HS Rats Genotyping Pipeline

Pipeline Summary Report Design

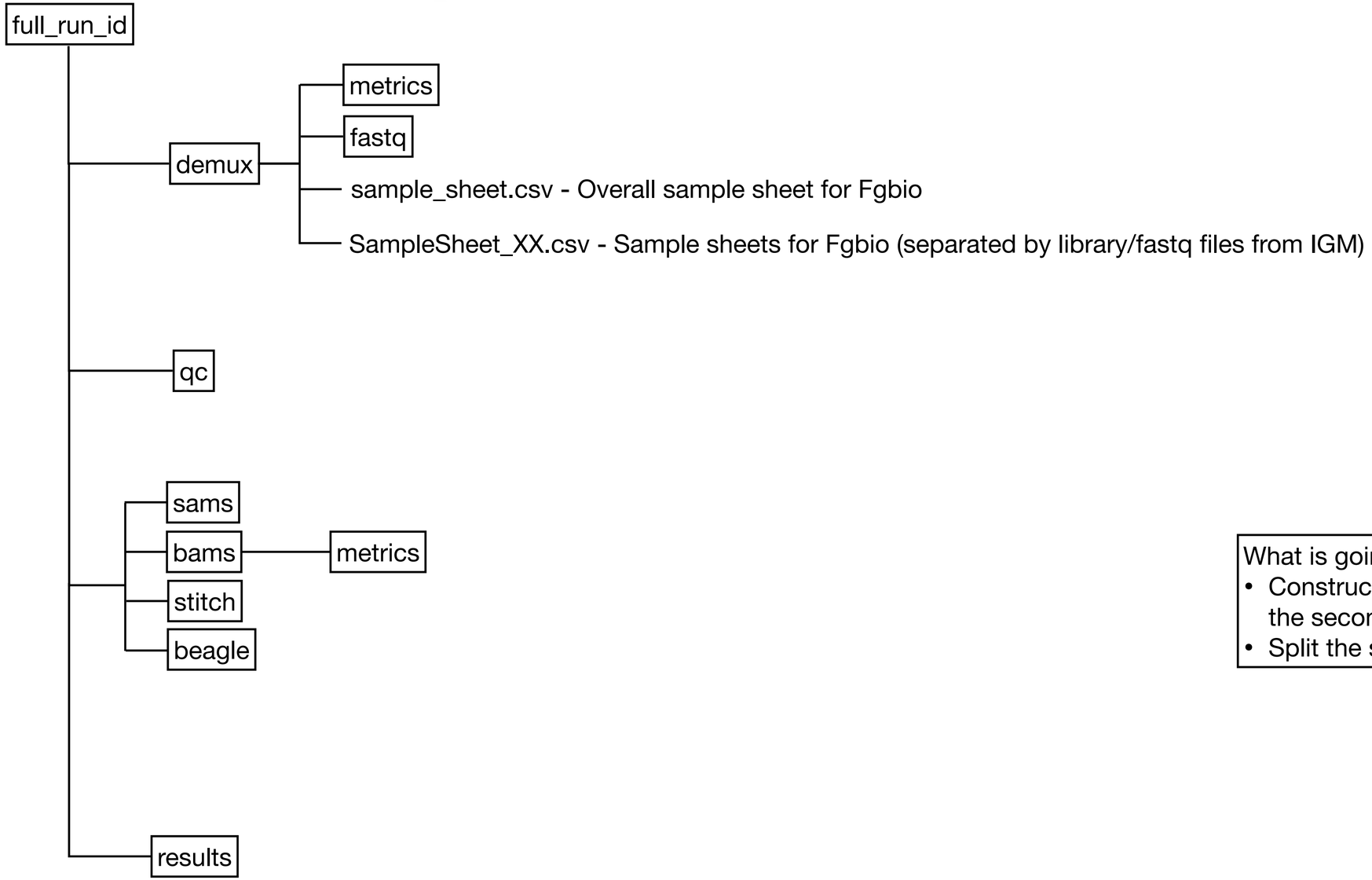
Denghui Chen
Dec. 11 2020

Pipeline Arguments

1. Your home directory
2. Directory where you want to keep all output files from the pipeline
3. The metadata for this flow cell
4. The directory where the sequencing files (fast.gz) for this flow cell locate
5. Reference genome for the alignment step
6. The directory where the reference panels for STITCH locate
7. The directory where the genetic maps for STITCH locate
8. The directory where you keep the code for this pipeline
9. The general name of this genotyping run (e.g. hs_rats_n1536) - need to find a clever way to make this automatic : /

