PALMER LAB
Behavioral Genetics of Mice, Rats and Men

Genes and Addiction
NIDA Center for GWAS in Outbred Rats

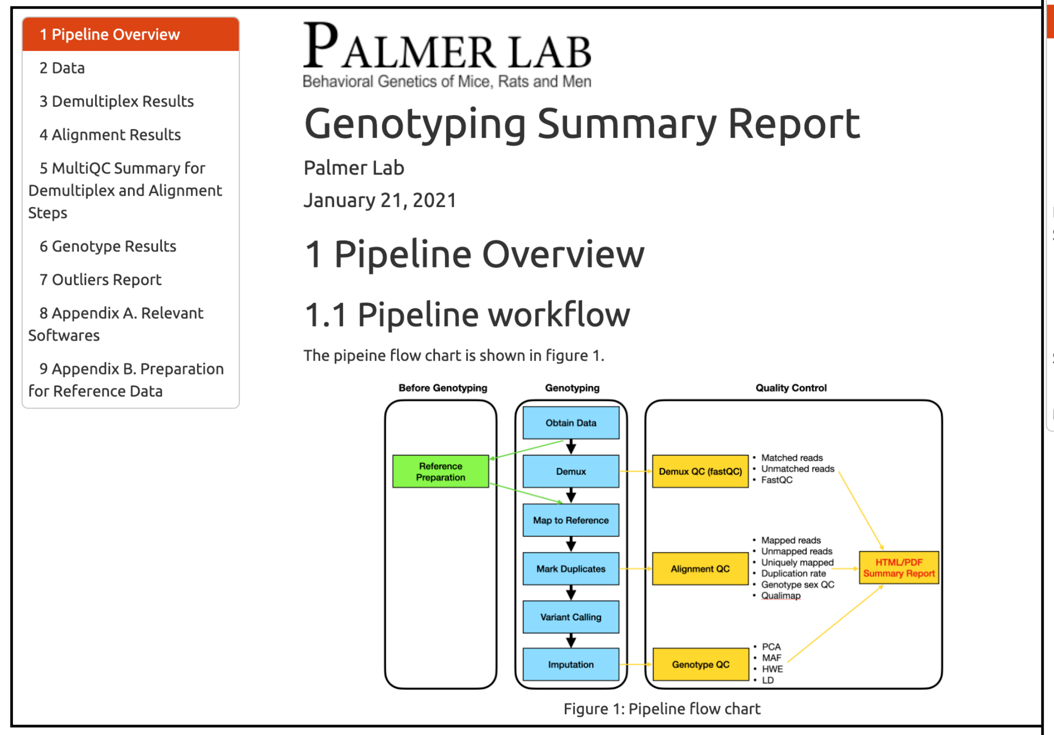
HS Rats Genotyping Pipeline

Pipeline Design

UCSD, Palmer Lab

Denghui Chen
Nov. 12 2021

Pipeline Overview



PALMER LAB
Behavioral Genetics of Mice, Rats and Men

Genotyping Summary Report

Palmer Lab

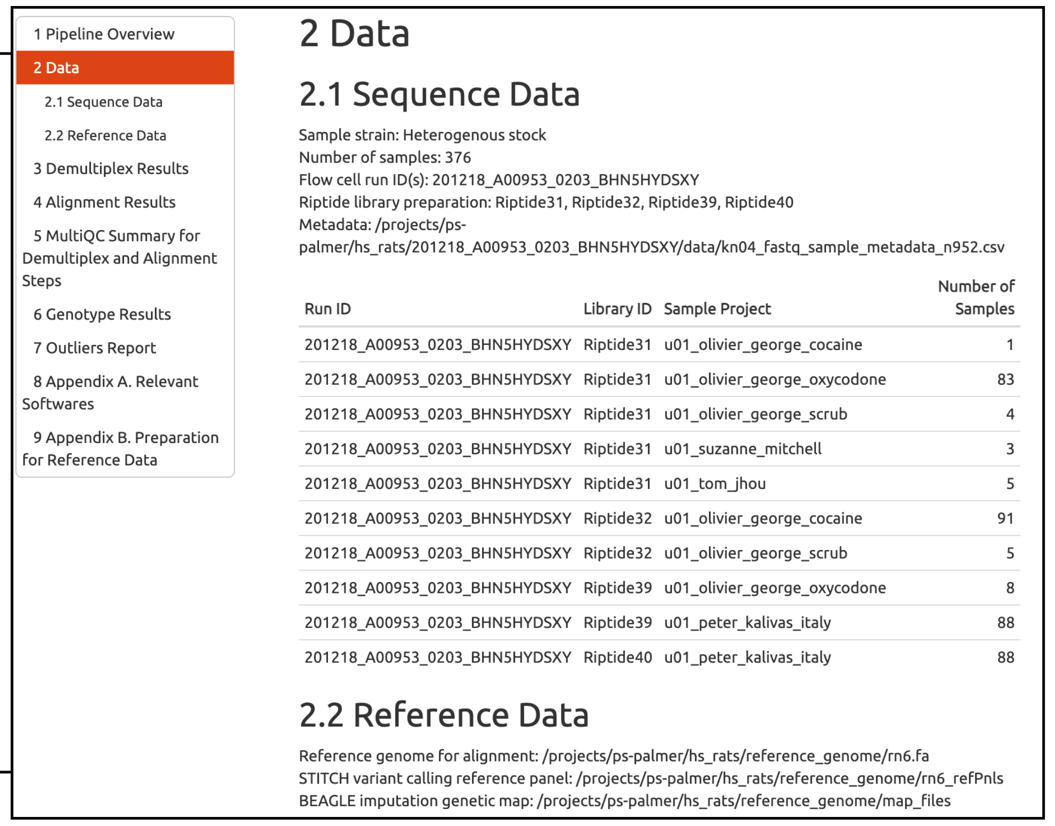
