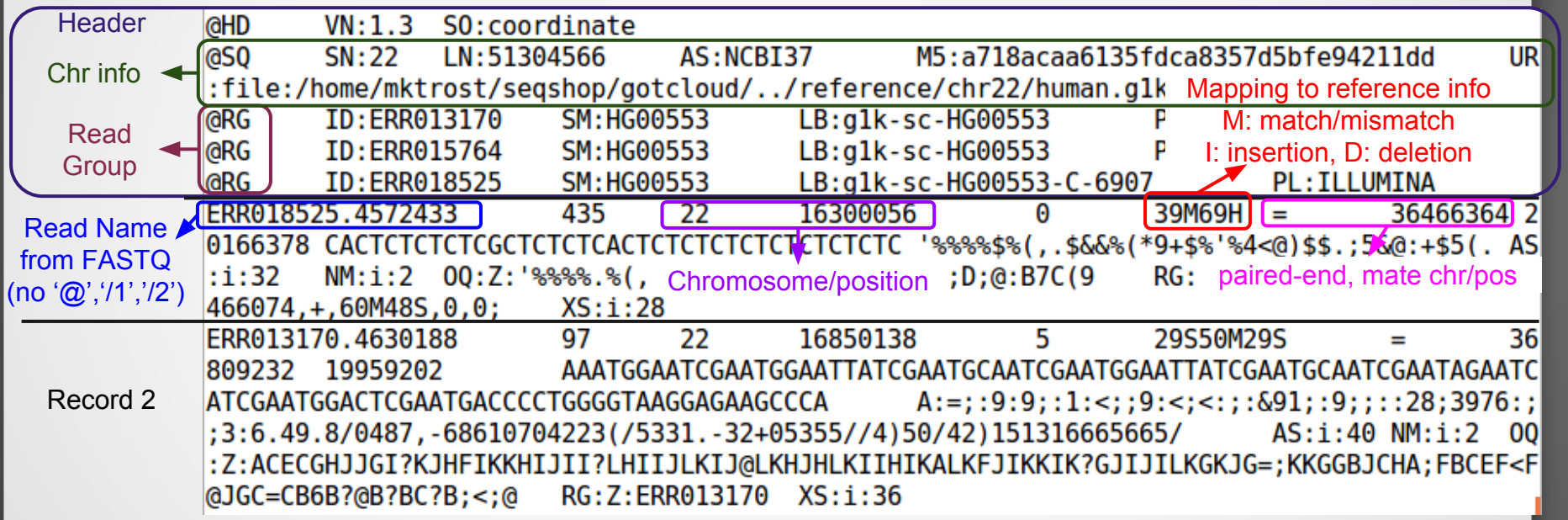
# SAM/BAM Records



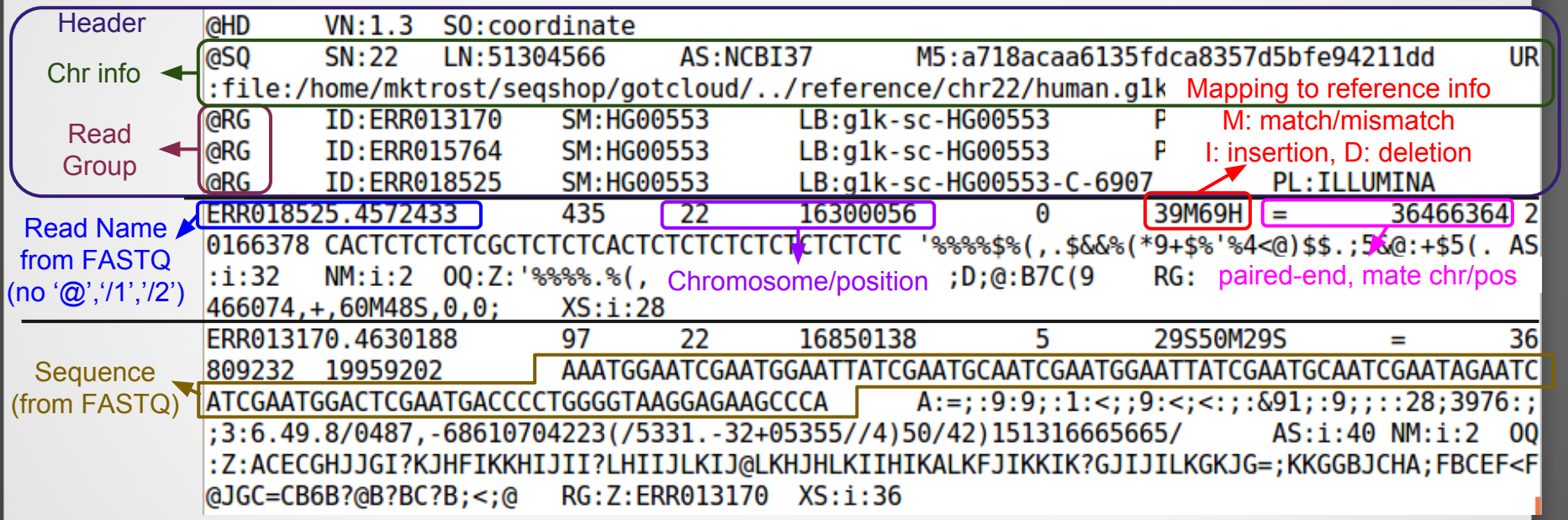
# SAM/BAM Records



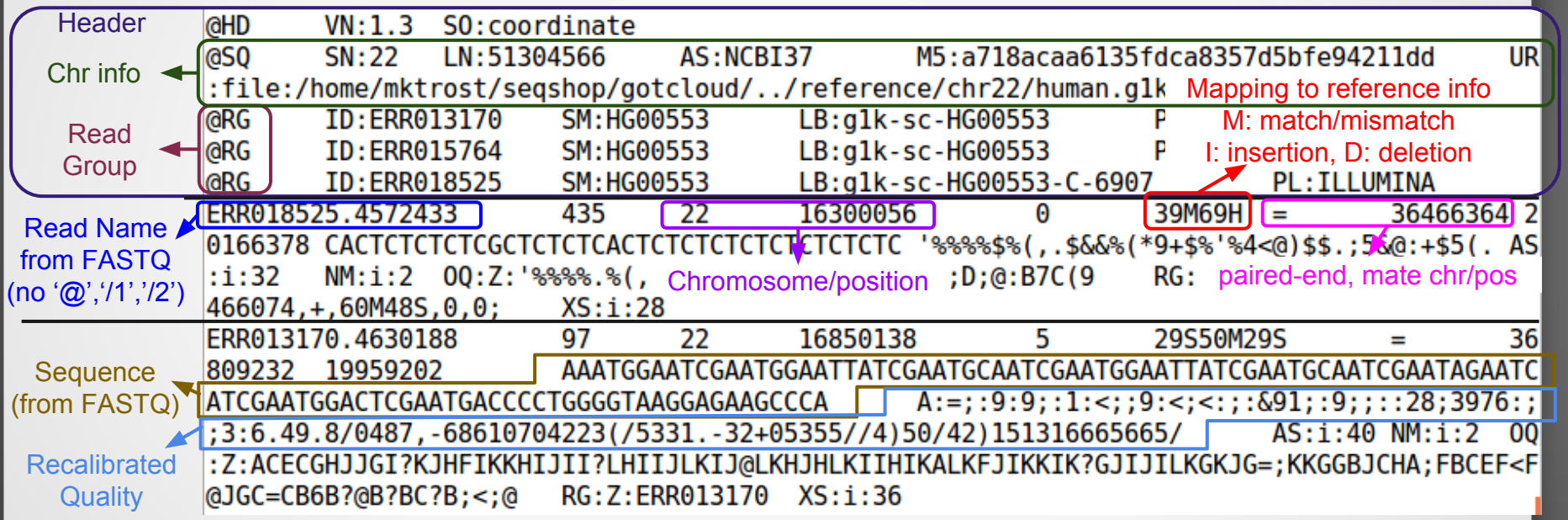
# SAM/BAM Records



# SAM/BAM Records



# SAM/BAM Records



# Viewing SAM/BAM Files

- Samtools *- what we will use today*
  - http://samtools.sourceforge.net/
  - view
    - read group, library, MAPQ >, region
  - tview
    - text alignment viewer - visualize reads by position
- BamUtil (Michigan tool)
  - http://genome.sph.umich.edu/wiki/BamUtil
  - Lot's of SAM/BAM tools
- Other tools
  - Picard, GATK, BamTools

