CCDG QC

Wenhan Lu

Data description

The Centers for Common Disease Genomics (CCDG) are a collaborative large-scale genome sequencing effort to comprehensively identify rare risk and protective variants contributing to multiple common disease phenotypes.

WGS data

136,959 samples from

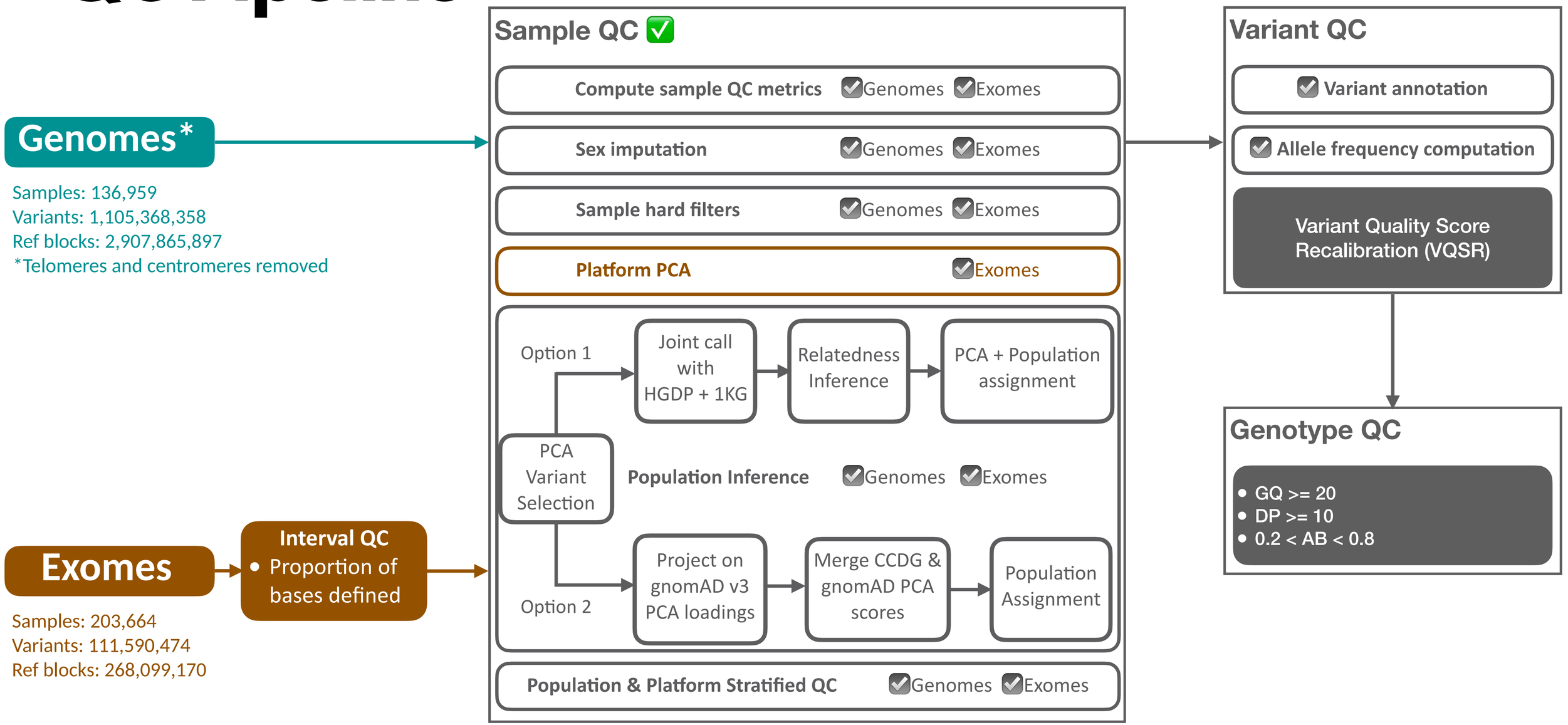
- Broad Institute: 16,862
- Baylor college: 35,493
- New York Genome Center (NYGC): 40,975
- Washington Univ. of St. Louis: 43,620

WES data

203,664 samples from

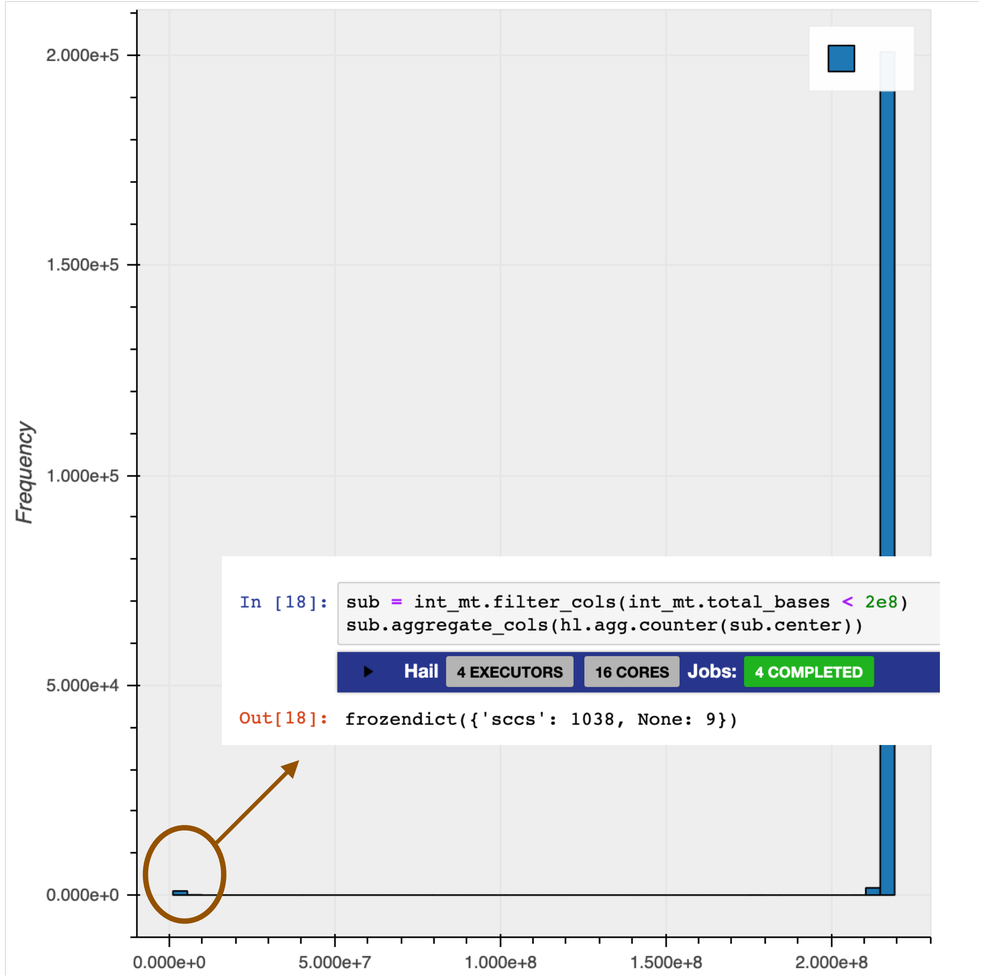
- Broad Institute
- Washington Univ. of St. Louis: 1,038

QC Pipeline

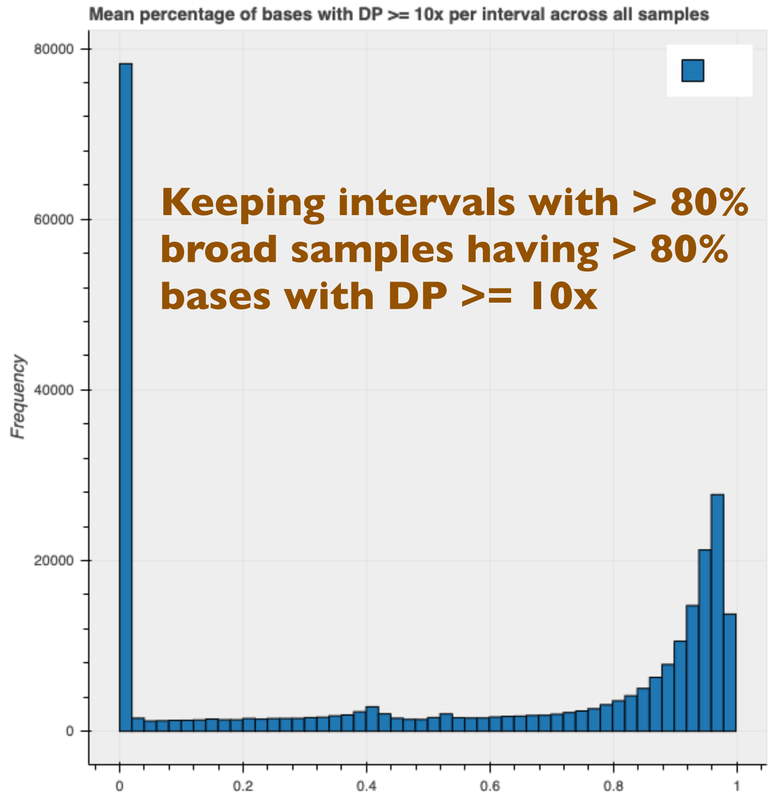

Genomes*
Samples: 136,959
Variants: 1,105,368,358
Ref blocks: 2,907,865,897
*Telomeres and centromeres removed
Exomes
Samples: 203,664
Variants: 111,590,474
Ref blocks: 268,099,170
Interval QC
• Proportion of bases defined
Sample QC ✅
Compute sample QC metrics Genomes Exomes
Sex imputation Genomes Exomes
Sample hard filters Genomes Exomes
Platform PCA Exomes
Option 1
Joint call with HGDP + 1KG
Relatedness Inference
PCA + Population assignment
PCA Variant Selection
Population Inference Genomes Exomes
Option 2
Project on gnomAD v3 PCA loadings
Merge CCDG & gnomAD PCA scores
Population Assignment
Population & Platform Stratified QC Genomes Exomes
Variant QC
Variant annotation
Allele frequency computation
Variant Quality Score Recalibration (VQSR)
Genotype QC
• GQ >= 20
• DP >= 10
• 0.2 < AB < 0.8

