pdac_demo

Section 1B: sample and variant−level genotyping quality control

Headline Counts

Metric	Value
Raw samples	1,461
Final samples	1,239
Sample retention	84.8%
Raw variants	430,000
Final variants	401,909
Variant retention	93.5%

Report Index

This report is produced from the output files generated by Steps 01−08. It combines the count table, filtering rationale, and QC figures into one portable PDF that can be reviewed before moving to PCA and association testing.

1. QC Pipeline Count Table

Filtering trajectory

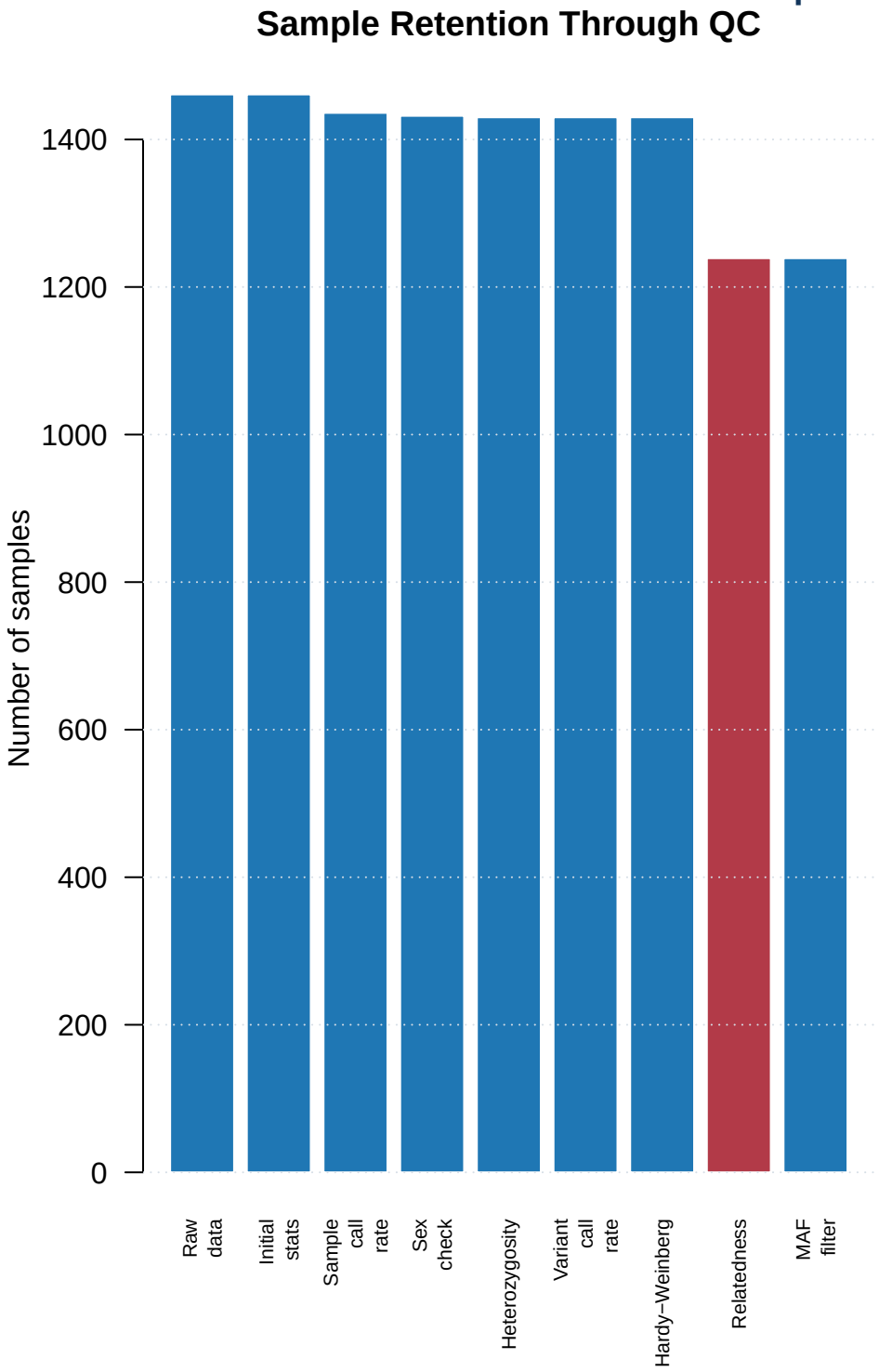
Step	Samples	Variants	Samples lost	Variants lost
00. Raw data	1,461	430,000	0	0
01. Initial stats	1,461	430,000	0	0
02. Sample call rate	1,436	430,000	25	0
03. Sex check	1,432	430,000	4	0
04. Heterozygosity	1,430	430,000	2	0
05. Variant call rate	1,430	421,520	0	8,480
06. Hardy−Weinberg	1,430	406,947	0	14,573
07. Relatedness	1,239	406,947	191	0
08. MAF filter	1,239	401,909	0	5,038

