Cell Genomics, Volume *2*

Supplemental information

Genome-wide risk prediction of common diseases across ancestries in one million people

Nina Mars, Sini Kerminen, Yen-Chen A. Feng, Masahiro Kanai, Kristi Läll, Laurent F. Thomas, Anne Heidi Skogholt, Pietro della Briotta Parolo, The Biobank Japan Project, FinnGen, Benjamin M. Neale, Jordan W. Smoller, Maiken E. Gabrielsen, Kristian Hveem, Reedik Mägi, Koichi Matsuda, Yukinori Okada, Matti Pirinen, Aarno Palotie, Andrea Ganna, Alicia R. Martin, and Samuli Ripatti

Table S1. Number of variants included in the LDpred polygenic risk scores (PRS). The table shows the number of variants used for calculating the PRS in each dataset shown in Figure 1. The table also shows the proportion (%) of variants out of the original LDpred-adjusted summary statistics used for calculating the PRS.

	CAD		**T2D**		**Breast cancer**		**Prostate cancer**	
	***p* = 0.003**		***p* = 0.003**		***p* = 0.03**		***p* = 0.01**	
	Variants	**%**	**Variants**	**%**	**Variants**	**%**	**Variants**	**%**
Original summary statistics	6 576 338	-	6 431 973	-	6 494 889	-	6 497 734	-
BioBank Japan	4 410 149	67.1	4 344 264	67.5	4 375 073	67.4	4 375 963	67.3
Estonian Biobank	6 259 397	95.2	6 429 286	100.0	6 490 064	99.9	6 492 877	99.9
FinnGen	6 068 083	92.3	6 320 939	98.3	6 375 257	98.2	6 377 882	98.2
HUNT	6 472 970	98.4	6 376 169	99.1	6 420 662	98.9	6 423 074	98.9
MGB Biobank								
European	5 911 460	89.9	5 796 975	90.1	5 806 787	89.4	5 804 880	89.3
African	5 086 035	77.3	4 881 847	75.9	4 918 230	75.7	4 917 955	75.7
UK Biobank								
European								
African / Caribbean	6 165 767	93.8	6 422 449	99.9	6 480 780	99.8	6 483 638	99.8
South Asian								

CAD = coronary artery disease, T2D = type 2 diabetes. *p* denotes the LDpred parameter for the fraction of causal variants in the selected PRS. The PRSs with the best discriminative capacity (measured with maximum area under the receiver-operator curve, AUC) were chosen based on an earlier FinnGen data freeze (DF4) with 176,899 individuals.

Table S2. Effect sizes, and case and control counts corresponding to Figure 1. Odds ratios (OR) with 95% confidence intervals (CI) are presented for 1-SD increase in the polygenic risk scores.

