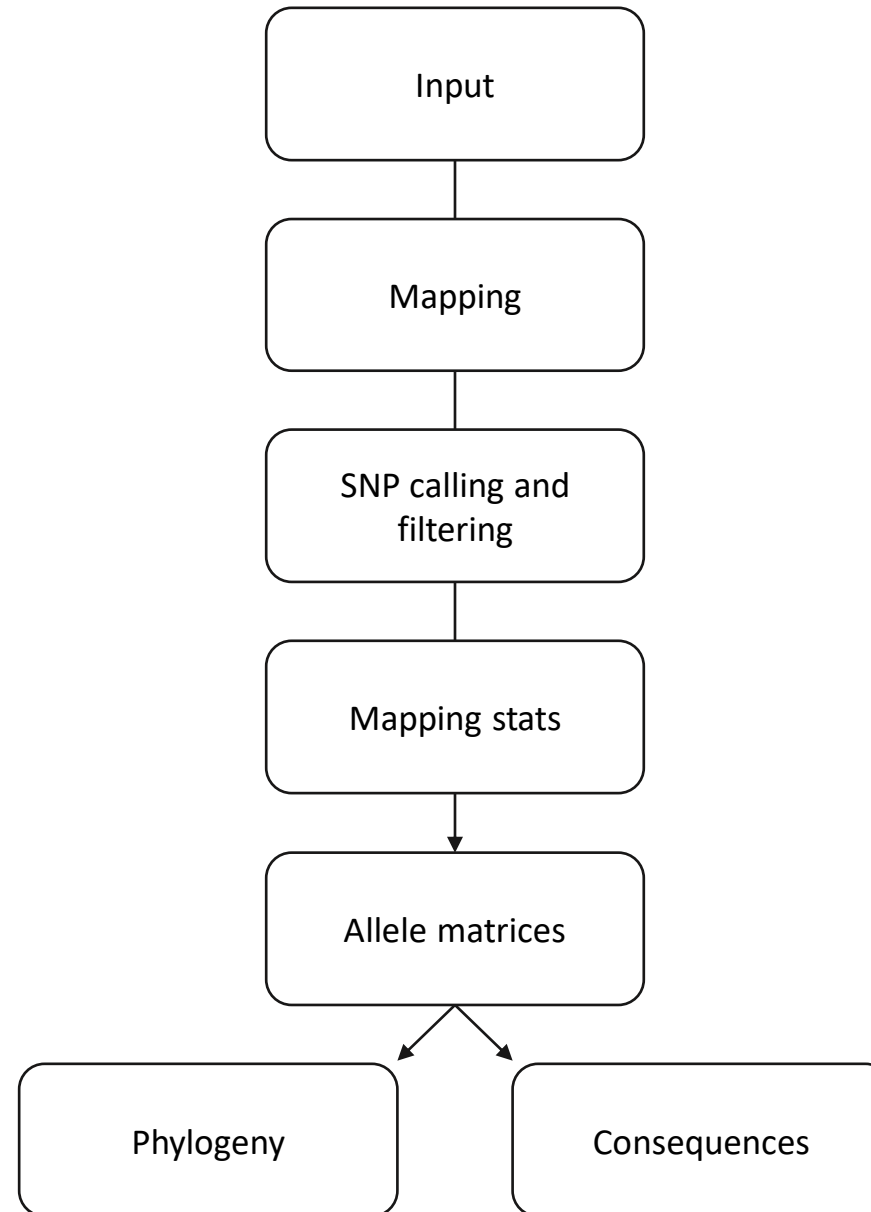