# High Quality BAMs from FASTQs

1. Map FASTQ reads to reference genome
   - Identify most likely chromosome/position for each read
2. Remove Duplicates
   - Remove duplicates so they are not counted multiple times as evidence for variant
3. Recalibrate Base Qualities
   - Improve qualities from sequencer

# High Quality BAMs from FASTQs

- Many tools & best practices to choose from
- Our solution:

## Genomes on the Cloud (GotCloud)

  - Sequence analysis pipelines
    - You don't need to know the details of individual components
    - Automates steps for you

# What is Genomes On The Cloud?

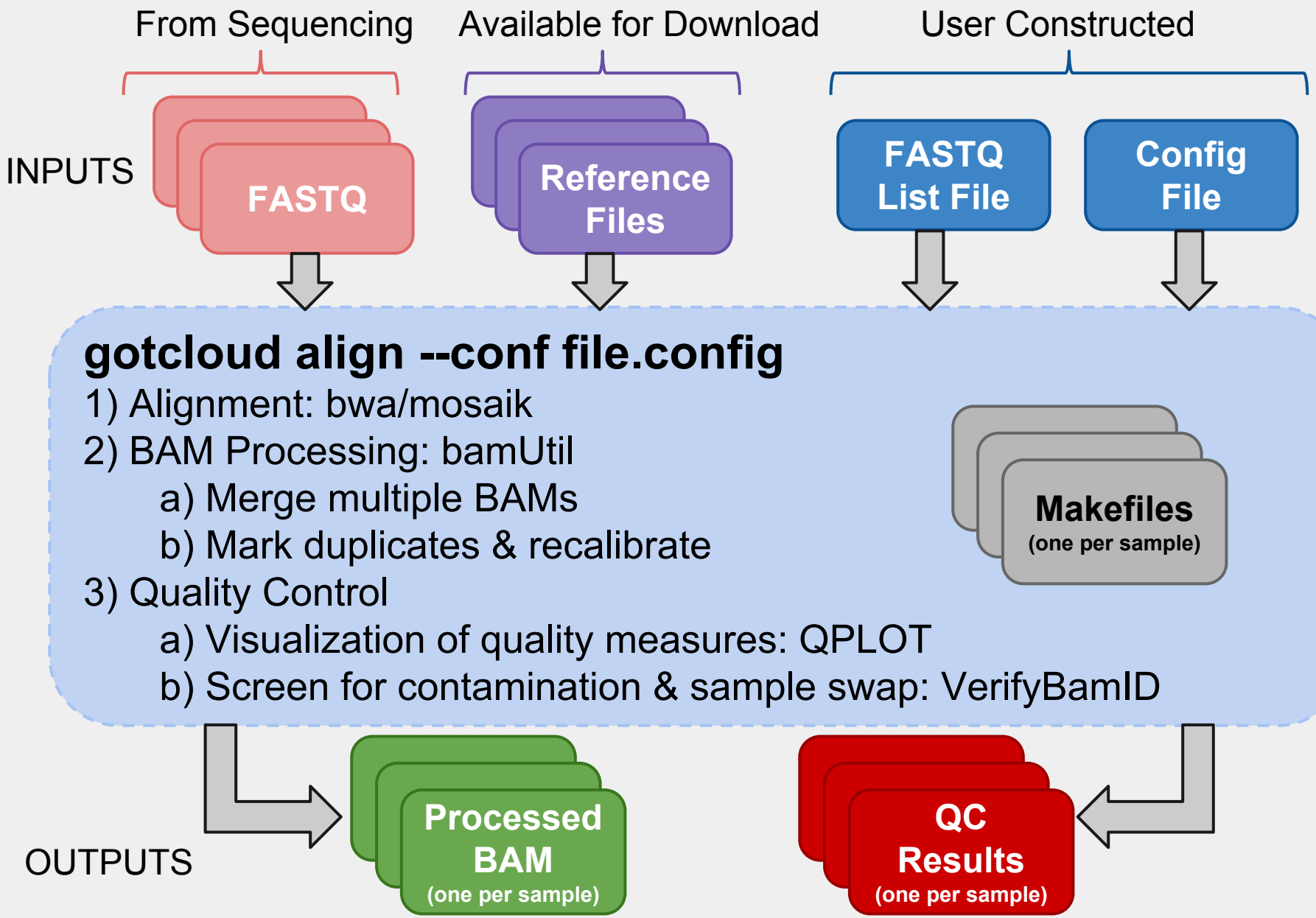
- ***Integrative*** **Alignment**, **QC**, Variant Calling, Phasing
- ***Seamless*** Requires only simple configuration files
- ***Robust*** ..against unexpected failures & stops
- ***Scalable*** ..to many thousands of genomes

