



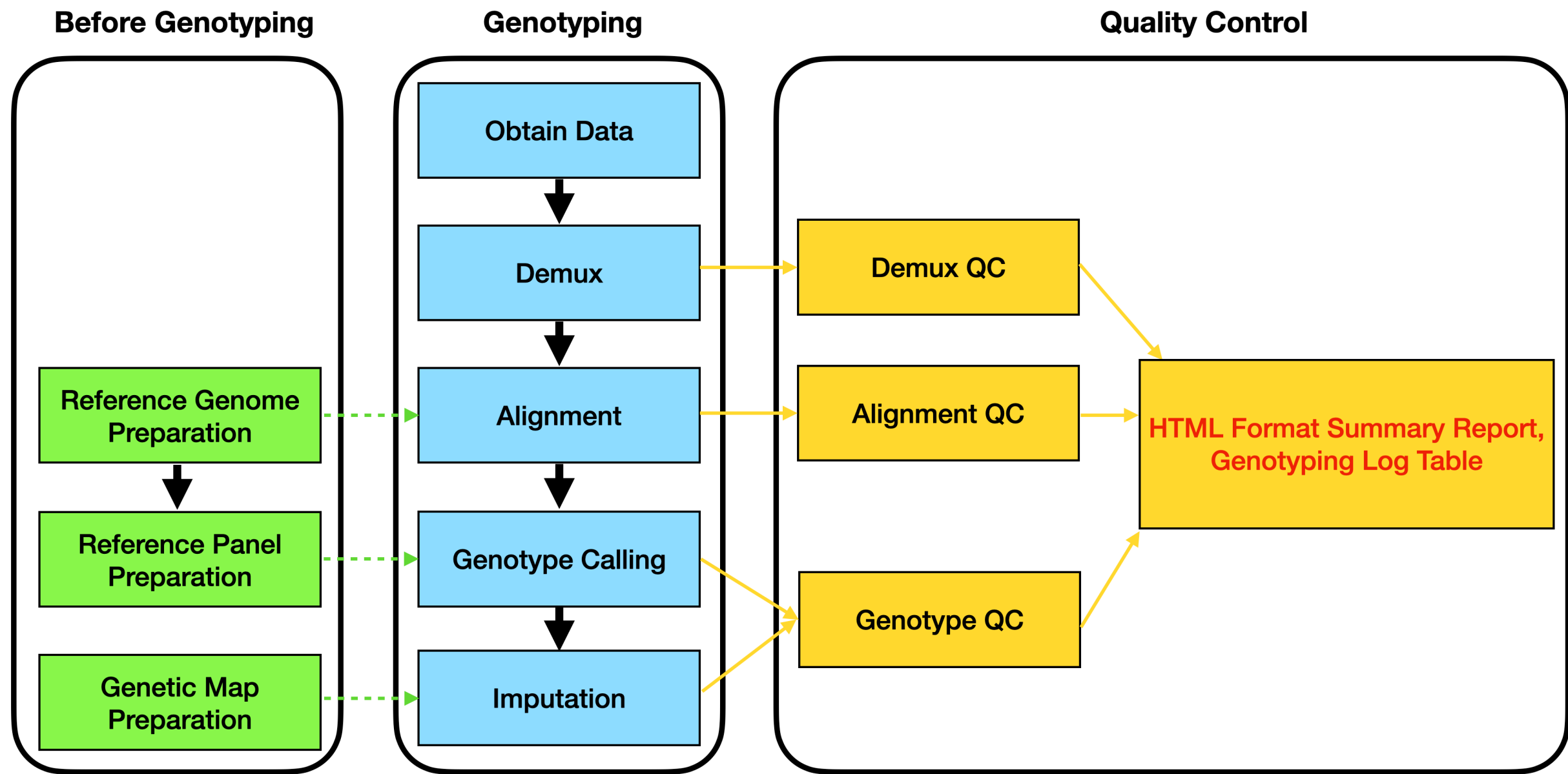
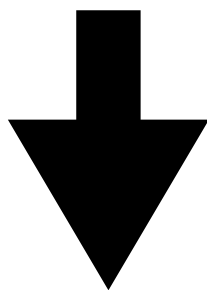
HS Rats Genotyping Pipeline

Pipeline Design

UCSD, Palmer Lab

Pipeline Overview

Submission Script



Flowchart

1 Pipeline Overview
2 Data
2.1 Sequence Data
2.2 Reference Data
3 Demultiplex Results
4 Alignment Results
5 MultiQC Summary for Demultiplex and Alignment Steps
6 Genotype Results
7 Outliers Report
8 Appendix A. Relevant Softwares
9 Appendix B. Preparation for Reference Data

PALMER LAB Behavioral Genetics of Mice, Rats and Men Genotyping Summary Report

Palmer Lab
January 21, 2021

1 Pipeline Overview

1.1 Pipeline workflow

The pipeline flow chart is shown in figure 1.

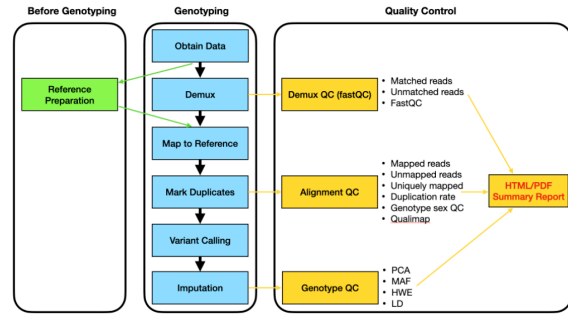


Figure 1: Pipeline flow chart

2 Data

2.1 Sequence Data

Sample strain: Heterogenous stock

Number of samples: 376

Flow cell run ID(s): 201218_A00953_0203_BHNSHYDSXY

Riptide library preparation: Riptide31, Riptide32, Riptide39, Riptide40

Metadata: /projects/ps-

palmer/hs_rats/201218_A00953_0203_BHNSHYDSXY/data/kn04_fastq_sample_metadata_n952.csv

