Practical Session 5

GotCloud Alignment with BWA

Jan 7th, 2014

STOM 2014 Workshop

Hyun Min Kang

University of Michigan, Ann Arbor

GOAL OF THIS SESSION

- Assuming that
  - The audiences know how to perform GWAS
  - But have very little experience in sequence data
- We want to learn..
  - Practical & detailed procedure of DNA sequencing
  - How to understand the basic file formats for the sequence data (FASTQ, BAM)
  - How to visualize sequence data to examine the reads aligned to particular genomic position
  - How to process the raw sequence reads to aligned sequences that are ready for variant calling
  - How to evaluate the quality of sequence data after running alignment pipeline

What the Raw Sequence Reads Look Like…

- QSEQ : Illumina's proprietary format
- FASTQ : Standard format for shotgun sequence data
  - Four lines per each fragment

```
@ERR3256:326:10:1561/1          (Typically cryptic read name)
AGCTGATAGCTAGCTATCTGACGAGCCCG   (Sequence read)
+                               ("+"-only line)
BGGG?GEGGGGFFGGG:GFF:EEB#####   (Quality scores)
```

  - The phred-scale quality score is represented as ASCII code of (33 + Phred-scale-quality)

What the Raw Sequence Reads Look Like…

- QSEQ : Illumina's proprietary format
- FASTQ : Standard format for shotgun sequence data
  - Four lines per each fragment

```
@ERR3256:326:10:1561/1          (Typically cryptic read name)
AGCTGATAGCTAGCTATCTGACGAGCCCG   (Sequence read)
+                               ("+"-only line)
BGGG?GEGGGGFFGGG:GFF:EEB#####   (Quality scores)
```

  - The phred-scale quality score is represented as ASCII code of (33 + Phred-scale-quality)
  - 32 is a special character ' ' (space), so the lowest quality score (Q=0) is represented as !(ASCII=33)

You Don't Need to Remember ASCII Table..

- Lower qualities are represented as special characters, or digits
  - ! (Q=0), " (Q=1), # (Q=3), + (Q=10), / (Q=14)
  - 0 (Q=15), 5 (Q=20), 9 (Q=24)
- Higher quality (>Q30) are mostly represented as upper case alphabets
  - : (Q=25), ? (Q=30), @ (Q=31)
  - A (Q=32), B (Q=33), G (Q=38)
- So