Step 0: Interval QC

Exomes only



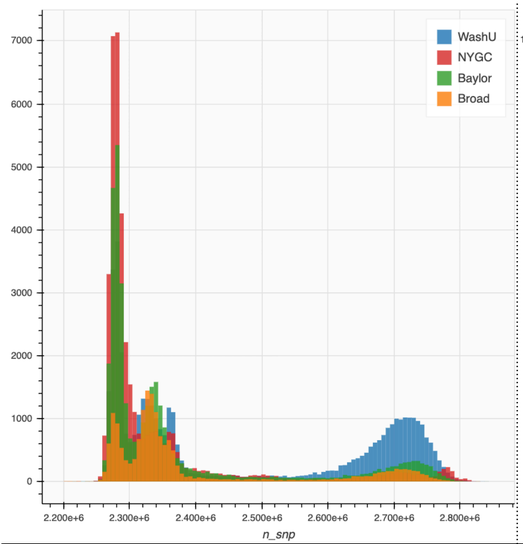
% bases > 10x per interval

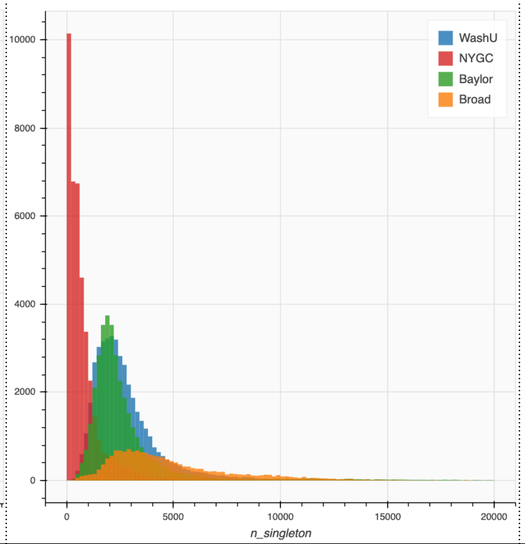


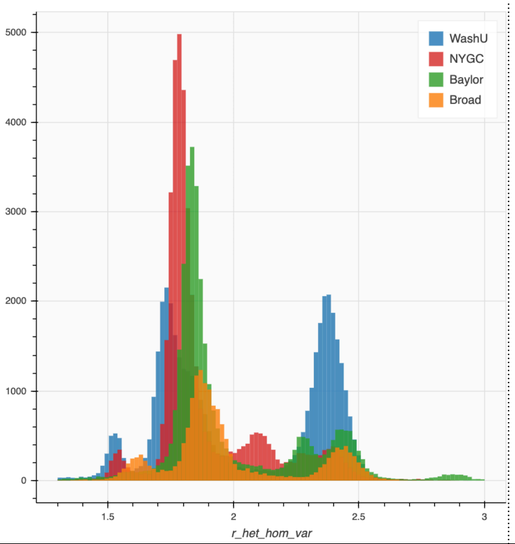
Sample QC metrics

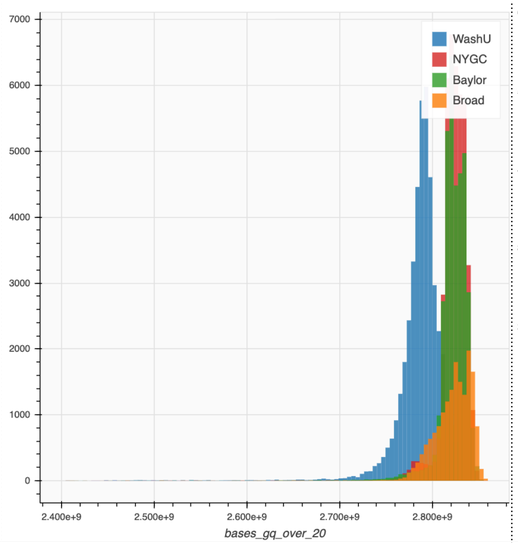
```
hl.vds.sample_qc(vds, …)
```

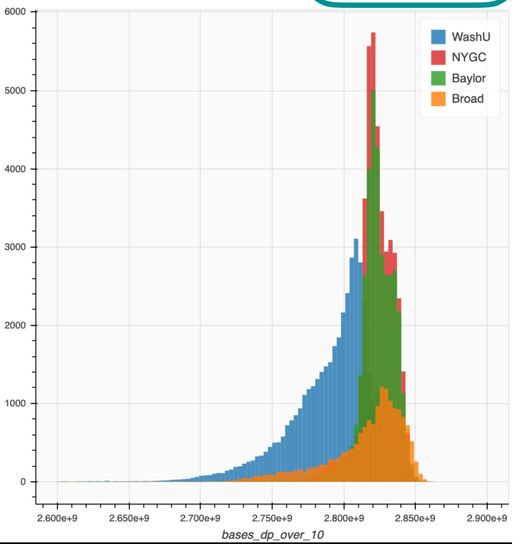
Genomes



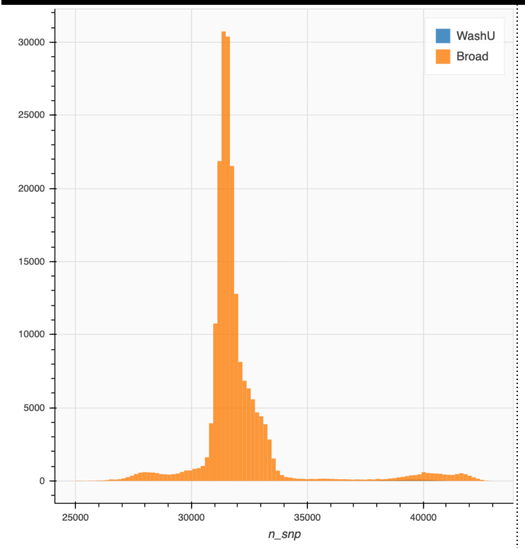


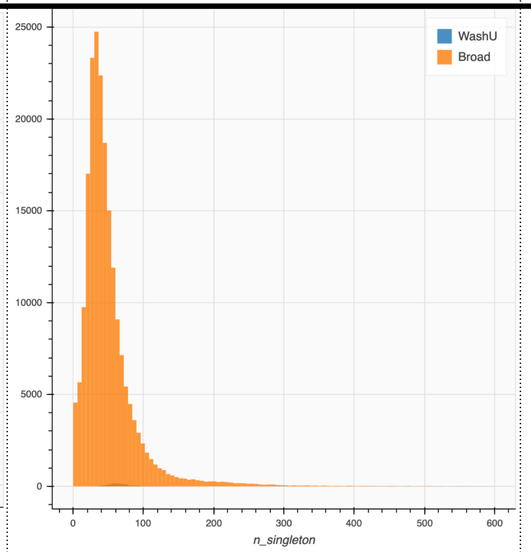


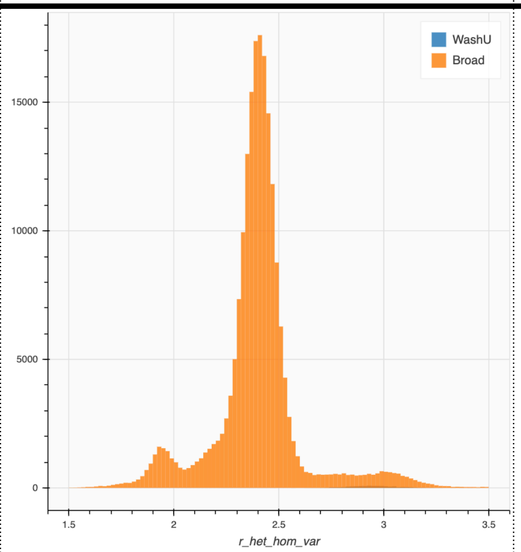


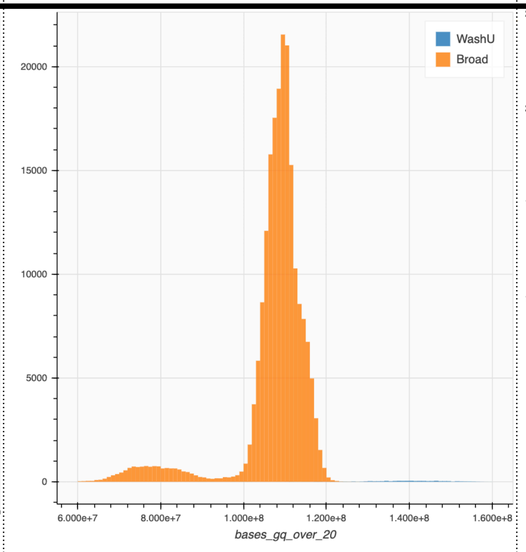


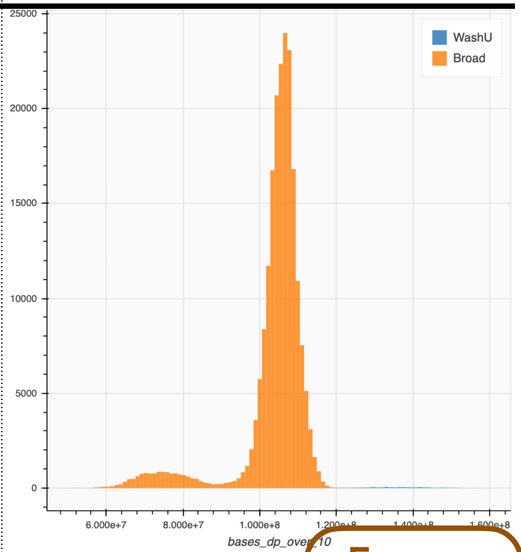
n_snp	n_singleton	r_het_hom_var	n_bases GQ >=20	n_bases DP >=10