- GotCloud also provides
  - Set of many useful software tools
  - Software library (C++) for sequence analysis

# GotCloud Pipelines

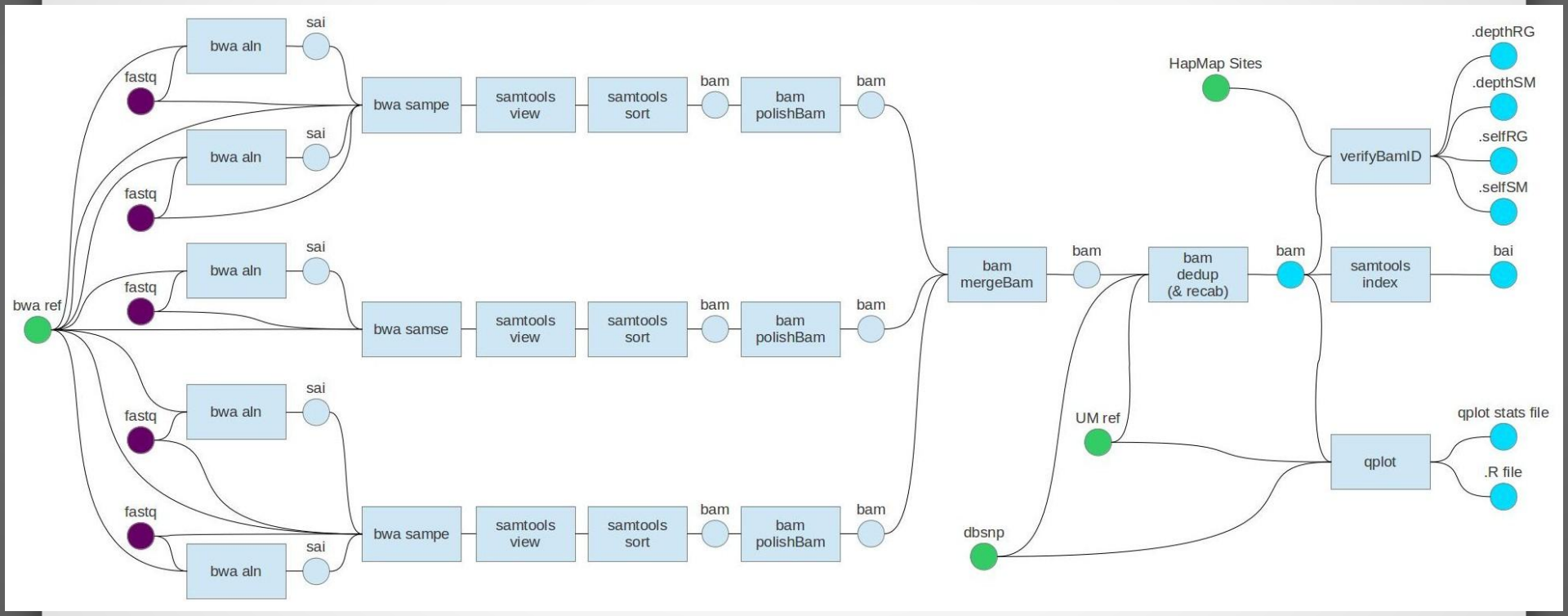
- Robust parallelization
  - Automatically partitions
    - Alignment: by sample
    - Others: by region
  - Takes advantage of clusters
    - Supports MOSIX, SLURM, SGE, PBS
    - Can setup a cluster on Amazon
      - Amazon Machine Image (AMI) available
  - via GNU make
    - Reliable and fault-tolerant
    - Restart where it stopped

# GotCloud Alignment Pipeline Overview



# What are all of the steps in GotCloud align?

For each sample (using BWA):




Aren’t you glad you don’t have to worry about and manage each step?