  AGCTGATAGCTAGCTATCTGACGAGCCCG

  BGGG?GEGGGGFFGGG:GFF:EEB#####

  AGCTgATAGCTAGCTAtCTGaCGAgcccg

Paired-End and Single-End FASTQs

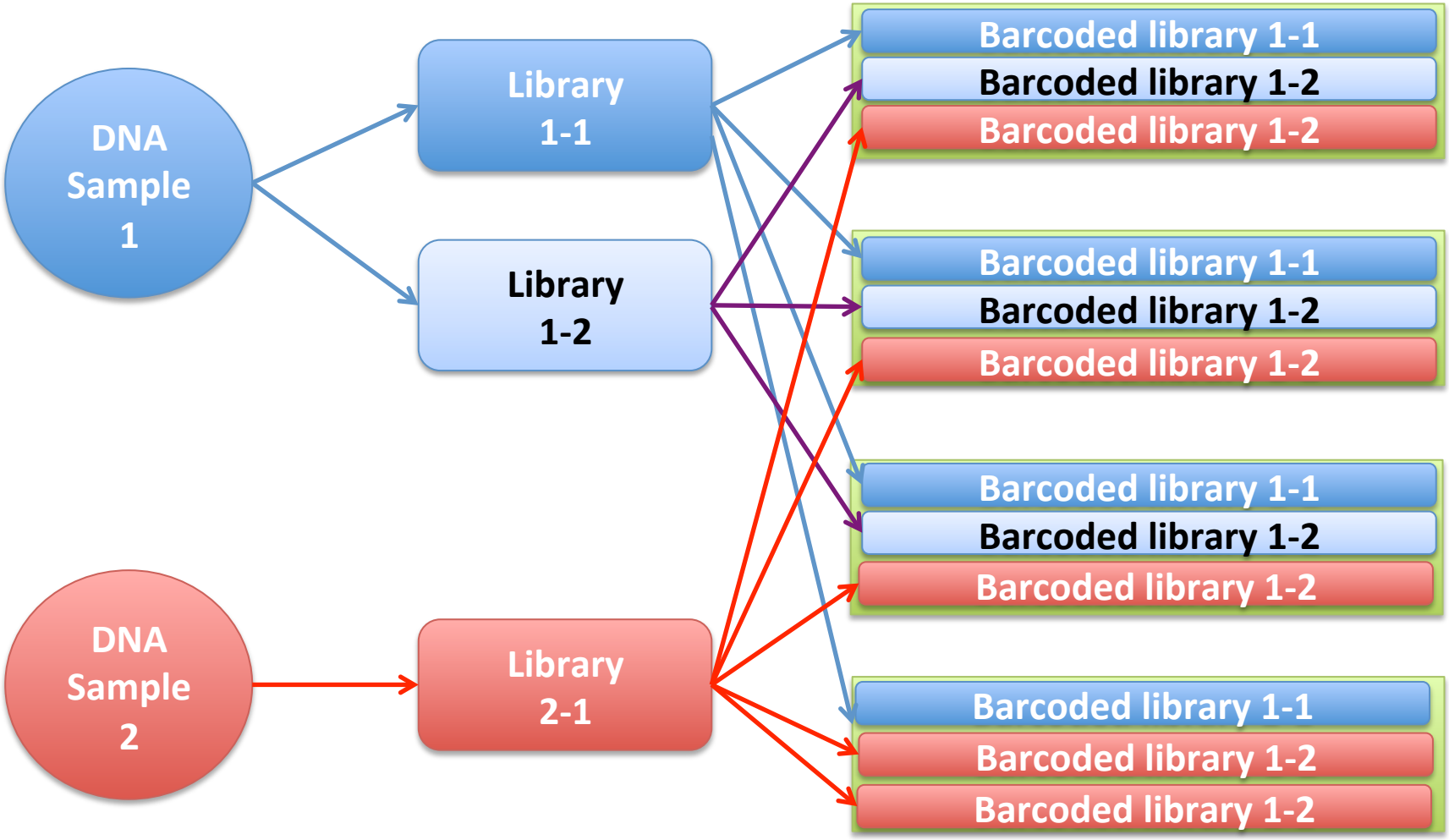
- Paired-end FASTQ files have corresponding pairs

```
% ls $S5/examples/fastq
SRR035022_1.fastq.gz  SRR035022_2.fastq.gz
SRR035024.fastq.gz    ...
```

- But some FASTQs may be single ended

- Q : Should paired-end and single-end reads be mapped differently?