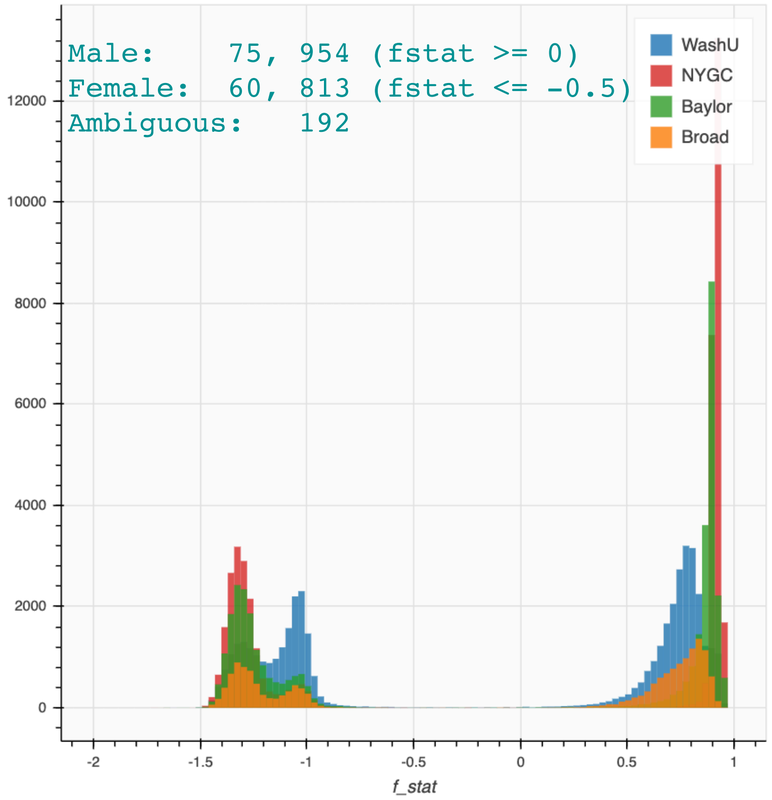


Exomes

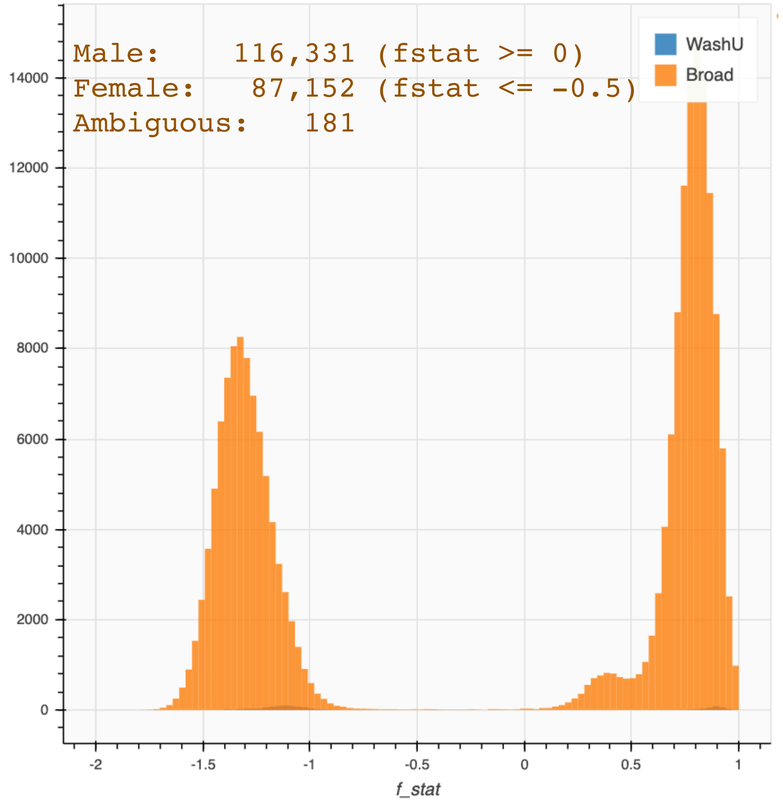
Sex imputation

`hl.vds.impute_sex_chromosome_ploidy(vds, …)`

Genomes



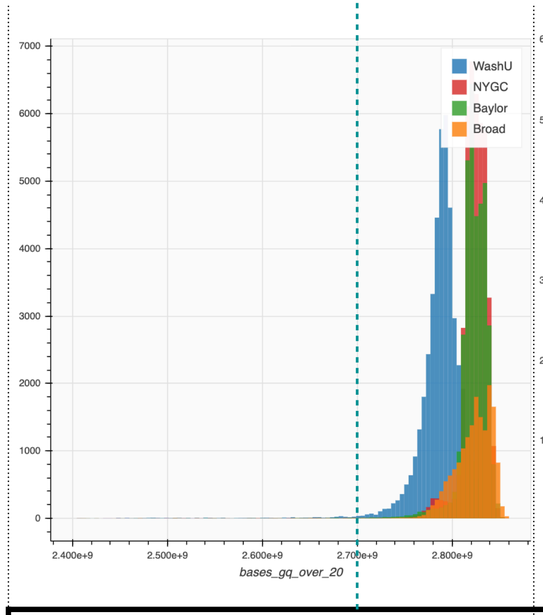
Exomes



Genomes

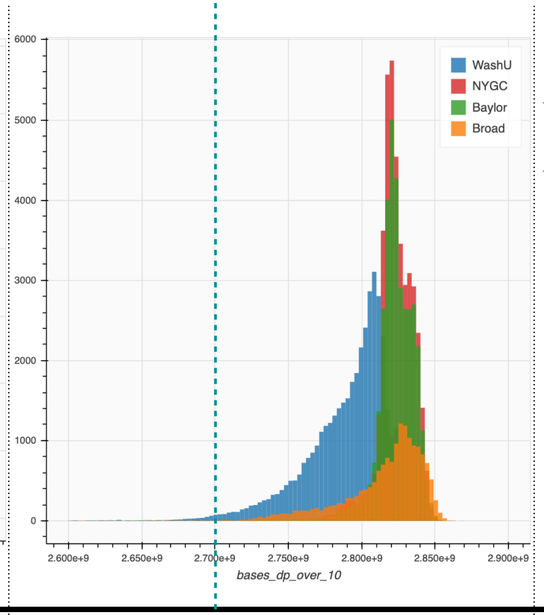
2,106 samples filtered

1,217 samples



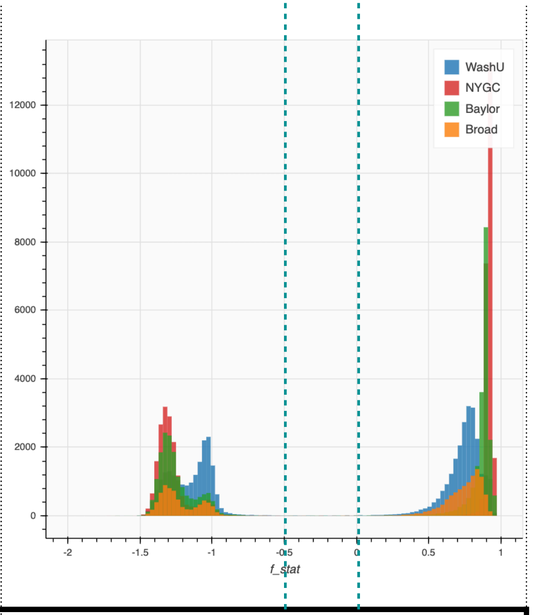
n_bases GQ >=20

1,605 samples



n_bases DP >=10

192 samples

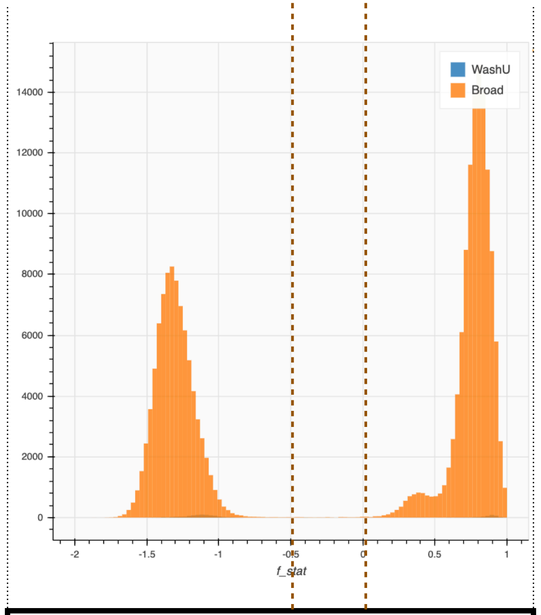


f_stat

Exomes

181 samples filtered

181 samples

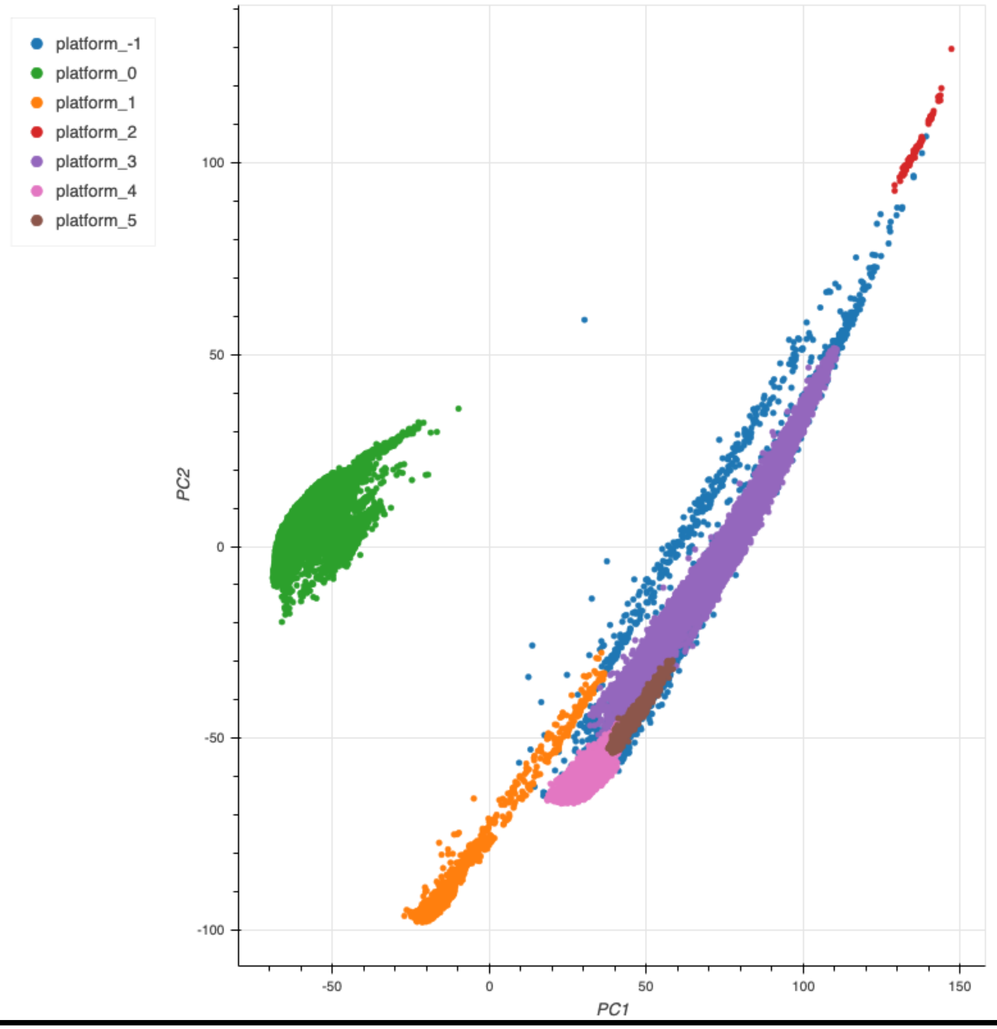


f_stat

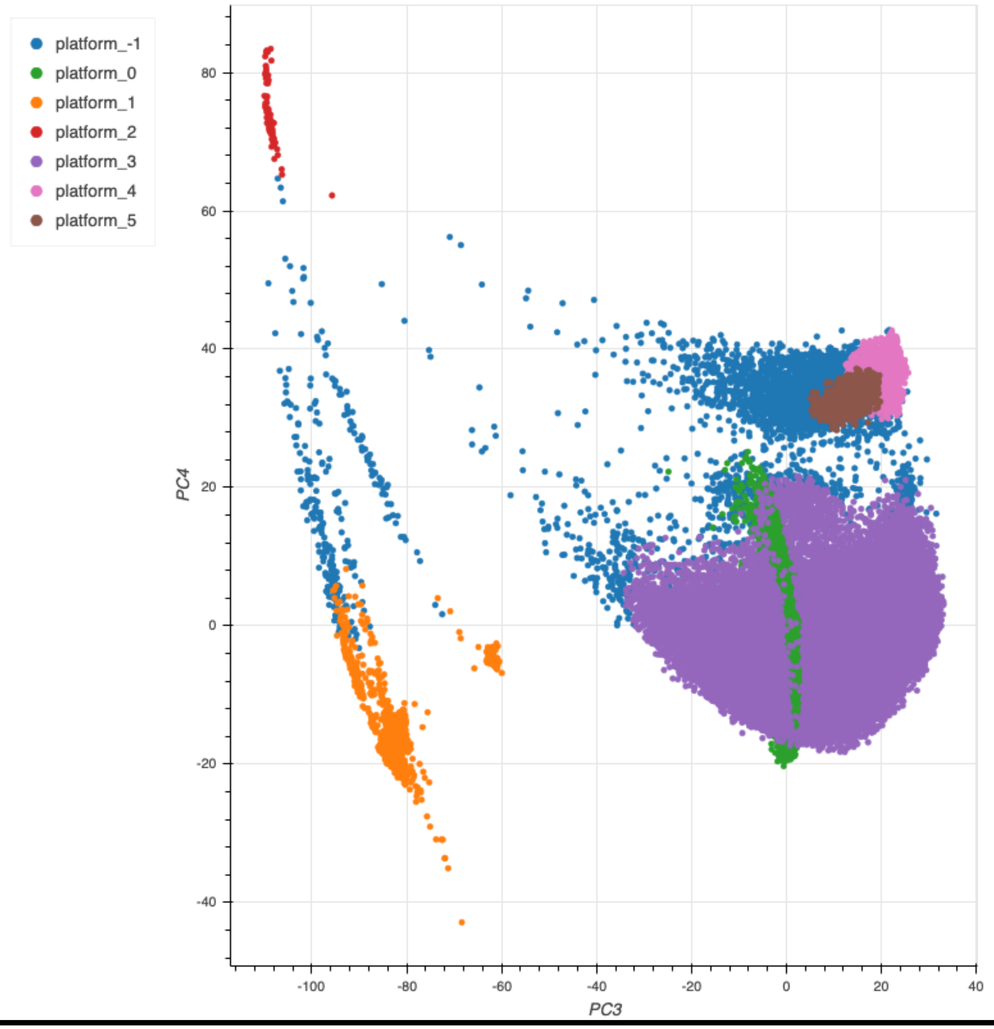
Compare to ‘true’ platforms

```
hl.vds.interval_coverage(vds, intervals=ht)
```

	True platforms															
	Custom Hybrid Selection _MTC	Exome Express	Exome Express v2	Express Somatic Human WES (Deep Coverage) v1	Express Somatic Human WES (Standard Coverage) v1	G4L WES + Array v1	G4L WES + Array v2	Nextera Exome	None	Standard Exome Sequencing	Standard Exome Sequencing v2	Standard Exome Sequencing v3	Standard Exome Sequencing v4	Standard Germline Exome v5	Standard Germline Exome v6	WashU
platform_-1	0	0	1	10	12	0	3	59	316	7	7	2257	93	885	1	0
platform_0	2	0	0	0	0	0	14	1	4470	1	7	5	2	136	105008	0
platform_1	0	96	100	0	0	0	0	0	1359	984	2825	0	0	0	0	0
platform_2	0	0	0	0	0	0	0	0	121	0	0	0	0	1	0	1035
platform_3	0	0	0	0	0	49	1385	16	803	4	23	0	0	76432	3	0
platform_4	0	0	0	0	0	62	0	1321	131	0	0	1098	1000	0	0	0
platform_5	0	0	0	0	10	0	0	0	0	0	0	1327	1	0	0	0

