Supplementary Materials

Jeremiah H. Li, Chase. A Mazur, Tomaz Berisa, Joseph K. Pickrell

Gencove Inc.

December 18, 2020

1 Supplementary Tables

Population	Super Population	n
ASW	AFR	12
ACB	AFR	8
ESN	AFR	8
GWD	AFR	8
LWK	AFR	8
MSL	AFR	8
YRI	AFR	8
GBR	EUR	24
IBS	EUR	12
TSI	EUR	12
CEU	EUR	6
FIN	EUR	6

Supplementary Table 1: Number of unique cell lines assayed from each population for the EUR and AFR super populations.

Experiment	Samples	Passing QC	Unique EUR	Unique AFR	Mapped coverage	Eff. Coverage
A	360	351	60	60	0.668±0.318	0.412 ± 0.188
B	360	350	60	60	1.248±0.263	0.717 ± 0.180
C	30	30	1	0	1.201±0.289	0.526 ± 0.112
D	60	58	30	30	1.260±0.0892	1.240 ± 0.106
E	360	358	60	60	NA	NA

Supplementary Table 2: Breakdown of experiments conducted as described in main text. For experiments A, B, and E, 120 unique cell lines were run in triplicate. For experiment C, NA12878 was run 30x in replicate. For experiment C, a subset of 30 for each superpopulation was run once. Mapped coverage was computed as the number of mapped bases divided by 3.3e9. Coverage is reported as the mean ± standard deviation.

Supplementary Table 3: This table shows the number of samples for each cell line in each experiment which passed QC. A dot indicates that the particular cell line was not part of the experiment.

Cell line	Population	Super population	Exp. A	Exp. B	Exp. C	Exp. D	Exp. E
HG01879	ACB	AFR	3	3	.	1	3
HG01886	ACB	AFR	3	3	.	.	3
HG01889	ACB	AFR	3	3	.	1	3
HG01890	ACB	AFR	3	3	.	.	3
HG02009	ACB	AFR	3	3	.	1	3
HG02010	ACB	AFR	3	3	.	.	3
HG02107	ACB	AFR	3	3	.	1	3
HG02282	ACB	AFR	3	3	.	1	3

Supplementary Table 3: This table shows the number of samples for each cell line in each experiment which passed QC. A dot indicates that the particular cell line was not part of the experiment. *(continued)*

Cell line	Population	Super population	Exp. A	Exp. B	Exp. C	Exp. D	Exp. E
HG02568	GWD	AFR	3	3	.	1	3
HG02759	GWD	AFR	3	3	.	1	3
HG02760	GWD	AFR	3	3	.	.	3
HG02878	GWD	AFR	3	3	.	1	3
HG02879	GWD	AFR	3	3	.	1	3
HG02881	GWD	AFR	3	3	.	.	3
HG02882	GWD	AFR	3	3	.	.	3
HG02922	ESN	AFR	3	3	.	1	3
HG02938	ESN	AFR	3	3	.	.	3
HG02941	ESN	AFR	3	3	.	1	3
HG02982	GWD	AFR	3	3	.	1	3
HG03052	MSL	AFR	3	3	.	.	3
HG03054	MSL	AFR	3	3	.	1	3
HG03060	MSL	AFR	3	3	.	1	3
HG03063	MSL	AFR	3	3	.	1	3
HG03077	MSL	AFR	3	3	.	1	3
HG03082	MSL	AFR	3	3	.	.	3
HG03095	MSL	AFR	3	3	.	.	3
HG03127	ESN	AFR	3	3	.	1	3
HG03202	ESN	AFR	3	3	.	.	3
HG03298	ESN	AFR	3	3	.	1	3
HG03382	MSL	AFR	3	3	.	.	3
HG03520	ESN	AFR	3	3	.	1	3
HG03521	ESN	AFR	3	3	.	.	3
NA18504	YRI	AFR	3	3	.	.	3
NA18912	YRI	AFR	3	3	.	1	3
NA19017	LWK	AFR	3	3	.	1	3
NA19041	LWK	AFR	3	3	.	.	3
NA19119	YRI	AFR	3	3	.	.	3
NA19131	YRI	AFR	3	3	.	.	3
NA19152	YRI	AFR	3	3	.	.	3
NA19204	YRI	AFR	3	3	.	.	3
NA19238	YRI	AFR	3	3	.	.	3
NA19239	YRI	AFR	3	3	.	.	3
NA19307	LWK	AFR	3	3	.	.	3
NA19308	LWK	AFR	3	3	.	.	3
NA19310	LWK	AFR	3	3	.	1	3

Supplementary Table 3: This table shows the number of samples for each cell line in each experiment which passed QC. A dot indicates that the particular cell line was not part of the experiment. *(continued)*

Cell line	Population	Super population	Exp. A	Exp. B	Exp. C	Exp. D	Exp. E
NA19350	LWK	AFR	3	3	.	1	3
NA19351	LWK	AFR	3	3	.	1	3
NA19471	LWK	AFR	3	3	.	1	3
NA19625	ASW	AFR	3	3	.	1	3
NA19914	ASW	AFR	3	3	.	.	3
NA19922	ASW	AFR	3	3	.	.	3
NA19923	ASW	AFR	3	3	.	1	3
NA19984	ASW	AFR	3	3	.	.	3
NA20298	ASW	AFR	3	3	.	1	3
NA20320	ASW	AFR	3	3	.	.	3
NA20351	ASW	AFR	3	3	.	.	3
NA20357	ASW	AFR	3	3	.	1	3
NA20359	ASW	AFR	3	3	.	.	3
NA20362	ASW	AFR	3	3	.	1	2
NA20412	ASW	AFR	3	3	.	.	3
HG00096	GBR	EUR	2	0	.	.	3
HG00101	GBR	EUR	3	3	.	.	3
HG00102	GBR	EUR	3	3	.	1	3
HG00105	GBR	EUR	3	3	.	1	3
HG00107	GBR	EUR	3	3	.	1	3
HG00108	GBR	EUR	3	3	.	.	3
HG00110	GBR	EUR	3	3	.	.	3
HG00111	GBR	EUR	3	3	.	1	3
HG00116	GBR	EUR	3	3	.	.	3
HG00119	GBR	EUR	3	3	.	.	3
HG00131	GBR	EUR	3	3	.	1	3
HG00132	GBR	EUR	3	3	.	1	3
HG00145	GBR	EUR	3	3	.	.	3
HG00160	GBR	EUR	3	3	.	1	3
HG00234	GBR	EUR	3	3	.	.	3
HG00239	GBR	EUR	3	3	.	1	3
HG00240	GBR	EUR	3	3	.	.	3
HG00242	GBR	EUR	3	3	.	1	3
HG00244	GBR	EUR	3	3	.	.	3
HG00250	GBR	EUR	3	3	.	1	3
HG00251	GBR	EUR	3	3	.	1	3
HG00255	GBR	EUR	3	3	.	.	3

Supplementary Table 3: This table shows the number of samples for each cell line in each experiment which passed QC. A dot indicates that the particular cell line was not part of the experiment. *(continued)*

Cell line	Population	Super population	Exp. A	Exp. B	Exp. C	Exp. D	Exp. E
HG00262	GBR	EUR	3	3	.	.	3
HG00268	FIN	EUR	3	3	.	1	3
HG00330	FIN	EUR	3	3	.	.	3
HG00360	FIN	EUR	3	3	.	.	3
HG00364	FIN	EUR	3	3	.	1	3
HG00365	FIN	EUR	2	3	.	0	3
HG00384	FIN	EUR	3	3	.	.	3
HG01500	IBS	EUR	3	3	.	1	3
HG01697	IBS	EUR	3	3	.	.	3
HG01699	IBS	EUR	3	3	.	1	3
HG01756	IBS	EUR	3	3	.	1	3
HG01757	IBS	EUR	3	3	.	.	3
HG01761	IBS	EUR	3	3	.	.	3
HG01762	IBS	EUR	3	3	.	1	3
HG01770	IBS	EUR	3	3	.	1	3
HG01771	IBS	EUR	3	3	.	.	3
HG01777	IBS	EUR	2	3	.	1	3
HG01781	IBS	EUR	3	3	.	1	3
HG01784	IBS	EUR	3	3	.	.	3
HG02215	GBR	EUR	3	3	.	1	3
NA06985	CEU	EUR	3	3	.	.	3
NA11930	CEU	EUR	3	3	.	.	3
NA12342	CEU	EUR	3	3	.	1	3
NA12842	CEU	EUR	3	3	.	1	3
NA12843	CEU	EUR	3	3	.	.	2
NA12878	CEU	EUR	3	3	30	1	3
NA20502	TSI	EUR	3	3	.	1	3
NA20503	TSI	EUR	3	3	.	.	3
NA20507	TSI	EUR	3	3	.	1	3
NA20513	TSI	EUR	3	2	.	1	3
NA20514	TSI	EUR	3	3	.	.	3
NA20798	TSI	EUR	3	3	.	.	3
NA20802	TSI	EUR	0	0	.	0	3
NA20803	TSI	EUR	3	3	.	.	3
NA20804	TSI	EUR	3	3	.	.	3
NA20805	TSI	EUR	3	3	.	1	3
NA20821	TSI	EUR	3	3	.	.	3
NA20822	TSI	EUR	0	0	.	.	3

		Experiment		
Coverage	Super Pop.	A	B	D
Nominal	AFR	0.6409 ± 0.1175	1.246 ± 0.2410	1.256 ± 0.08499
Nominal	EUR	0.7281 ± 0.5148	1.215 ± 0.1715	1.264 ± 0.09485
Effective	AFR	0.3712 ± 0.08586	0.6691 ± 0.1672	1.236 ± 0.08772
Effective	EUR	0.4740 ± 0.2977	0.7507 ± 0.1574	1.235 ± 0.1238

Supplementary Table 4: Nominal and effective coverages (each cell is the mean ± standard deviation) of the representative cohorts for each superpopulation for each experiment.