	Disease	OR	95% CI	p-value for test of heterogeneity	Number of cases	Number of controls
Figure 1, Panel A						
MGB Biobank, African	CAD	1.10	0.96-1.26	0.06	285	1 250
UK Biobank, African / Caribbean	CAD	1.32	1.13-1.54		169	7 459
BioBank Japan	CAD	1.32	1.30-1.34		29 080	149 646
UK Biobank, South Asian	CAD	1.41	1.30-1.53		740	6 888
European (pooled estimate)	CAD	1.54	1.53-1.55		-	-
MGB Biobank, African	T2D	1.24	1.09-1.42	7.38e-06	660	875
UK Biobank, African / Caribbean	T2D	1.46	1.32-1.62		691	6 656
BioBank Japan	T2D	1.37	1.36-1.39		40 121	137 024
UK Biobank, South Asian	T2D	1.66	1.55-1.79		1 120	6 145
European (pooled estimate)	T2D	1.62	1.61-1.64		-	-
MGB Biobank, African	Breast cancer	0.90	0.69-1.17	0.03	64	879
UK Biobank, African / Caribbean	Breast cancer	1.12	0.93-1.35		132	4 210
BioBank Japan	Breast cancer	1.25	1.21-1.28		5 316	69 629
UK Biobank, South Asian	Breast cancer	1.47	1.23-1.75		139	3 375
European (pooled estimate)	Breast cancer	1.49	1.47-1.51		-	-
MGB Biobank, African	Prostate cancer	1.19	0.91-1.55	0.001	80	512
UK Biobank, African / Caribbean	Prostate cancer	1.35	1.14-1.61		199	3 077
BioBank Japan	Prostate cancer	1.69	1.64-1.74		5 192	90 773
UK Biobank, South Asian	Prostate cancer	2.21	1.73-2.81		72	4 042
European (pooled estimate)	Prostate cancer	1.89	1.86-1.92		-	-
Figure 1, Panel B						
MGB Biobank, European	CAD	1.35	1.29 - 1.40	3.55e-28	3 206	22 490
Estonian Biobank	CAD	1.47	1.43 - 1.52		5 064	105 533
FinnGen	CAD	1.53	1.50 - 1.55		25 706	232 696
HUNT	CAD	1.44	1.40 - 1.48		6 594	62 827
UK Biobank, European	CAD	1.64	1.61 - 1.67		17 986	325 690
MGB Biobank, European	T2D	1.46	1.41 - 1.51	3.48e-35	5 182	20 514
Estonian Biobank	T2D	1.55	1.51 - 1.59		7 066	103 531
FinnGen	T2D	1.58	1.56 - 1.60		37 001	213 319
HUNT	T2D	1.64	1.60 - 1.69		5 228	64 191
UK Biobank, European	T2D	1.78	1.75 - 1.81		13 616	326 173

MGB Biobank, European	Breast cancer	1.45	1.38 - 1.54		1 513	12 139
Estonian Biobank	Breast cancer	1.45	1.37 - 1.53		1 379	73 053
FinnGen	Breast cancer	1.48	1.45 - 1.51	0.63	11 573	134 561
HUNT	Breast cancer	1.50	1.43 - 1.58		1 731	35 053
UK Biobank, European	Breast cancer	1.50	1.47 - 1.53		11 075	173 498
MGB Biobank, European	Prostate cancer	1.66	1.57 - 1.76		1 593	10 451
Estonian Biobank	Prostate cancer	1.79	1.68 - 1.91		1 202	34 963
FinnGen	Prostate cancer	1.96	1.91 - 2.01	2.91e-07	8 709	103 559
HUNT	Prostate cancer	1.80	1.72 - 1.88		2 224	30 413
UK Biobank, European	Prostate cancer	1.91	1.86 – 1.96		7 429	151 674
Figure 1, Panel C						
Early settlement	CAD	1.54	1.51-1.58		12 487	131 981
Borderline	CAD	1.51	1.45-1.56	0.56	4 809	42 888
Late settlement	CAD	1.54	1.50-1.59		6 837	51 283
Early settlement	T2D	1.59	1.56-1.62		19 937	119 799
Borderline	T2D	1.55	1.51-1.60	0.32	6 636	39 561
Late settlement	T2D	1.59	1.55-1.63		9 045	47 429
Early settlement	Breast cancer	1.49	1.45-1.53		6 866	75 151
Borderline	Breast cancer	1.48	1.42-1.55	0.70	2 098	25 506
Late settlement	Breast cancer	1.46	1.40-1.52		2 260	29 856
Early settlement	Prostate cancer	1.93	1.87-1.99		5 161	57 290
Borderline	Prostate cancer	2.09	1.97-2.22	0.07	1 451	18 642
Late settlement	Prostate cancer	1.95	1.84-2.06		1 651	24 353

CAD = coronary artery disease, T2D = type 2 diabetes. In Panel A, ORs from Panel B are combined by random effects meta-analysis to the European pooled estimate; In Panel C, out of 258,402 in FinnGen, 8,117 individuals were excluded, comprising 3,157 born abroad, 4,304 born in regions ceded to Soviet, 182 born in Åland Islands, and 474 with missing data. Detailed information of the Finnish regions in Panel C provided in supplementary methods. P-value for heterogeneity was calculated based on Cochran's heterogeneity statistic.