January 21, 2021

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

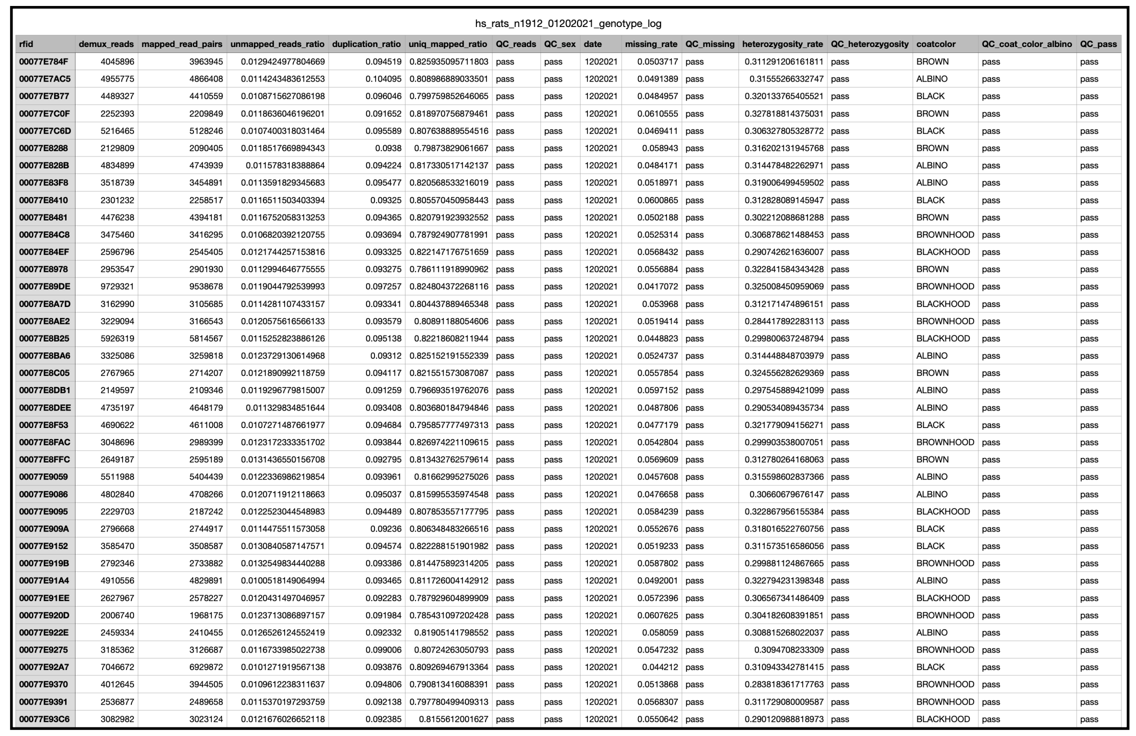
Sample strain: Heterogenous stock
Number of samples: 376
Flow cell run ID(s): 201218_A00953_0203_BHN5HYDSXY
Riptide library preparation: Riptide31, Riptide32, Riptide39, Riptide40
Metadata: /projects/ps-palmer/hs_rats/201218_A00953_0203_BHN5HYDSXY/data/kn04_fastq_sample_metadata_n952.csv