Super Population	A	B	D	E
AFR	0.9456 (0.9392-0.9527)	0.9533 (0.9472-0.9578)	0.9609 (0.958-0.9628)	0.9214 (0.9087-0.9313)
EUR	0.9626 (0.9585-0.9676)	0.9694 (0.9663-0.9723)	0.9731 (0.9708-0.9751)	0.9663 (0.9612-0.969)

Supplementary Table 5: Mean and interquartile range (in parentheses) NRC for filtered SNPs by experiment and super population.

Super Population	A	B	D	E
AFR	0.8577 (0.8462-0.8681)	0.8737 (0.8631-0.8823)	0.8875 (0.8819-0.89)	0.7998 (0.7787-0.8137)
EUR	0.8662 (0.8551-0.8767)	0.8823 (0.8761-0.8876)	0.8908 (0.8865-0.8952)	0.8549 (0.8487-0.8616)

Supplementary Table 6: Mean and interquartile range (in parentheses) NRC for unfiltered indels by experiment and super population.

Super Population	A	B	D	E
AFR	0.9256 (0.9193-0.9316)	0.9286 (0.9236-0.9338)	0.9303 (0.9263-0.9344)	0.9144 (0.9015-0.9238)
EUR	0.9379 (0.9326-0.9447)	0.9398 (0.9345-0.9451)	0.9389 (0.9359-0.9451)	0.948 (0.943-0.9535)

Supplementary Table 7: Mean and interquartile range (in parentheses) NRC for filtered indels by experiment and super population.

Super Population	A	B	D	E
AFR	0.9943 (0.9937-0.9949)	0.9956 (0.995-0.996)	0.9967 (0.9965-0.9968)	0.9903 (0.989-0.9912)
EUR	0.9964 (0.996-0.997)	0.9974 (0.9972-0.9976)	0.998 (0.9978-0.9981)	0.9957 (0.9952-0.9959)

Supplementary Table 8: Mean and interquartile range (in parentheses) overall concordance for unfiltered SNPs by experiment and super population.

Super Population	A	B	D	E
AFR	0.9973 (0.997-0.9976)	0.9976 (0.9973-0.9978)	0.9979 (0.9977-0.998)	0.9967 (0.9962-0.9971)
EUR	0.9985 (0.9983-0.9987)	0.9987 (0.9985-0.9988)	0.9988 (0.9987-0.9989)	0.9988 (0.9986-0.9989)

Supplementary Table 9: Mean and interquartile range (in parentheses) overall concordance for filtered SNPs by experiment and super population.

Super Population	A	B	D	E
AFR	0.9716 (0.9689-0.9737)	0.9749 (0.9728-0.9767)	0.9777 (0.9766-0.9784)	0.9596 (0.9552-0.9626)
EUR	0.9766 (0.9745-0.9786)	0.9795 (0.9784-0.9805)	0.981 (0.9803-0.9819)	0.9746 (0.9734-0.976)

Supplementary Table 10: Mean and interquartile range (in parentheses) overall concordance for unfiltered indels by experiment and super population.

Super Population	A	B	D	E
AFR	0.9873 (0.9863-0.9883)	0.9874 (0.9865-0.9883)	0.9873 (0.9865-0.9881)	0.9879 (0.9865-0.9888)
EUR	0.9909 (0.9902-0.9918)	0.9908 (0.99-0.9916)	0.9904 (0.99-0.9914)	0.9933 (0.9929-0.994)

Supplementary Table 11: Mean and interquartile range (in parentheses) overall concordance for filtered indels by experiment and super population.

Super Population	Non-Reference Allele Frequency	A	B	D	E
AFR	0.005	0.6331	0.6624	0.6916	0.5141
AFR	0.015	0.8443	0.8781	0.9091	0.6734
AFR	0.025	0.8700	0.9023	0.9305	0.7056
AFR	0.035	0.8822	0.9127	0.9396	0.7272
AFR	0.045	0.8860	0.9161	0.9420	0.7417
AFR	0.055	0.8887	0.9181	0.9437	0.7598
AFR	0.065	0.8888	0.9182	0.9432	0.7696
AFR	0.075	0.8909	0.9204	0.9451	0.7825
AFR	0.085	0.8937	0.9226	0.9468	0.7930
AFR	0.095	0.8947	0.9232	0.9478	0.8002
AFR	0.125	0.8971	0.9254	0.9497	0.8178
AFR	0.175	0.9012	0.9288	0.9528	0.8372
AFR	0.225	0.9053	0.9316	0.9543	0.8457
AFR	0.375	0.9153	0.9378	0.9583	0.8623
AFR	0.750	0.9538	0.9651	0.9778	0.9235
EUR	0.005	0.4087	0.4603	0.5030	0.3782
EUR	0.015	0.7124	0.7756	0.8225	0.6822
EUR	0.025	0.7870	0.8376	0.8719	0.7091
EUR	0.035	0.8325	0.8762	0.9023	0.7583
EUR	0.045	0.8569	0.8967	0.9190	0.7935
EUR	0.055	0.8730	0.9101	0.9303	0.8159
EUR	0.065	0.8803	0.9154	0.9353	0.8298
EUR	0.075	0.8877	0.9227	0.9417	0.8424
EUR	0.085	0.8969	0.9296	0.9480	0.8556
EUR	0.095	0.8986	0.9315	0.9494	0.8575
EUR	0.125	0.9071	0.9389	0.9559	0.8739
EUR	0.175	0.9158	0.9453	0.9612	0.8900
EUR	0.225	0.9207	0.9484	0.9629	0.9026
EUR	0.375	0.9303	0.9533	0.9657	0.9143
EUR	0.750	0.9651	0.9762	0.9829	0.9590

Supplementary Table 12: Mean non-reference concordance by minor allele frequency bin in 1KGP3 for EUR and AFR cohorts by experiment. The values in the Minor Allele Frequency column denote the midpoint of the respective bin. These are the values plotted in Figure 2

Minor Allele Frequency	A	D	B	E
0.005	0.6184	0.7235	0.6478	0.6184
0.015	0.8532	0.9261	0.8874	0.7328
0.025	0.8733	0.9383	0.9057	0.7536
0.035	0.8798	0.9418	0.9099	0.7684
0.045	0.8801	0.9433	0.9116	0.7802
0.055	0.8793	0.9414	0.9098	0.7889
0.065	0.8784	0.9404	0.9080	0.7956
0.075	0.8799	0.9426	0.9111	0.8008
0.085	0.8824	0.9443	0.9132	0.8058
0.095	0.8824	0.9450	0.9139	0.8080
0.125	0.8854	0.9470	0.9169	0.8181
0.175	0.8907	0.9500	0.9212	0.8321
0.225	0.8949	0.9511	0.9240	0.8412
0.375	0.8996	0.9517	0.9266	0.8481

Supplementary Table 13: Mean r^2s by minor allele frequency bin in 1KGP3 for the AFR cohort by experiment. The values in the Minor Allele Frequency column denote the midpoint of the respective bin. The bin edges are therefore at 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.15, 0.2, 0.25, 0.5. These are the values plotted in the left pane of Figure 3.

Minor Allele Frequency	A	D	B	E
0.005	0.4522	0.5896	0.5002	0.5172
0.015	0.7496	0.8614	0.8031	0.7940
0.025	0.8137	0.8938	0.8535	0.8160
0.035	0.8492	0.9165	0.8836	0.8485
0.045	0.8669	0.9296	0.9001	0.8650
0.055	0.8778	0.9363	0.9095	0.8750
0.065	0.8852	0.9395	0.9143	0.8823
0.075	0.8925	0.9456	0.9234	0.8855
0.085	0.8988	0.9503	0.9295	0.8900
0.095	0.9030	0.9529	0.9334	0.8936
0.125	0.9119	0.9592	0.9411	0.9003
0.175	0.9205	0.9640	0.9483	0.9098
0.225	0.9240	0.9649	0.9506	0.9158
0.375	0.9248	0.9630	0.9494	0.9168

Supplementary Table 14: Mean r^2s by minor allele frequency in 1KGP3 for the EUR cohort by experiment. The values in the Minor Allele Frequency column denote the midpoint of the respective bin. The bin edges are therefore at 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.15, 0.2, 0.25, 0.5. These are the values plotted in the right pane of Figure 3.

Super population	Experiment A	 D	 B	 E
AFR	0.8917 (0.8614-0.9895)	0.9488 (0.9558-1)	0.9208 (0.9079-0.9986)	0.8302 (0.7635-0.9784)
EUR	0.9159 (0.907-0.998)	0.9596 (0.993-1)	0.9429 (0.9514-0.9999)	0.9074 (0.907-0.9987)

Supplementary Table 15: Mean and interquartile range (in parentheses) imputation r^2s for each representative cohort for each experiment by superpopulation cohort at biallelic SNPs above 5% minor allele frequency in the 1KGP3.

		Experiment			
Super population	Trait	A	B	D	E
AFR	Breast Cancer	0.8867	0.9353	0.9745	0.8943
EUR	Breast Cancer	0.9054	0.9471	0.9698	0.9486
AFR	CAD	0.8924	0.9195	0.9380	0.8726
EUR	CAD	0.9570	0.9832	0.9894	0.9721

Supplementary Table 16: Squared Pearson correlation coefficient (r^2) between PRS estimates from imputed dosages vs. true PRS as calculated off 1KGP3 genotypes by superpopulation and trait.