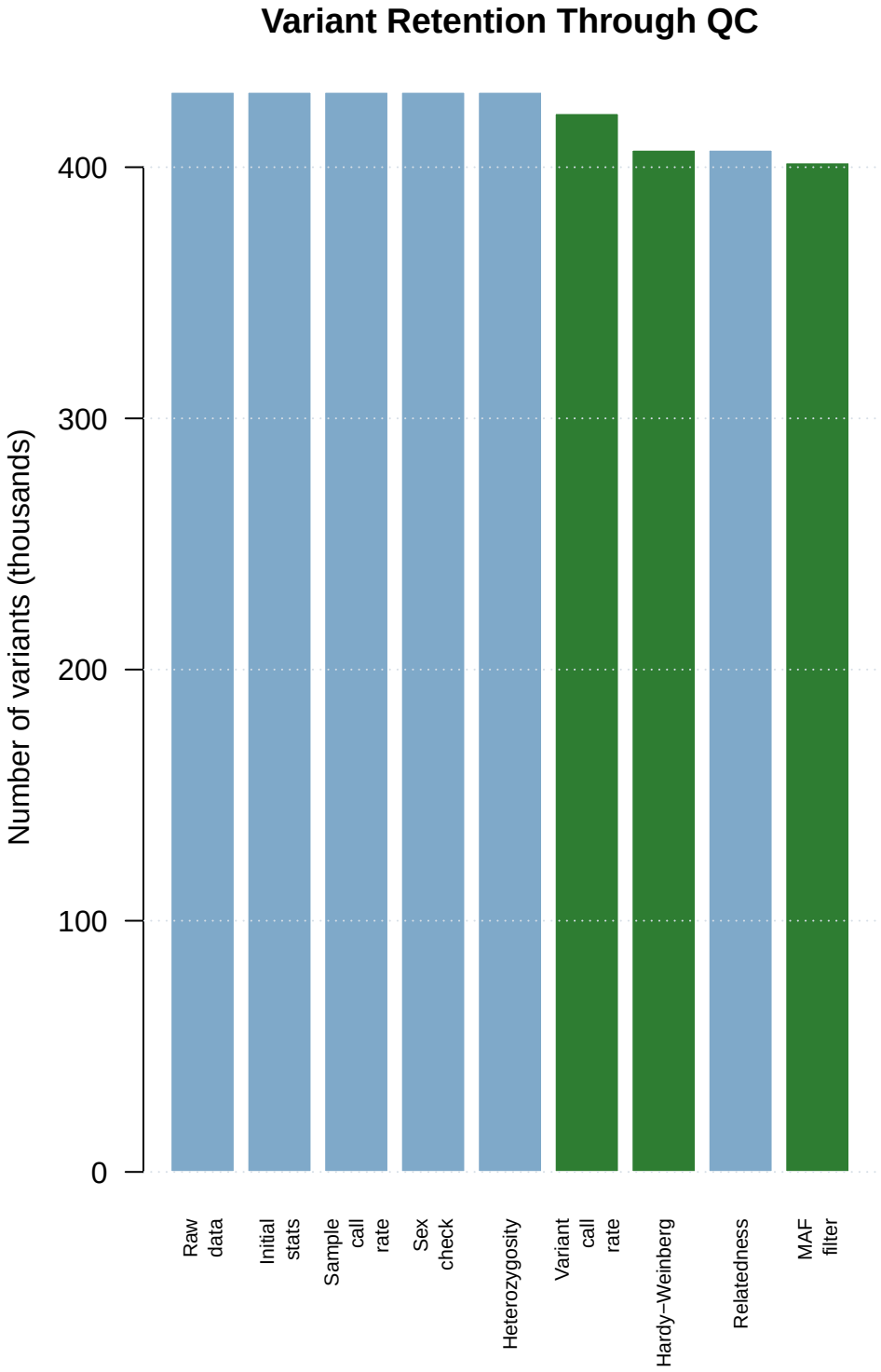
Total sample loss: 222 (15.2%)

Total variant loss: 28,091 (6.5%)

Interpretation: sample−level filters remove low−quality or non−independent individuals; variant−level filters remove unreliable markers before downstream population structure and association analyses.