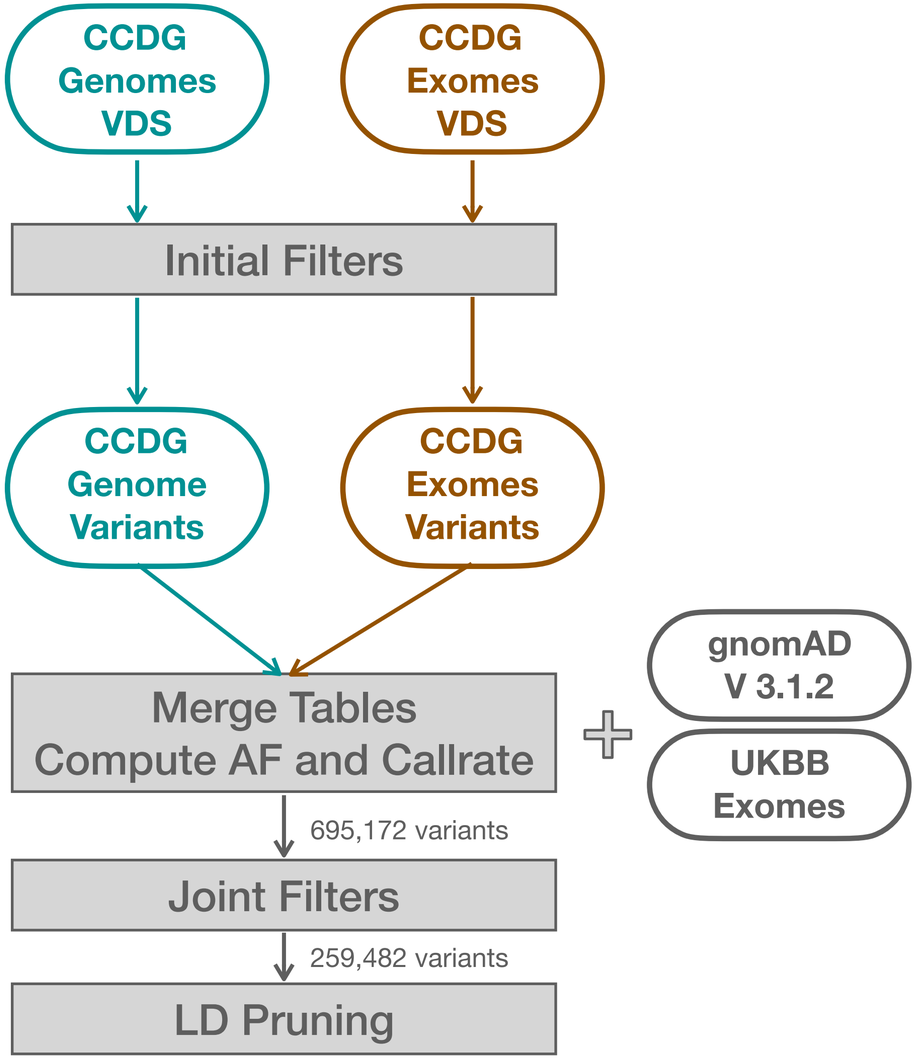


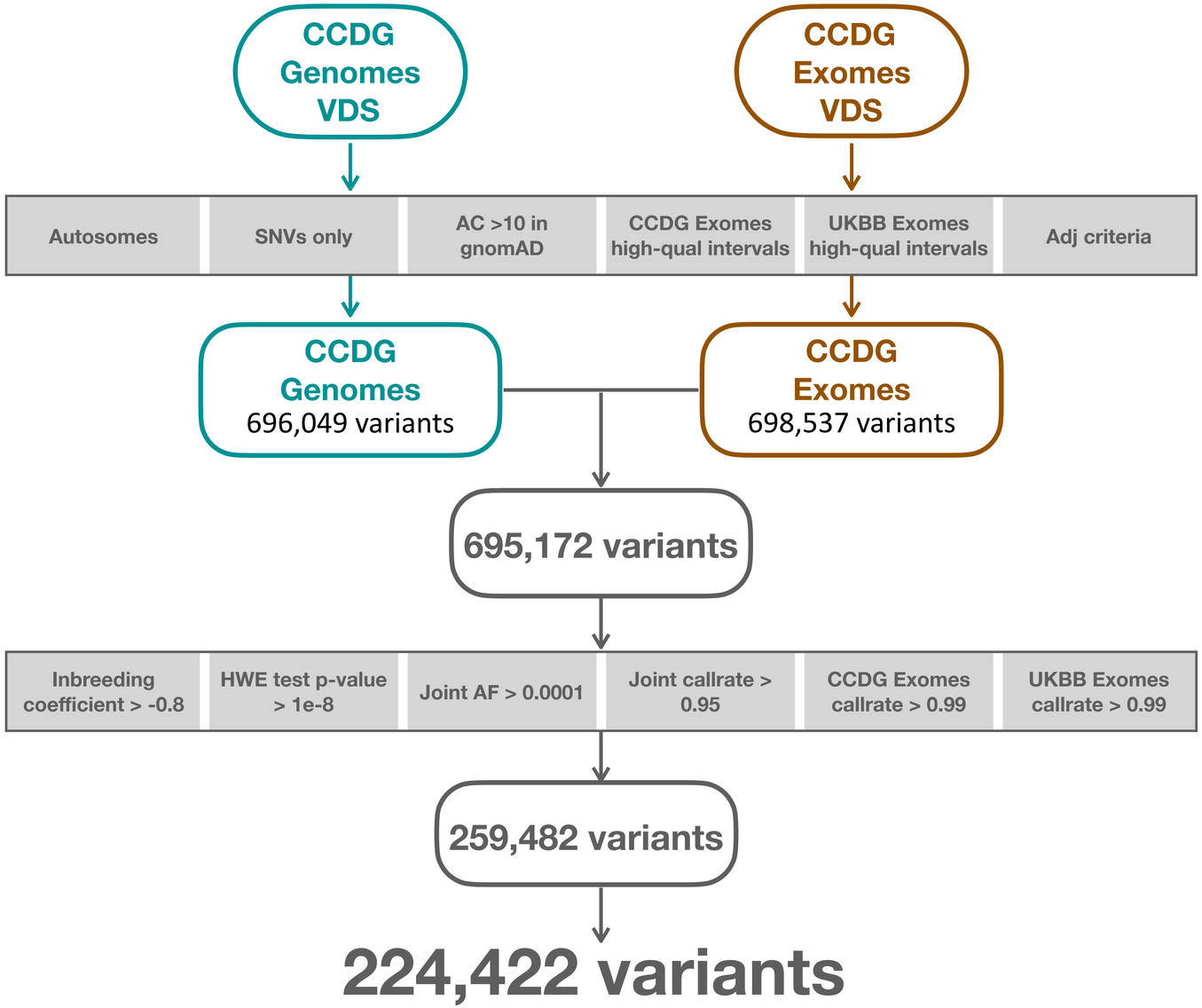
PC1 vs. PC2



PC3 vs. PC4

PCA variant selection ✅

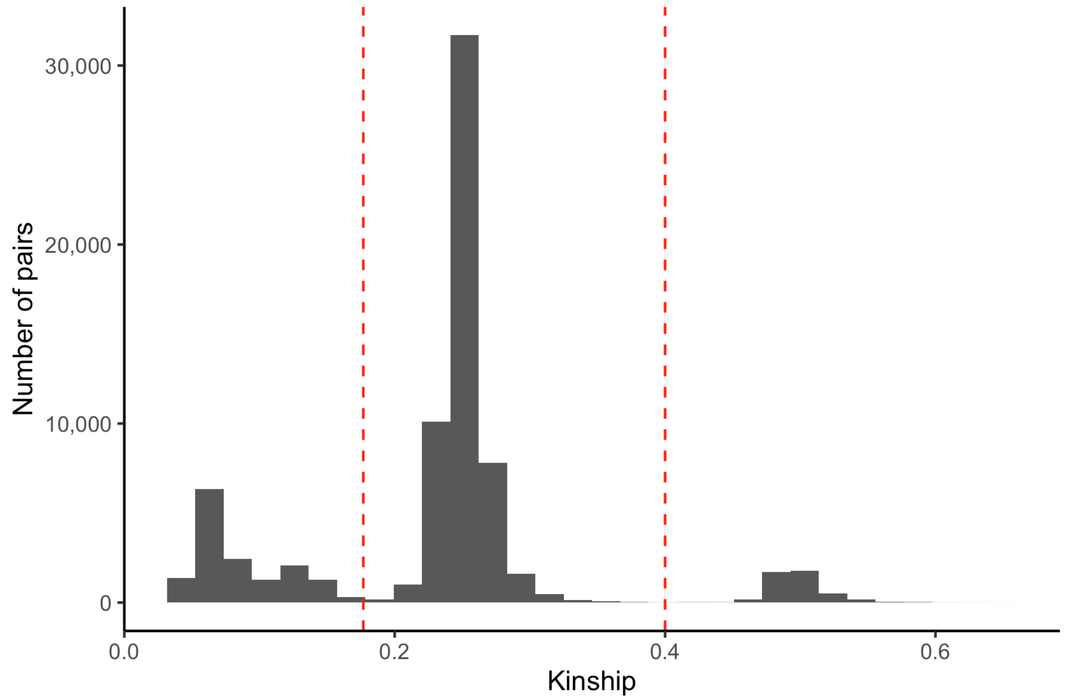




Relatedness Inference

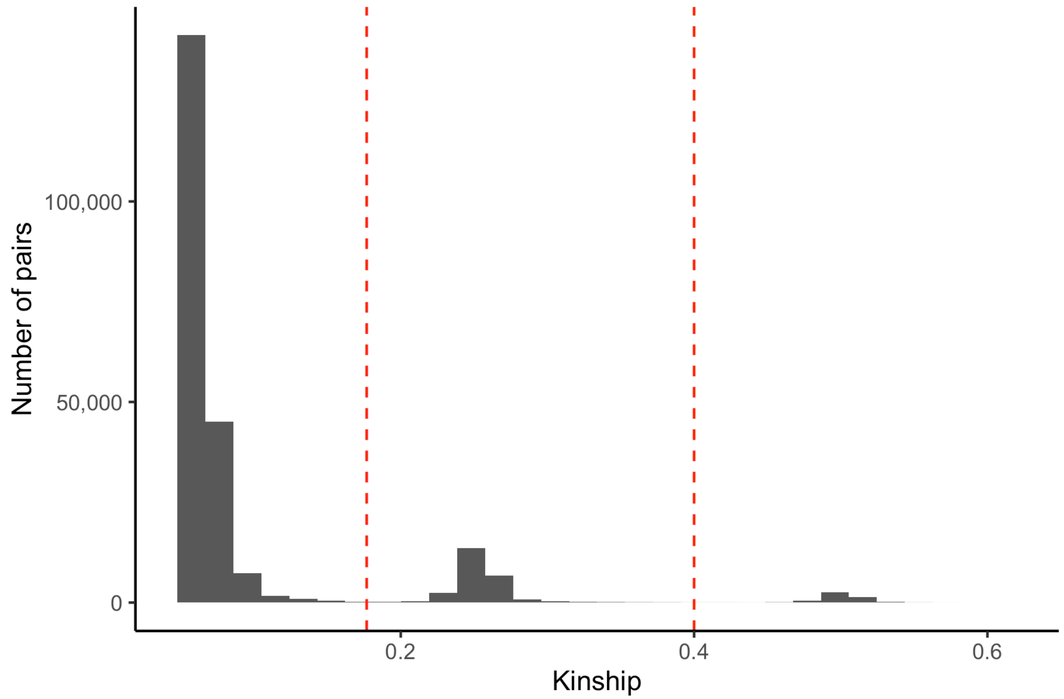
min_kinship = 0.05

Genomes



28,072 related samples to remove

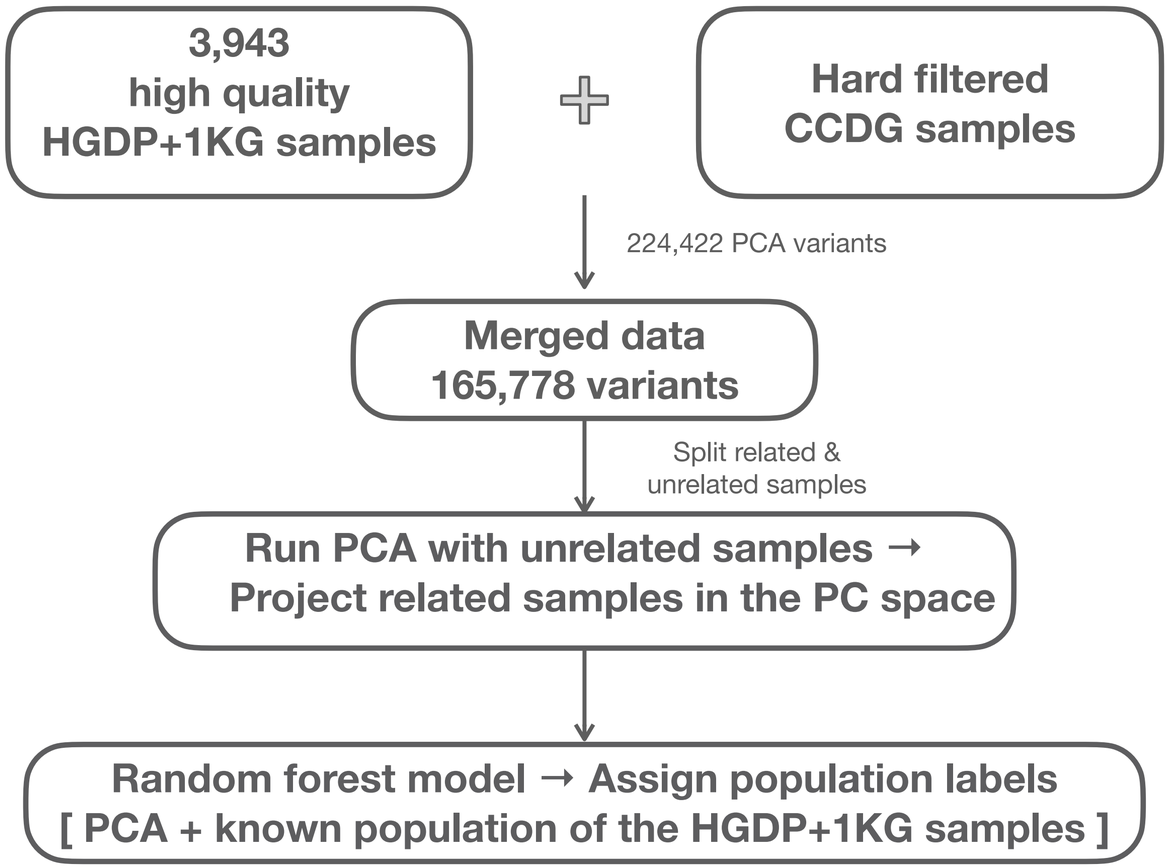
Exomes



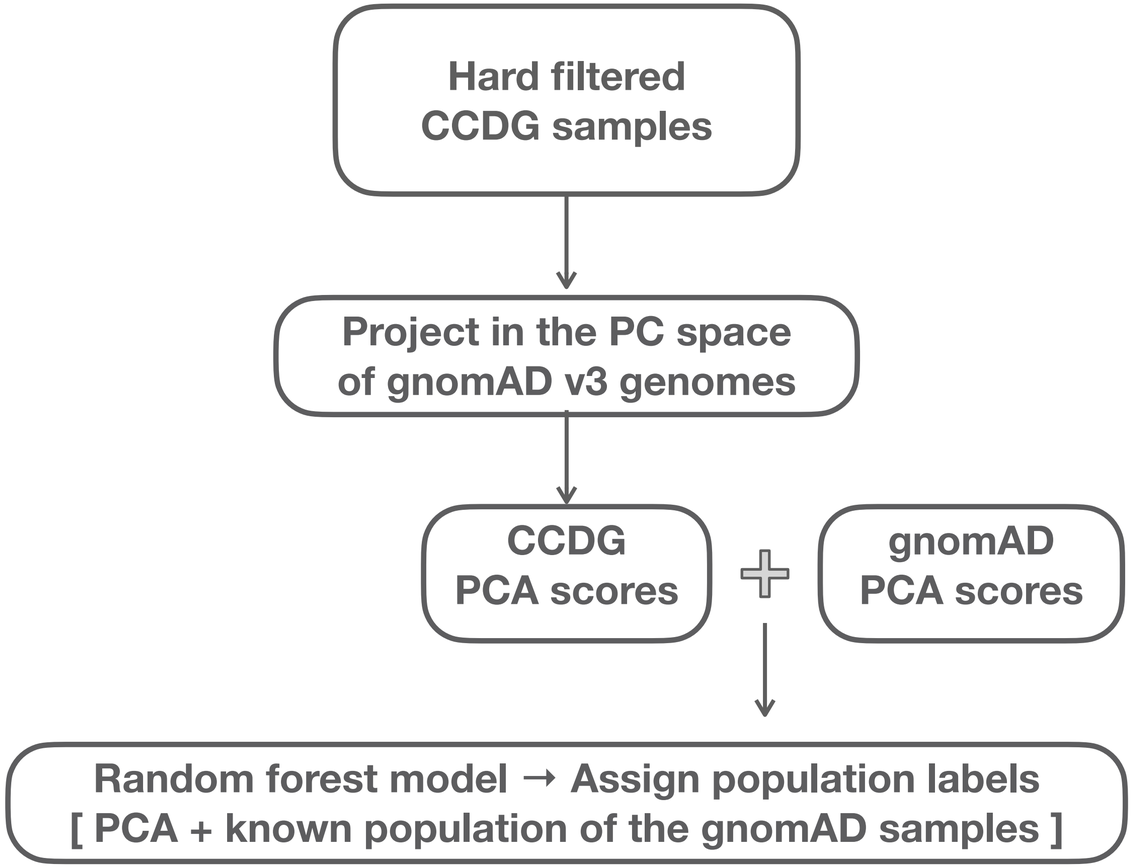