Super population	Trait	MSE	SE of MSE
AFR	Breast Cancer	0.0400	0.0075
EUR	Breast Cancer	0.0195	0.0035
AFR	CAD	0.0076	0.0010
EUR	CAD	0.0116	0.0017

Supplementary Table 17: MSE in array-based PRS estimates and the standard error of the MSE. The MSE is calculated by first taking the average of the squared error in PRS estimates among replicates of a given cell line and trait, and then averaging over all cell lines within a population for a given trait.

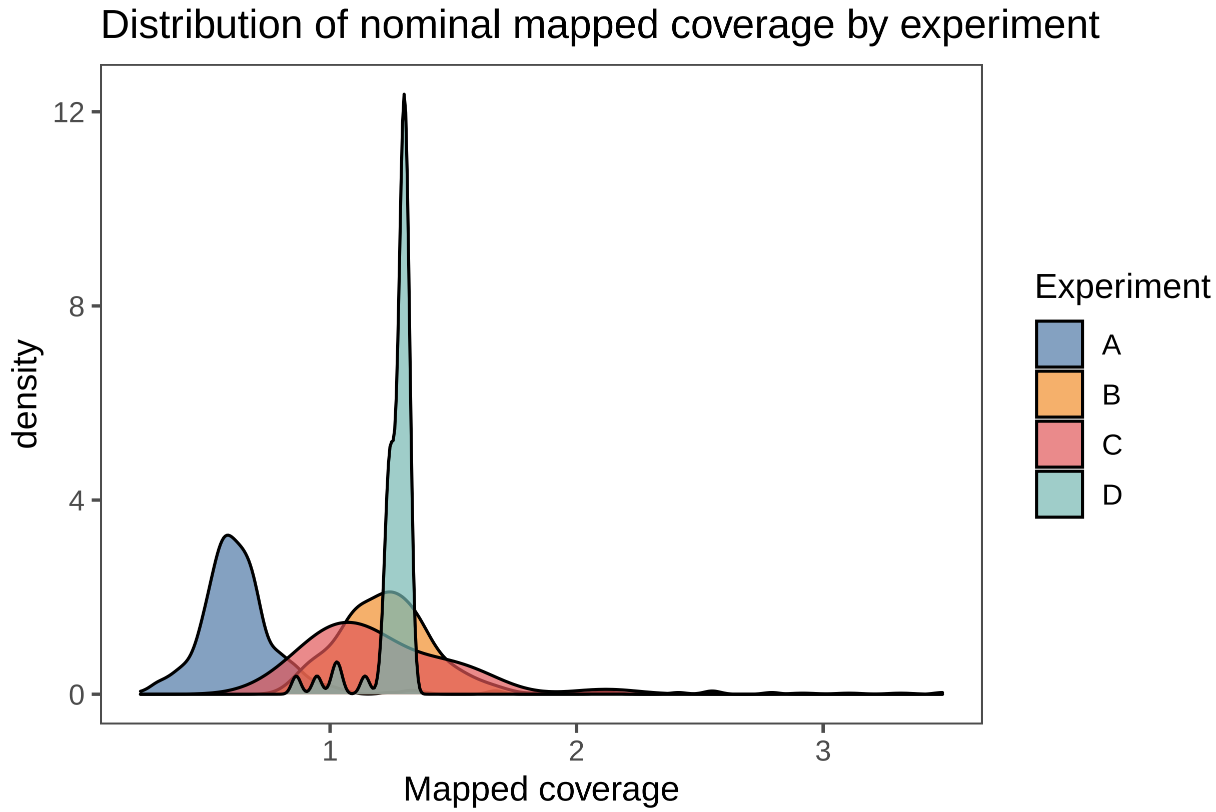
Experiment	Trait	Super population	MSE	SE of MSE	MSE fold decrease
A	Breast Cancer	AFR	0.0429	0.0055	0.9316
A	Breast Cancer	EUR	0.0347	0.0059	0.5626
B	Breast Cancer	AFR	0.0247	0.0032	1.6175
B	Breast Cancer	EUR	0.0197	0.0026	0.9884
C	Breast Cancer	EUR	0.0411	NA	0.4749
D	Breast Cancer	AFR	0.0112	0.0043	3.5671
D	Breast Cancer	EUR	0.0180	0.0051	1.0862
A	CAD	AFR	0.0065	0.0010	1.1629
A	CAD	EUR	0.0124	0.0017	0.9350
B	CAD	AFR	0.0049	0.0008	1.5331
B	CAD	EUR	0.0051	0.0006	2.2675
C	CAD	EUR	0.0073	NA	1.6011
D	CAD	AFR	0.0027	0.0006	2.7985
D	CAD	EUR	0.0027	0.0011	4.2722

Supplementary Table 18: Mean squared error (MSE) in sequence-based PRS estimates and the standard error (SE) of the MSE. The MSE is calculated by first taking the average of the squared error in PRS estimates among replicates of a given cell line and trait, and then averaging over all cell lines within a population for a given trait. MSE fold decrease is the x-fold decrease in MSE compared to the array counterpart. For instance, the MSE for experiment D's CAD PRS estimates for the AFR cohort was approximatedly four-fold the MSE of the array-based AFR CAD estimates (Supplementary Table 17).

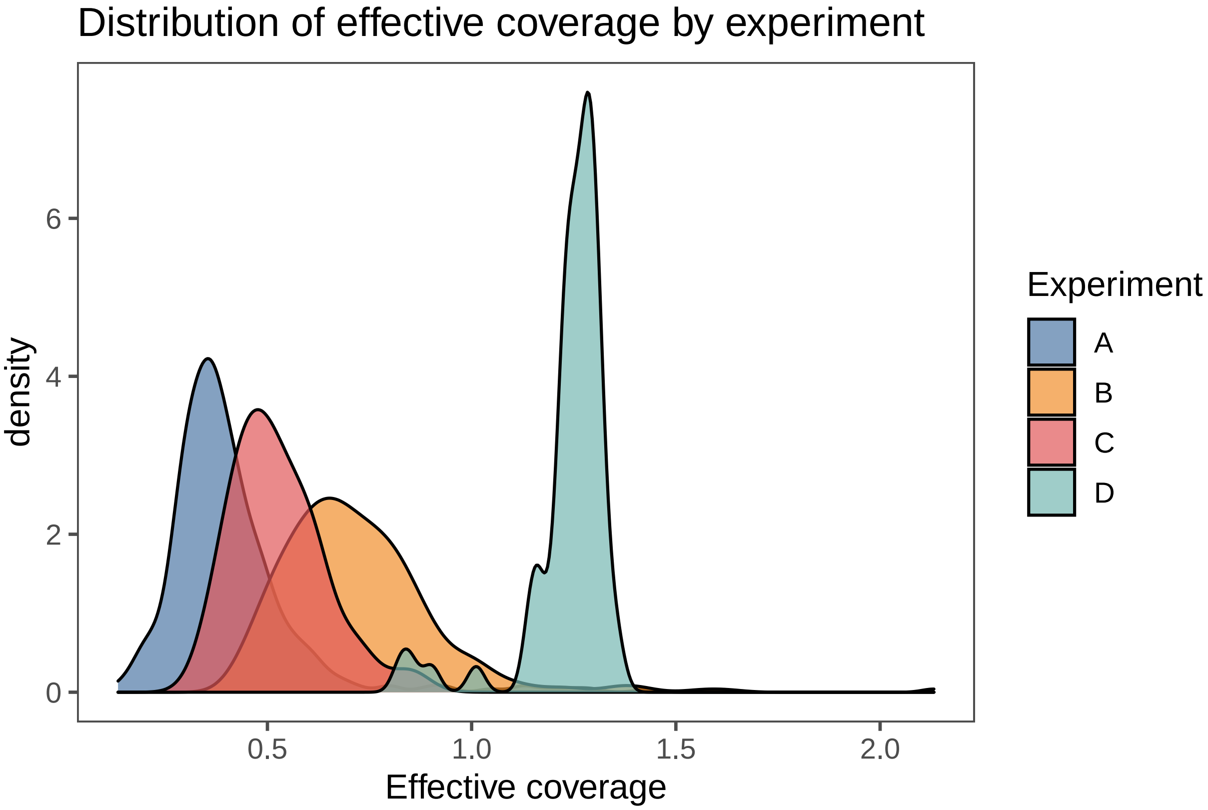
Super population	Trait	Experiment	p value
EUR	CAD	A	0.625
EUR	BRCA	A	**0.00621**
EUR	CAD	B	**4.699e-08**
EUR	BRCA	B	0.936
EUR	CAD	C	**0.0406**
EUR	BRCA	C	**0.0352**
EUR	CAD	D	**2.45e-08**
EUR	BRCA	D	0.768
AFR	CAD	A	0.308
AFR	BRCA	A	0.672
AFR	CAD	B	**0.00426**
AFR	BRCA	B	**0.00472**
AFR	CAD	D	**6.077e-07**
AFR	BRCA	D	**1.311e-05**

Supplementary Table 19: P-values from Welch two-sample t-tests for differences in means of the squared errors in PRS estimates between each sequence-based experiment and the corresponding squared errors in PRS estimates from the array-based estimates. Bolded values are significant at a level $\alpha = 0.05$.