Typical Sequencing Experiments Today

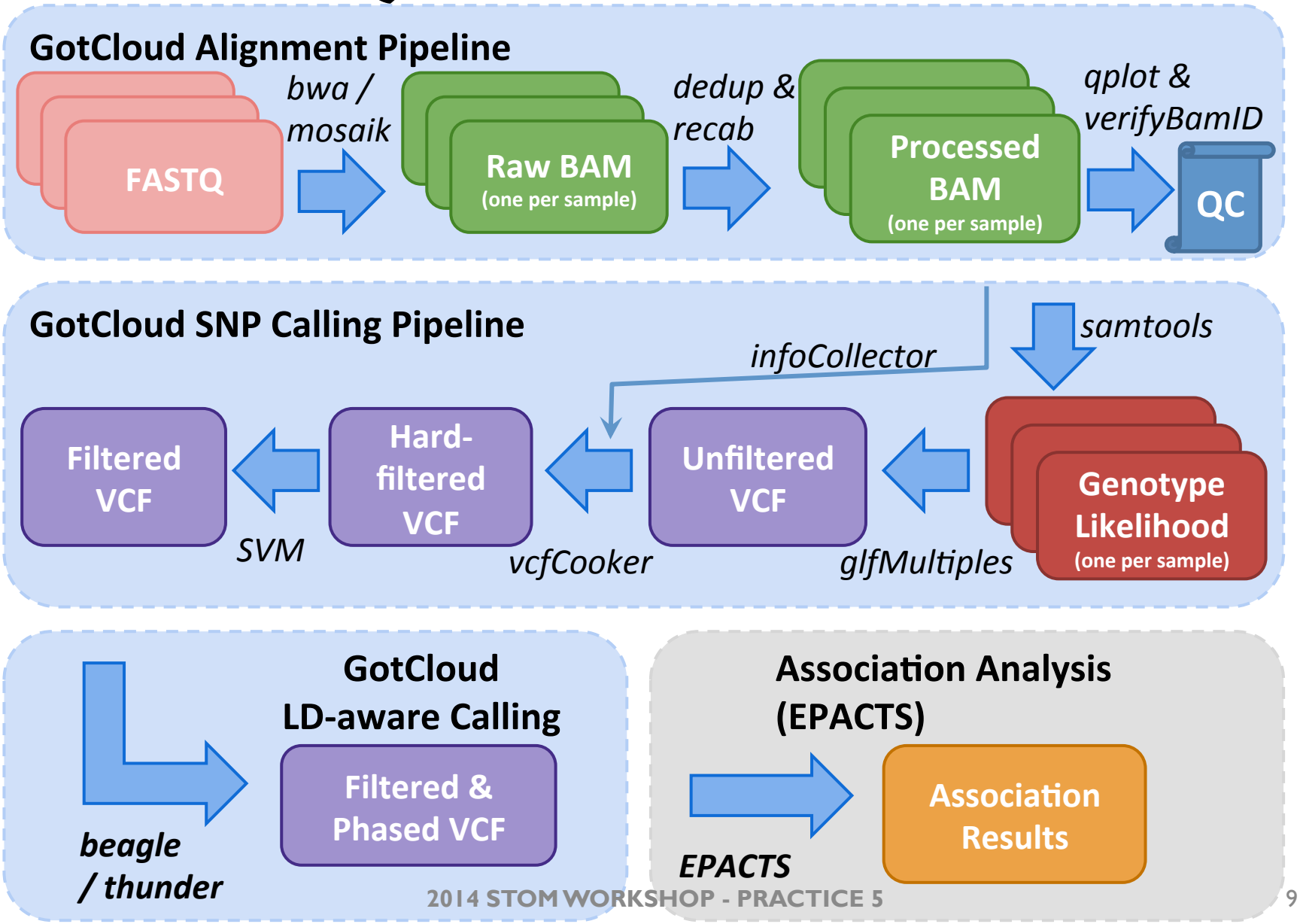


Multiplexed samples are typically de-multiplexed and de-barcoded at sequencing center

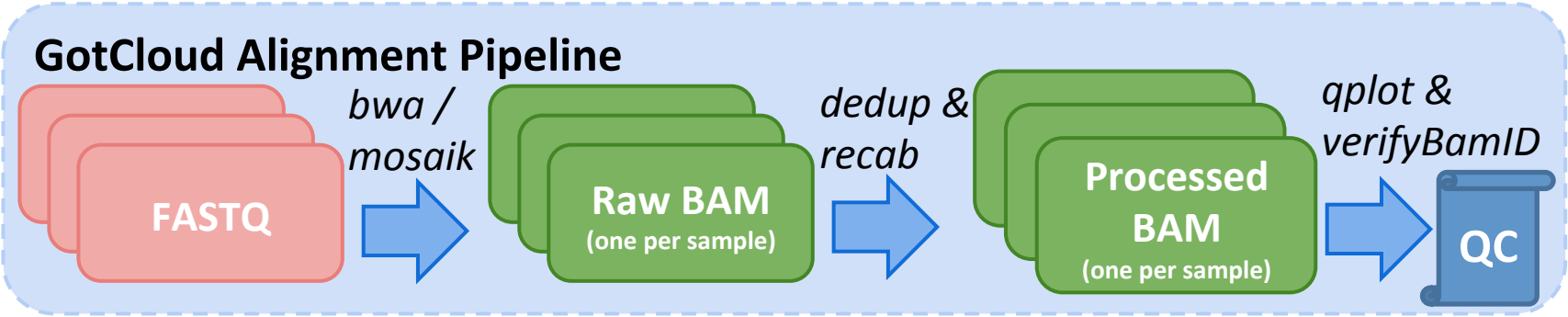
Typical Sequencing Experiments Today

- What are the advantages and disadvantages of using multiple library preparation of the same sample?
- What are the advantages and disadvantages of barcoding the library and multiplex with other samples across multiple lanes?
- What would happen if there are errors during de-multiplexing?

GenomesOnTheCloud (GotCloud) Sequence Analysis Pipeline



GotCloud Alignment Pipeline



WHY GOTCLOUD?