Run ID	Library ID	Sample Project	Number of Samples
201218_A00953_0203_BHNSHYDSXY	Riptide31	u01_olivier_george_cocaine	1
201218_A00953_0203_BHNSHYDSXY	Riptide31	u01_olivier_george_oxycodone	83
201218_A00953_0203_BHNSHYDSXY	Riptide31	u01_olivier_george_scrub	4
201218_A00953_0203_BHNSHYDSXY	Riptide31	u01_suzanne_mitchell	3
201218_A00953_0203_BHNSHYDSXY	Riptide31	u01_tom_jhou	1
201218_A00953_0203_BHNSHYDSXY	Riptide32	u01_olivier_george_cocaine	91
201218_A00953_0203_BHNSHYDSXY	Riptide32	u01_olivier_george_scrub	5
201218_A00953_0203_BHNSHYDSXY	Riptide39	u01_olivier_george_oxycodone	8
201218_A00953_0203_BHNSHYDSXY	Riptide39	u01_peter_kalivas_italy	88
201218_A00953_0203_BHNSHYDSXY	Riptide40	u01_peter_kalivas_italy	88

2.2 Reference Data

Reference genome for alignment: /projects/ps-palmer/hs_rats/reference_genome/rn6.fa

STITCH variant calling reference panel: /projects/ps-palmer/hs_rats/reference_genome/rn6_refPnls

BEAGLE imputation genetic map: /projects/ps-palmer/hs_rats/reference_genome/map_files

rs_id	demux_reads	mapped_reads	unmapped_reads	duplication_ratio	uniq_mapped_ratio	QC_reads	QC_data	missing_rate	QC_missing	heterozygosity_rate	QC_heterozygosity	coatscolor	QC_coat_color	albino	QC_albino
0007E7E9F	4645066	3963545	0.0129424877404693	0.004919	0.802893900711603	pass	pass	1202021	0.003717	pass	0.31191208191811	pass	BROWN	pass	pass
0007E7EAC	4860719	4896426	0.0114263484129261	0.104065	0.808964663291902	pass	pass	1202021	0.049789	pass	0.31162636323747	pass	ALBINO	pass	pass
0007E7E37	4780227	4412550	0.0108715627066198	0.008468	0.79970368304005	pass	pass	1202021	0.044807	pass	0.320133784050021	pass	BLACK	pass	pass
0007E7E0F	2382093	2202840	0.0118620046198021	0.001602	0.81889708878461	pass	pass	1202021	0.061005	pass	0.327818914370201	pass	BROWN	pass	pass
0007E7E0D	1245465	9104246	0.0101001402164041	0.005050	0.810160899505416	pass	pass	1202021	0.048415	pass	0.306027085007072	pass	BLACK	pass	pass
0007E7E36	3129639	2090425	0.0118517668684343	0.0038	0.7987369061687	pass	pass	1202021	0.058843	pass	0.316020131945758	pass	BROWN	pass	pass
0007E7E3B	4843469	4743309	0.01578318388864	0.00424	0.8173203171413737	pass	pass	1202021	0.048171	pass	0.314478482326781	pass	ALBINO	pass	pass
0007E7E2F	3518739	3443481	0.0115091824534863	0.00477	0.80266332149219	pass	pass	1202021	0.018971	pass	0.319049494456502	pass	ALBINO	pass	pass
0007E7E1D	2321232	2238517	0.0118011026302384	0.00020	0.805275602890443	pass	pass	1202021	0.000060	pass	0.312160389143401	pass	BLACK	pass	pass
0007E7E41	4176238	4284181	0.0116702038431263	0.00406	0.802079182932552	pass	pass	1202021	0.002188	pass	0.30211208881028	pass	BROWN	pass	pass
0007E7E3C	3475400	3414295	0.010803392120705	0.00088	0.787824862781891	pass	pass	1202021	0.002514	pass	0.3068738021488433	pass	BROWNHOOD	pass	pass
0007E7E4F	2965766	2944505	0.011443219132816	0.00020	0.821741710211608	pass	pass	1202021	0.006403	pass	0.296716261050007	pass	BLACKHOOD	pass	pass
0007E7E76	2905547	2901900	0.011268464575505	0.00075	0.786115918998962	pass	pass	1202021	0.005884	pass	0.3208410843434028	pass	BROWN	pass	pass
0007E7E0C	9729201	9538678	0.0118044792339963	0.00757	0.824804372288116	pass	pass	1202021	0.0417072	pass	0.320008402650609	pass	BROWNHOOD	pass	pass
0007E7EAD	2160900	9105665	0.0114081071623157	0.00041	0.80437894645438	pass	pass	1202021	0.003968	pass	0.312171474586151	pass	BLACKHOOD	pass	pass
0007E7E42	522904	9185453	0.0120570194681138	0.00279	0.82881180540508	pass	pass	1202021	0.019414	pass	0.284417892281312	pass	BROWNHOOD	pass	pass
0007E7E35	2020219	5815657	0.01102028886126	0.005138	0.80218026011644	pass	pass	1202021	0.0448023	pass	0.29800007246734	pass	BLACKHOOD	pass	pass
0007E7E8A	3320585	3208918	0.0123729135844968	0.00512	0.805152191552326	pass	pass	1202021	0.004737	pass	0.31448884720978	pass	ALBINO	pass	pass
0007E7E3E	2170785	2714207	0.0121848026110708	0.004173	0.821551570881887	pass	pass	1202021	0.003704	pass	0.326560582020089	pass	BROWN	pass	pass
0007E7E81	2149507	2109846	0.0118026778815007	0.001259	0.79686851982076	pass	pass	1202021	0.009752	pass	0.2975468842109	pass	ALBINO	pass	pass
0007E7E8E	4735107	4684179	0.01120834851641	0.00408	0.803802164784846	pass	pass	1202021	0.0487008	pass	0.290253084502754	pass	ALBINO	pass	pass
0007E7E7A	4688402	4611006	0.0107071487861977	0.00046	0.79680777847213	pass	pass	1202021	0.047179	pass	0.31707938188071	pass	BLACK	pass	pass
0007E7E7C	9148908	2895309	0.012317233361192	0.00044	0.80287121158815	pass	pass	1202021	0.054504	pass	0.298003093070701	pass	BROWNHOOD	pass	pass
0007E7E7F	2849187	2385189	0.013438055167038	0.00701	0.8134325762578614	pass	pass	1202021	0.0056009	pass	0.31278008168683	pass	BROWN	pass	pass
0007E7E69	9511888	5443439	0.0122036882188541	0.00061	0.81862860274008	pass	pass	1202021	0.0407808	pass	0.31936898337387	pass	ALBINO	pass	pass
0007E7E6E	4822380	4782026	0.012071912118865	0.00027	0.819885026919448	pass	pass	1202021	0.0478608	pass	0.32660078761741	pass	ALBINO	pass	pass
0007E7E66	2229703	2182242	0.0122023544446803	0.00488	0.807863057177795	pass	pass	1202021	0.0040238	pass	0.322867958105384	pass	BLACKHOOD	pass	pass
0007E7E8A	2790608	2744811	0.0114473311527008	0.00026	0.80034842005116	pass	pass	1202021	0.002878	pass	0.318816027607156	pass	BLACK	pass	pass
0007E7E7B	3584617	3695847	0.0115840074919711	0.00476	0.82238915961862	pass	pass	1202021	0.0818035	pass	0.311079105888008	pass	BLACK	pass	pass
0007E7E78	2782546	2773982	0.013258384342088	0.00088	0.814475830314206	pass	pass	1202021	0.0487602	pass	0.298881124847868	pass	BROWNHOOD	pass	pass
0007E7E1A	4810556	4823891	0.010501814004404	0.00485	0.811728034142812	pass	pass	1202021	0.0400051	pass	0.322787423138548	pass	ALBINO	pass	pass
0007E7E1E	2627807	2573237	0.0120411487168957	0.00290	0.794789804898008	pass	pass	1202021	0.0072396	pass	0.306847341486459	pass	BLACKHOOD	pass	pass
0007E7E0D	2002140	1888175	0.012271208867152	0.00196	0.784431807203428	pass	pass	1202021	0.0070505	pass	0.304162082818148	pass	BROWNHOOD	pass	pass
0007E7E3E	2486034	2410455	0.018602814362419	0.00032	0.81905141798852	pass	pass	1202021	0.006059	pass	0.308815088202037	pass	ALBINO	pass	pass
0007E7E75	3180260	3106887	0.011873982027278	0.00006	0.8072440300793	pass	pass	1202021	0.0547302	pass	0.308470823320329	pass	BROWNHOOD	pass	pass
0007E7E37	7104827	6891872	0.0101071918487138	0.00070	0.800084697812564	pass	pass	1202021	0.0447032	pass	0.21040324032784418	pass	BLACK	pass	pass
0007E7E91	4512565	3944005	0.0109612338311837	0.04806	0.794813416088191	pass	pass	1202021	0.0138688	pass	0.283818381717763	pass	BROWNHOOD	pass	pass
0007E7E90	2358877	2488508	0.002138	0.787780468408313	pass	pass	pass	1202021	0.006007	pass	0.311720080005887	pass	BROWNHOOD	pass	pass
0007E7E3C	3048507	3023124	0.010167020402116	0.00280	0.8155815001687	pass	pass	1202021	0.005042	pass	0.29012908818673	pass	BLACKHOOD	pass	pass