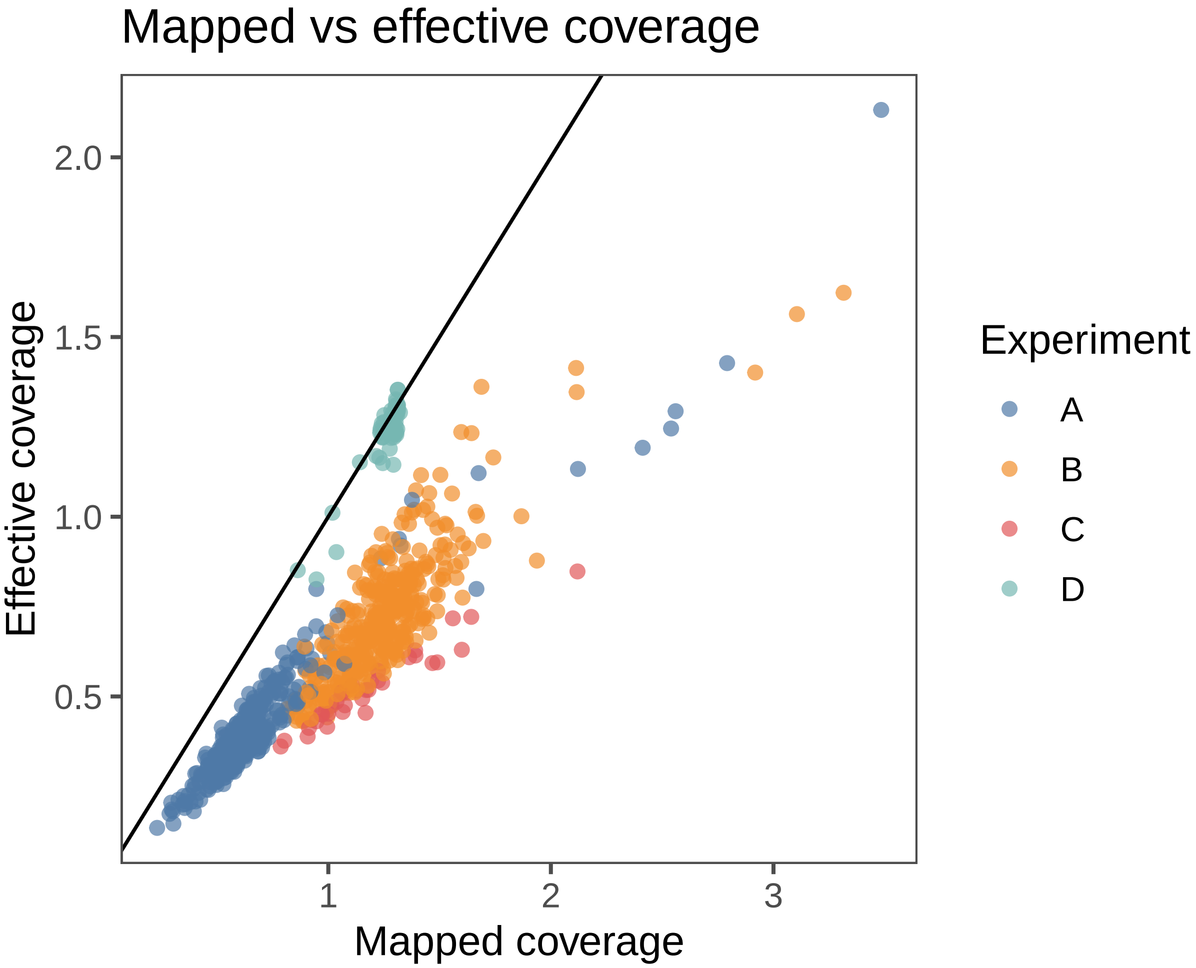
2 Figures



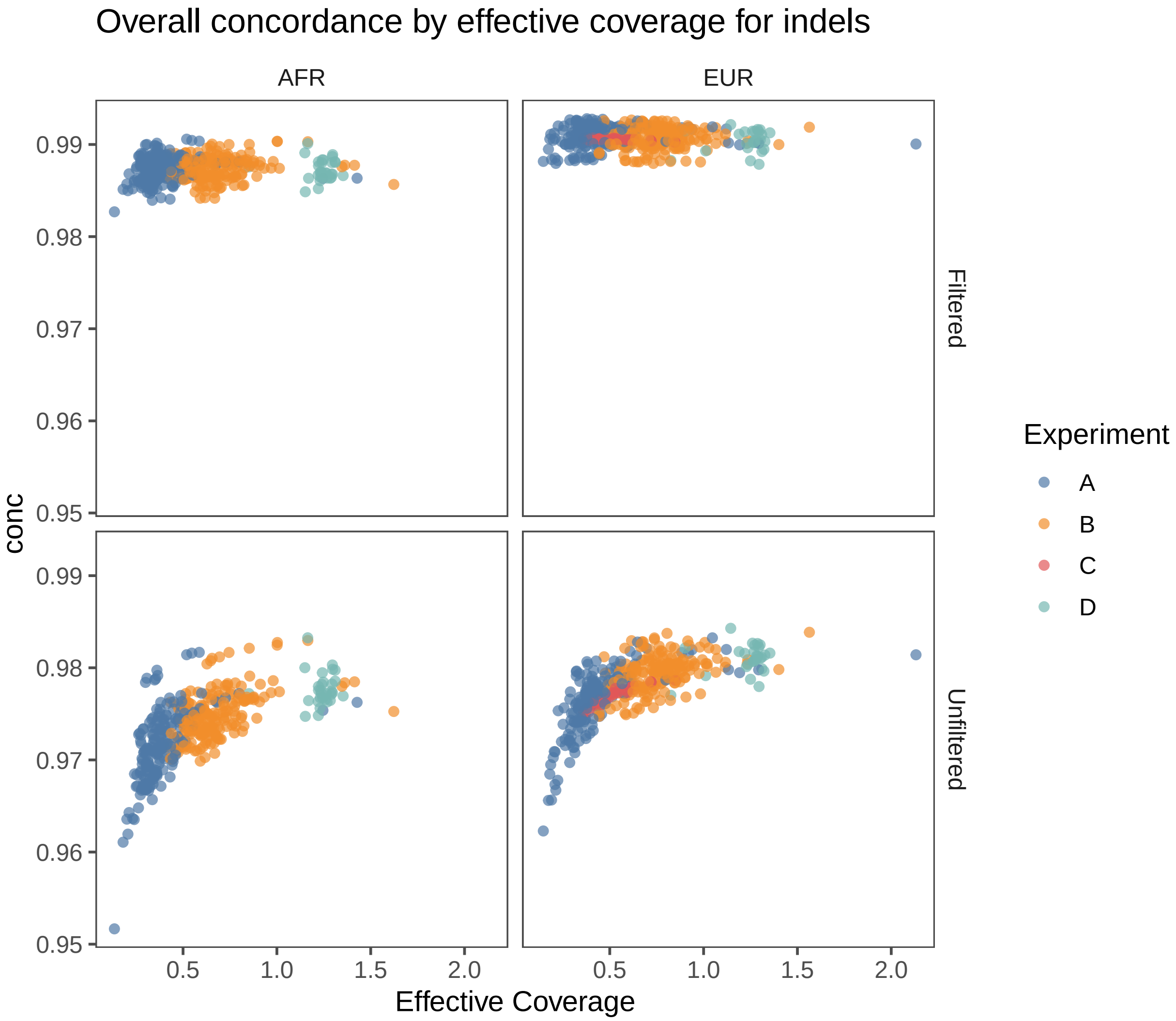
Supplementary Figure 1: Distribution of mapped coverages by experiment.



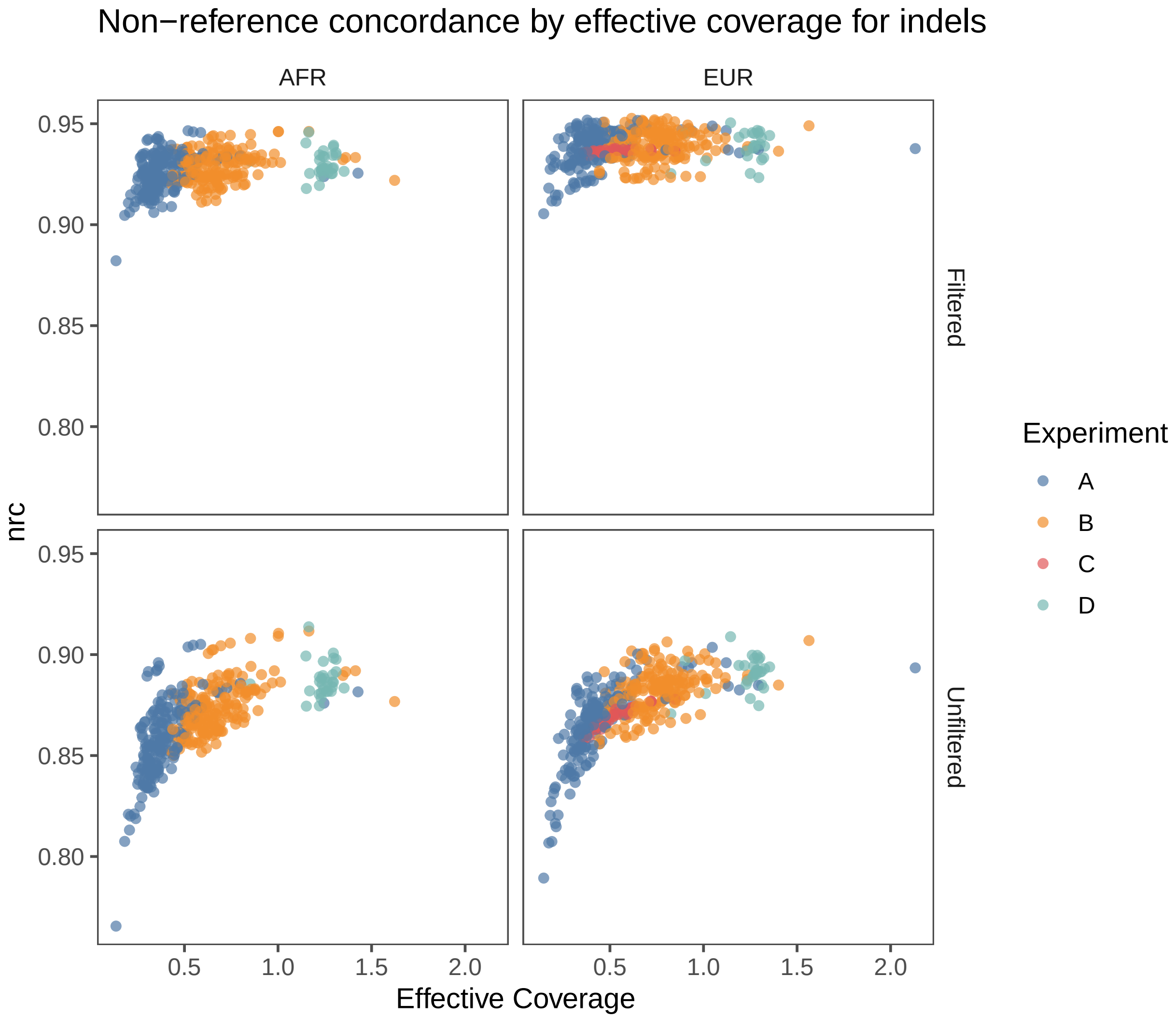
Supplementary Figure 2: Distribution of effective coverages by experiment.