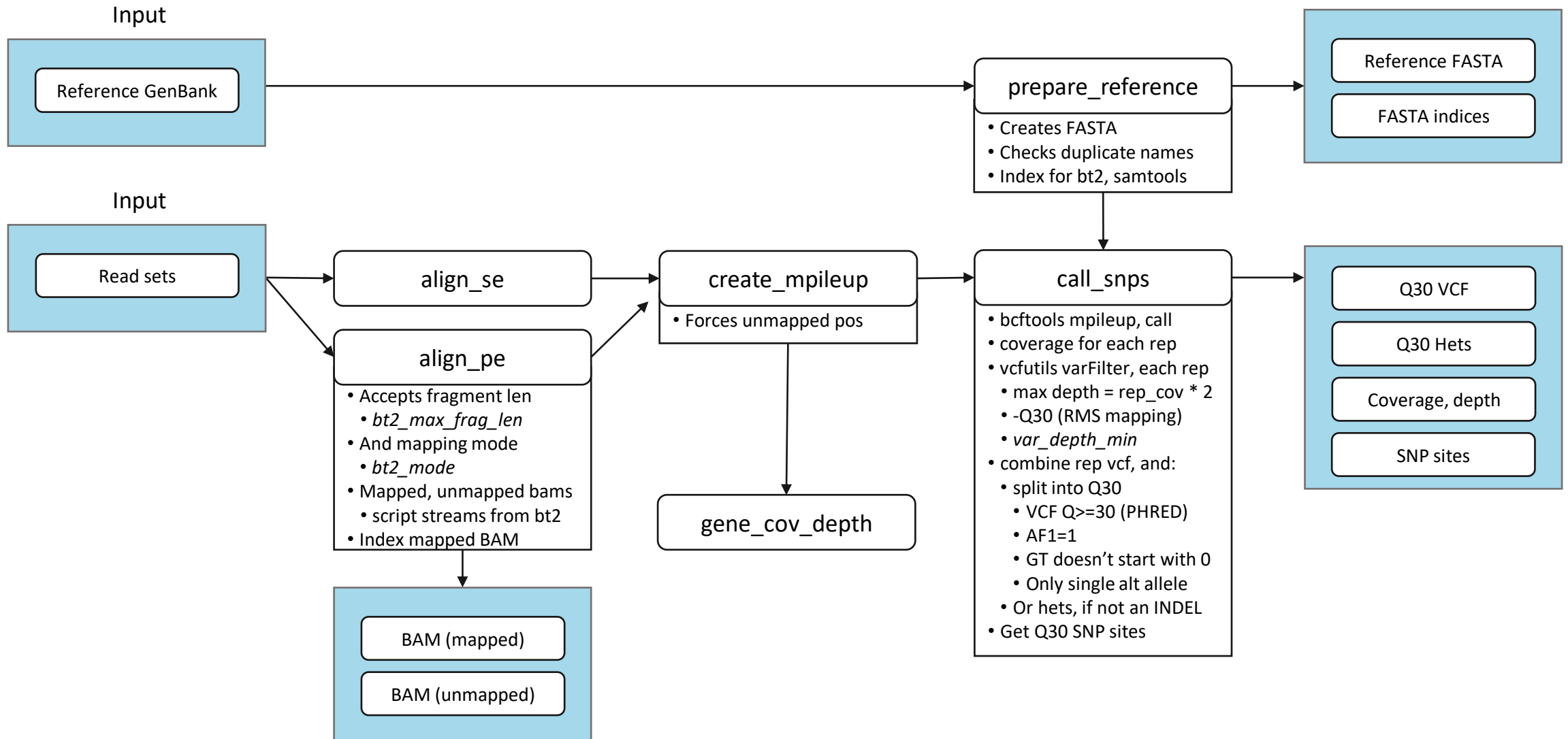