- Easy to learn & run
  - All-in-one package for sequence analysis pipeline
  - You don't have to know the details of individual component
- Robust parallelization
  - Automatic partition of multi-sample jobs for parallelization
  - Reliable and fault-tolerant parallelization via GNU make
    - Restart from where it stopped upon unexpected crash
- Cloud & Cluster-friendly
  - Supports multiple clusters such as MOSIX, Slurm& SGE
  - Amazon instances allows running large-scale jobs without having your own cluster

Getting Started With Example Dataset

- Examples from the 1000 Genomes Project…
  - Even for one sample, FASTQ files are huge
    - >10GB per sample, ~1 day for finishing the alignment
    - Not feasible for simultaneous
  - For the workshop, small FASTQ files are prepared
    - Most sequence reads should be mapped to a specific genomic region around *CFTR*
- We want to learn..
  - How to translate the raw sequence reads to the aligned, and post-processed sequenced, ready to analyze

Setting Up Computing Environment

- To see the files for the session, type

```
ls /data/stom2014/session5/
```

  – If you see any errors, please let me know now!

- For convenience, let's set some variables

```
export S5=/data/stom2014/session5
mv ~/out ~/out_session2
mkdir ~/out
```

Input Files For This Session

- Check if FASTQ files are accessible
  `ls $S5/examples/`

- Inputs needed for `gotcloud align`
  - FASTQ files under `$S5/examples/fastq`
  - Reference genome and indices (chr7 only)
    `$S5/examples/chr7Ref`
  - Index of FASTQ

OVERVIEW OF GOTCLOUD EXAMPLES

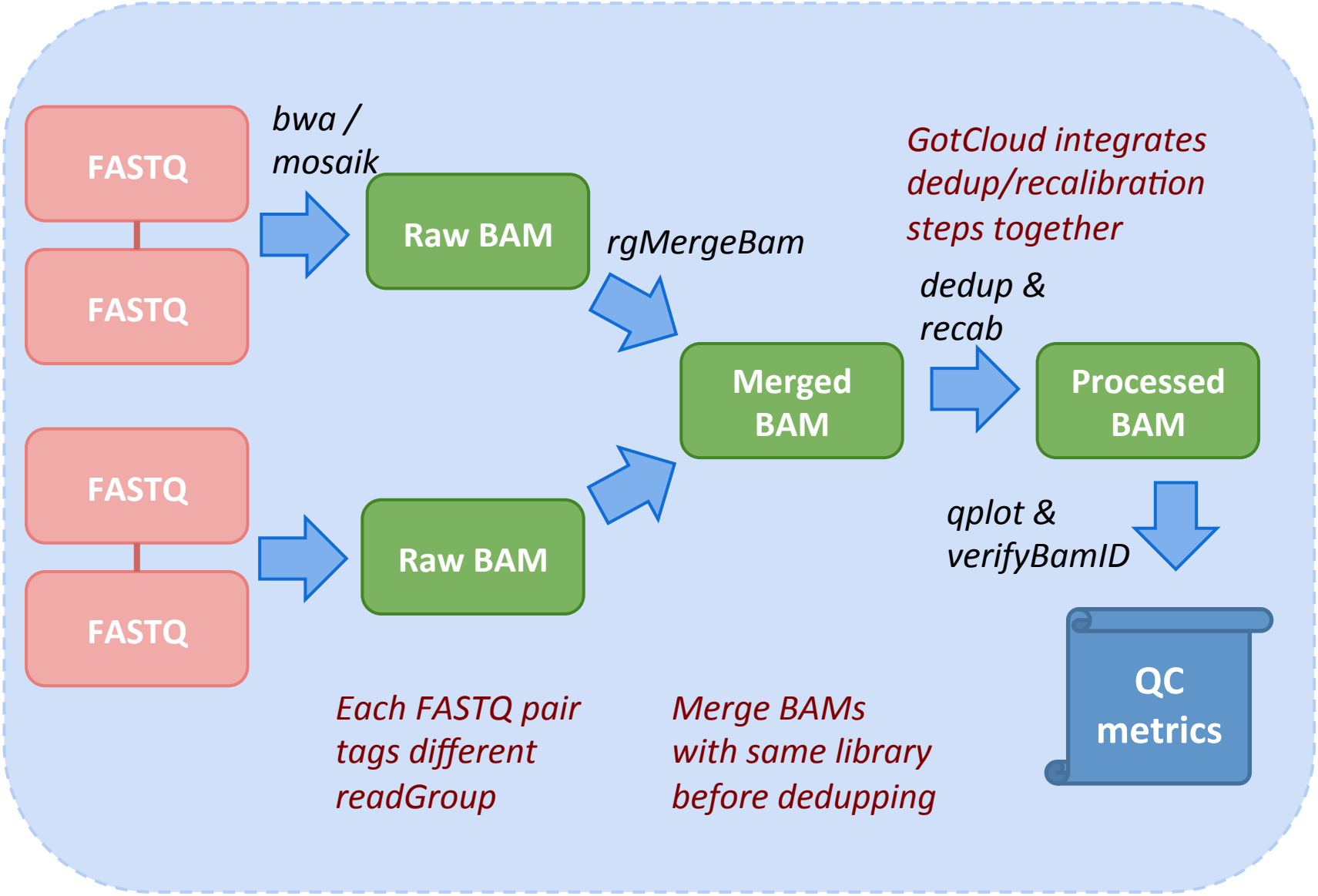
```
% ls $S5/examples
  chr7Ref      (Partial reference genome for the
example)
  fastq        (FASTQ files from 1000G)
  index        (Contains index and configuration files)
```