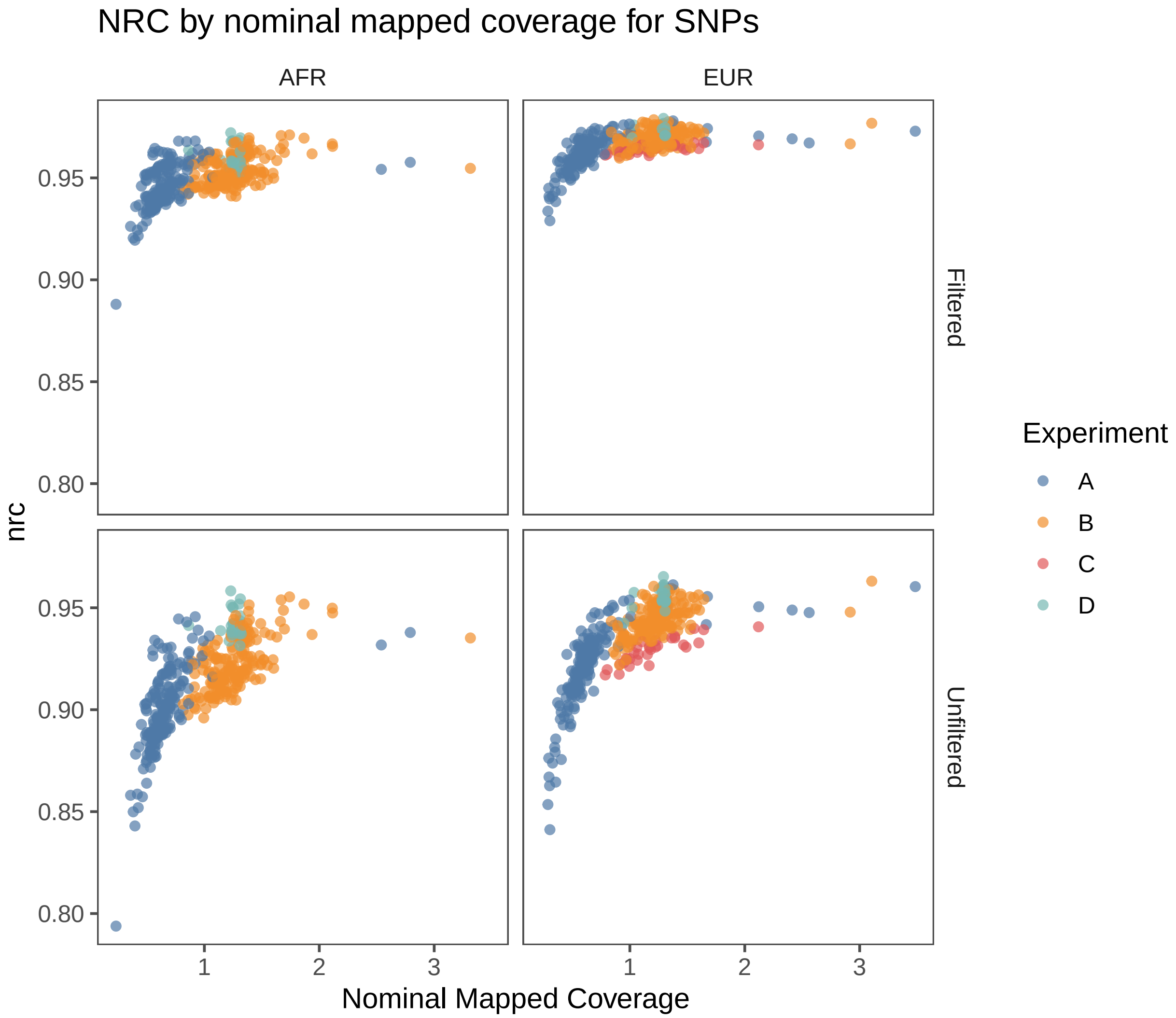
Supplementary Figure 3: Mapped deduplicated coverage vs. effective coverage as described in main manuscript.



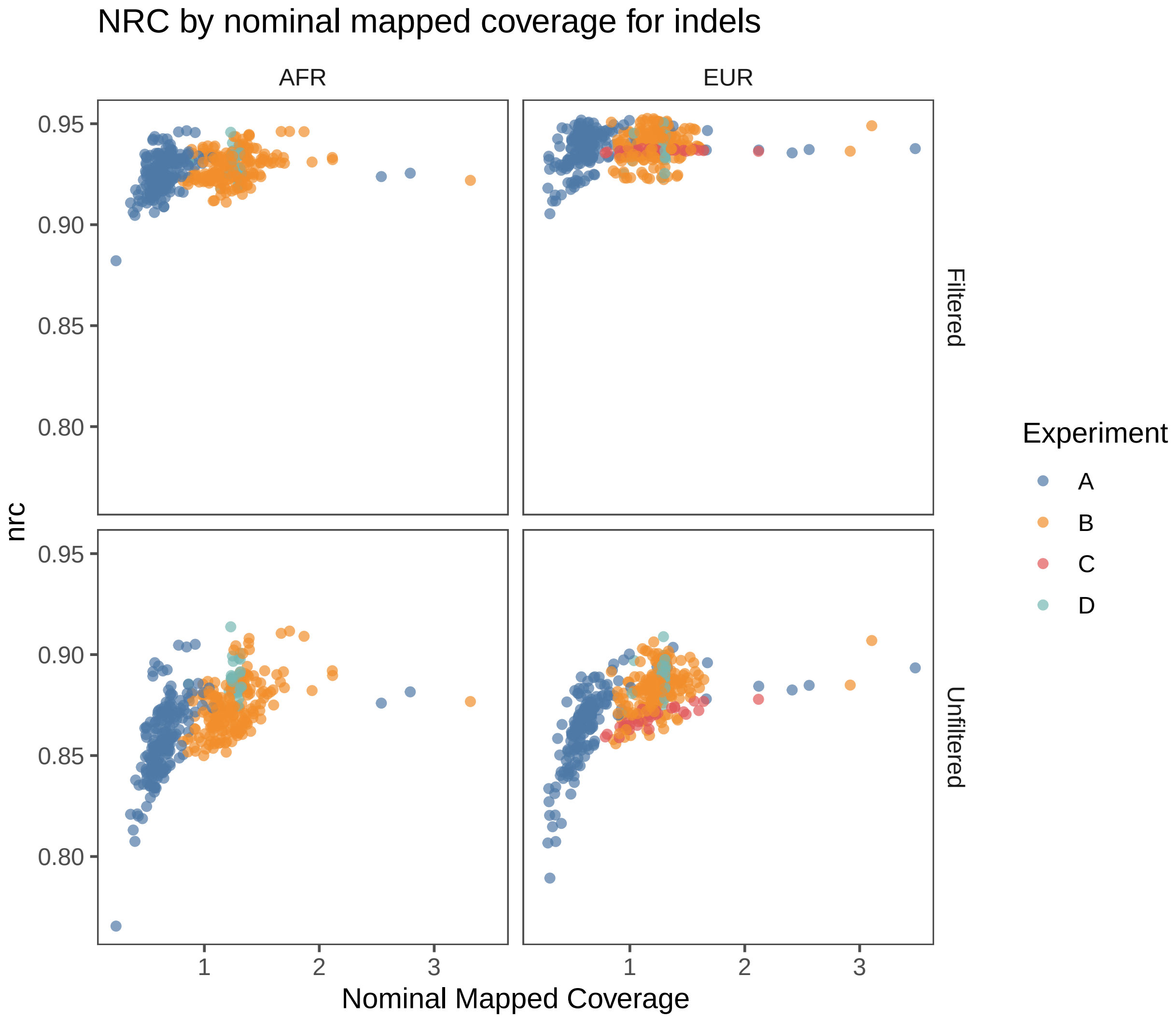
Supplementary Figure 4: Overall concordance for SNPs by effective coverage for both filtered and unfiltered variant calls.