After 9. The directories of previous runs' bam files

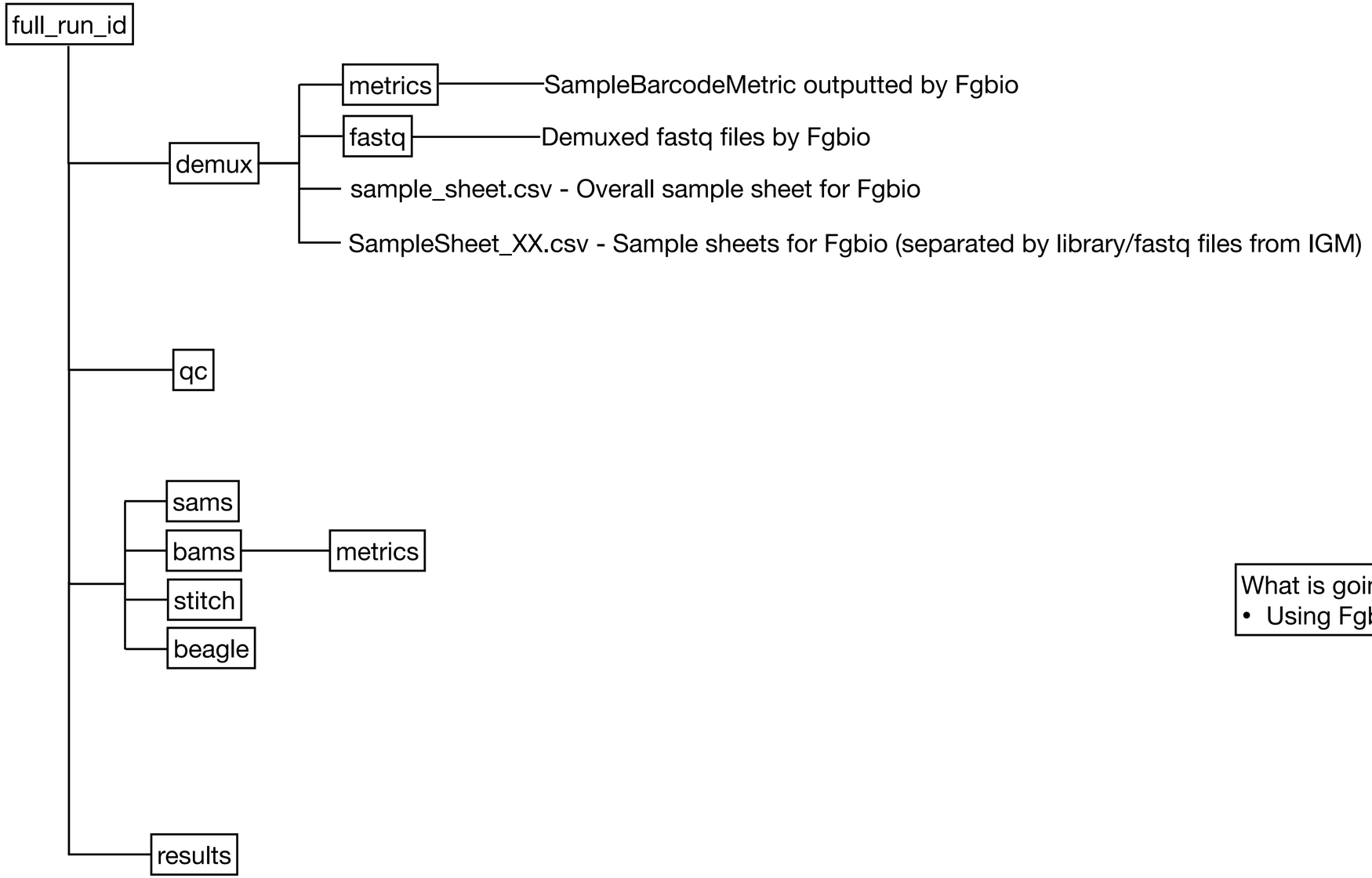
Step 1 - Preparation



What is going on:

- Construct the basic structure of the directory from the second line of the Pipeline Argument file
- Split the sample sheet for each library prep for Fgbio

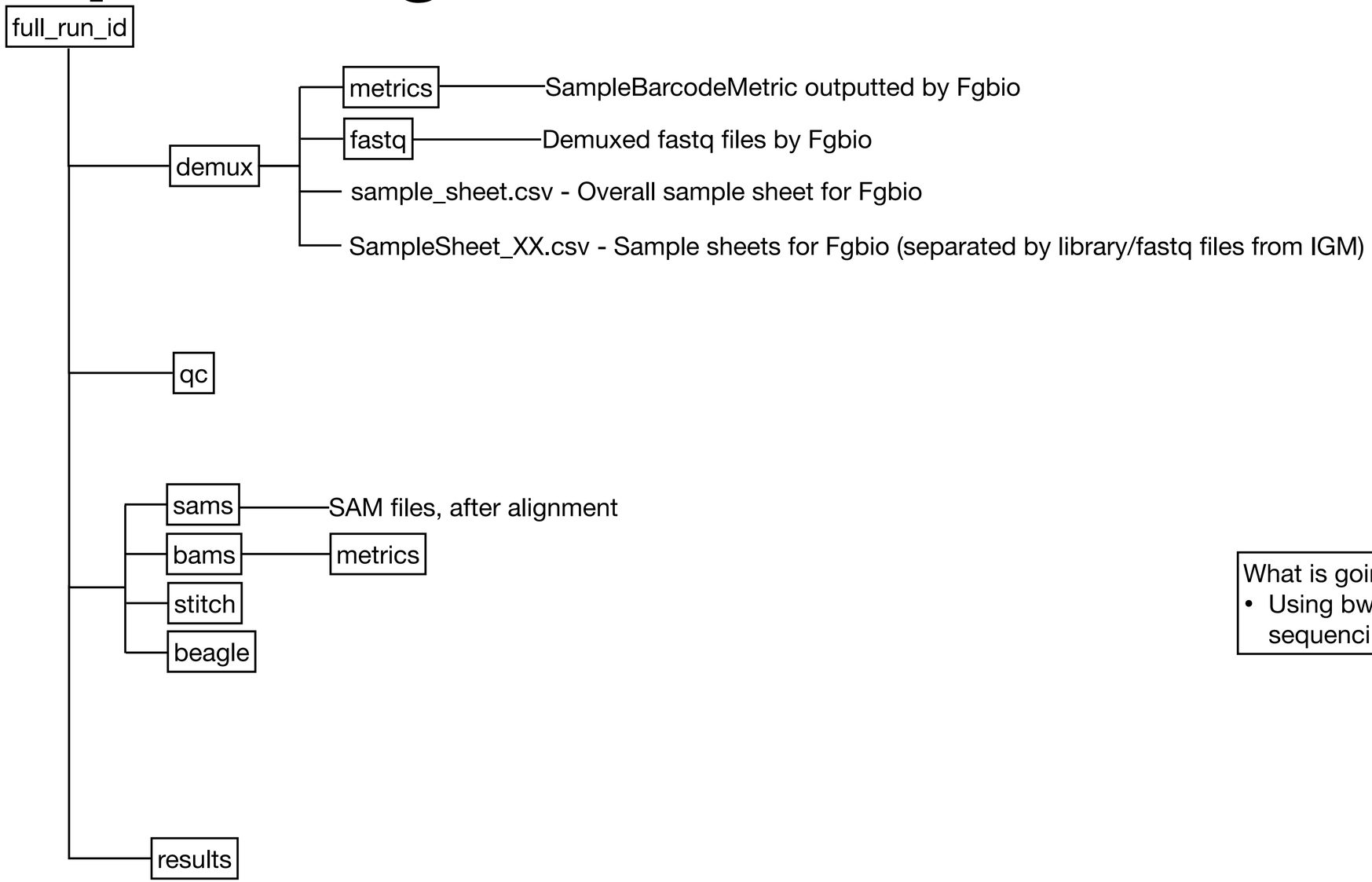
Step 2 - Demux



What is going on:

- Using Fgbio to demultiplex the fastq files

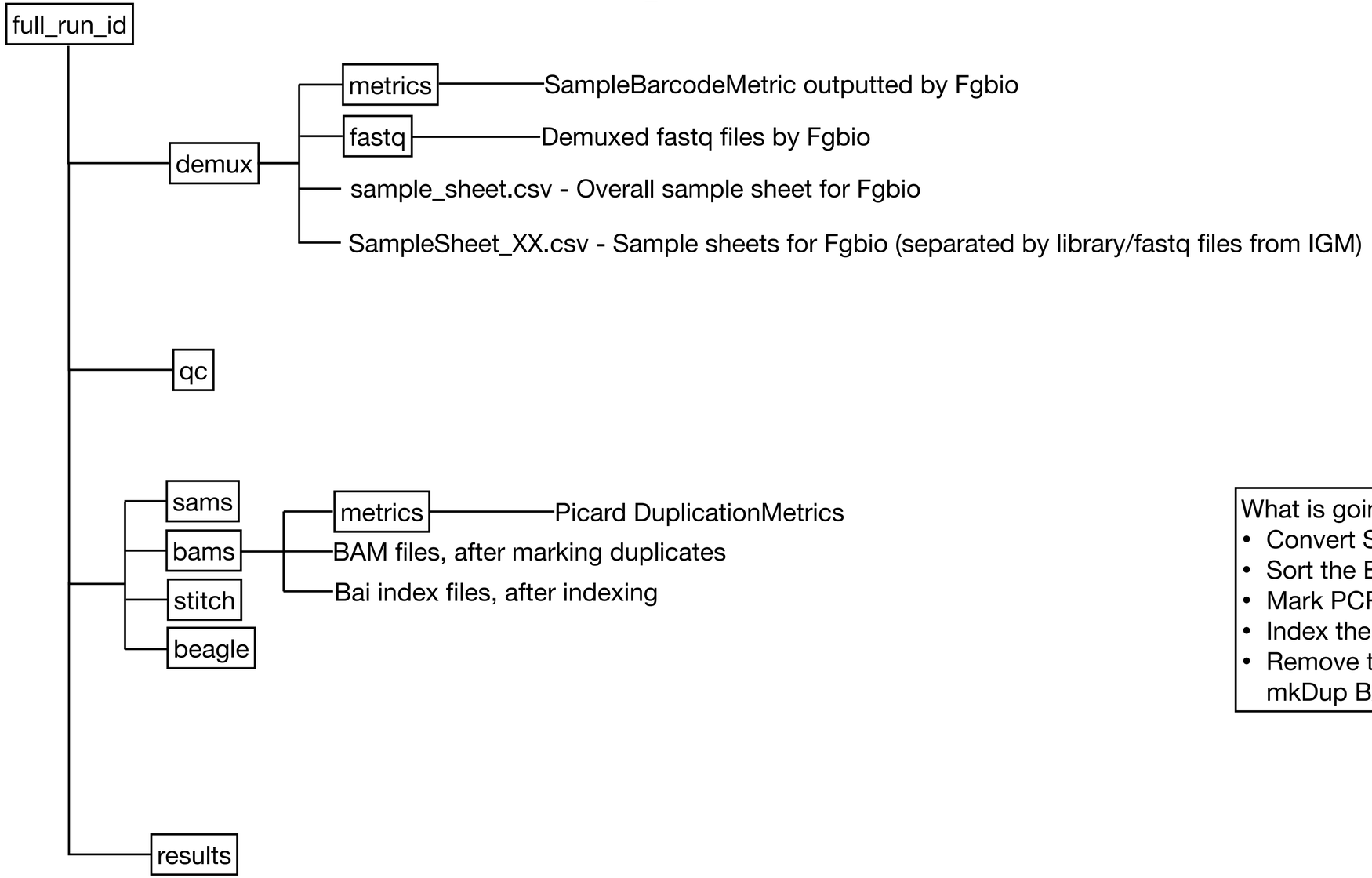
Step 3 - Alignment



What is going on:

- Using bwa mem to map the demultiplexed sequencing reads to reference genome

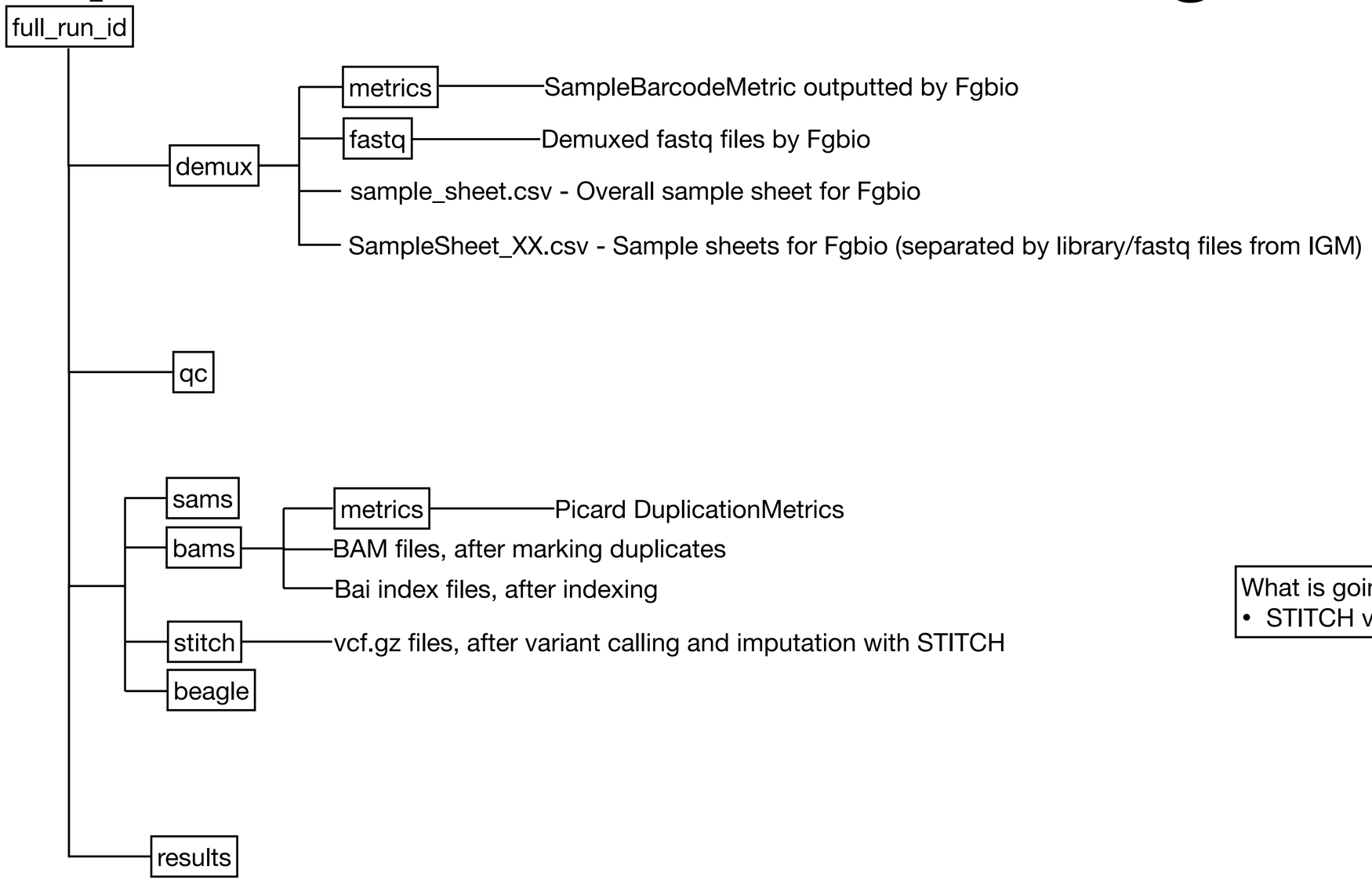
Step 4 - Mark Duplicates



What is going on:

- Convert SAM files to BAM files
- Sort the BAM files
- Mark PCR duplicates
- Index the marked duplicates BAM files
- Remove the SAM files, unsorted BAM files, and non-mkDup BAM files

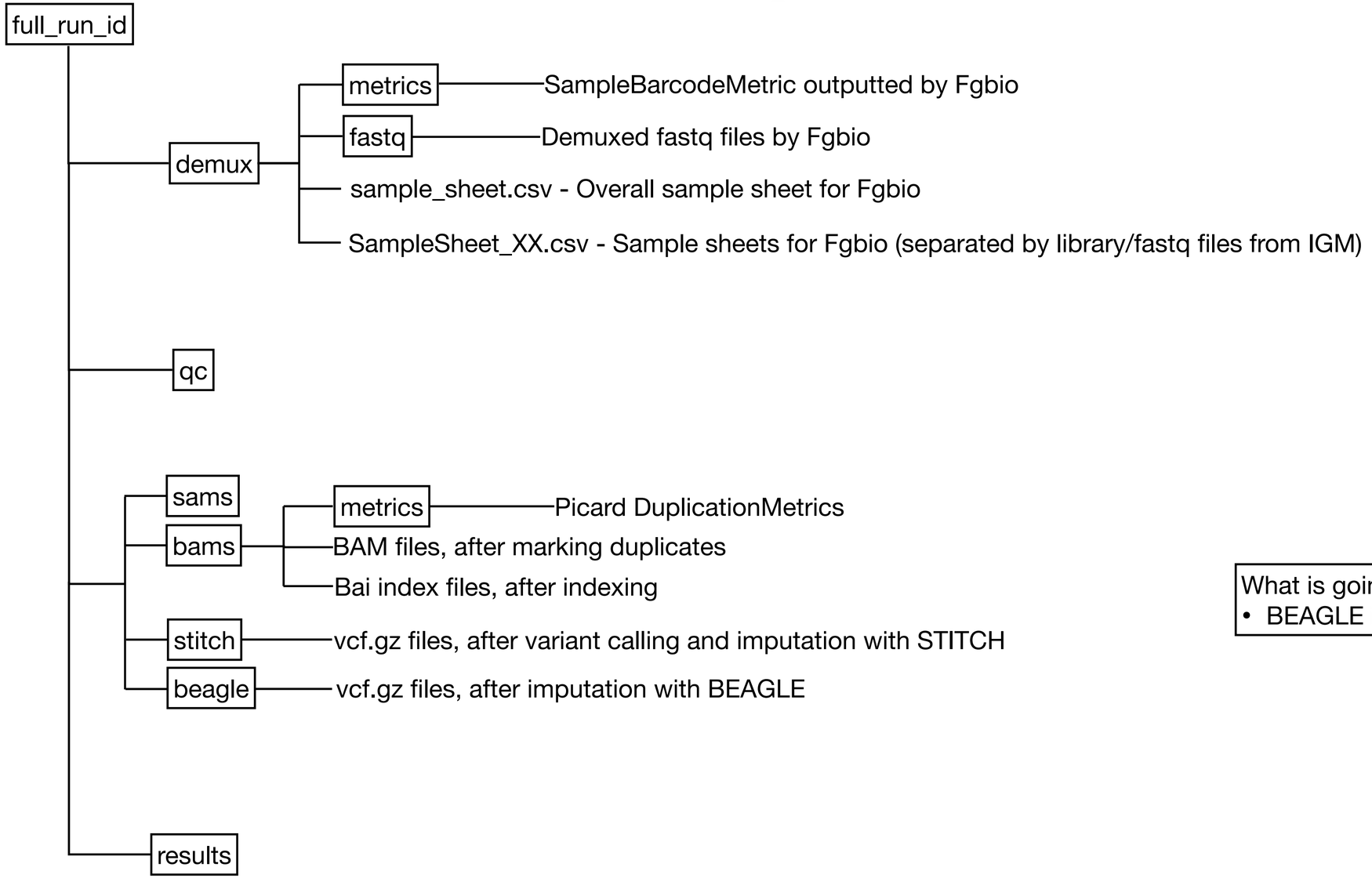
Step 5 - STITCH Variant Calling



What is going on:

- STITCH variant calling

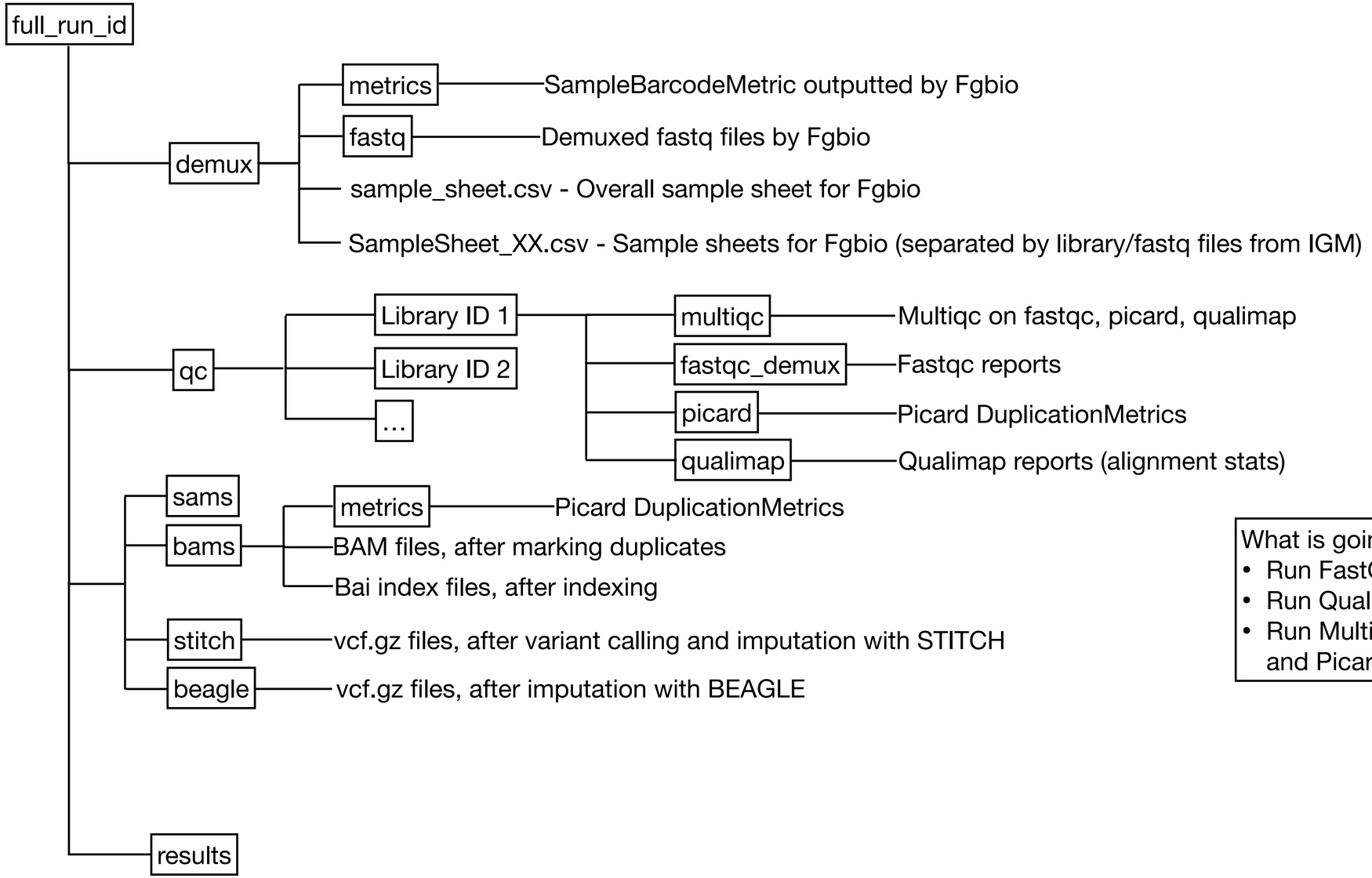
Step 6 - BEAGLE Imputation



What is going on:

- BEAGLE imputation

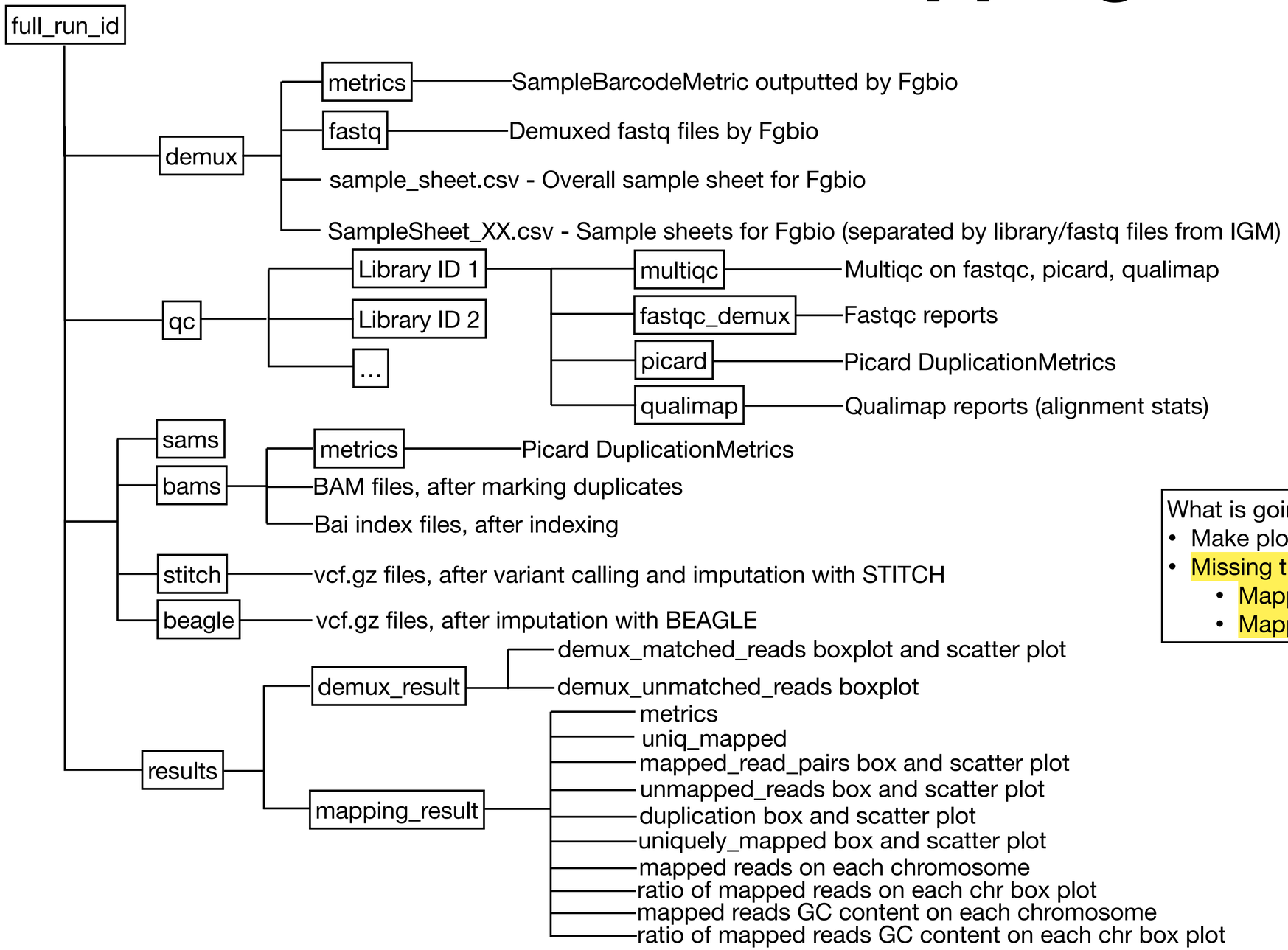
Result 1 - MultiQC



What is going on:

- Run FastQC on each library's fastq files
- Run Qualimap on each library's mapped bam files
- Run MultiQC on each library's FastQC, Qualimap, and Picard DuplicationMetrics results

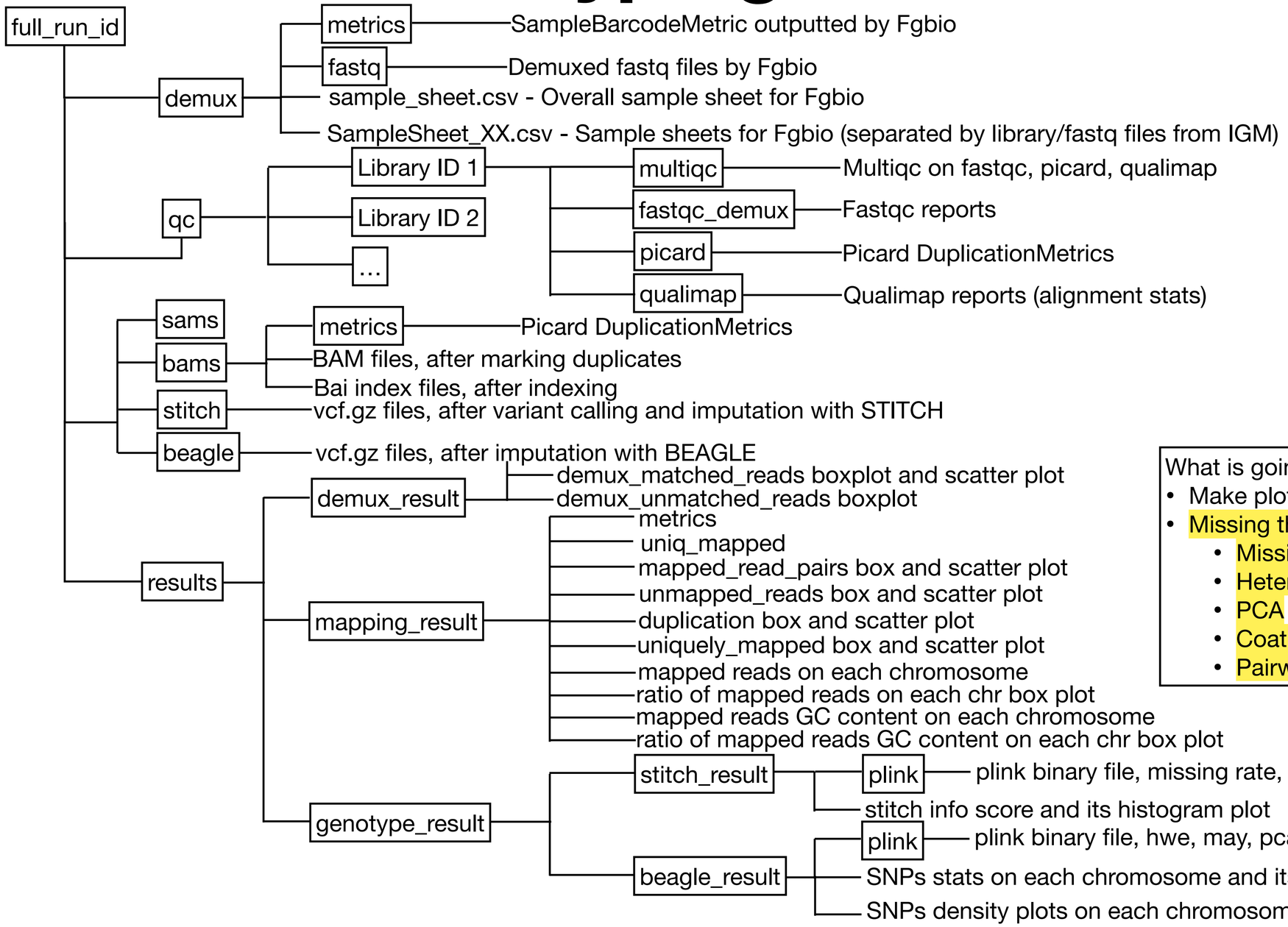
Result 2 - Demux and Mapping Results



What is going on:

- Make plots for demux result and mapping result
- Missing the code for:
  - Mapping coverage histogram/box plot
  - Mapping quality histogram