Note that the all example files are derived from the 1000 Genomes project, focusing on the 500kb region around CFTR

Questions before processing sequence reads

- Are FASTQ files separated by sample or by lane?
  - If multiple samples are sequenced in one lane, make sure that they are de-multiplexed.
- Single-ended or paired end sequencing?
- Was each sample prepared by single library prep?
  - If there are multiple library preps per sample, duplicate removal between different libraries is unnecessary.
- Are all samples processed with the same pipeline?
  - If different pipeline is used for each different sample, batch effect may arise due to pipeline differences
  - Use the same configuration of pipeline within the same project.

GotCloud Align : A little more detailed view



You need to provide input data..

- Input for **`gotcloud align`** requires at least..
  - The full path of your FASTQ files
  - Sample name corresponding to each FASTQ file
    - Each sample may have multiple FASTQ file
    - Each FASTQ file must belong only to one sample
- And you need to create an index file
  - **`MERGE_NAME`** for sample name
  - **`FASTQ1`** for first-pair (or single-ended) FASTQ
  - **`FASTQ2`** for second-pair FASTQ (or empty if single-ended)

Useful columns for the index file (optional)

- Add information for each set of sequence reads
  - **`RGID`** : Assign specific name for each FASTQ
  - **`SAMPLE`** : Sample ID
    - If empty, **`MERGE_NAME`** will be used.
  - **`LIBRARY`** : Library name
    - Important for dedupping
    - Different libraries
  - **`CENTER`** : Sequencing center name (e.g. UWCMG)
  - **`PLATFORM`** : Sequencing Platform (e.g. ILLUMINA)

ALIGNMENT: PREPARING INDEX FILE

```
% cat $S5/examples/index/chr7.CFTR.fastq.index
```

MERGE_NAME	FASTQ1	FASTQ2	RGID	SAMPLE	LIBRARY	CENTER	PLATFORM	
NA06984	fastq/SRR035022.fastq.gz		.	SRR035022		NA06984	Solexa-16556	BI
NA06984	fastq/SRR035022_1.fastq.gz	fastq/SRR035022_2.fastq.gz		SRR035022				NA
NA06984	fastq/SRR035023.fastq.gz		.	SRR035023		NA06984	Solexa-16556	BI
NA06984	fastq/SRR035023_1.fastq.gz	fastq/SRR035023_2.fastq.gz		SRR035023				NA
NA06984	fastq/SRR035024.fastq.gz		.	SRR035024		NA06984	Solexa-16556	BI
NA06984	fastq/SRR035024_1.fastq.gz	fastq/SRR035024_2.fastq.gz		SRR035024				NA
NA06984	fastq/SRR035025.fastq.gz		.	SRR035025		NA06984	Solexa-16556	BI
NA06984	fastq/SRR035025_1.fastq.gz	fastq/SRR035025_2.fastq.gz		SRR035025				NA
NA06984	fastq/SRR035026.fastq.gz		.	SRR035026		NA06984	Solexa-16556	BI
NA06984	fastq/SRR035026_1.fastq.gz	fastq/SRR035026_2.fastq.gz		SRR035026				NA
NA06984	fastq/SRR035027.fastq.gz		.	SRR035027		NA06984	Solexa-16556	BI
NA06984	fastq/SRR035027_1.fastq.gz	fastq/SRR035027_2.fastq.gz		SRR035027				NA
NA06984	fastq/SRR035669.fastq.gz		.	SRR035669		NA06984	Solexa-16556	BI
NA06984	fastq/SRR035669_1.fastq.gz	fastq/SRR035669_2.fastq.gz		SRR035669				NA
NA12878	fastq/SRR622461.fastq.gz		.	SRR622461		NA12878	Illumina_NA12878	
NA12878	fastq/SRR622461_1.fastq.gz	fastq/SRR622461_2.fastq.gz		SRR622461				NA