19,805 related samples to remove

Two PCA options

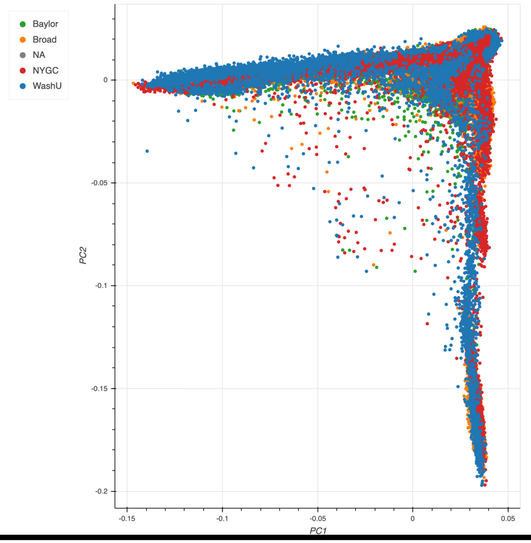
HGDP +1KG

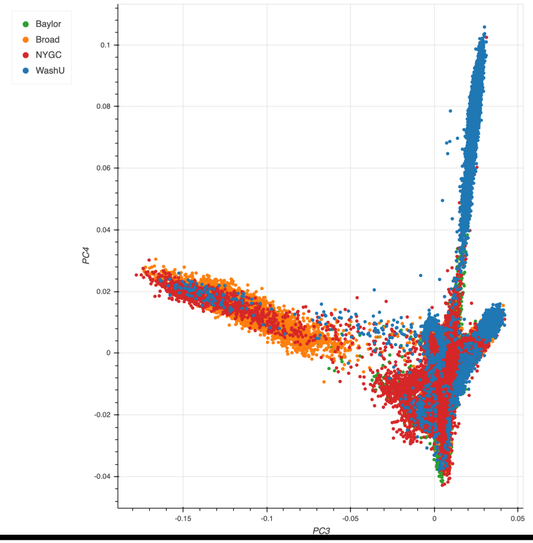


gnomAD



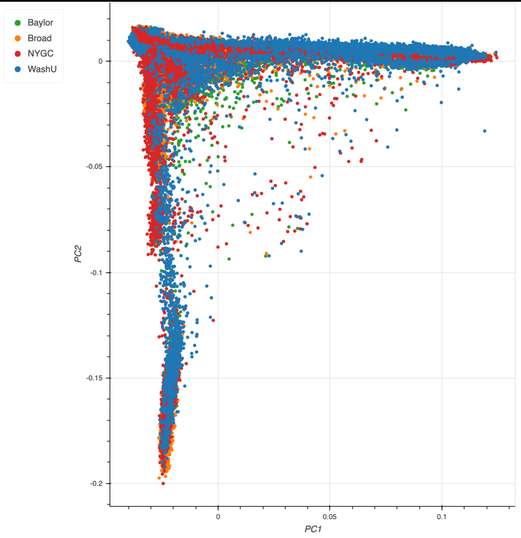
HGDP + 1KG

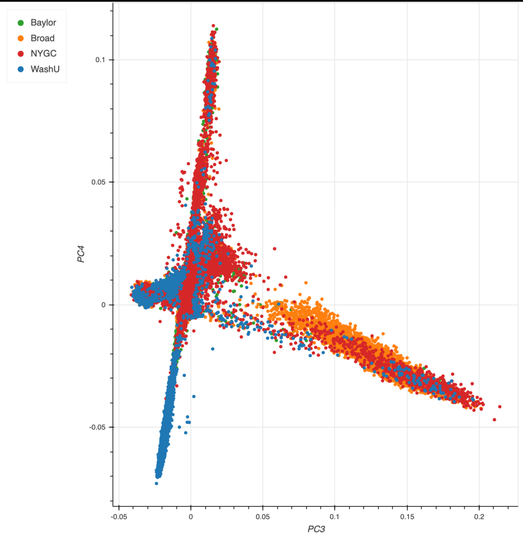




PC1 vs. PC2

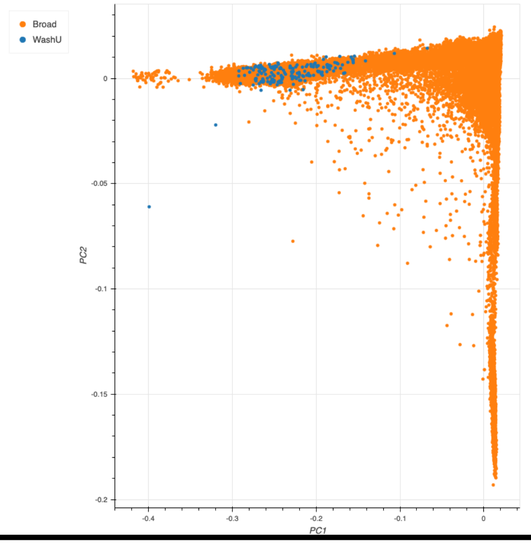
PC3 vs. PC4

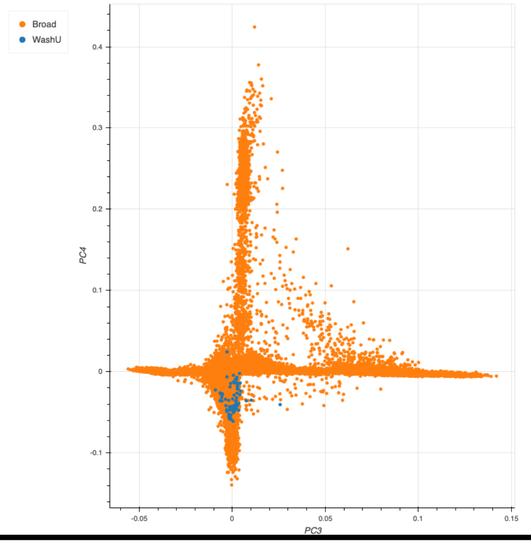




gnomAD

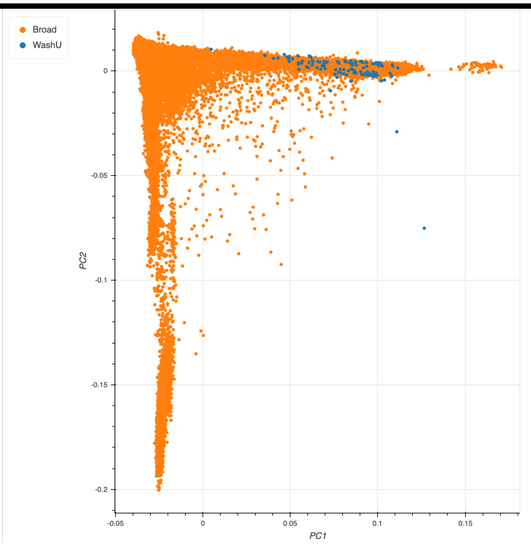