# User Created Input: FASTQ List 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 Created Input: FASTQ List File

* Spacing adjusted for easier reading
(just a single tab between columns)

Header Row

| SAMPLE | FASTQ1 | FASTQ2 |
|---|---|---|
| HG00551 | fastq/HG00551.SRR190851.fastq | . |
| HG00551 | fastq/HG00551.SRR190851_1.fastq | fastq/HG00551.SRR190851_2.fastq |
| HG00553 | fastq/HG00553.ERR013170.fastq | . |
| HG00553 | fastq/HG00553.ERR013170_1.fastq | fastq/HG00553.ERR013170_2.fastq |
| HG00553 | fastq/HG00553.ERR015764.fastq | . |
| HG00553 | fastq/HG00553.ERR015764_1.fastq | fastq/HG00553.ERR015764_2.fastq |
| HG00553 | fastq/HG00553.ERR018525.fastq | . |
| HG00553 | fastq/HG00553.ERR018525_1.fastq | fastq/HG00553.ERR018525_2.fastq |
| HG00640 | fastq/HG00640.ERR013174.fastq | . |
| HG00640 | fastq/HG00640.ERR013174_1.fastq | fastq/HG00640.ERR013174_2.fastq |
| HG00640 | fastq/HG00640.ERR015768.fastq | . |
| HG00640 | fastq/HG00640.ERR015768_1.fastq | fastq/HG00640.ERR015768_2.fastq |
| HG00640 | fastq/HG00640.ERR018527.fastq | . |
| HG00640 | fastq/HG00640.ERR018527_1.fastq | fastq/HG00640.ERR018527_2.fastq |
| HG00641 | fastq/HG00641.SRR069531.fastq | . |
| HG00641 | fastq/HG00641.SRR069531_1.fastq | fastq/HG00641.SRR069531_2.fastq |

# User Created Input: FASTQ List File

* Spacing adjusted for easier reading
(just a single tab between columns)

Header Row

| SAMPLE | FASTQ1 | FASTQ2 |
|---|---|---|
| HG00551 | fastq/HG00551.SRR190851.fastq | . |
| HG00551 | fastq/HG00551.SRR190851_1.fastq | fastq/HG00551.SRR190851_2.fastq |
| HG00553 | fastq/HG00553.ERR013170.fastq | . |
| HG00553 | fastq/HG00553.ERR013170_1.fastq | fastq/HG00553.ERR013170_2.fastq |
| HG00553 | fastq/HG00553.ERR015764.fastq | . |
| HG00553 | fastq/HG00553.ERR015764_1.fastq | fastq/HG00553.ERR015764_2.fastq |
| HG00553 | fastq/HG00553.ERR018525.fastq | . |
| HG00553 | fastq/HG00553.ERR018525_1.fastq | fastq/HG00553.ERR018525_2.fastq |
| HG00640 | fastq/HG00640.ERR013174.fastq | . |
| HG00640 | fastq/HG00640.ERR013174_1.fastq | fastq/HG00640.ERR013174_2.fastq |
| HG00640 | fastq/HG00640.ERR015768.fastq | . |
| HG00640 | fastq/HG00640.ERR015768_1.fastq | fastq/HG00640.ERR015768_2.fastq |
| HG00640 | fastq/HG00640.ERR018527.fastq | . |
| HG00640 | fastq/HG00640.ERR018527_1.fastq | fastq/HG00640.ERR018527_2.fastq |
| HG00641 | fastq/HG00641.SRR069531.fastq | . |
| HG00641 | fastq/HG00641.SRR069531_1.fastq | fastq/HG00641.SRR069531_2.fastq |

Groups all FASTQs for a sample in a single BAM

# User Created Input: FASTQ List File

* Spacing adjusted for easier reading
(just a single tab between columns)

Header Row

| SAMPLE | FASTQ1 | FASTQ2 |
|---|---|---|
| HG00551 | fastq/HG00551.SRR190851.fastq | . |
| HG00551 | fastq/HG00551.SRR190851_1.fastq | fastq/HG00551.SRR190851_2.fastq |
| HG00553 | fastq/HG00553.ERR013170.fastq | . |
| HG00553 | fastq/HG00553.ERR013170_1.fastq | fastq/HG00553.ERR013170_2.fastq |
| HG00553 | fastq/HG00553.ERR015764.fastq | . |
| HG00553 | fastq/HG00553.ERR015764_1.fastq | fastq/HG00553.ERR015764_2.fastq |
| HG00553 | fastq/HG00553.ERR018525.fastq | . |
| HG00553 | fastq/HG00553.ERR018525_1.fastq | fastq/HG00553.ERR018525_2.fastq |
| HG00640 | fastq/HG00640.ERR013174.fastq | . |
| HG00640 | fastq/HG00640.ERR013174_1.fastq | fastq/HG00640.ERR013174_2.fastq |
| HG00640 | fastq/HG00640.ERR015768.fastq | . |
| HG00640 | fastq/HG00640.ERR015768_1.fastq | fastq/HG00640.ERR015768_2.fastq |
| HG00640 | fastq/HG00640.ERR018527.fastq | . |
| HG00640 | fastq/HG00640.ERR018527_1.fastq | fastq/HG00640.ERR018527_2.fastq |
| HG00641 | fastq/HG00641.SRR069531.fastq | . |
| HG00641 | fastq/HG00641.SRR069531_1.fastq | fastq/HG00641.SRR069531_2.fastq |