The index file contains map between sample name and FASTQ files

ALIGNMENT: PREPARING CONFIGURATION FILE

```
% cat $S5/examples/index/chr7.CFTR.align.conf

INDEX_FILE = index/chr7.CFTR.fastq.index
####################
# References
REF_DIR = chr7Ref
AS = NCBI37
REF = $(REF_DIR)/hs37d5.chr7.fa
DBSNP_VCF =  $(REF_DIR)/dbsnp_135.b37.chr7.CFTR.vcf.gz
HM3_VCF = $(REF_DIR)/hapmap_3.3.b37.sites.chr7.CFTR.vcf.gz
```

Configuration file contains the reference to the index file, reference genome, and other resources

Running GotCloud Alignment Pipeline

```
% $S5/gotcloud/gotcloud align
  --conf $S5/examples/index/chr7.CFTR.align.conf
  --outDir ~/out/align --baseprefix $S5/examples
```

```
File sizes of 32 FASTQ input files referenced in '/data/stom2014/session5/examples/index/chr7
Size of BAMs from aligner will be about 0.01 GB
Intermediate files from snpcaller will be about 0.01 GB
Final VCF output from snpcaller will be about 0.00 GB
Be sure you have enough space to hold all this data
Created /home/hmkang/out/align/Makefiles/align_NA06984.Makefile
Created /home/hmkang/out/align/Makefiles/align_NA12878.Makefile
-----------------------------------------------------------------
Waiting while samples are processed...
Processing finished in 45 secs with no errors reported
36.28user 7.55system 0:45.51elapsed 96%CPU (0avgtext+0avgdata 1016784maxresident)k
0inputs+87648outputs (0major+4394628minor)pagefaults 0swaps
```

Outcome of alignment pipeline

- Aligned & processed sequence reads

```
% ls ~/out/align/bams
NA06984.recal.bam          NA06984.recal.bam.bai.done
NA06984.recal.bam.metrics  NA12878.recal.bam
NA12878.recal.bam.bai.done NA12878.recal.bam.metrics
NA06984.recal.bam.bai      NA06984.recal.bam.done
NA06984.recal.bam.qemp     NA12878.recal.bam.bai
NA12878.recal.bam.done     NA12878.recal.bam.qemp
```

- QC & summary statistics

```
% ls ~/out/align/QCFiles/
NA06984.genoCheck.depthRG  NA06984.genoCheck.selfRG
NA06984.qplot.R            NA12878.genoCheck.depthSM
NA12878.genoCheck.selfSM   NA12878.qplot.stats
…
```