Directory Structure

flowcell	# bam files list for the samples in this genotyping run
sampleNames_file	# sample name file for the samples in this genotyping run
demux	
metrics	# SampleBarcodeMetric outputted by Fgbio
fastqc	# Demuxed fastqc files by Fgbio
sample_sheet.csv	# Overall sample sheet for Fgbio
SampleSheet_XX.csv	# Sample sheets for Fgbio (separated by library/fastqc files from IGM)
qc	
Library ID	
multique	# Multiqc on fastqc, picard, qualimap
fastqc_demux	# Fastqc reports
picard	# Picard DuplicationMetrics
qualimap	# Qualimap reports (alignment stats)
...	
sams	# Temporary directory for sam files
bams	
metrics	# Picard DuplicationMetrics
XX.bam	# BAM files, bai index files, after marking duplicates (XX means using Sample_ID as the file name prefix)
stitch	# vcf.gz files, after variant calling and imputation with STITCH
beagle	# vcf.gz files, after imputation with BEAGLE
results	
demux_result	# Only contains the demux results for this flow cell
demux_barcode_metrics	
after_demux_demux_reads.png	
mapping_result	# Only contains the mapping results for this flow cell
mkDup_metrics	
mapped_chr	
after_mkDup_mapped_reads.png	
after_mkDup_duplication_rate.png	
after_mkDup_unmapped_rate.png	
after_mkDup_mapped_reads_per_chr.png	
after_mkDup_percent_mapped_per_chr.png	
after_mkDup_QC_sex.png	
after_mkDup_QC_mapped_reads_threshold_1M.csv	
after_mkDup_QC_sex_mapped_reads_percent.csv	
genotype_result	# Contains the demux, mapping, genotype results for all flow cells in this genotyping run (XXX means the prefix; format: Heterogenous-stock_n672_11012021)
XXX_metadata.csv	
XXX_demux_barcode_metrics	
after_demux_demux_reads.png	
XXX_mkDup_metrics	
XXX_mapped_chr	
after_mkDup_mapped_reads.png	
after_mkDup_duplication_rate.png	
after_mkDup_unmapped_rate.png	
after_mkDup_mapped_reads_per_chr.png	
after_mkDup_percent_mapped_per_chr.png	
after_mkDup_QC_sex.png	
after_mkDup_QC_mapped_reads_threshold_1M.csv	
after_mkDup_QC_sex_mapped_reads_percent.csv	
stitch_result	
stitch_INFO	
after_stitch_SNPs_INFO_histogram.png	
after_stitch_chrXX_SNPs_density_plot.png	
after_stitch_number_of_SNPs_per_chr.csv	
after_stitch_QC_mapped_reads_threshold_1M.csv	
after_stitch_QC_sample_missing_rate_threshold_1Qpercent.csv	
after_stitch_sample_missing_vs_mapped_reads.png	
after_stitch_QC_sample_het_rate_threshold_3std.csv	
after_stitch_sample_het_vs_missing.png	
after_stitch_SNPs_missing_vs_maf.png	
after_stitch_poly_SNPs_missing_vs_maf.png	
plink	
plink_make-bed, missing, het, freq results	
beagle_result	
SNPs	
after_beagle_chrXX_SNPs_density_plot.png	
after_beagle_number_of_SNPs_per_chr.csv	
after_beagle_SNPs_hwe_vs_maf.png	
after_beagle_poly_SNPs_hwe_vs_maf.png	
after_beagle_sample_PC2_vs_PC1_Library_ID.png	
after_beagle_sample_PC2_vs_PC1_Sample_Project.png	
after_beagle_sample_PC2_vs_PC1_Family.png	
after_beagle_QC_coatcolor_albino_snp_1_151097606.csv	
plink	
plink_make-bed, hardy, freq, pca, snp 1:151097606 results	
ref2	
...	