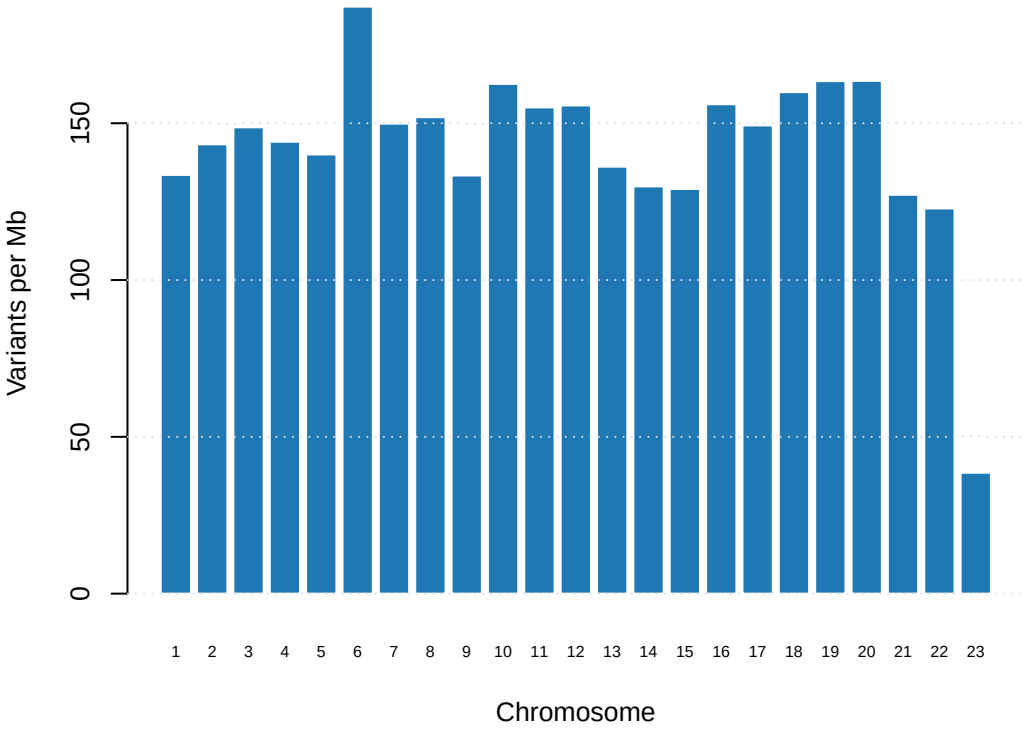
2. Sample and Variant Retention



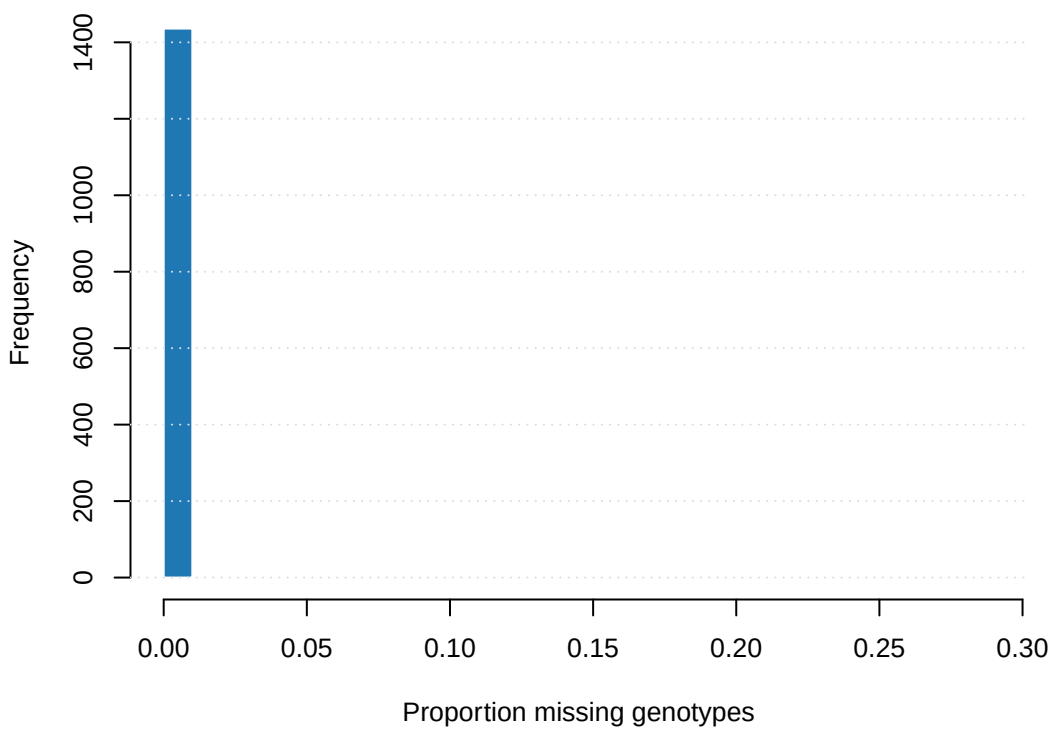


3. Initial QC Distributions

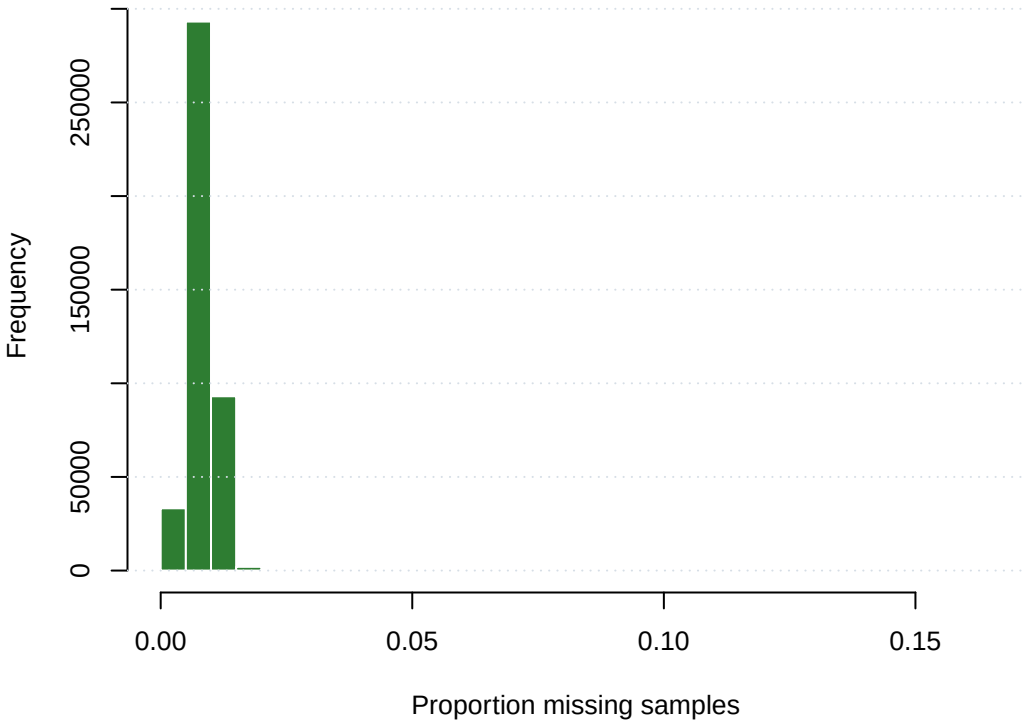
Variant Density by Chromosome



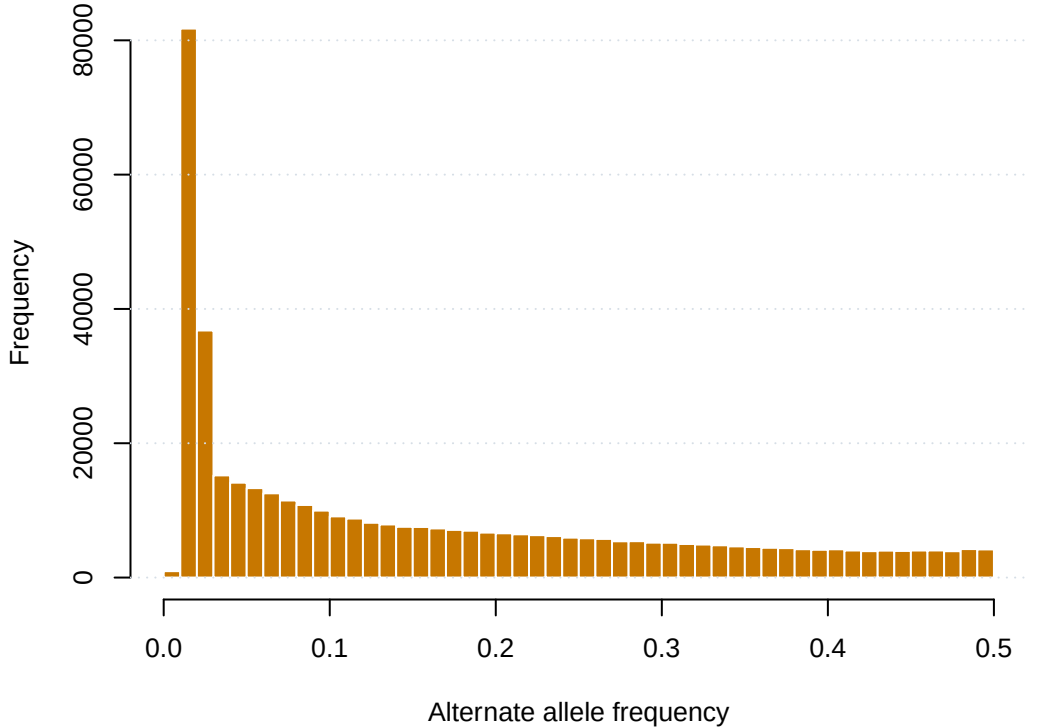
Sample Missingness



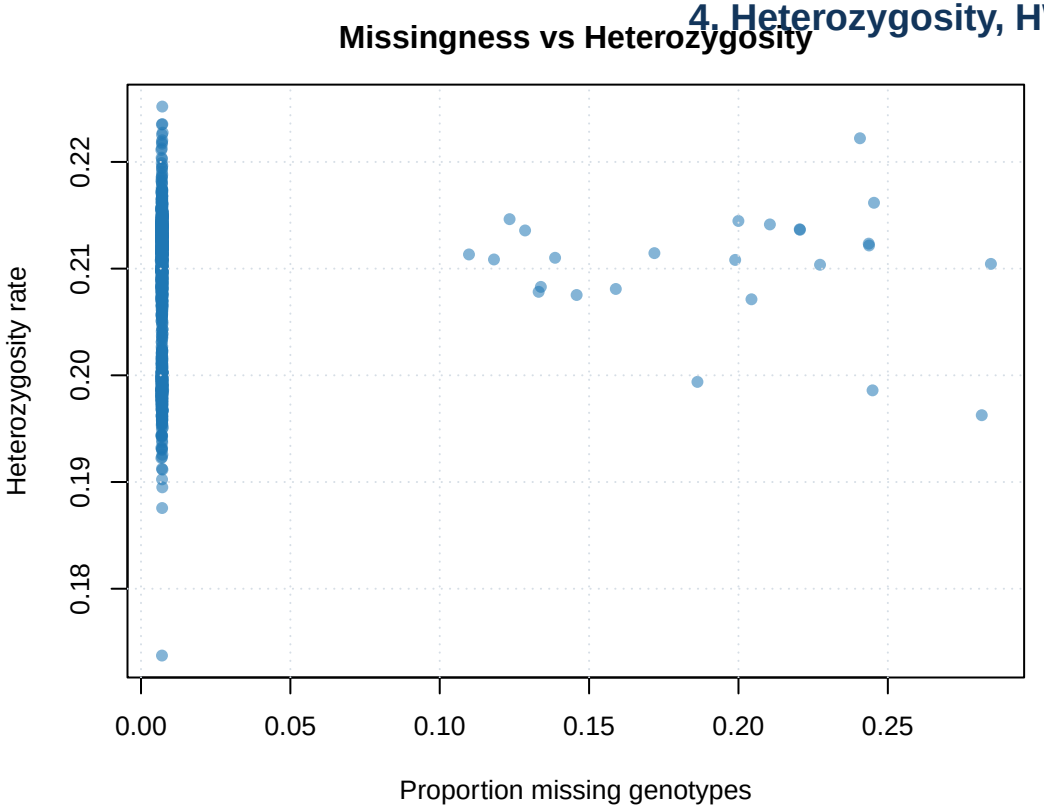
Variant Missingness

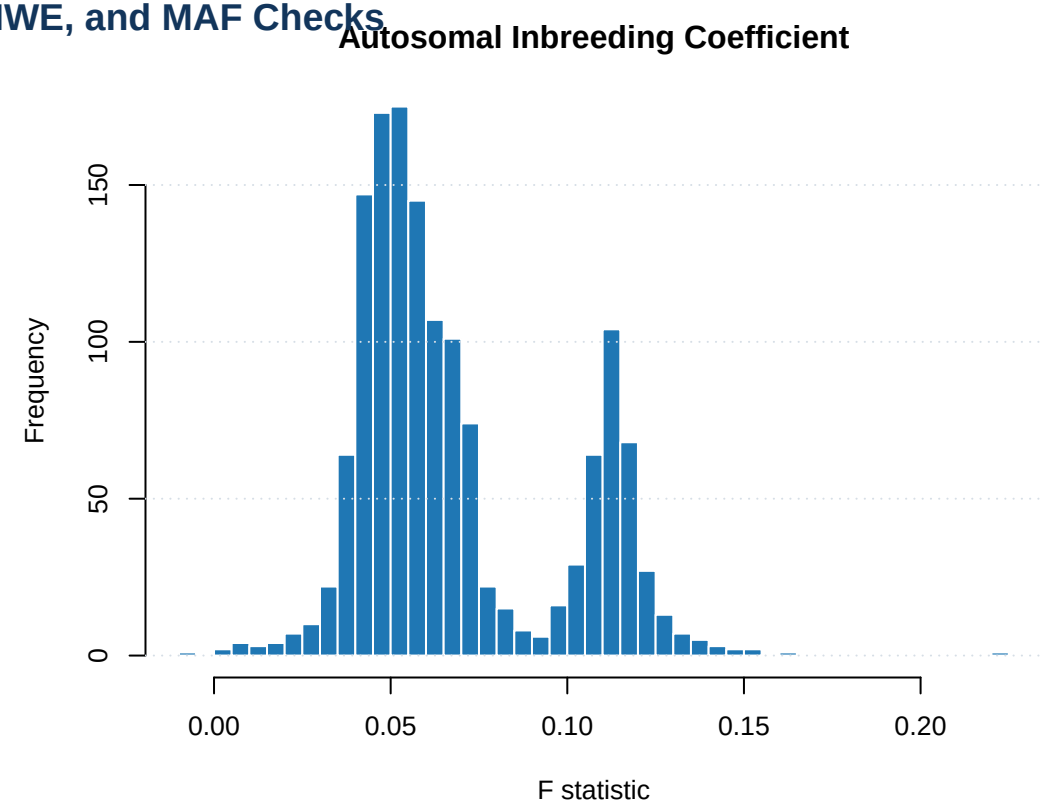


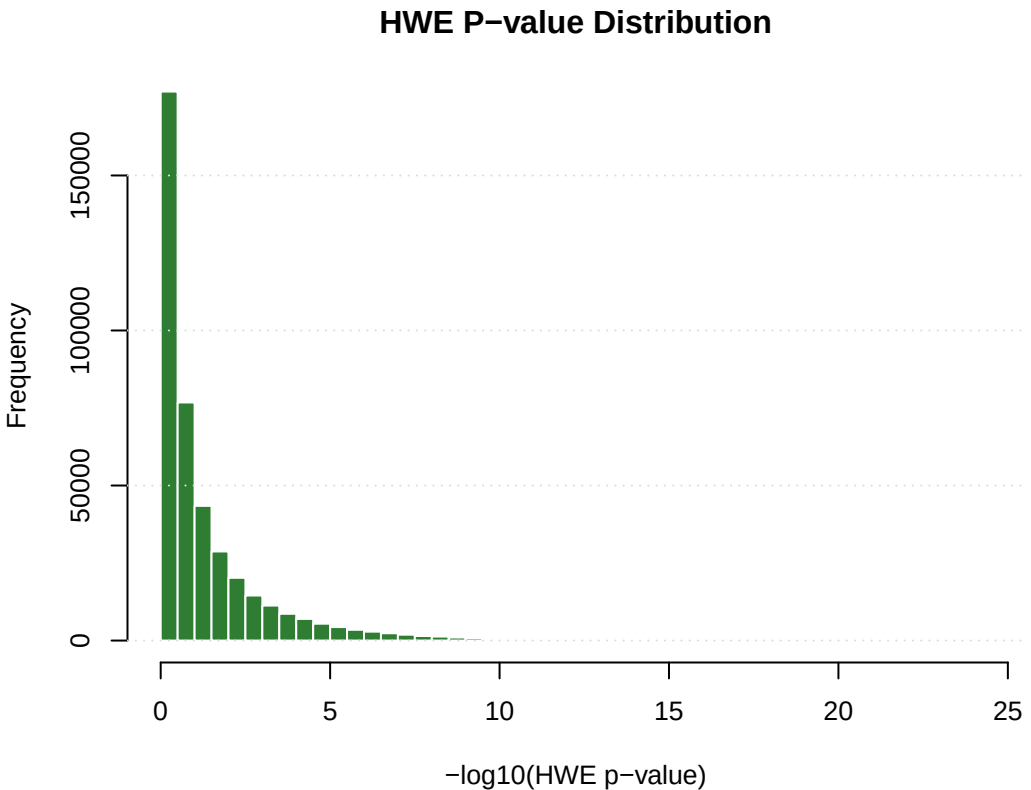
Initial Allele Frequencies

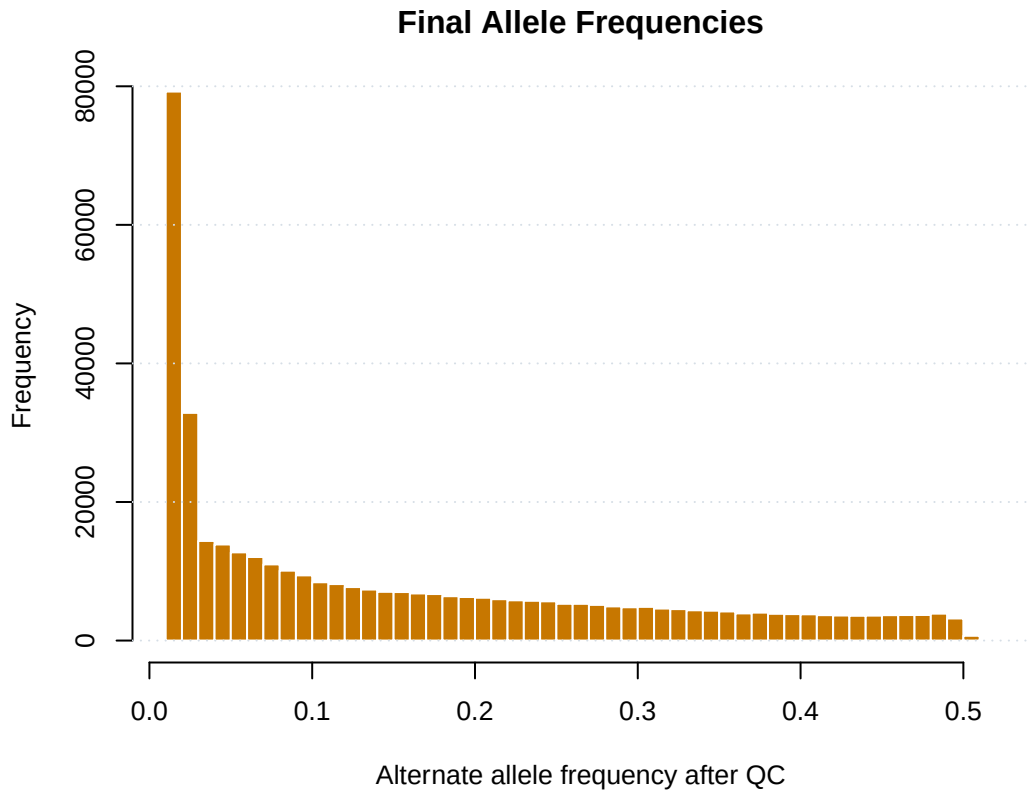


4. Heterozygosity, HWE, and MAF Checks









5. Relatedness Pruning Diagnostics

KING + PI_HAT

Headline diagnostics

Metric	Value
samples_before_relatedn...	1430
king_pruning_pairs_gt_0...	284