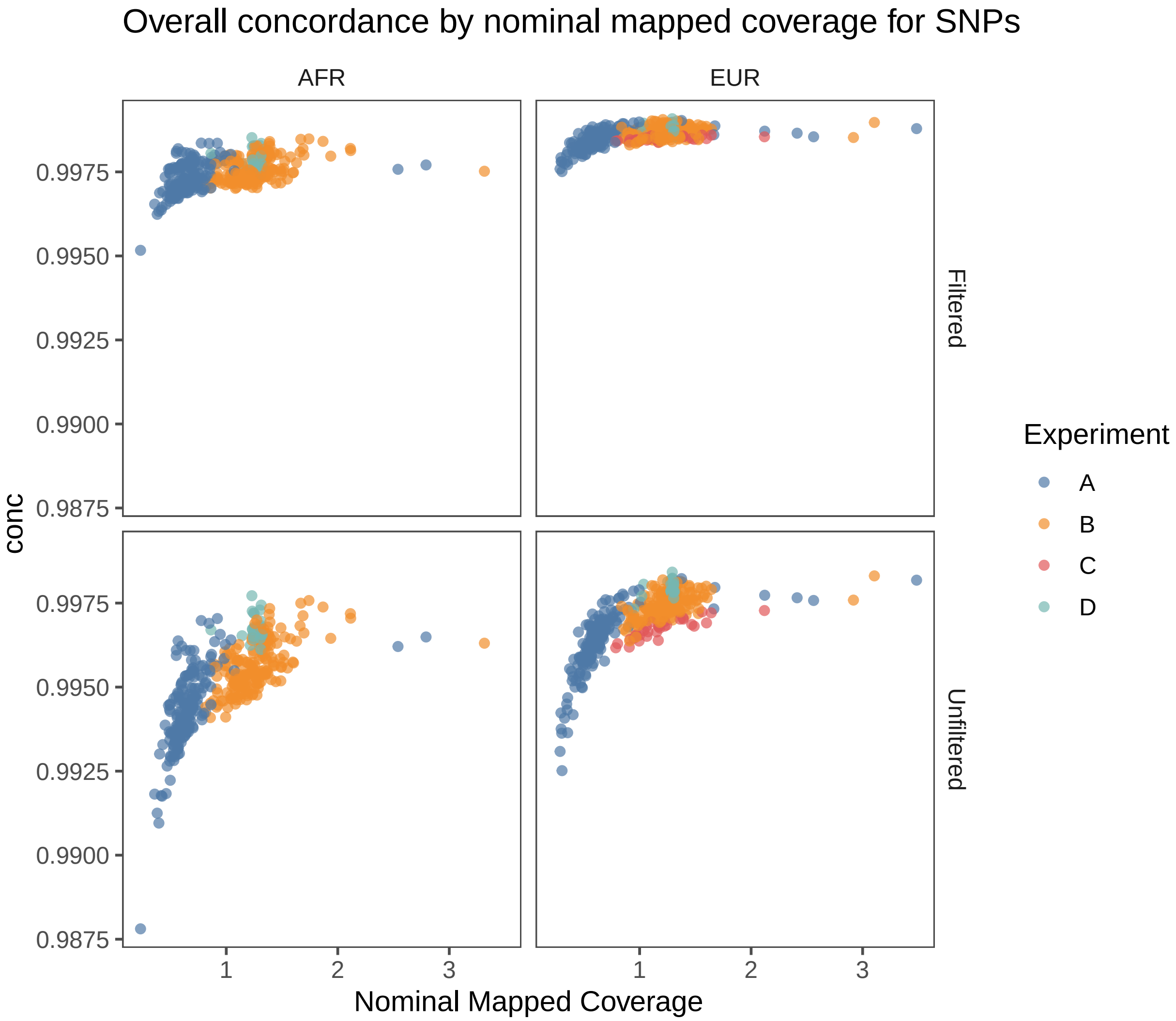
Supplementary Figure 5: Non-reference concordance for indels by effective coverage for both filtered and unfiltered variant calls.



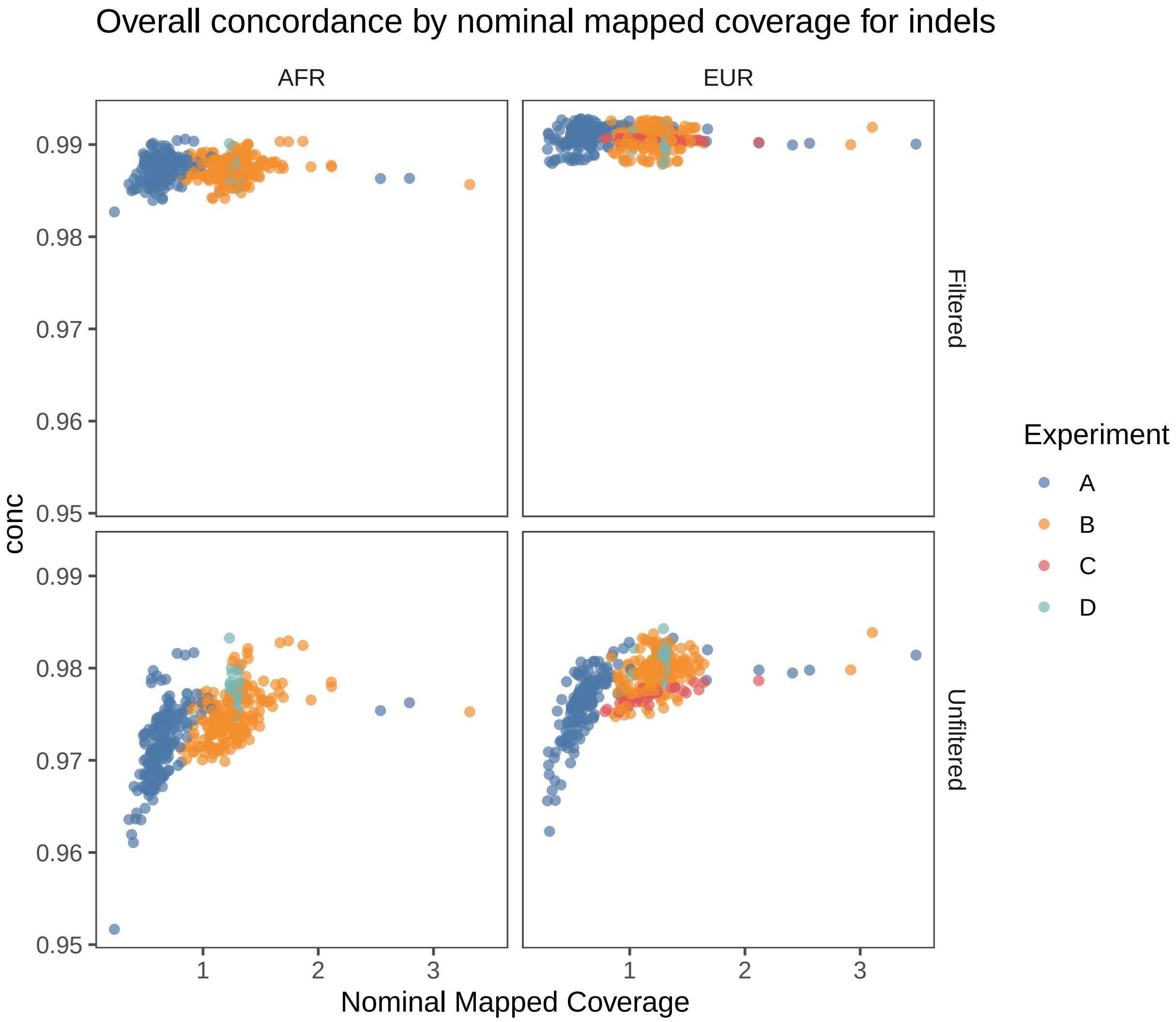
Supplementary Figure 6: Non-reference concordance for SNPs by nominal coverage for both filtered and unfiltered variant calls.



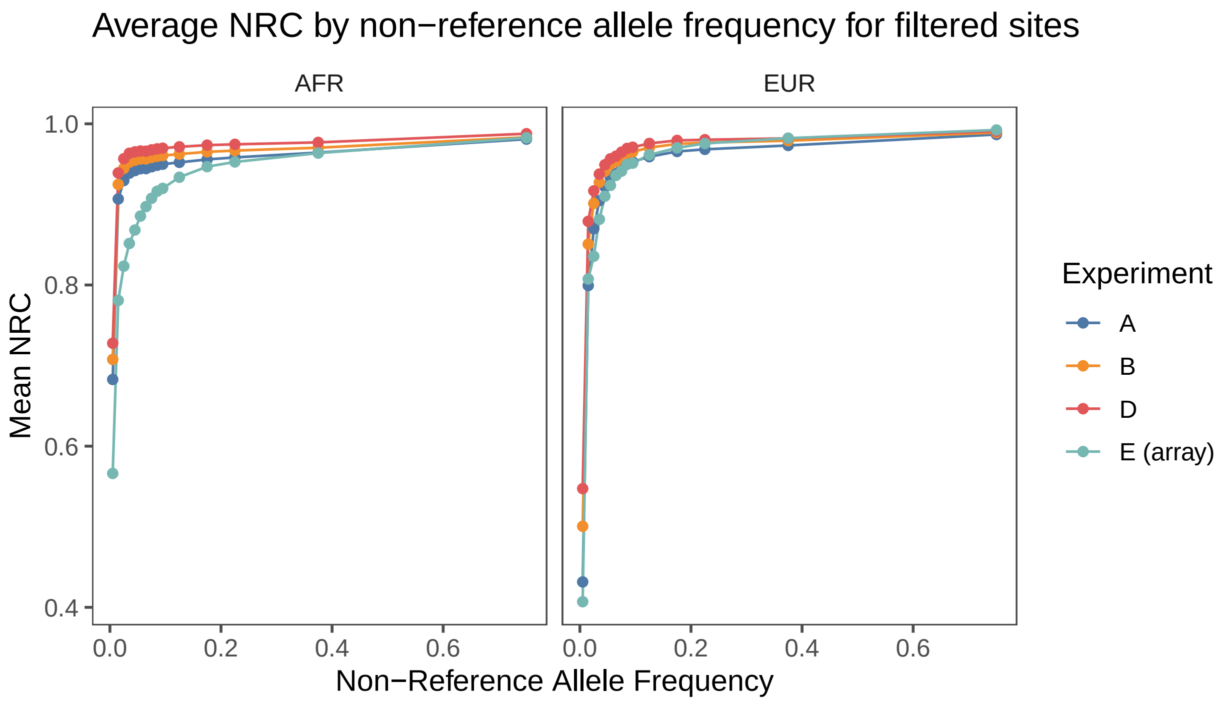
Supplementary Figure 7: Non-reference concordance for indels by nominal coverage for both filtered and unfiltered variant calls.