Pipeline Required Documents

pipeline_arguments

Line 1: Flow cell directory

Line 2: Flow cell metadata

Line 3: Sequencing data directory

Line 4: Reference genome

Line 5: Reference panels for STITCH

Line 6: Genetic map for BEAGLE

Line 7: Directory where you keep the code for the pipeline

previous_flow_cells_metadata

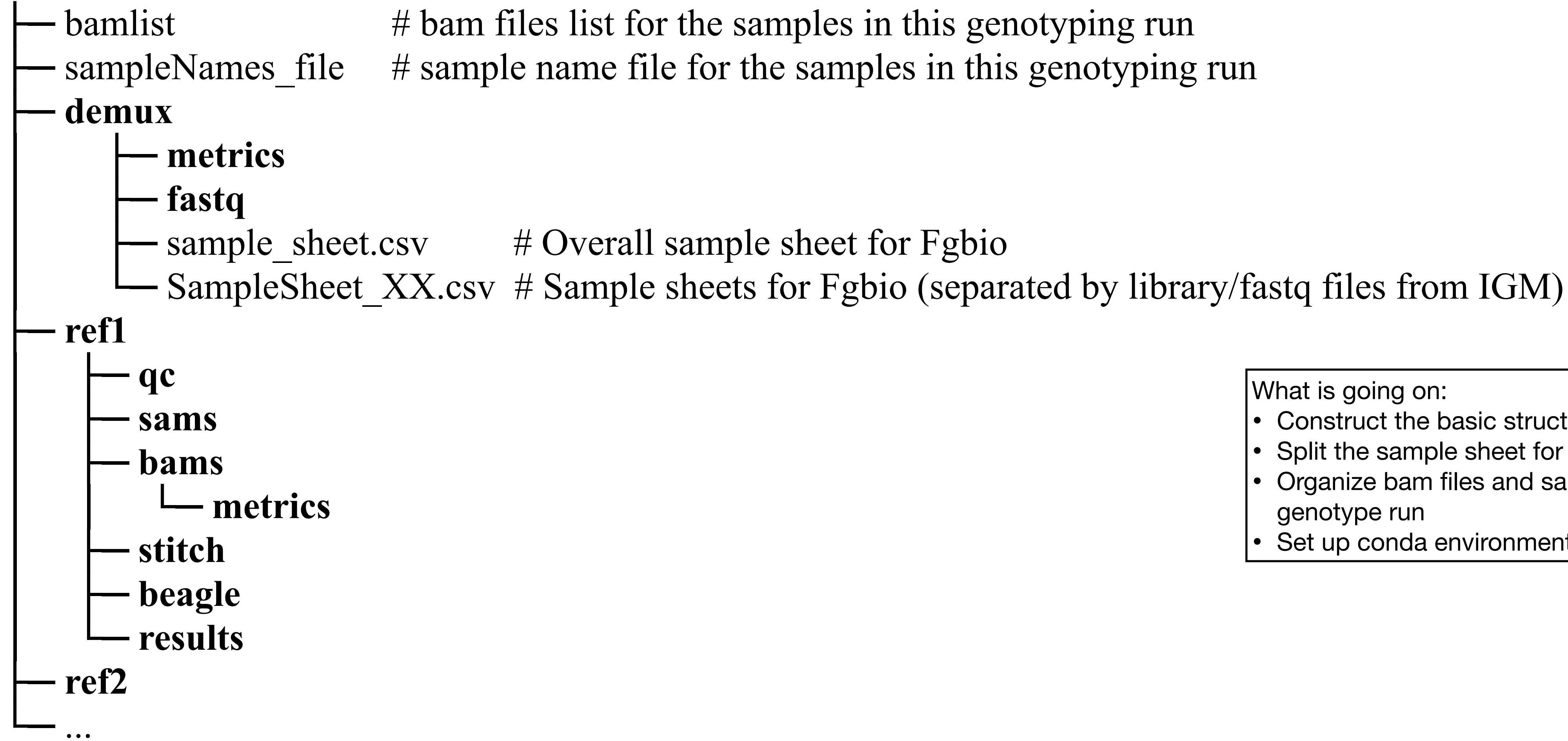
Paths to previous flow cells' metadata.

previous_flow_cells_bams

Paths to previous flow cells' BAM files.