unique_samples_in_king_...	448
relatedness_components	164
largest_component_samples	6
largest_component_pairs	5
plink2_king_cutoff_remo...	167
phenotype_aware_removed...	191
phenotype_aware_removed...	13.4
controls_removed	154

Largest relatedness components

component id	n samples	n pairs	n controls	n cases	n unknown	max kinship	mean kinship
66	6	5	5	1	0	0.251522	0.24772
45	4	3	2	2	0	0.25497	0.248981
1	3	2	0	3	0	0.248696	0.248002
10	3	2	3	0	0	0.250308	0.24945
100	3	2	2	1	0	0.247884	0.247528
102	3	2	2	1	0	0.250526	0.250471

KING related−pair summary

Relationship	Pairs	Unique samples
Duplicate or twin	0	0
First degree	284	448
Second degree	4	8
Third degree	10	20
No third degree or clos...		964
Total samples before re...		1430

KING vs PI_HAT agreement

Metric	Value
total_pairs_in_either_r...	55337
pairs_reported_by_both	298
same_category	294
different_category_chec...	4
reported_by_king_only	0
reported_by_pihat_only	55039
pruning_pairs	284
pruning_pairs_with_pihat	284

Showing first 8 rows.

6. Methods Notes and Interpretation

Review before downstream analysis

QC thresholds

Level	Check	Threshold	Rationale
Sample	Call rate	−−mind 0.02	Remove low−quality samples
Sample	Sex check	X chromosome F statistic	Detect swaps or metadat...
Sample	Heterozygosity	F +/− 3 SD	Detect contamination/ou...
Sample	Relatedness	KING kinship > 0.1875	Preserve independence; ...