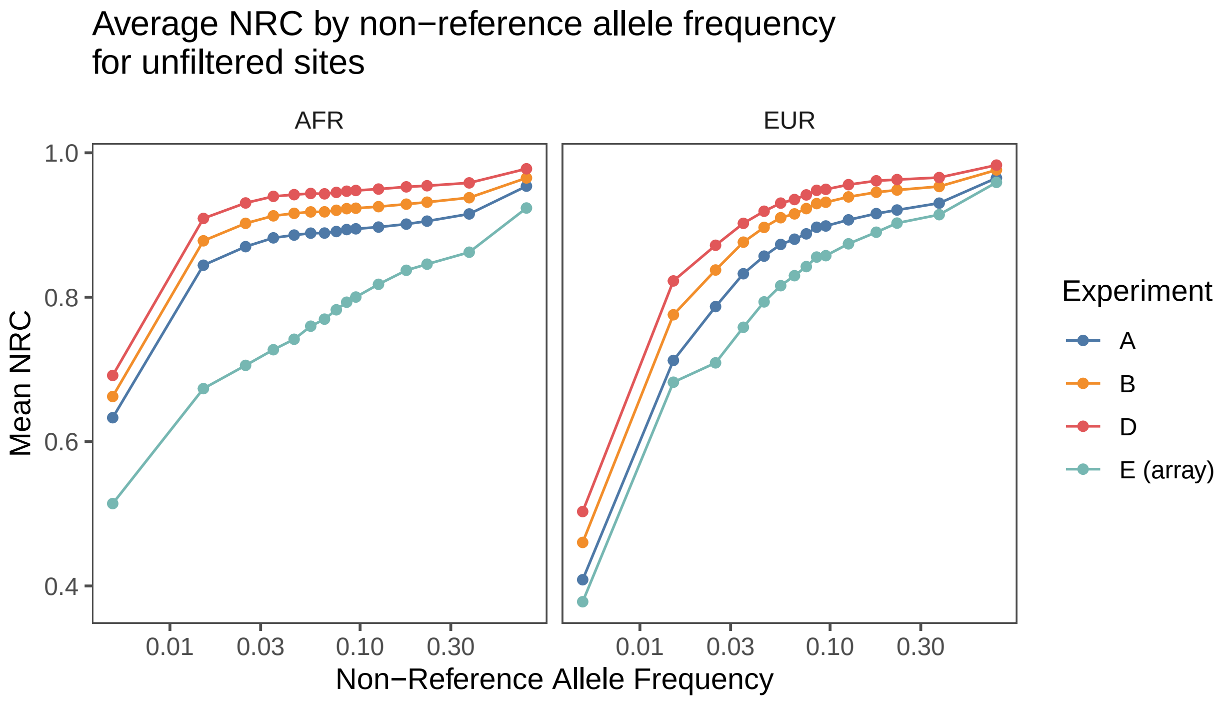
Supplementary Figure 8: Overall concordance for SNPs by nominal coverage for both filtered and unfiltered variant calls.



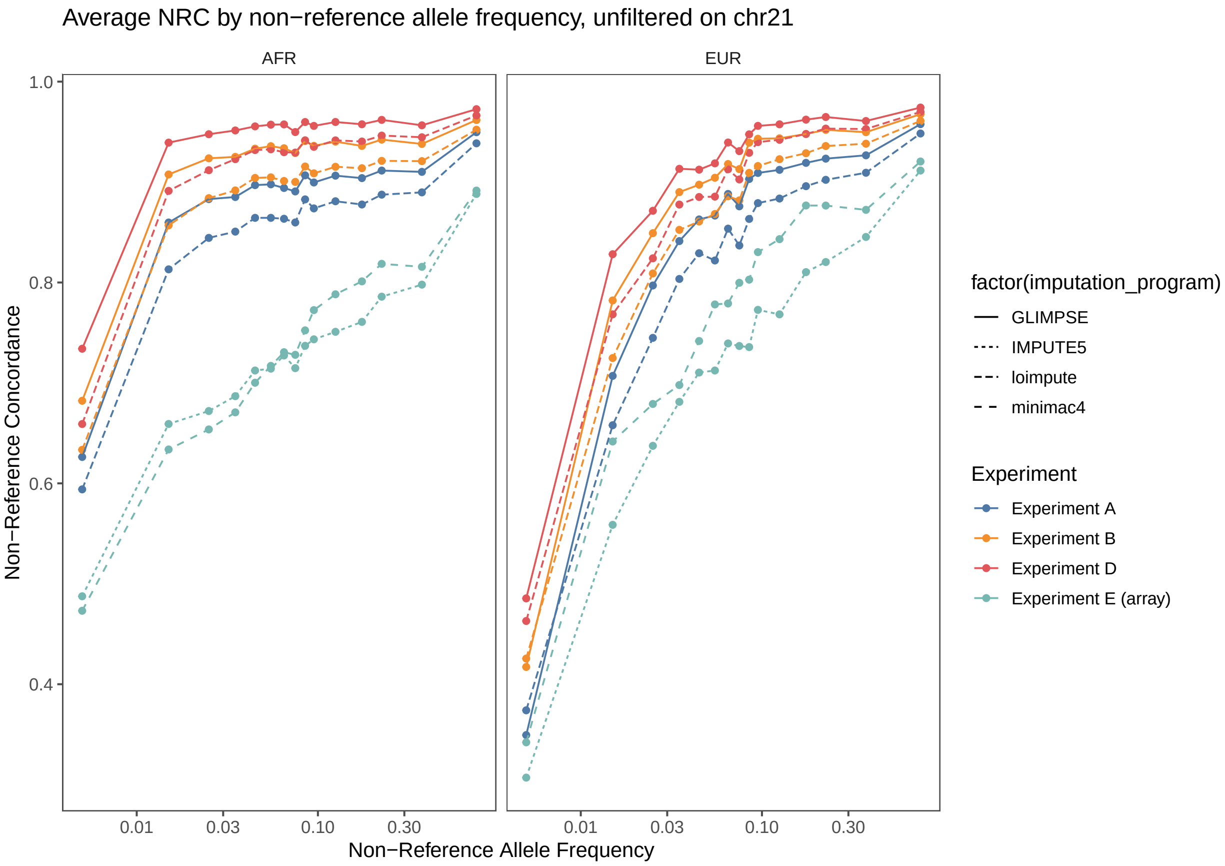
Supplementary Figure 9: Overall concordance for indels by nominal coverage for both filtered and unfiltered variant calls.



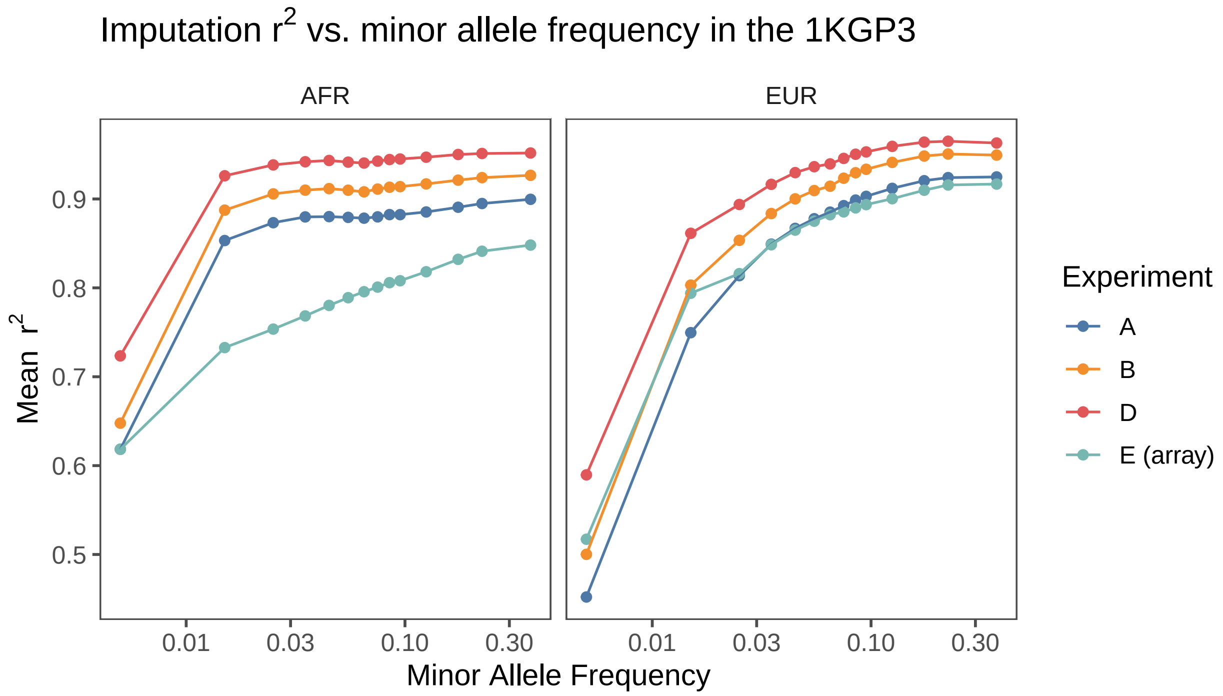
Supplementary Figure 10: Non-reference concordance for SNPs at filtered variant calls by non-reference allele frequency in the 1000 Genomes Phase 3 callset.