Table S3. Comparison of polygenic risk scores (PRS) in UK Biobank. Related to Figure 2, the table shows a comparison of PRSs developed with different methodologies. The decreases in effect sizes were calculated from regression estimates (log odds). The number of cases and controls in each category is listed in Table 1.

	OR	95% CI	Decrease in effect size compared to European ancestry	Decrease in effect size compared to PRS-CS in European ancestry	Decrease in effect size compared to PRS-CS in South Asian ancestry	Decrease in effect size compared to PRS-CS in African / Caribbean ancestry
Coronary artery disease						
Limited-variant PRS						
European	1.41	1.39-1.43	Ref	64 %		
South Asian	1.34	1.23-1.46	85 %		61 %	
African / Caribbean	1.18	0.96-1.46	49 %			63 %
LDpred PRS						
European	1.64	1.61-1.67	Ref	93 %		
South Asian	1.41	1.30-1.53	69 %		71 %	
African / Caribbean	1.32	1.13-1.54	56 %			104 %
PRS-CS PRS						
European	1.70	1.68-1.73	Ref	Ref		
South Asian	1.61	1.48-1.75	90 %		Ref	
African / Caribbean	1.30	1.12-1.52	56 %			Ref
Type 2 diabetes						
Limited-variant PRS						
European	1.69	1.66-1.72	Ref	92 %		
South Asian	1.61	1.50-1.74	91 %		98 %	
African / Caribbean	1.35	1.22-1.49	57 %			89 %
LDpred PRS						
European	1.78	1.75-1.81	Ref	101 %		
South Asian	1.66	1.55-1.79	88 %		105 %	
African / Caribbean	1.46	1.32-1.62	65 %			113 %
PRS-CS PRS						
European	1.77	1.74-1.80	Ref	Ref		
South Asian	1.63	1.51-1.75	85 %		Ref	
African / Caribbean	1.40	1.25-1.55	58 %			Ref
Breast cancer						
Limited-variant PRS						
European	1.64	1.61-1.67	Ref	86 %		
South Asian	1.36	1.14-1.62	62 %		65 %	
African / Caribbean	1.34	1.13-1.60	60 %			70 %
LDpred PRS						
European	1.50	1.47-1.53	Ref	71 %		
South Asian	1.47	1.23-1.75	95 %		81 %	
African / Caribbean	1.12	0.93-1.35	28 %			27 %
PRS-CS PRS						
European	1.77	1.74-1.81	Ref	Ref		
South Asian	1.61	1.35-1.92	83 %		Ref	
African / Caribbean	1.53	1.27-1.84	74 %			Ref
Prostate cancer						
Limited-variant PRS						
European	2.20	2.14-2.25	Ref	104 %		
South Asian	2.06	1.60-2.64	92 %		77 %	
African / Caribbean	1.72	1.46-2.02	69 %			151 %
LDpred PRS						
European	1.91	1.86-1.96	Ref	85 %		
South Asian	2.21	1.73-2.81	123 %		85 %	
African / Caribbean	1.35	1.14-1.61	47 %			84 %
PRS-CS PRS						
European	2.14	2.09-2.19	Ref	Ref		
South Asian	2.54	1.98-3.26	123 %		Ref	
African / Caribbean	1.43	1.21-1.69	47 %			Ref

Table S4. Information on genome-wide association study (GWAS) summary statistics. Information on GWAS used for constructing the polygenic risk scores in Figure 1.