Run ID	Library ID	Sample Project	Number of Samples
201218_A00953_0203_BHN5HYDSXY	Riptide31	u01_olivier_george_cocaine	1
201218_A00953_0203_BHN5HYDSXY	Riptide31	u01_olivier_george_oxycodone	83
201218_A00953_0203_BHN5HYDSXY	Riptide31	u01_olivier_george_scrub	4
201218_A00953_0203_BHN5HYDSXY	Riptide31	u01_suzanne_mitchell	3
201218_A00953_0203_BHN5HYDSXY	Riptide31	u01_tom_jhou	5
201218_A00953_0203_BHN5HYDSXY	Riptide32	u01_olivier_george_cocaine	91
201218_A00953_0203_BHN5HYDSXY	Riptide32	u01_olivier_george_scrub	5
201218_A00953_0203_BHN5HYDSXY	Riptide39	u01_olivier_george_oxycodone	8
201218_A00953_0203_BHN5HYDSXY	Riptide39	u01_peter_kalivas_italy	88
201218_A00953_0203_BHN5HYDSXY	Riptide40	u01_peter_kalivas_italy	88

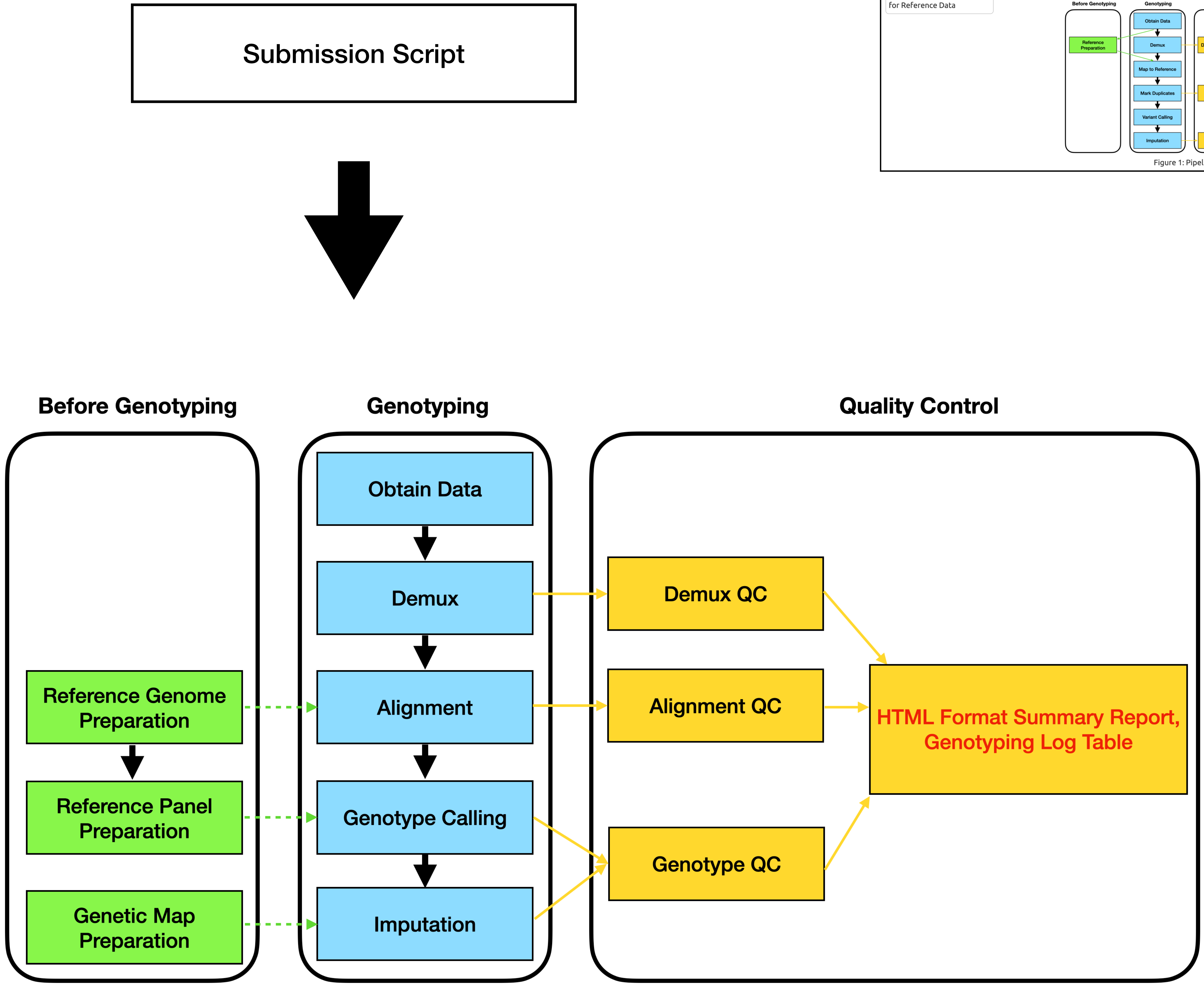
2.2 Reference Data

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

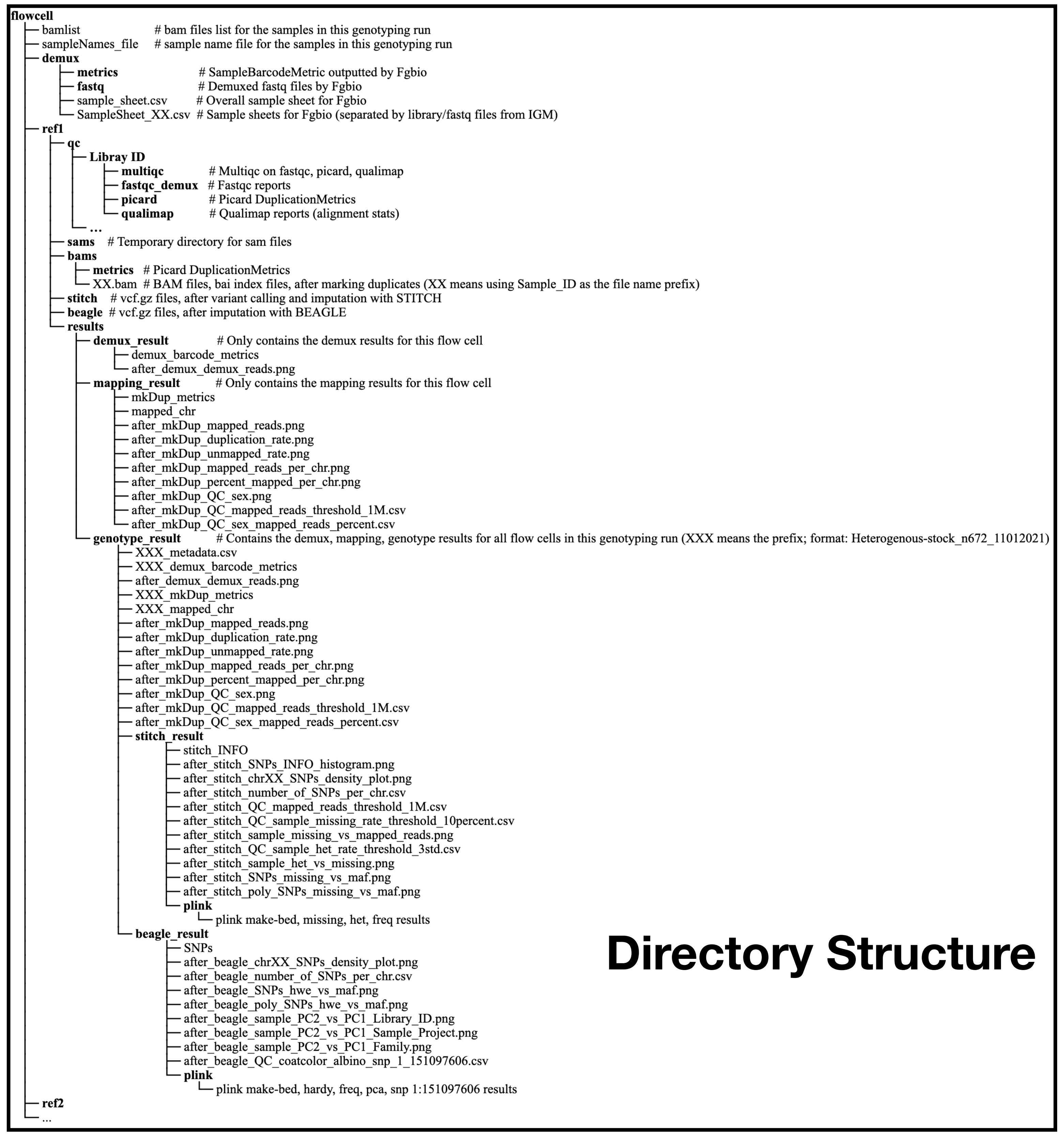
Summary Report



Genotyping Log Table



Flowchart



```
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
│   │   ├─ Libray ID
│   │   │   ├─ multiqc       # 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_10percent.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
└─ ...
```

Directory Structure

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
- 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**
  - **fastq**
  - 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**
  - **bams**
    - **metrics**
  - **stitch**
  - **beagle**
  - **results**
- **ref2**
- …

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
- 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**
  - **bams**
    - **metrics**
  - **stitch**
  - **beagle**
  - **results**
- **ref2**
- ...

What is going on:
- Using Fgbio to demultiplex the fastq files

Genotyping Step 3 - Alignment

```
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
│   ├── beagle
│   └── results
├── ref2
└── ...
```

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
- 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** # vcf.gz files, after imputation with BEAGLE
  - **results**
- **ref2**
- ...

What is going on:
- BEAGLE imputation

Quality Control Step 1 - MultiQC

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**
    - **Libray ID**
      - **multiqc** # 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**
- **ref2**
- ...

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

Quality Control Step 2 - Demux and Mapping Results

```
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
│   │   ├── Libray ID
│   │   │   ├── multiqc         # 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
├── ref2
└── ...
```

What is going on:

- Make plots for demux result and mapping result

Quality Control Step 3 - Genotyping Results

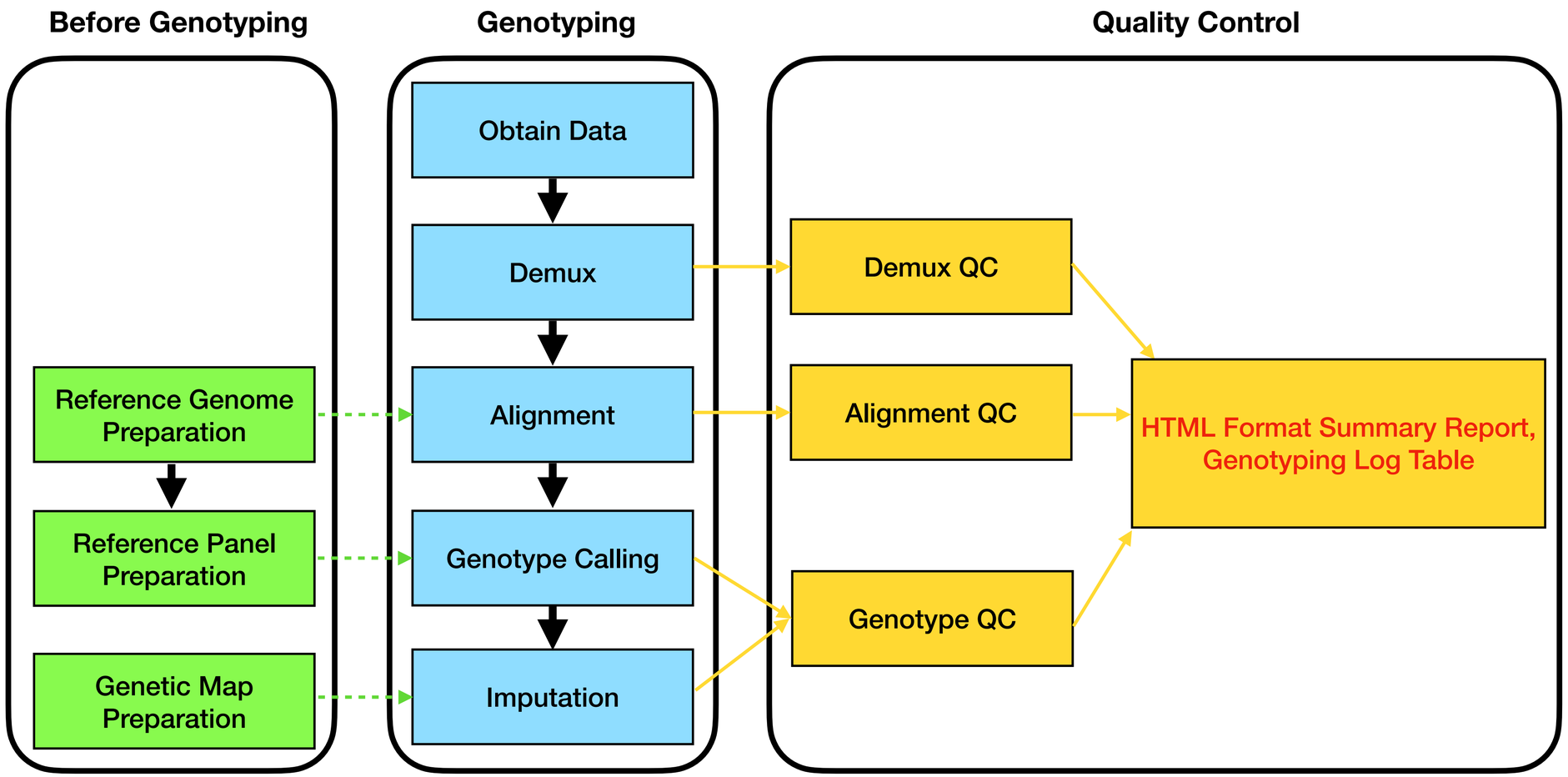
flowcell
- bamlist # bam files list for the samples in this genotyping run
- sampleNames_file # sample name file for the samples in this genotyping run
- **demux**
  - **metrics** # SampleBarcodeMetric outputted by Fgbio
  - **fastq** # Demuxed fastq files by Fgbio
  - sample_sheet.csv # Overall sample sheet for Fgbio
  - SampleSheet_XX.csv # Sample sheets for Fgbio (separated by library/fastq files from IGM)
- **ref1**
  - **qc**