Genotyping Step 1 - Preparation

flowcell

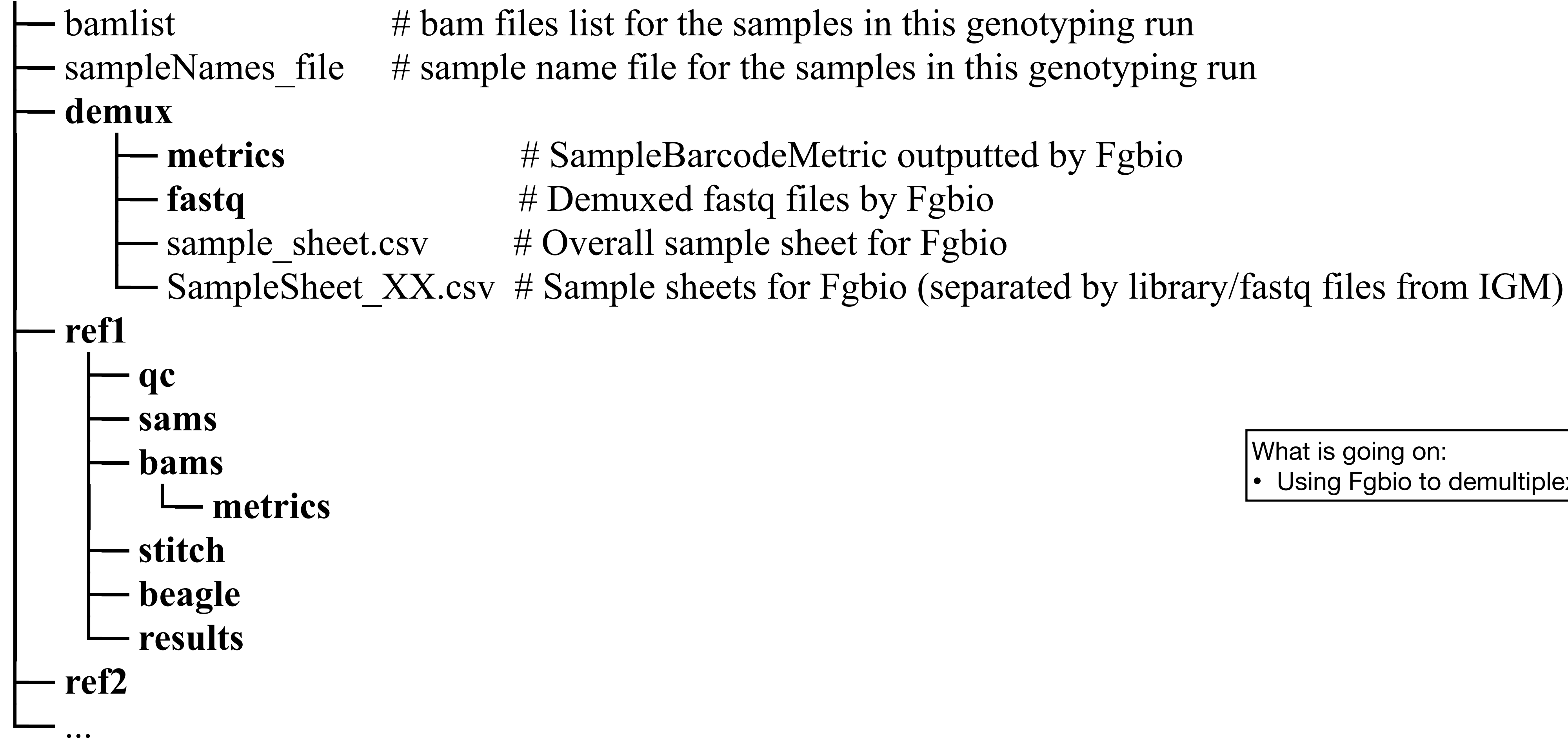


What is going on:

- Construct the basic structure of the directory
- Split the sample sheet for each library prep for Fgbio
- Organize bam files and sample names for the genotype run
- Set up conda environment for the pipeline

Genotyping Step 2 - Demux

flowcell

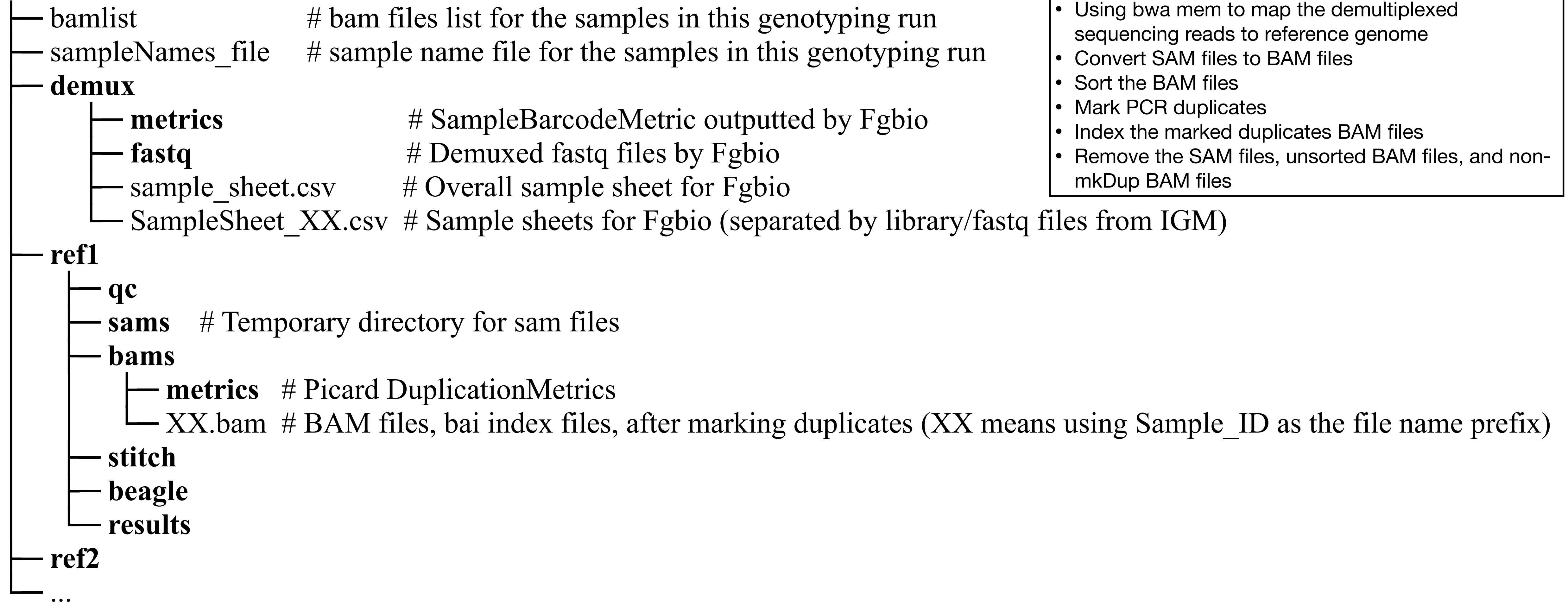


What is going on:

- Using Fgbio to demultiplex the fastq files

Genotyping Step 3 - Alignment

flowcell



What is going on:

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

Genotyping Step 4 - STITCH Genotype Calling

flowcell

— bamlist # bam files list for the samples in this genotyping run
— sampleNames_file # sample name file for the samples in this genotyping run

demux

— metrics # SampleBarcodeMetric outputted by Fgbio
— fastq # Demuxed fastq files by Fgbio
— sample_sheet.csv # Overall sample sheet for Fgbio
— SampleSheet_XX.csv # Sample sheets for Fgbio (separated by library/fastq files from IGM)

ref1

— qc
— sams # Temporary directory for sam files
— bams
 — metrics # Picard DuplicationMetrics
 — XX.bam # BAM files, bai index files, after marking duplicates (XX means using Sample_ID as the file name prefix)
— stitch # vcf.gz files, after variant calling and imputation with STITCH
— beagle
— results