Disease	GWAS	Ethnicity	N Cases / N Controls	Proportion of test datasets overlapping with GWAS
Coronary artery disease	Nikpay et al. https://doi.org/10.1038/ng.3396	European 77%, 13% South Asian, 6% East Asian, 4% other	60,801 / 123,504	5.9% of Estonian Biobank, 2.0% of FinnGen
Type 2 diabetes	Scott et al. https://doi.org/10.2337/db16-1253	European	26,676 / 132,532	7.5% of Estonian Biobank
Breast cancer	Michailidou et al. https://doi.org/10.1038/nature24284	European 89%, East Asian 11%	137,045 / 119,078	No overlap detected
Prostate cancer	Schumacher et al. https://doi.org/10.1038/s41588-018-0142-8	European	46,939 / 27,910	No overlap detected

Figure S1. Impact of LDpred parameter choice. Effect sizes across ancestries in UK Biobank with the different default fractions of causal variants with LDpred. Odds ratios (OR) with 95% confidence intervals (CI) are shown for 1-SD increase in the polygenic risk scores. The fraction of causal variants used in the main analyses in Figure 1 are bolded. The number of cases and controls in each category is listed in Table 1.

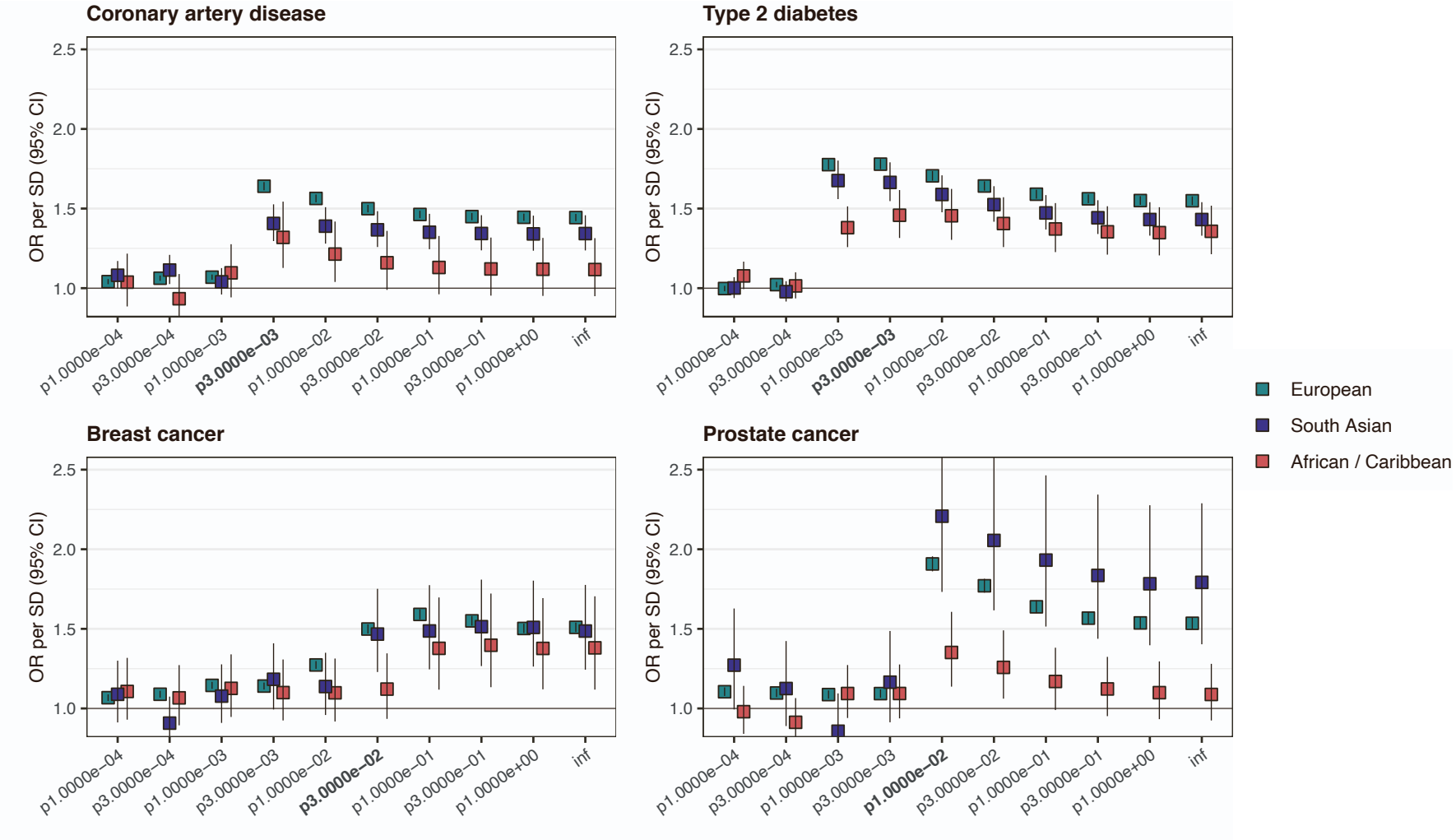
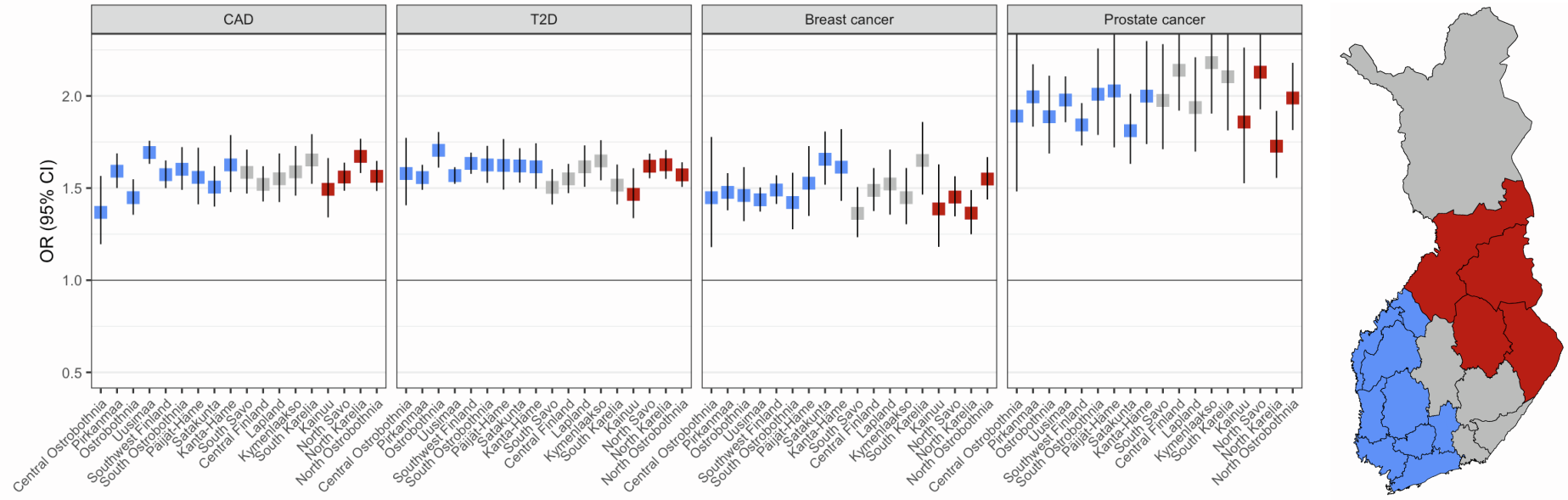


Figure S2. Detailed effect size comparison across early- and late-settlement regions in Finland. The figure shows detailed results by region within the settlement regions shown in Figure 1 panel C, using the same PRSs as in Figure 1. The early-settlement region is shown in blue, the late-settlement region in red, and the borderline region in gray.



OR = odds ratio, CAD = coronary artery disease, T2D = type 2 diabetes. Regions are based on data on birthplace. Out out of 258,402 individuals in FinnGen, 8,117 individuals excluded, including 3,157 born abroad, 4,304 born in regions ceded to Soviet, 182 born in Åland Islands (not shown in the map due to the exclusion; excluded due to low sample size), and 474 with missing data.