Implementation

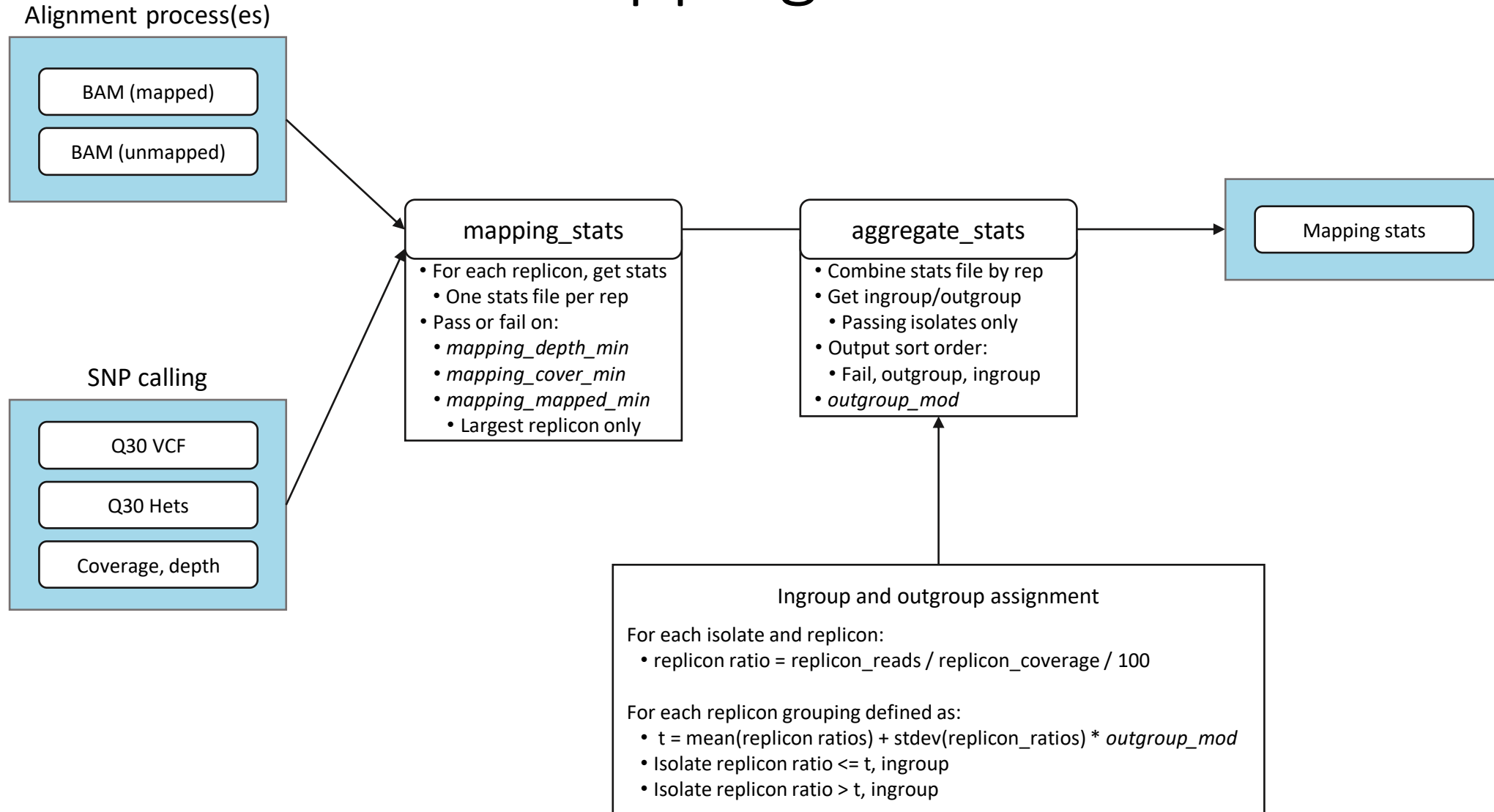
High level overview

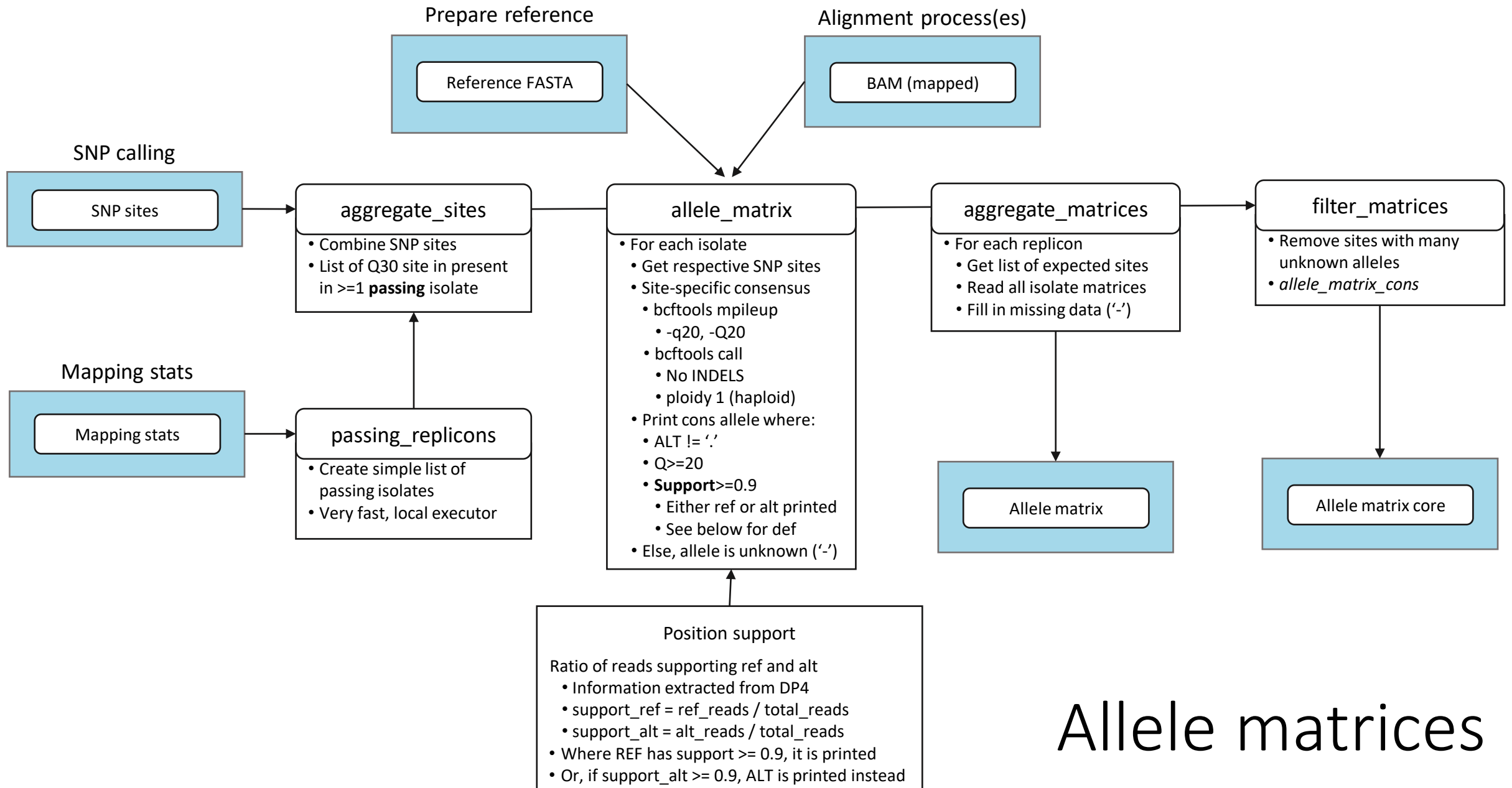


Mapping and SNP calling



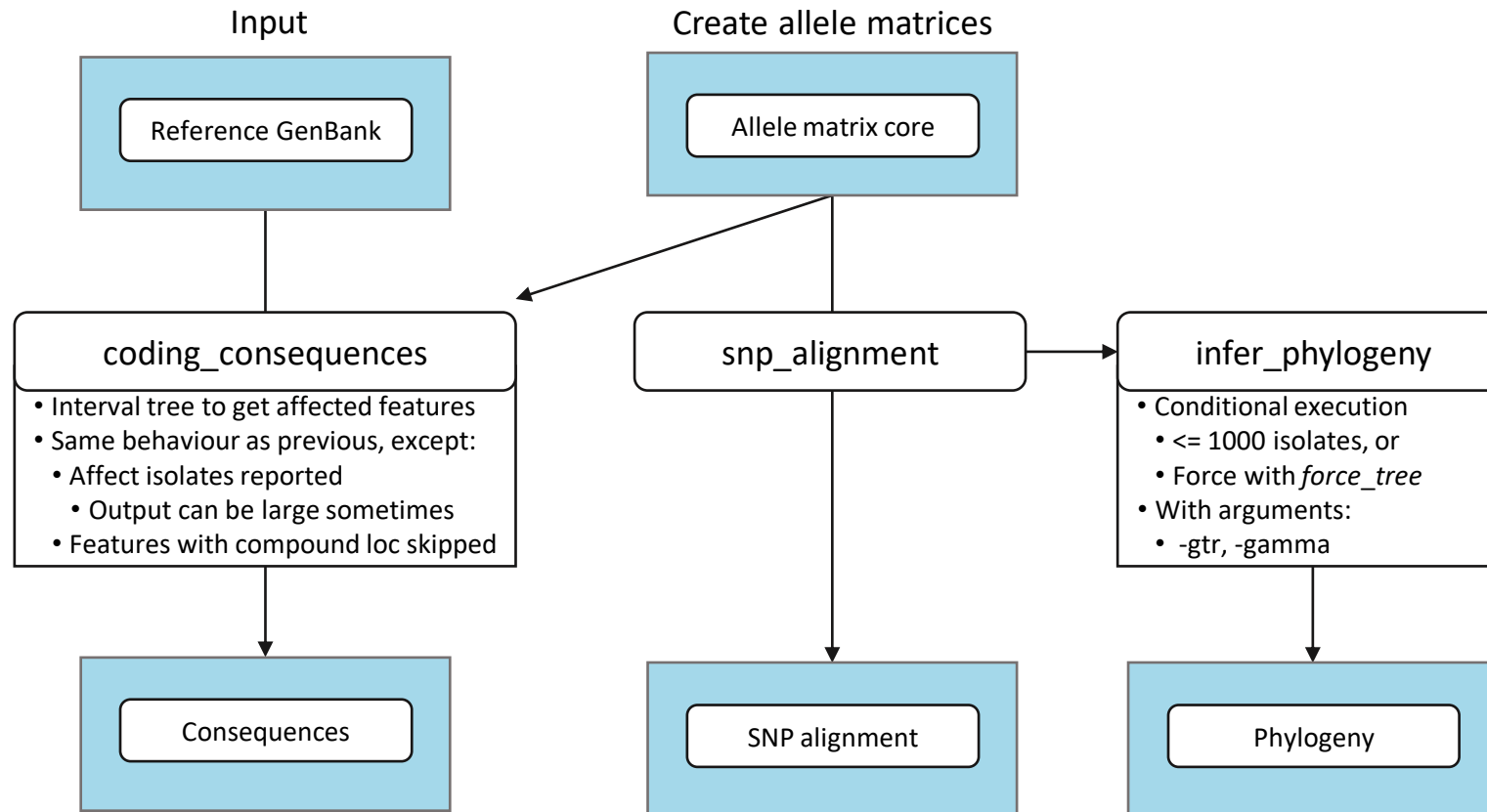
Mapping stats



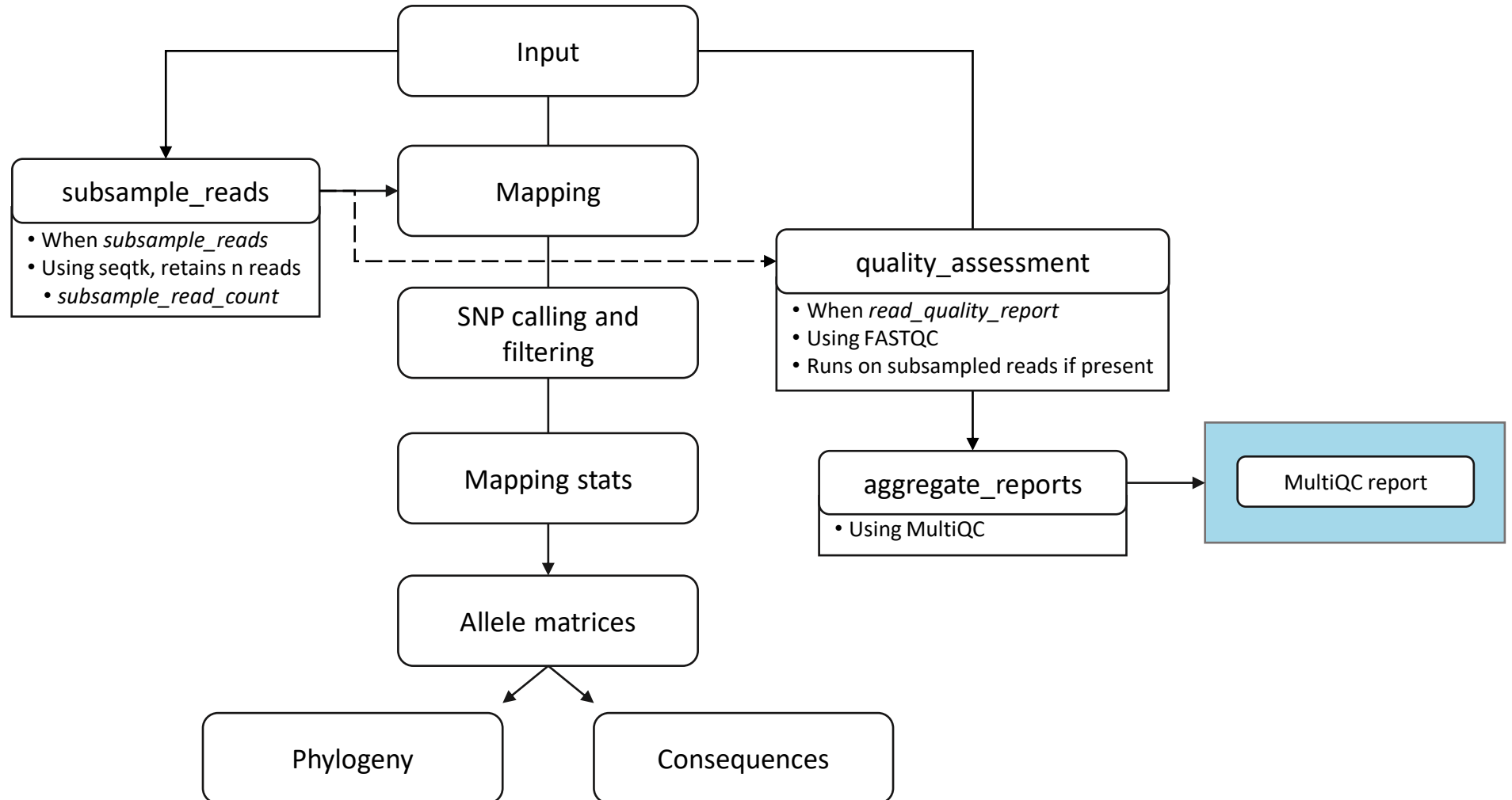


Allele matrices

Consequences, phylogeny



Optional stages



Merge run settings

merge_run: enables a merge run

output_dir: output directory

previous_run_dir: data to merge

merge_ignore_errors: ignore merge data validation errors

Merge runs

Input

Previous run dir

Read sets

Reference Genbank

validate_merge_data

- Initial basic check that data exists
- Check run configuration matches
 - Same thresholds, mapping params, etc
- Check reference matches
 - Reference name, replicon names
 - Replicon sequence md5 hashes
- Isolates present in expected outputs
 - Using BAM files as ground truth
- Check no collisions in isolate namespace

Data merged