ref2

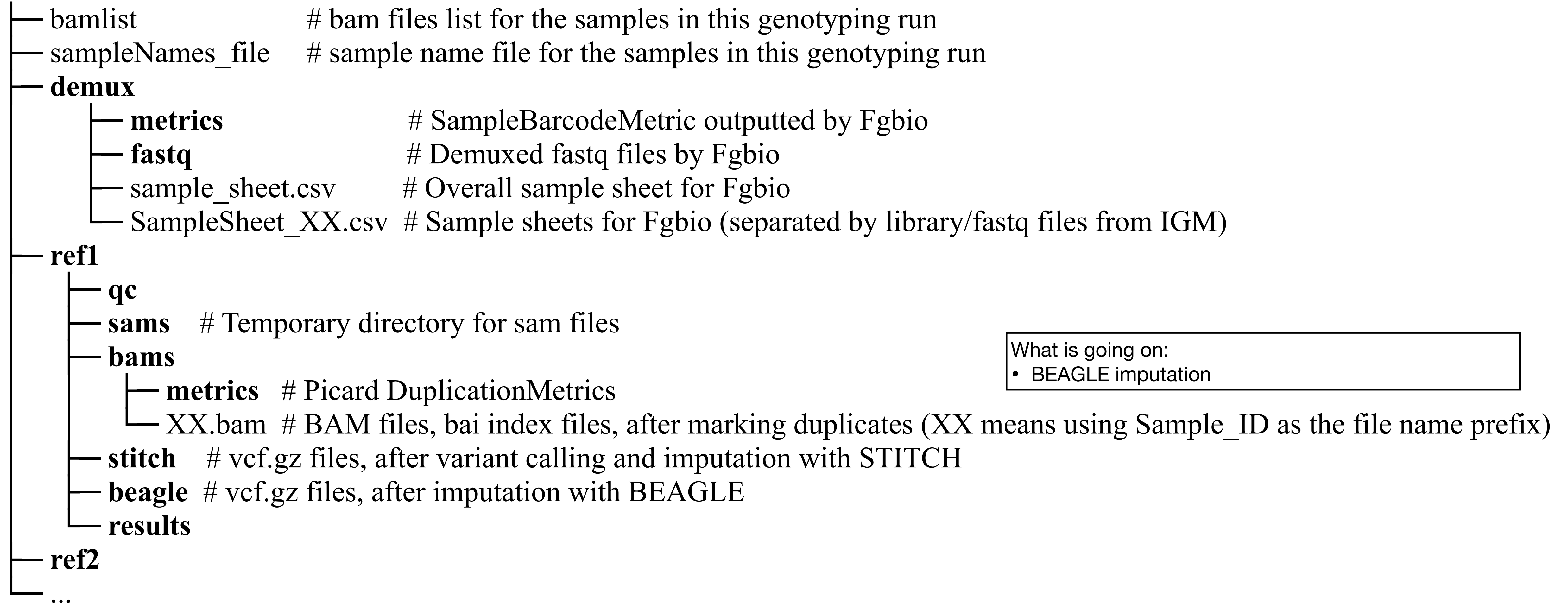
...

What is going on:

- STITCH variant calling

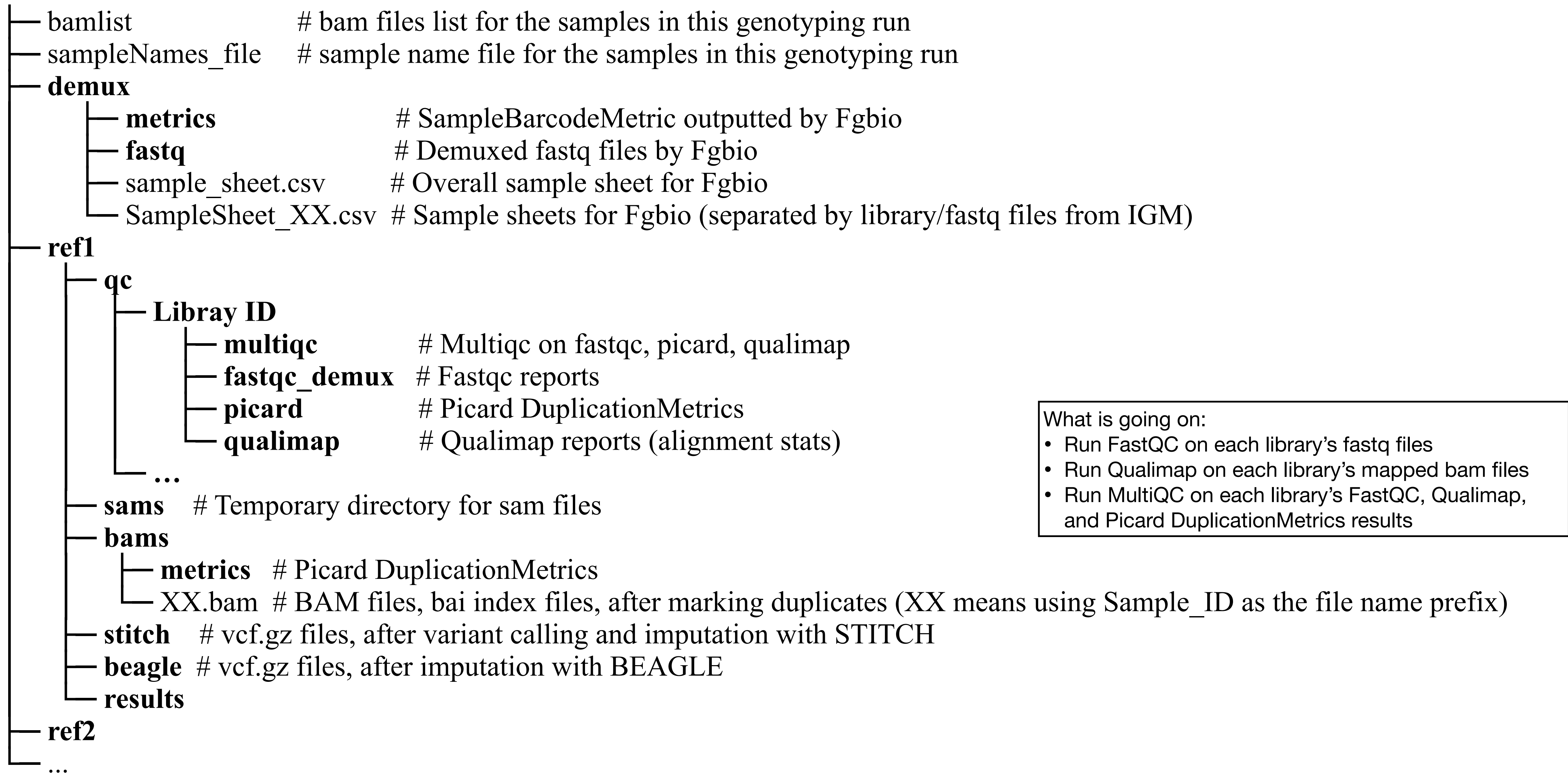
Genotyping Step 5 - BEAGLE Imputation

flowcell

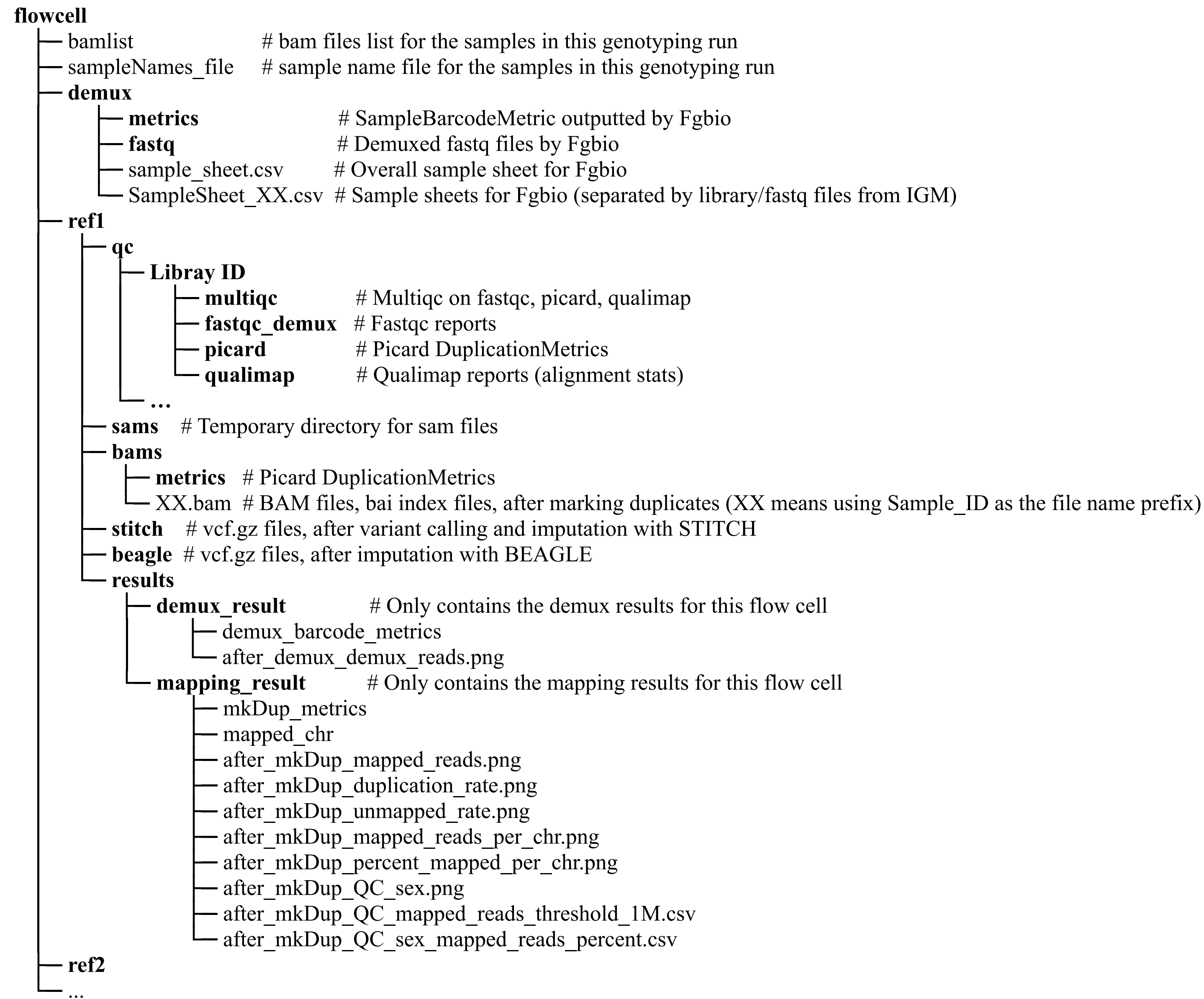


Quality Control Step 1 - MultiQC

flowcell



Quality Control Step 2 - Demux and Mapping Results



What is going on:

- Make plots for demux result and mapping result

Quality Control Step 3 - Genotyping Results



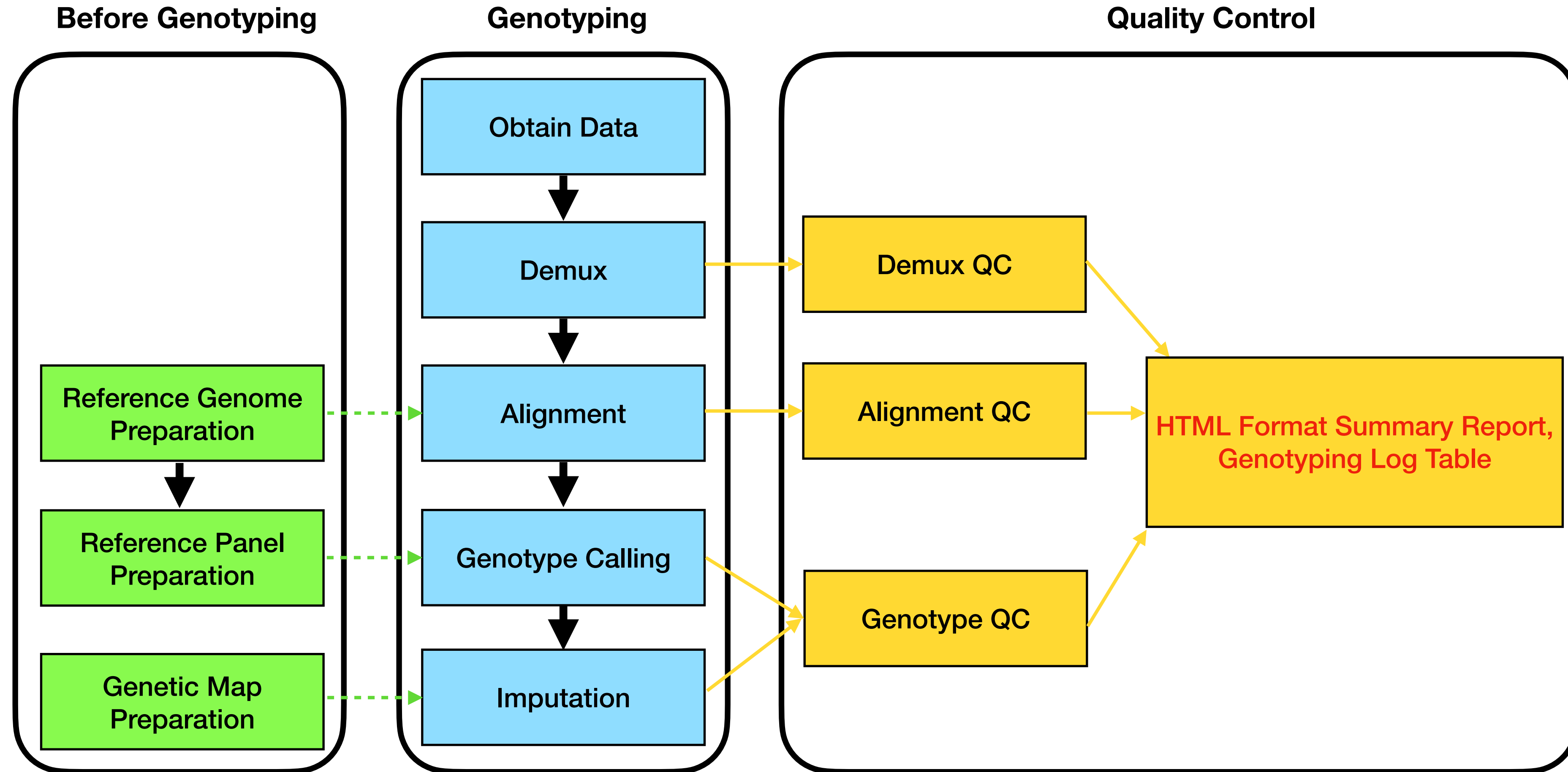
What is going on:

- Make plots for genotypes result

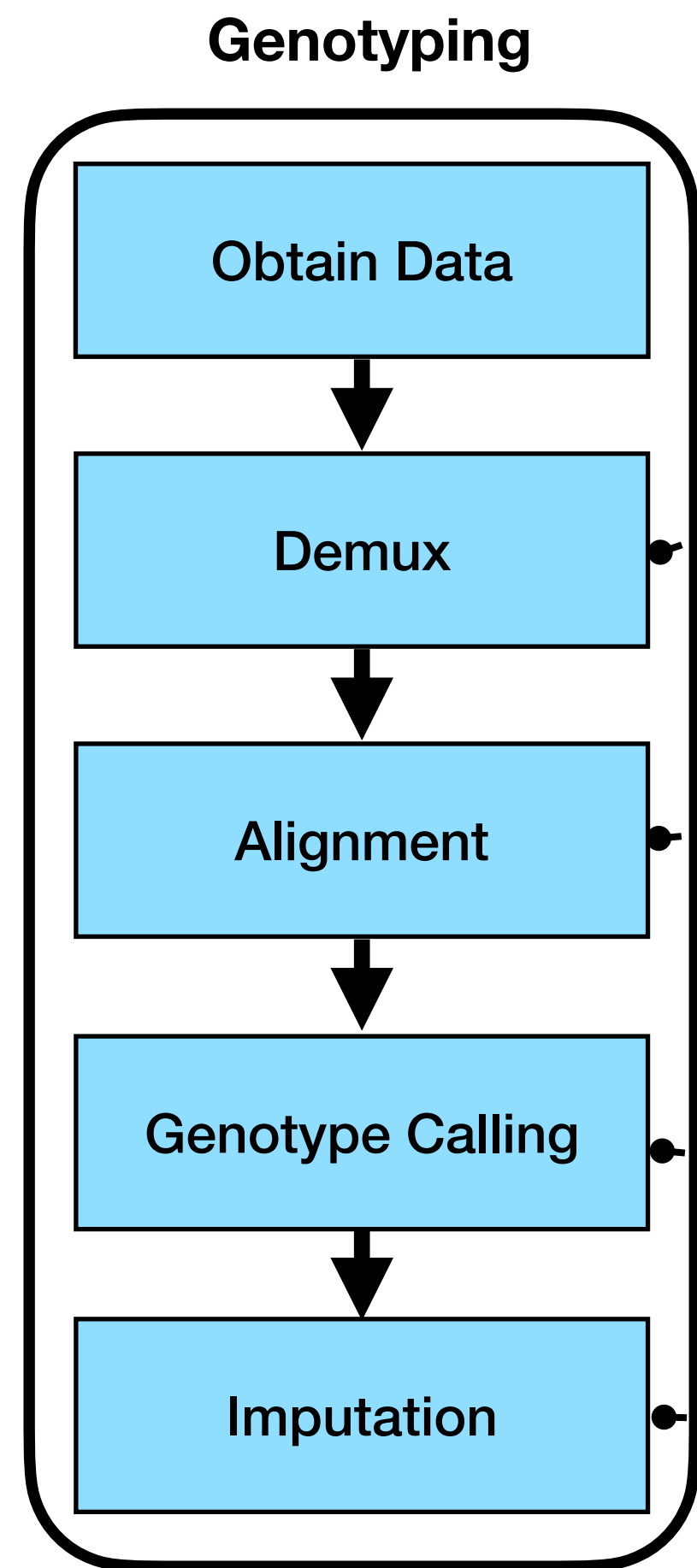
Directory Structure



Pipeline Overview



Genotyping Session Flowchart



```
java -jar 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 \  
  --metrics ${out_path}/${fastq_temp}demux_barcode_metrics.txt --output ${out_path}
```

```
bwa mem -aM \  
  -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} ${f}_R1.fastq.gz ${f}_R2.fastq.gz > ${f}.sam &  
  
samtools view -h -b -t ${reference_data} -o ${f}.bam ${f}.sam
```

```
samtools sort -o ${out_path}/bams/${f}_sorted.bam ${out_path}/bams/${f}.bam
```

```
java -jar picard.jar MarkDuplicates \  
  --REMOVE_DUPLICATES false --INPUT ${f}_sorted.bam --ASSUME_SORTED true \  
  --METRICS_FILE ${f}_sorted_mkDup_metrics.txt --OUTPUT ${f}_sorted_mkDup.bam &  
  
samtools index ${f}_sorted_mkDup.bam ${out_path}/bams/${f}_sorted_mkDup.bai
```

```
STITCH(buffer = 1e+6, method = "diploid", reference_haplotype_file = refHap,  
        reference_legend_file = refLgd, K = 16, niterations = 1, nGen = 100)
```

```
java -jar beagle.27Jan18.7e1.jar \  
  gt=chr${ch}_stitch.vcf.gz gprobs=true ne=150 \  
  map=${genetic_map}/chr${ch}_beagle.map out=chr${ch}_bgl
```