HGDP + 1KG

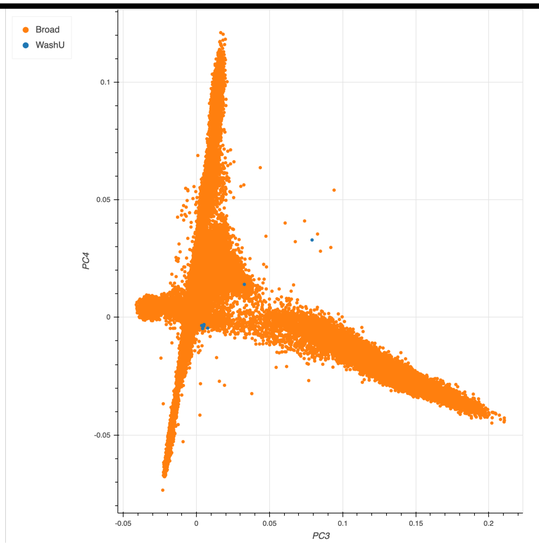




PC1 vs. PC2

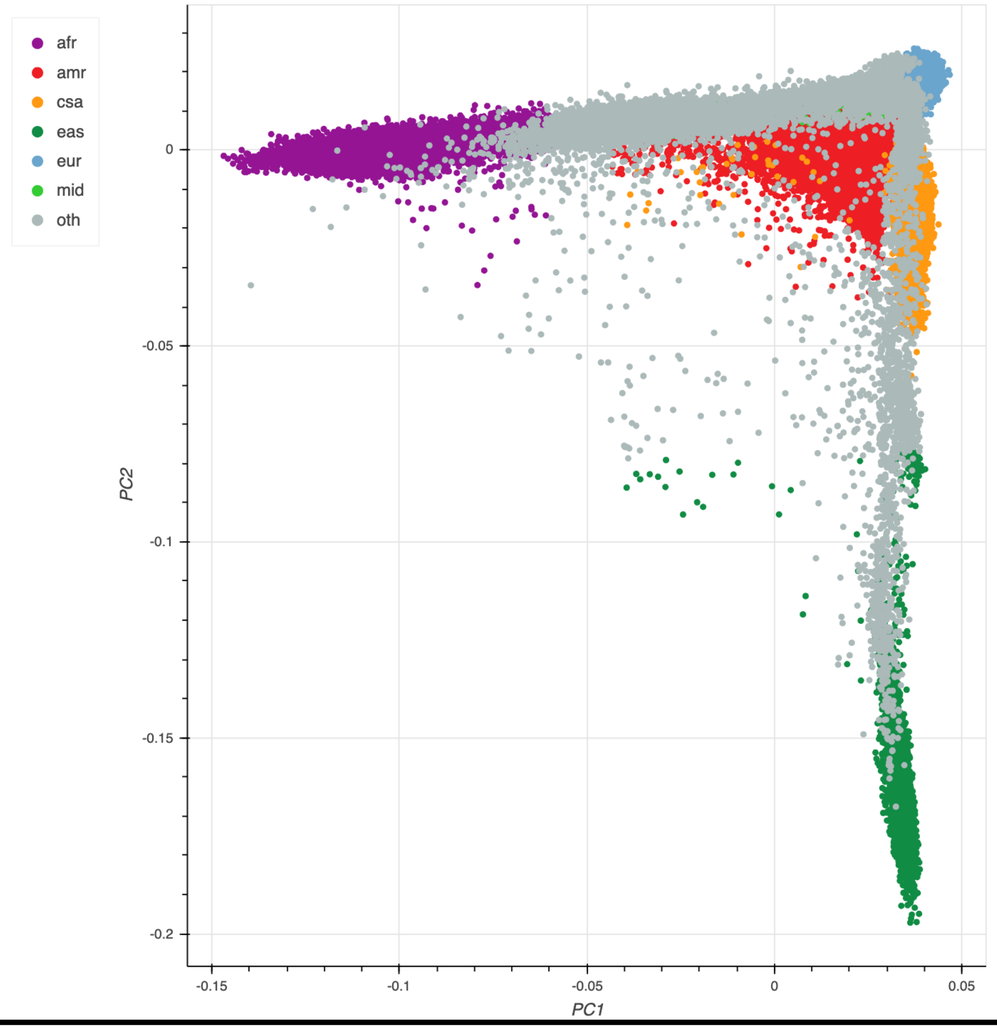
PC3 vs. PC4



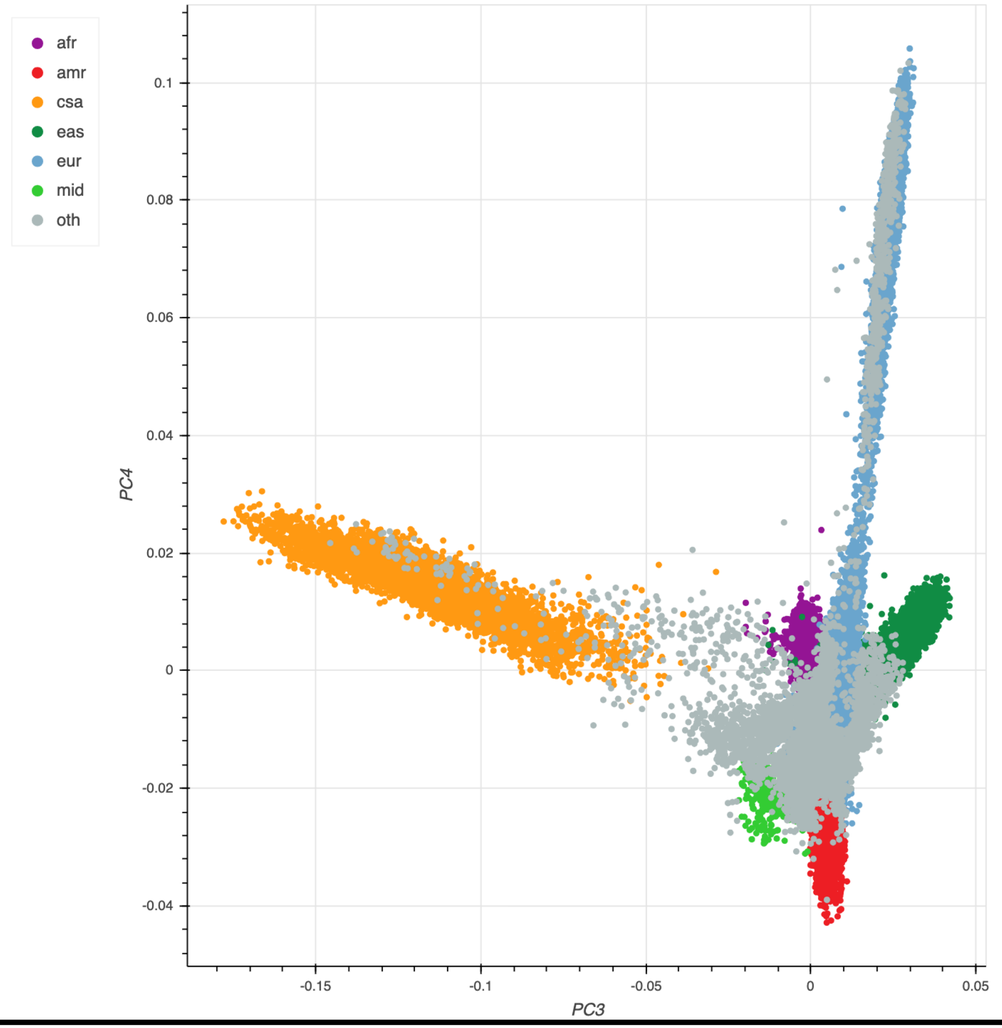


gnomAD

min_prob=0.9



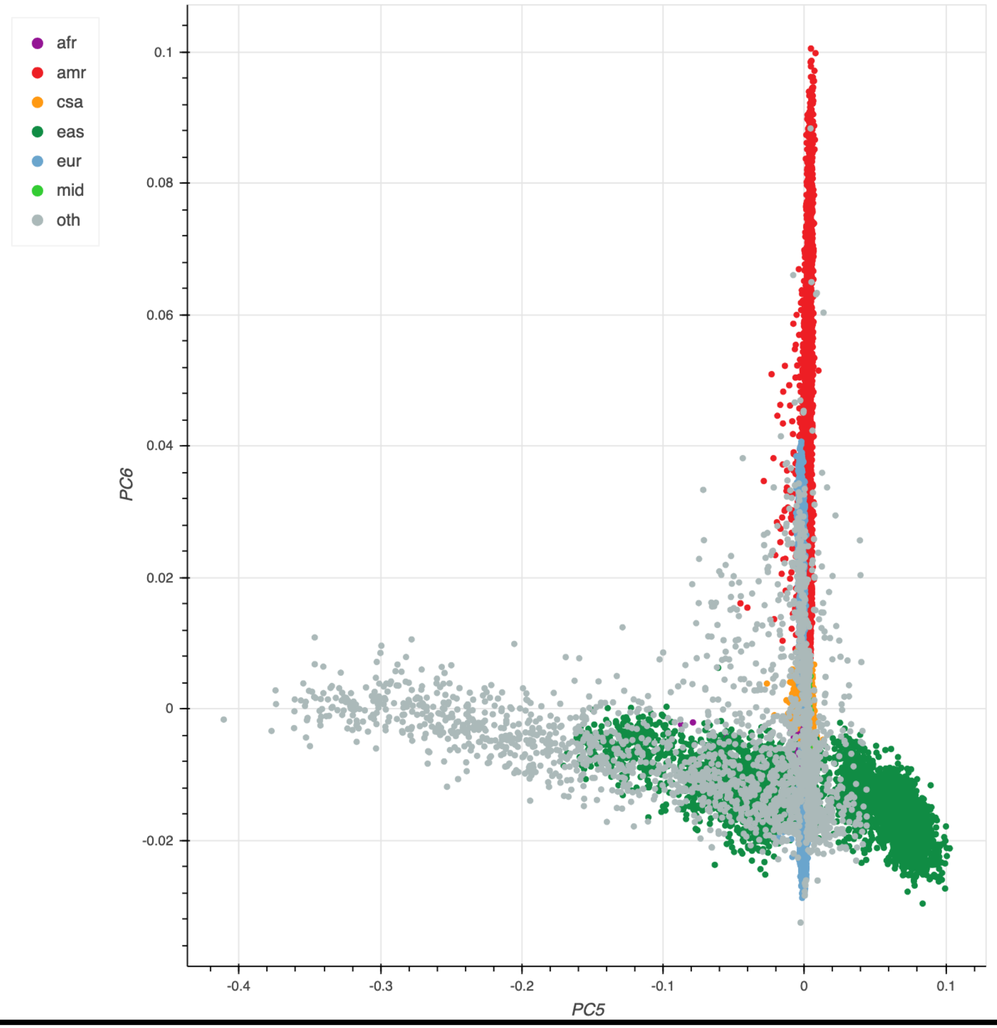
PC1 vs. PC2



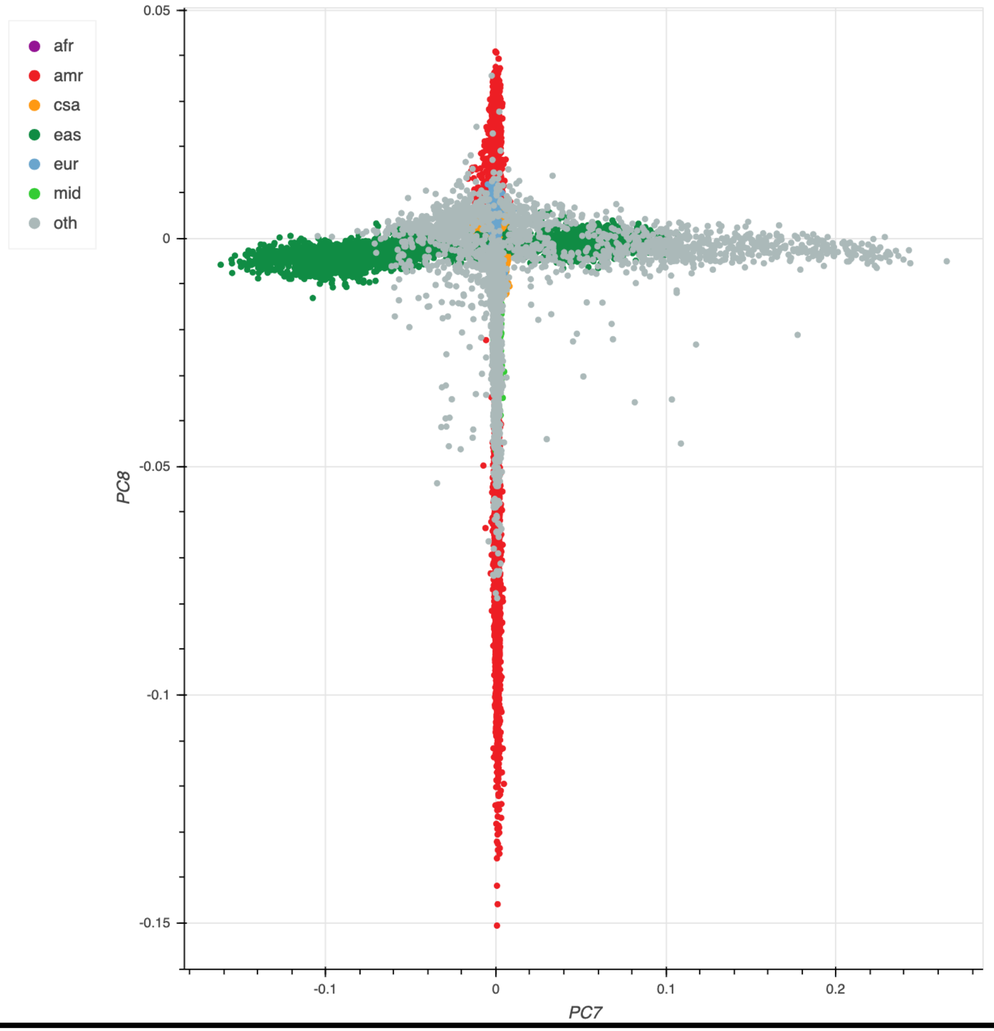
PC3 vs. PC4

HGDP + 1KG

min_prob=0.9



PC5 vs. PC6



PC7 vs. PC8

gnomAD population + sequencing center	Sample QC metrics	HGDP + 1KG population + sequencing center
4,258	n_het	2,305
812	n_hom_var	773
1,116	n_singleton	1,134
518	n_snp	474
544	n_insertion	509
340	n_deletion	173
526	n_transition	461
503	n_transversion	453
660	r_ti_tv	636
2,071	r_het_hom_var	1,796
623	r_insertion_deletion	232
972	bases_gq_over_20	919
2,294	bases_dp_over_10	2,216
1,174	inbreeding	808
9,102	**Total failed**	**7,040**