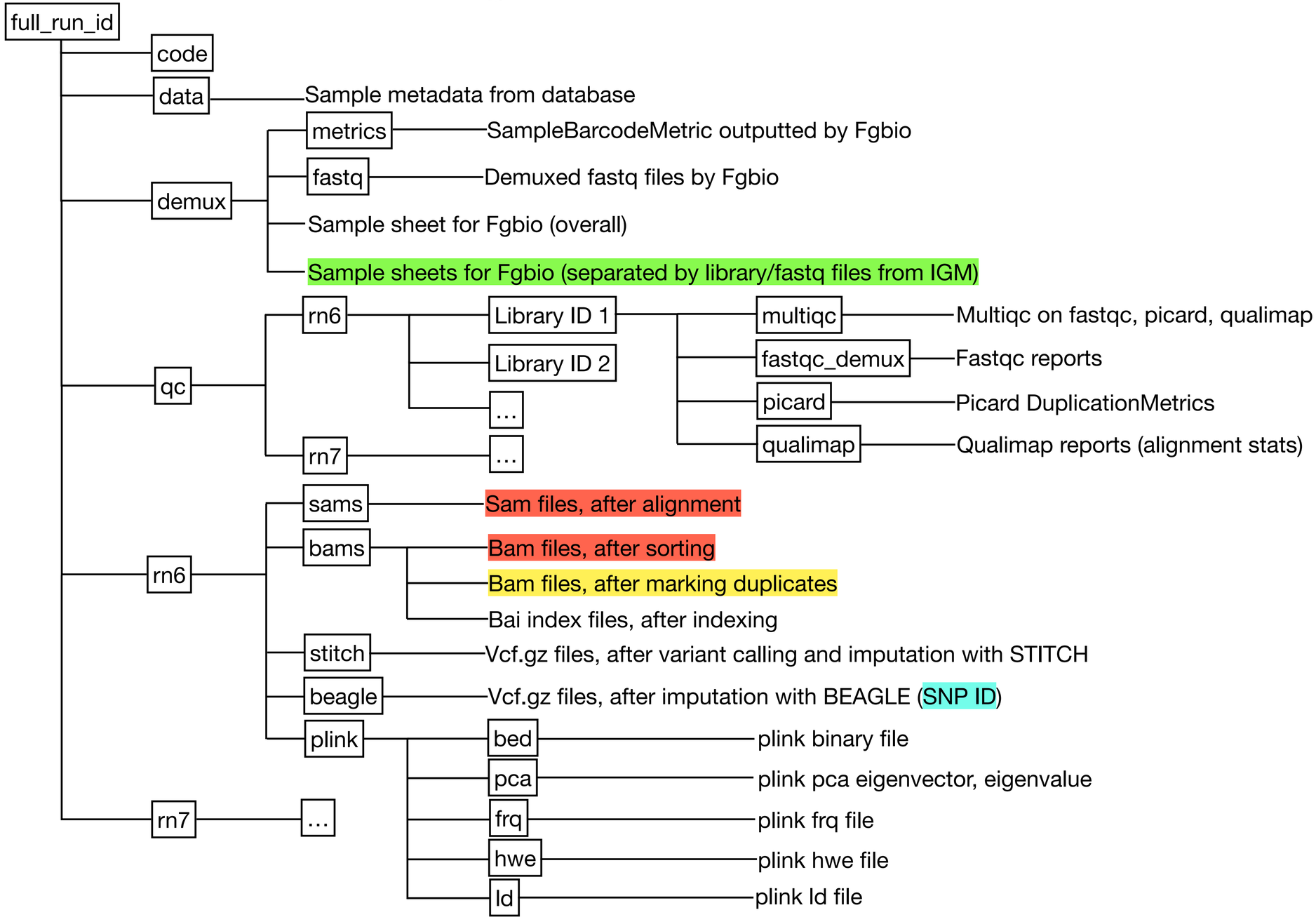
Result 3 - Genotyping Results



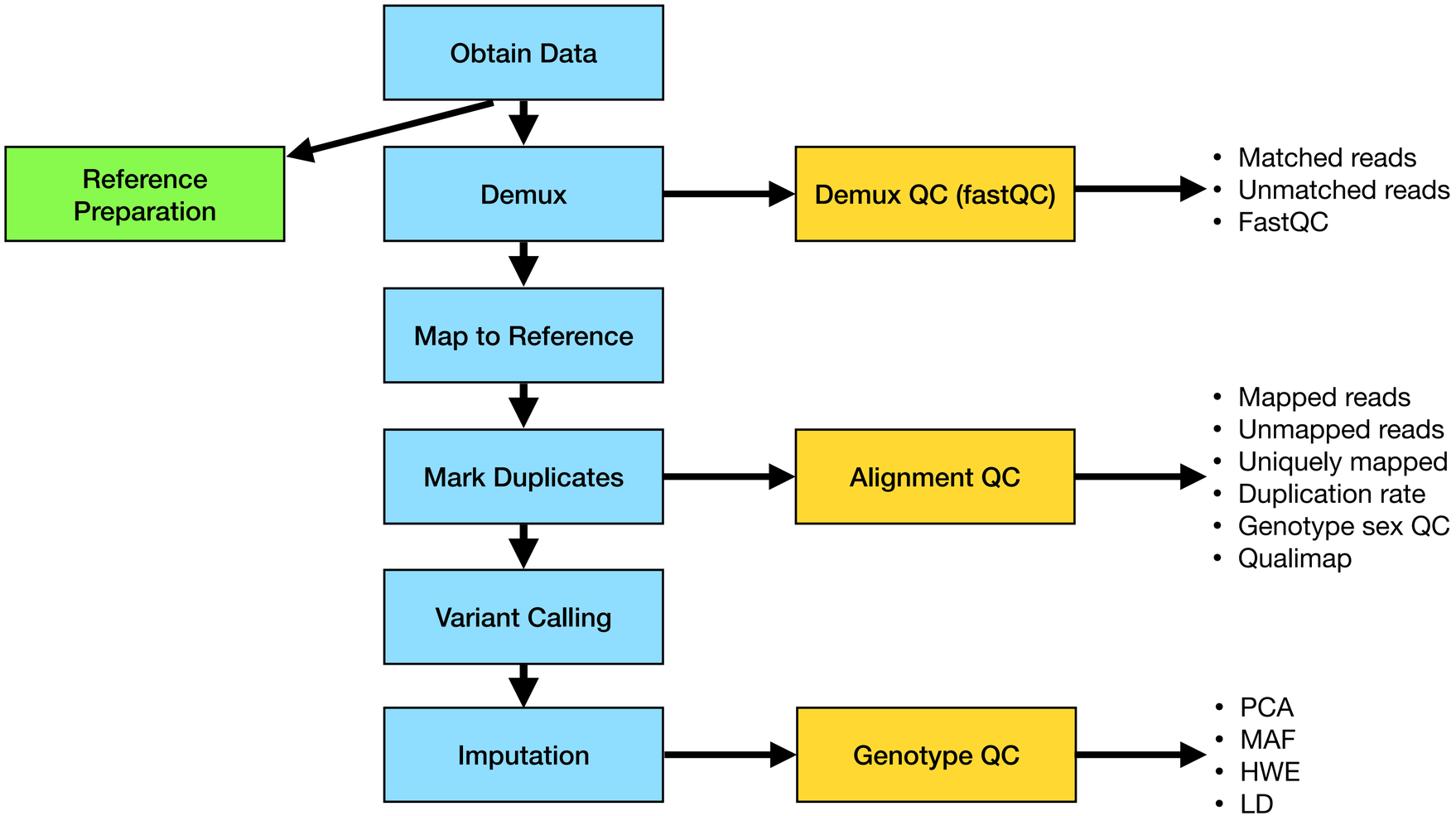
What is going on:

- Make plots for genotypes result
- Missing the code for:
  - Missing rate vs. # of reads
  - Heterozygosity rate vs. missing rate
  - PCA
  - Coat color QC
  - Pairwise concordance check

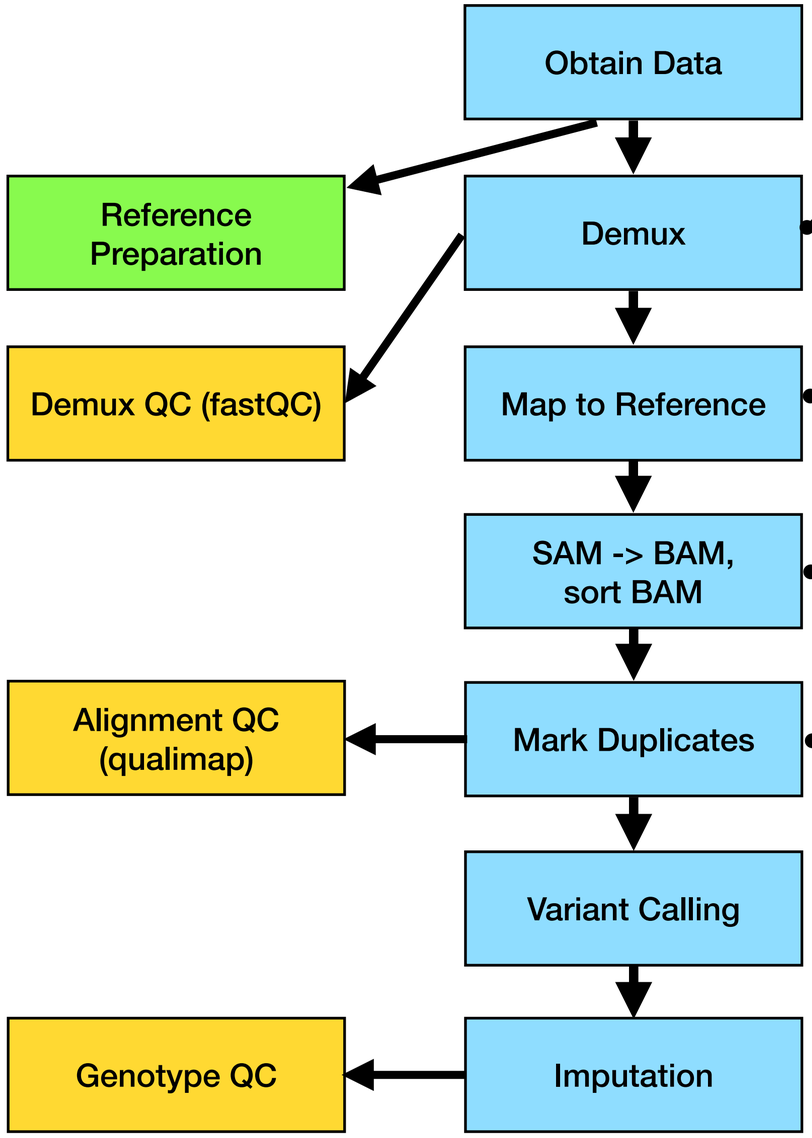
TSCC Directory Structure



Pipeline Flowchart



Pipeline Flowchart




```
java -Xmx40G -XX:+AggressiveOpts -XX:+AggressiveHeap \
    -jar /projects/ps-palmer/software/local/src/fgbio-1.2.0/fgbio-1.2.0.jar DemuxFastqs \
    --inputs ${pre_demux_fastq_R1} ${pre_demux_fastq_R2} \
    --metadata ${sample_sheet} \
    --read-structures 8B12M+T 8M+T \
    --output-type=Fastq \
    --threads $ncpu \
    --output ${out_path}/demux/fastq \
    --metrics ${out_path}/demux/metrics/${fastq_temp}demux_barcode_metrics.txt
```

```
/projects/ps-palmer/software/local/src/bwa-0.7.12/bwa mem -aM -t 2\
 -R "@RG\tID:${instrument_name}.${run_id}.${flowcell_id}.${flowcell_lane}\tLB:${library_id}
\tPL:ILLUMINA\tSM:${sample_id}\tPU:${flowcell_id}.${flowcell_lane}.${sample_barcode}" \
 ${reference_data} ${demux_data}/${f}_R1.fastq.gz \
 ${demux_data}/${f}_R2.fastq.gz > ${out_path}/sams/${f}.sam &
```

```
/projects/ps-palmer/software/local/src/samtools-1.10/samtools view -h -b \
  -t ${reference_data} -o ${out_path}/bams/${f}.bam ${mapped_data}/${f}.sam

/projects/ps-palmer/software/local/src/samtools-1.10/samtools sort -m 30G \
  -o ${out_path}/bams/${f}_sorted.bam ${out_path}/bams/${f}.bam
```

```
java -Xmx20G -XX:+AggressiveOpts -XX:+AggressiveHeap\
  -jar /projects/ps-palmer/software/local/src/picard-2.23.3/picard.jar MarkDuplicates \
  --INPUT ${out_path}/bams/${f}_sorted.bam \
  --REMOVE_DUPLICATES false \
  --ASSUME_SORTED true \
  --METRICS_FILE ${out_path}/bams/metrics/${f}_sorted_mkDup_metrics.txt \
  --OUTPUT ${out_path}/bams/${f}_sorted_mkDup.bam &

/projects/ps-palmer/software/local/src/samtools-1.10/samtools index \
  ${out_path}/bams/${f}_sorted_mkDup.bam ${out_path}/bams/${f}_sorted_mkDup.bai
```

Pipeline Code Flowchart