Gene cov, depth

Mapping stats

Channels merged

BAMs (mapped)

Mapping stats

FastQC

Allele matrices

SNP Sites

Process Overview

Symlinked

BAMs (mapped)

VCFs (Q30, hets)

FastQC

Input

Mapping

SNP calling and filtering

Mapping stats

Allele matrices

Phylogeny

Consequences

Testing

Dataset simulator

Create readsets from single specification file and reference

Isolate spec section

- Name
- Mean depth
- Read length and outer length (for PE)
- Unmappable proportion of reads

Example:

```
# Defaults
metric value type
mean_depth 15 int
outer_length 750 int
read_length 250 int
# Readsets
isolate_name data
## Coding consequences test isolate
isolate_1 read_type:pe
isolate_2 read_type:pe
## Mapping stats pass/ fail test isolates
## Pass
isolate_6 unmapped:0.5;read_type:pe
isolate_7 unmapped:0.45;read_type:pe
isolate_8 mean_depth:11;read_type:pe
isolate_9 mean_depth:25;read_type:pe
isolate_10 mean_depth:11;unmapped:0.45;read_type:pe
```

Modification spec section

- Homozygous SNPs
- Heterozygous SNPs, at specific ratio
- INDELs
- Low quality positions, appear as '-' in allele matrix

Example:

```
# Variants
isolate_name replicon type data
## SNPs to testing coding consequences - start and stop codons, and various others
## Non-synonymous
isolate_1 contig_1 hom position:1158;alt:a;note:sul1_G2R
isolate_1 contig_1 hom position:1644;alt:g;note:sul1_T164A
isolate_1 contig_1 hom position:2081;alt:t;note:sul1_*309Y
isolate_1 contig_1 hom position:2092;alt:t;note:secA_M1L
## Heterozygous SNPs
## Both different to reference allele
isolate_3 contig_1 het position:672;alt_1:t;alt_2:c;ratio:0.5
isolate_3 contig_1 het position:3181;alt_1:t;alt_2:c;ratio:0.4
## Site with <5% unknown - retained (1/22 unknown (4.5%))
isolate_18 contig_1 hom position:4820;alt:g;note:secD_S5A
isolate_19 contig_1 low_quality position:4820
## Site with >5% unknown - filtered (2/22 unknown (9.1%))
isolate_20 contig_1 hom position:4850;alt:a;note:secD_L15M filtered
isolate_21 contig_1 low_quality position:4850
isolate_22 contig_1 low_quality position:4850
```

Automated testing

Tests expected output data from pipeline run on simulated dataset

- Fast turn around for debug loop
- Runtime for 37 isolates on 31 kpb reference (two reps) of ~60 seconds

Compares data specification file to:

- Mapping stats
- Allele tables
- Consequences

Merge run testing approach is to:

- Split simulated dataset into two groups
- Execute a typical run the first group
- Execute merge run with second group
- Run automated comparison

Secondary to simulated datasets, we can test previously known good runs

- Requires previous run output
- Script compares all major output files
- Performs ordering of data for correct comparison