Variant	Call rate	−−geno 0.05	Remove poorly genotyped...
Variant	Hardy−Weinberg	−−hwe 1e−6 in controls	Remove likely genotypin...
Variant	Minor allele frequency	−−maf 0.01	Keep low−frequency sign...

Review notes

− Rare cancer setting: cases are precious, so phenotype−aware relatedness pruning is used.

− KING is used for final pruning because it is robust for relationship inference in GWAS QC; PI_HAT is reported for review and teaching.

− A large relatedness loss should trigger a review of recruitment design, family structure, and downstream relatedness−aware association options.

− The final dataset from Step 08 is the starting point for ancestry analysis, PCA, and association testing.

7. Plain−Text Summary Appendix

pdac_demo_09_qc_summary.txt

```
PDAC GWAS Genotyping Quality Control Summary
============================================

Dataset: pdac_demo
Analysis date: Thu Jun 25 17:46:45 WEDT 2026
Genome build: GRCh38

QC Pipeline Summary
−−−−−−−−−−−−−−−−−−−

Step                          Samples    Variants   Samples lost   Variants lost
00. Raw data                  1461      430000     0              0
01. Initial stats             1461      430000     0              0
02. Sample call rate          1436      430000     25             0
03. Sex check                 1432      430000     4              0
04. Heterozygosity            1430      430000     2              0
05. Variant call rate         1430      421520     0              8480
06. Hardy−Weinberg            1430      406947     0              14573
07. Relatedness (KING)        1239      406947     191             0
08. MAF filter (>=1%)         1239      401909     0              5038

Final dataset: 1239 samples and 401909 variants
Total sample loss: 222 (15.2%)
Total variant loss: 28091 (6.5%)
Sample retention: 84.8%
Variant retention: 93.5%

QC Thresholds Applied
−−−−−−−−−−−−−−−−−−−−

Sample−level:
− Sample call rate: −−mind 0.02
− Sex check: X chromosome F−statistic discordance
− Heterozygosity: autosomal F +/− 3 SD
```

Appendix truncated in PDF; see the full text summary file for complete notes.