Sequencing center summary

Population	Total	Male	Female	Related samples	Failed HGDP + 1KG stratified QC
Baylor	35,493	19,059	16,395	2,723	1,251
Broad	16,862	10,127	6,703	1,052	1,713
NYGC	40,975	23,125	17,829	18,813	1,304
WashU	43,620	23,639	19,881	5,480	2,771
None	9	4	5	4	1
Total	**136,959**	**75,954**	**60,813**	**28,072**	**7,040**

HGDP +1KG population summary

Population	Total	Male	Female	Related samples	Failed HGDP + 1KG stratified QC
AFR	27,003	11,098	15,905	3,177	1,957
AMR	23,035	11,082	11,953	4,007	1,239
CSA	6,771	4,868	1,903	596	1,283
EAS	5,149	2,636	2,513	902	365
EUR	53,171	34,279	18,892	13,795	1,148
MID	509	301	208	175	58
Oth	19,215	10,408	8,807	5,420	990
None (Hard filtered)	2,106	1,282	632		
Total	**136,959**	**75,954**	**60,813**	**28,072**	**7,040**

gnomAD population + Imputed platform	Sample QC metrics	HGDP + 1KG population + Imputed platform
1,983	n_het	3,886
279	n_hom_var	896
1,333	n_singleton	1,246
369	n_snp	2,386
164	n_insertion	533
72	n_deletion	238
322	n_transition	2,299
319	n_transversion	1,819
716	r_ti_tv	717
1,570	r_het_hom_var	2,192
85	r_insertion_deletion	80
1,025	bases_gq_over_20	1,006
1,334	bases_dp_over_10	1,329
216	inbreeding	836
5,115	**Total failed**	**7,090**

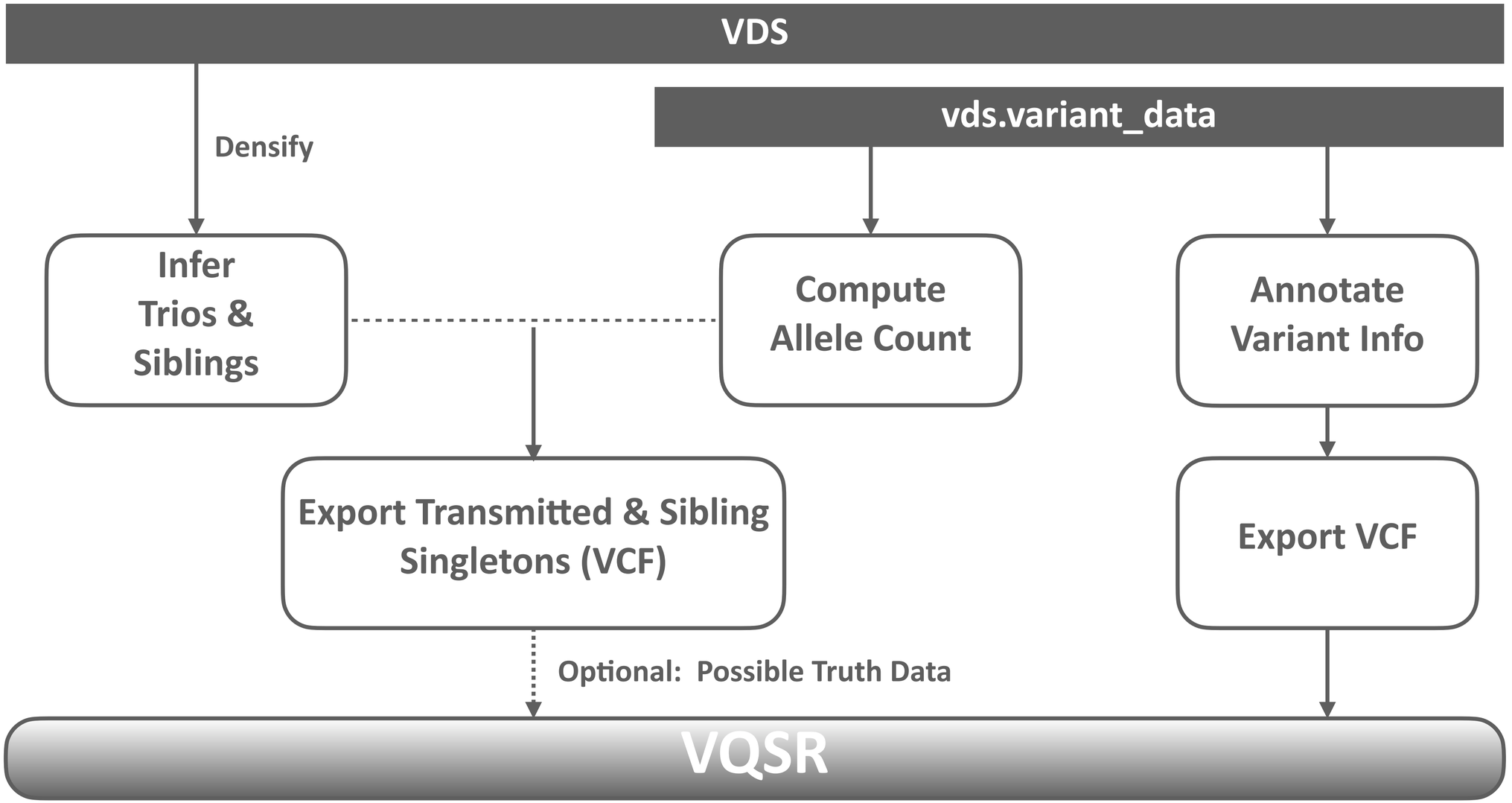
Platform (min_cluster_size = 500) summary

Platform	Total	Failed HGDP + 1KG stratified QC
-1	3,651	165
0	109,646	3,845
1	5,364	410
2	1,157	67
3	78,715	2,515
4	3,612	69
5	1,338	19
Total	**203,664**	**7,090**

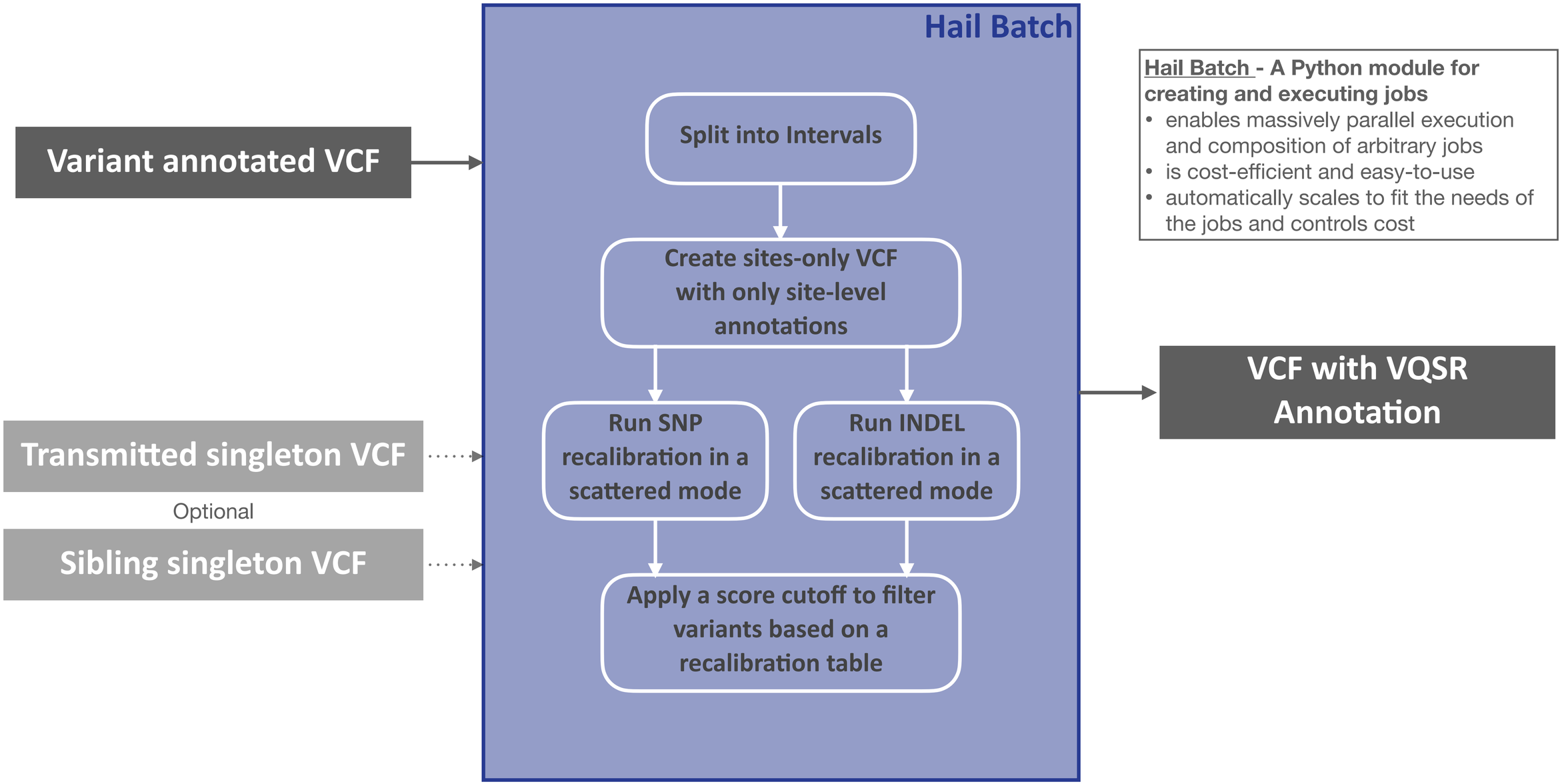
Sequencing center summary

Population	Total	Male	Female	Related samples	Failed HGDP + 1KG stratified QC
Broad	202,626	116,029	86,419	19,785	7,032
WashU	1,038	302	733	20	58
Total	**203,664**	**116,331**	**87,152**	**19,805**	**7,090**

Variant QC annotations



VQSR 👉 on Hail Batch !



Hail Batch - A Python module for creating and executing jobs

- enables massively parallel execution and composition of arbitrary jobs
- is cost-efficient and easy-to-use
- automatically scales to fit the needs of the jobs and controls cost