PEEKING THE OUTPUT BAMS

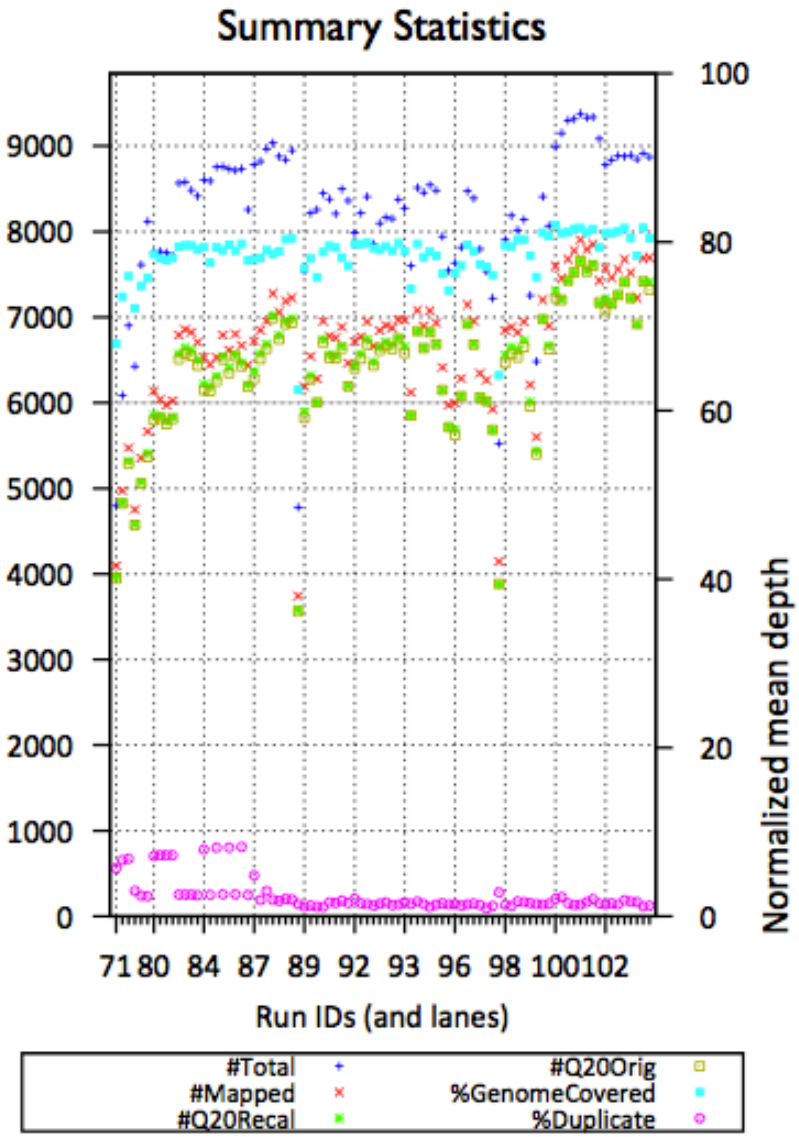
```
% samtools view -h ~/out/align/bams/
NA06984.recal.bam | less
```

```
@SQ     SN:7    LN:159138663    AS:NCBI37       M5:618366e953d6aaad97dbe4777c29375e     UR:file:/data/stom2014/session
5/examples/chr7Ref/hs37d5.chr7.fa
@RG     ID:SRR035022    SM:NA06984      LB:Solexa-16556 CN:BI   PL:ILLUMINA
@PG     ID:bwa  PN:bwa  VN:0.6.1-r104
@RG     ID:SRR035023    SM:NA06984      LB:Solexa-16556 CN:BI   PL:ILLUMINA
@RG     ID:SRR035024    SM:NA06984      LB:Solexa-16556 CN:BI   PL:ILLUMINA
@RG     ID:SRR035025    SM:NA06984      LB:Solexa-16556 CN:BI   PL:ILLUMINA
@RG     ID:SRR035026    SM:NA06984      LB:Solexa-16556 CN:BI   PL:ILLUMINA
@RG     ID:SRR035027    SM:NA06984      LB:Solexa-16556 CN:BI   PL:ILLUMINA
@RG     ID:SRR035669    SM:NA06984      LB:Solexa-16556 CN:BI   PL:ILLUMINA
SRR035669.18390646      163     7       3853391 0       76M     =       3853627 312     CCTCTCCAGCACCTGTTGTTTCCTGACTTT
TTAATGATTGCCATTCTAACTGGTGTGAGATGATATCTCATAGTGG    :4545425525435538637853554557677566536751245644452450)423/20+303341-
12,40.4.    AM:i:0  MD:Z:76 NM:i:0  OQ:Z:?EFCC?CEDFGEGIHGEHHEDHJKIIHKFEFFHLIJIIFJJJKIEILIKHKJJAJIJJIJEJIIGJIIGHCDDFCC
      RG:Z:SRR035669  SM:i:0  XT:A:R  X0:i:37 XM:i:0  XO:i:0  XG:i:0
SRR035669.18390646      83      7       3853627 0       76M     =       3853391 -312    GTAGATTCTGGATATTAGCCCTTTGTCAGA
TGAGTAGGTTGCGAAAATTTTCTCCCATGTTGTAGGTTGCCTGTTC    -43366255345,687533468765566$465657645974'6999899967+558865945874697
4577595:    AM:i:0  MD:Z:76 NM:i:0  OQ:Z:BBECDHFGIJIIFIKIKJJJILKKHJIK2IKIKHIKJIKJJ?IEEEIKKKIGEJJJIJHKJHHKIHJIIIGIBDBA
      RG:Z:SRR035669  SM:i:0  XT:A:R  X0:i:26 X1:i:81 XM:i:0  XO:i:0  XG:i:0
SRR035022.12548671      69      7       5501368 0       35M41S  =       5501368 0       AGGCTCATTTTTGTAATTTTAGAAGGGACA
TGGTCTTACCATATTGGCCAGGCGGGGTCCAACGAGTGGTTAAGCG    ###################################################################
########    OQ:Z:##################################################################      RG:Z:SRR035022  XC
:i:35
```