Groups all FASTQs for a sample in a single BAM

Multiple FASTQs for 1 sample

# User Created Input: FASTQ List File

* Spacing adjusted for easier reading
(just a single tab between columns)

Header Row

| SAMPLE | FASTQ1 | FASTQ2 |
|---|---|---|
| HG00551 | fastq/HG00551.SRR190851.fastq | . |
| HG00551 | fastq/HG00551.SRR190851_1.fastq | fastq/HG00551.SRR190851_2.fastq |
| HG00553 | fastq/HG00553.ERR013170.fastq | . |
| HG00553 | fastq/HG00553.ERR013170_1.fastq | fastq/HG00553.ERR013170_2.fastq |
| HG00553 | fastq/HG00553.ERR015764.fastq | . |
| HG00553 | fastq/HG00553.ERR015764_1.fastq | fastq/HG00553.ERR015764_2.fastq |
| HG00553 | fastq/HG00553.ERR018525.fastq | . |
| HG00553 | fastq/HG00553.ERR018525_1.fastq | fastq/HG00553.ERR018525_2.fastq |
| HG00640 | fastq/HG00640.ERR013174.fastq | . |
| HG00640 | fastq/HG00640.ERR013174_1.fastq | fastq/HG00640.ERR013174_2.fastq |
| HG00640 | fastq/HG00640.ERR015768.fastq | . |
| HG00640 | fastq/HG00640.ERR015768_1.fastq | fastq/HG00640.ERR015768_2.fastq |
| HG00640 | fastq/HG00640.ERR018527.fastq | . |
| HG00640 | fastq/HG00640.ERR018527_1.fastq | fastq/HG00640.ERR018527_2.fastq |
| HG00641 | fastq/HG00641.SRR069531.fastq | . |
| HG00641 | fastq/HG00641.SRR069531_1.fastq | fastq/HG00641.SRR069531_2.fastq |

'.' means single-end
filename means 2nd in pair

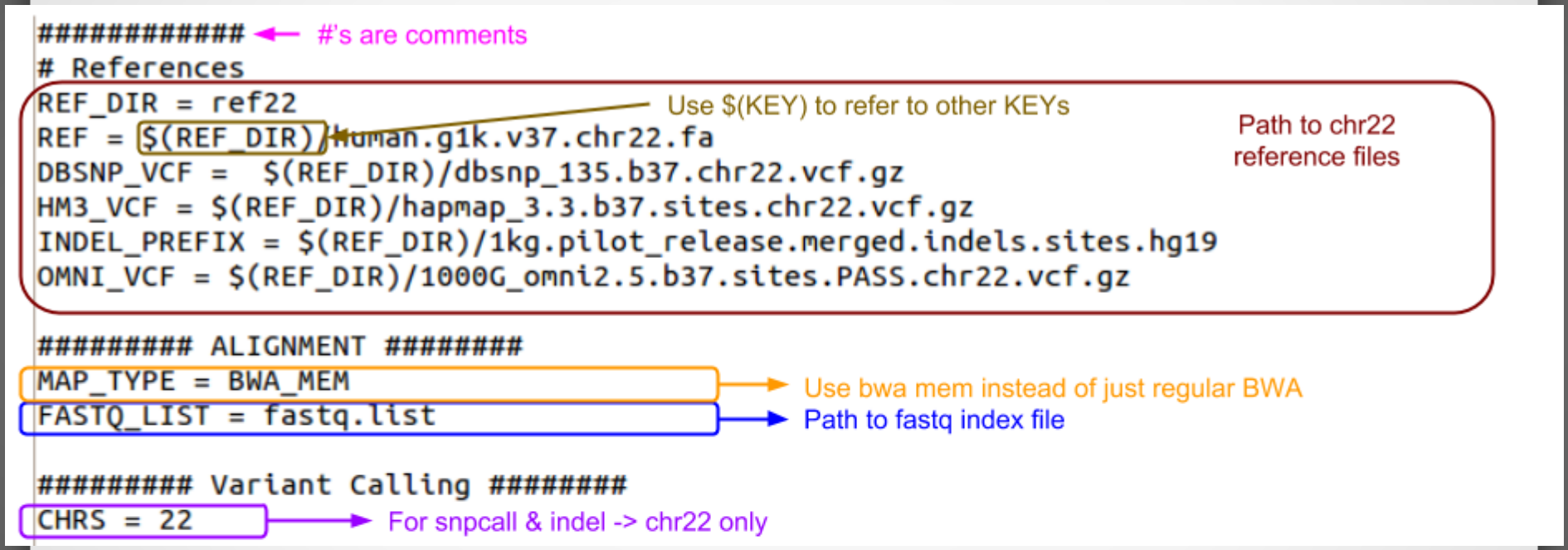
Groups all FASTQs for a sample in a single BAM