    - **Libray ID**
      - **multiqc** # 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_10percent.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**
- …

What is going on:
- Make plots for genotypes result

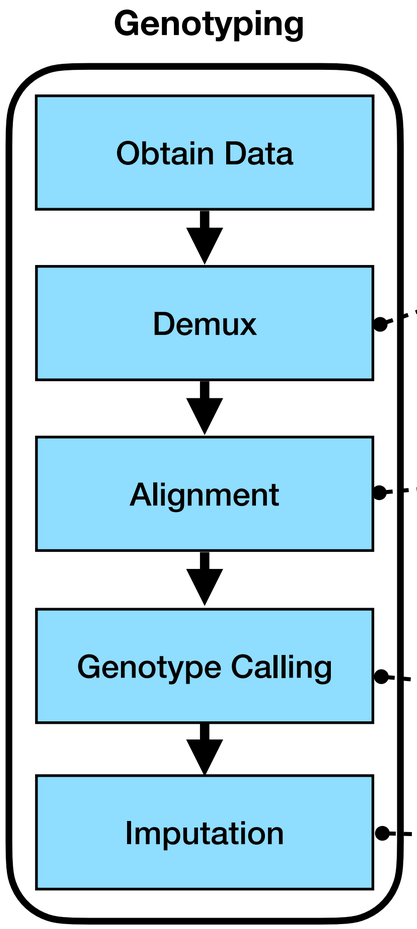
Directory Structure

- **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**
      - **Library ID**
        - **multiqc** # 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_10percent.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 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
```