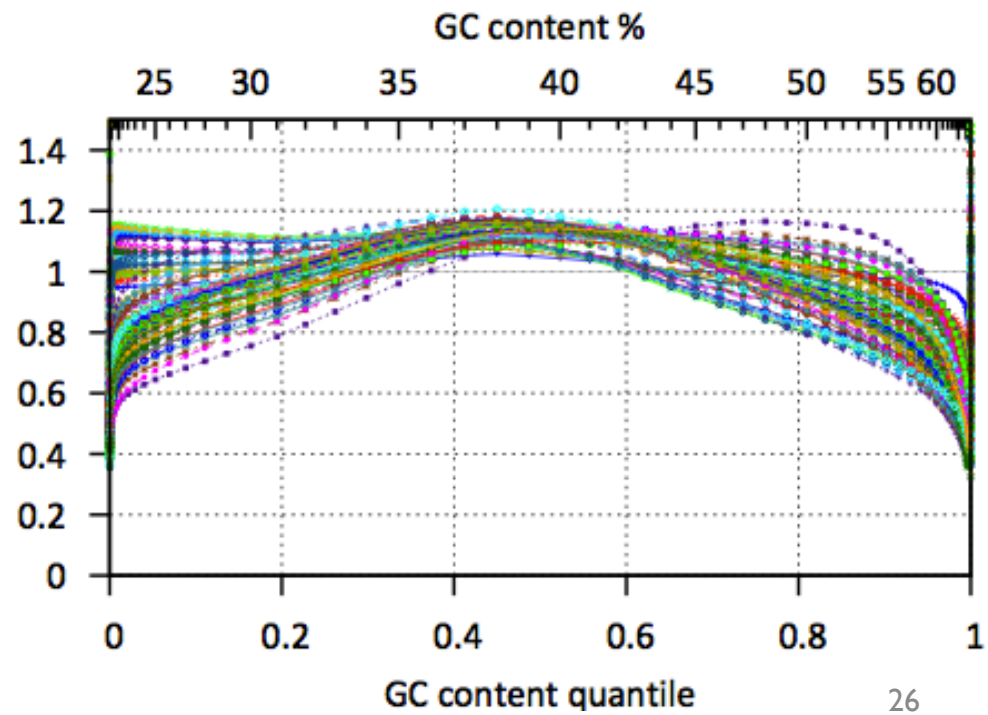
Understanding QC metrics

```
% cat ~/out/align/QCFiles/NA06984.qplot.stats
Stats\BAM  /home/yourid/out/align/bams/NA06984.recal.bam
TotalReads(e6) 0.06
MappingRate(%) 97.44
MapRate_MQpass(%) 97.44
TargetMapping(%)  0.00
ZeroMapQual(%) 0.56
MapQual<10(%)  0.70
PairedReads(%) 98.27
ProperPaired(%)   92.83
MappedBases(e9)   0.00
Q20Bases(e9)  0.00
Q20BasesPct(%) 95.27
MeanDepth  7.33
GenomeCover(%) 0.33
...
```

Quality Assessment with Qplot



- Qplot checks for
  - base quality distribution
  - GC bias
  - Insert size distribution
  - #Q20, coverage, %Duplicate, etc



UNDERSTANDING VERIFYBAMID OUTPUT

```
% cat ~/out/align/QCFiles/NA06984.genoCheck.selfSM
```

#SEQ_ID	RG	CHIP_ID	#SNPS	#READS	AVG_DP	FREEMIX	FREELK1	FREELK0
NA06984	ALL	NA	175	1038	5.93	0.00841	351.80	351.85

```
% cat ~/out/align/QCFiles/NA12878.genoCheck.selfSM
```

#SEQ_ID	RG	CHIP_ID	#SNPS	#READS	AVG_DP	FREEMIX	FREELK1	FREELK0
NA12878	ALL	NA	175	850	4.86	0.00000	146.59	146.59

- Neither data show significant evidence of contamination
- Results only based on 175 SNPs – may change when tested against larger number of SNPs

ANOTHER GOOD PIPELINE YOU CAN IMPLEMENT..

- 1000G provides a good reference of alignment pipeline
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/README.alignment_data
- These steps involves..
  - alignment with bwa
  - samtools sort, fixmate
  - GATK Indel realignment
  - GATK recalibration
  - BAQ (per-Base Aligment Quality) adjustment
  - Picard MarkDuplicate
- GotCloud pipeline is similar to, but not identical to 1000G pipeline

SUMMARY : PRACTICAL SESSION 5

- If we have a set of FASTQ files…
  - Can you align to reference sequence and process the BAMs ready for variant calling?
  - What do you need to achieve it?

- What kind of quality metrics are useful to assess the quality of aligned sequence reads?