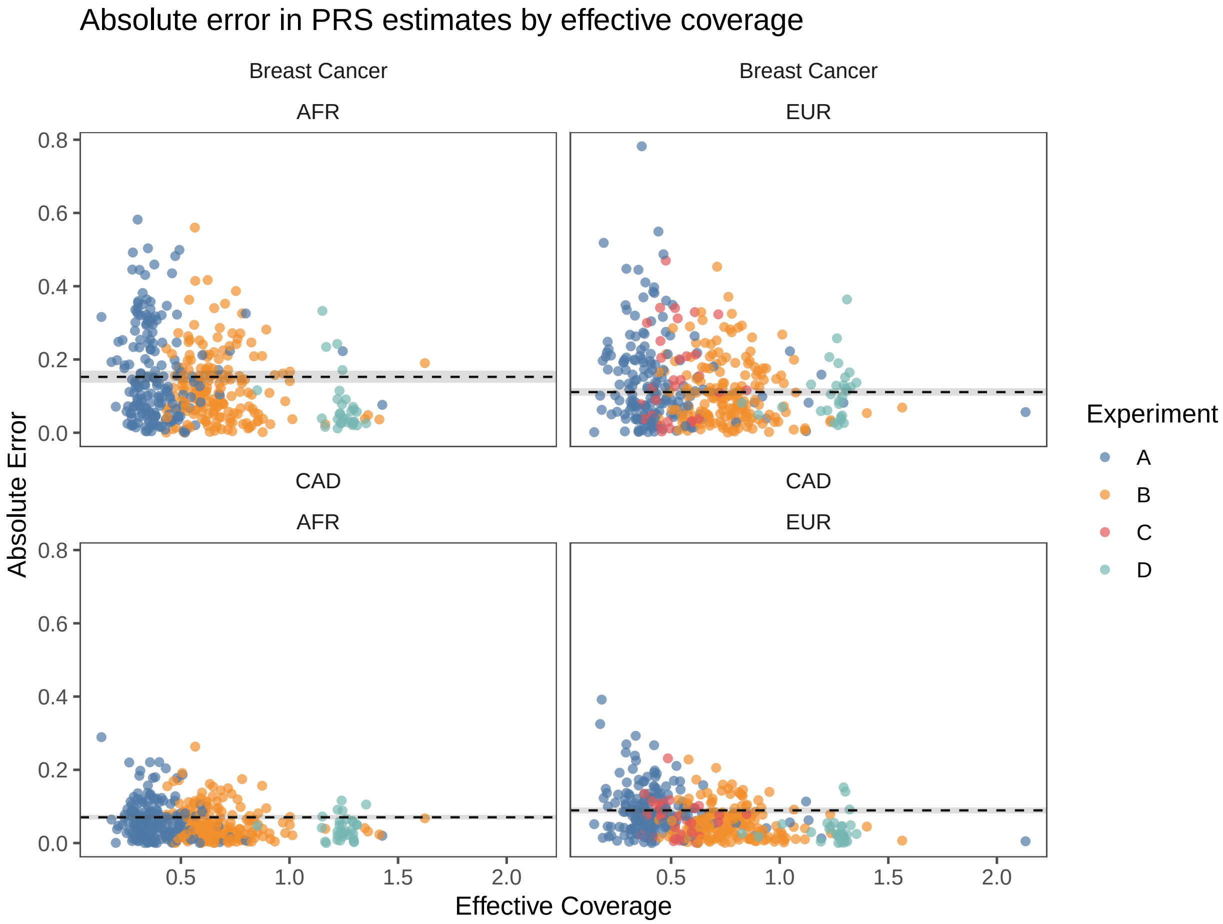
Supplementary Figure 11: Average non-reference concordance for unfiltered SNPs by superpopulation by non-reference allele frequency in 1KGP3. Same figure as Figure 2 on a log scale.



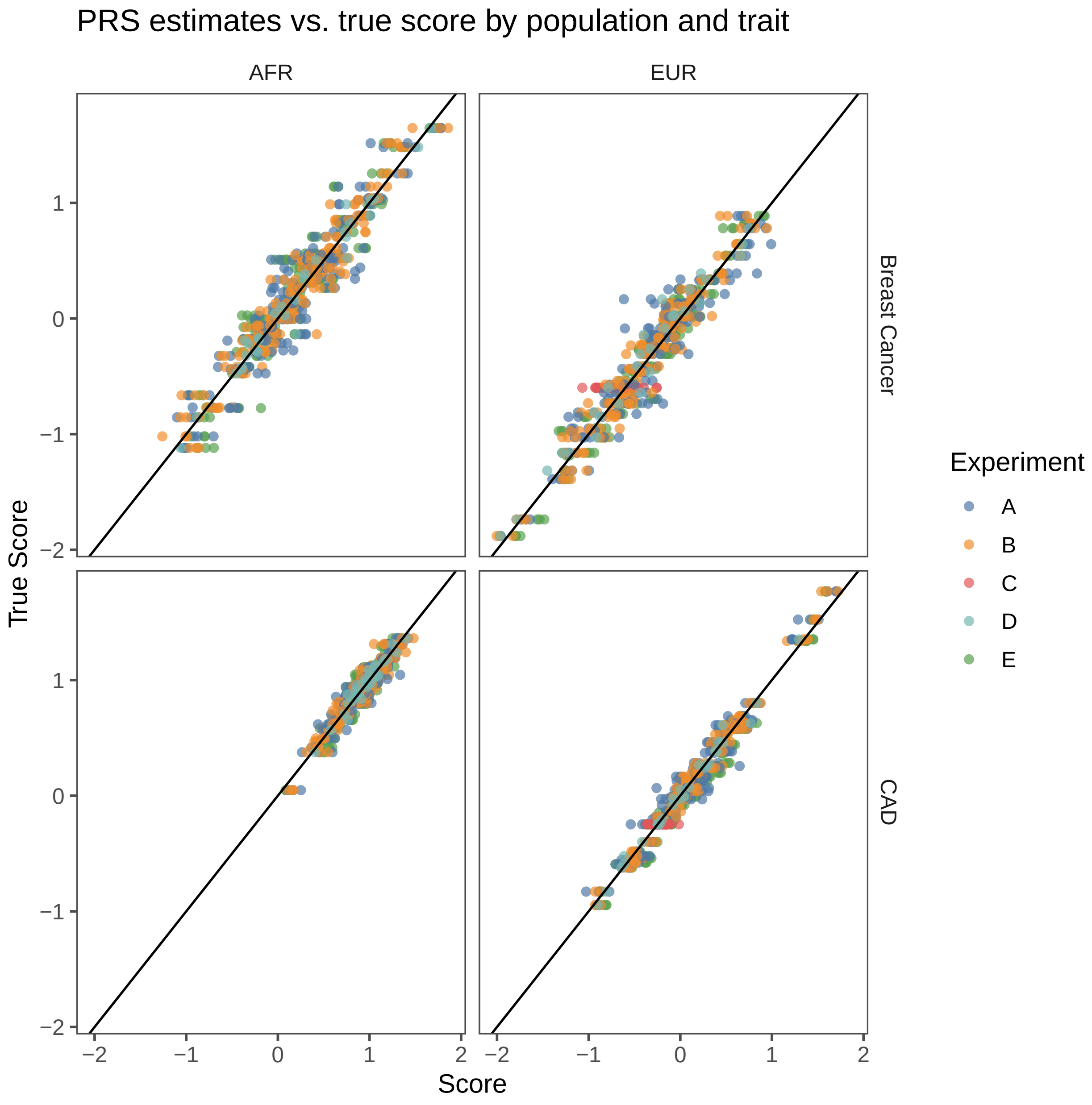
Supplementary Figure 12: Non-reference concordance for SNPs at unfiltered variant calls by non-reference allele frequency in the 1000 Genomes Phase 3 callset, by imputation algorithm and Experiment.



Supplementary Figure 13: Same as in Figure 3 but on a log scale. The values on the x-axis are the midpoints of the minor allele frequency bins. The bin edges are at 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.15, 0.2, 0.25, 0.5.



Supplementary Figure 14: Absolute errors in polygenic risk score estimates by effective coverage. The dotted horizontal lines denote the mean error in PRS estimation from Illumina GSA calls across cell lines replicates. The shaded area about each dotted horizontal line denotes the standard error of the mean of these quantities.



Supplementary Figure 15: Scatterplot of estimated PRS from imputed dosages vs the true PRS for all samples across experiments, populations, and traits. The PRS computed off genotypes in the 1KGP3 genotypes was considered the "truth" value for all cell lines.