Multiple FASTQs for 1 sample

# FASTQ List File: Optional RG Fields

| Field Name | Description | Default |
|---|---|---|
| RGID | Read Group ID - unique for each run | Derived from first line of FASTQ *or* incrementing numbers |
| LIBRARY | Separates FASTQs for a sample that were prepared separately | SAMPLE |
| PLATFORM | Sequencing Platform | ILLUMINA |
| CENTER | Sequencing Center<br>Useful if multiple centers | unknown |
| MERGE_NAME | Group FASTQs into a single BAM file<br>Only need if want multiple BAMs for a sample | SAMPLE |

# User Created Input: Configuration



# User Created Input: 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_MASK_FASTA = $(REF_DIR)/human_g1k_v37.chr22.mask.100.fasta
GENOMESTRIP_PLOIDY_MAP = $(REF_DIR)/humgen_g1k_v37_ploidy.chr22.map
```

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

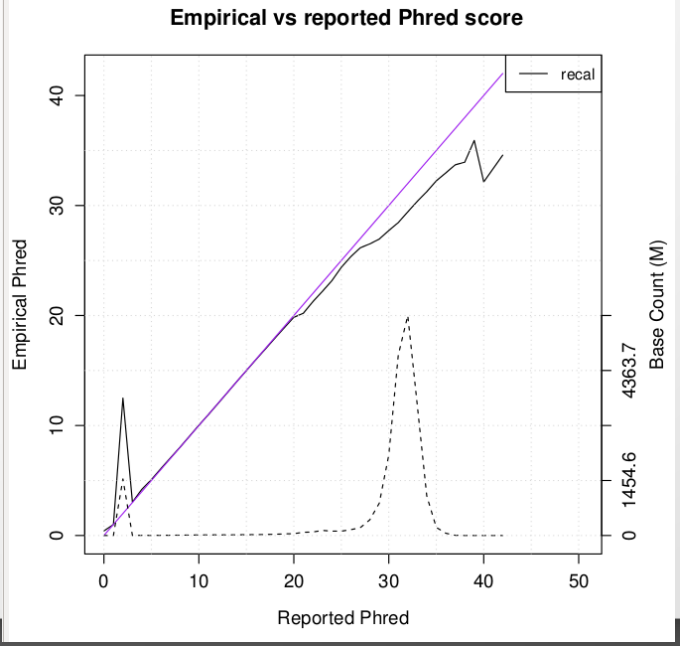
| #SEQ_ID | RG | CHIP_ID | #SNPS | #READS | AVG_DP | FREEMIX | FREELK1 | FREELK0 | FREE_RH | FREE_RA | CHIPMIX |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HG00551 | ALL | NA | 20056 | 3654 | 0.18 | 0.00000 | 955.99 | 955.99 | NA | NA | NA |

# 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(%)   0.52
MapQual<10(%)    0.67
PairedReads(%)   98.91
ProperPaired(%) 86.43
MappedBases(e9) 0.01
Q20Bases(e9)     0.01
Q20BasesPct(%)   89.10
MeanDepth        7.44
```



# 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