Published on *NHLBI Trans-Omics for Precision Medicine* (https://topmed.nhlbi.nih.gov)

DCC-harmonized phenotypes

Updated 10/28/2021

This page contains information on the DCC phenotype harmonization strategy and available DCC-harmonized phenotypes as of November 5, 2019 at 11:20:28 AM PST.

Contents

Back to top [17]

Available datasets

The following is a list of available datasets. Clicking on the dataset name will take you to more information about which phenotypes are included and the number of participants with non-missing information by study.

Dataset name	version	Date uploaded
Atherosclerosis events incident [4]	1	2019-10-31
Atherosclerosis events prior [5]	1	2019-10-31

Dataset name	version	Date uploaded
Demographic [6]	4	2019-10-29
Baseline Common Covariates [7]	3	2019-10-04
Sleep [8]	1	2019-10-04
Inflammation [9]	1	2019-04-19
Lipids [10]	3	2018-12-13
VTE [11]	1	2018-11-20
Blood Cell Count [12]	3	2018-10-12
Blood Pressure [13]	1	2018-08-27
Atherosclerosis [14]	1	2018-06-01

These datasets are split by study and uploaded to each TOPMed study's exchange area. They can be found under the "Provisional Files" tab and within the Phenotype/DCC/official folder. An example for one study is shown below:

```
topmed-dcc
  exchange
    phs000956_TOPMed_WGS_Amish
      Phenotype
        DCC
          official
            topmed_dcc_baseline_common_covariates_v1_phs000956.tar
            topmed_dcc_demographic_v1_phs000956.tar
```

The study-specific data files downloaded from the exchange areas can then be combined for cross-study analysis. Phenotypes for studies without a TOPMed exchange area will be uploaded once the exchange area is created at dbGaP.

The remainder of the information on this page is about phenotypes that have been harmonized using the DCC's official system. The DCC has also occasionally assisted with phenotype harmonization for phenotypes that are not tracked in our system. Other files containing phenotypes not shown on this page, or for the same phenotypes for studies not yet released on dbGaP, can be found in the `Phenotype/DCC/unofficial` folder.

Back to top [17]

Authorship guidelines

If you have used phenotypes that the DCC has harmonized in your analysis, please see authorship guidelines for DCC-harmonized phenotypes [18] for information about including DCC authors from the phenotype harmonization team.

Back to top [17]

Available phenotypes by dataset

For each phenotype, an associated **age at measurement** variable is also provided. For example, “weight_baseline_1” is body weight at the baseline exam and “age_at_weight_baseline_1” is the age of the participant at which that weight measurement was made. These age variables are not shown in the available phenotypes below but are a part of the datasets. The exception is for demographic phenotypes (e.g., sex, race, etc.), which do not have an associated age; they were derived primarily from baseline information, although later exams were used in some cases.

Back to top [17]

Atherosclerosis events incident

Phenotype	description
angina_incident_1	An indicator of whether a subject had an angina event (that was verified by adjudication or by medical professionals) during the follow-up period.
cabg_incident_1	An indicator of whether a subject had a coronary artery bypass graft (CABG) procedure (that was verified by adjudication or by medical professionals) during the follow-up period.
cad_followup_start_age_1	Age of subject at the start of the follow-up period during which atherosclerosis events were reviewed and adjudicated.
chd_death_definite_1	An indicator of whether the cause of death was determined by medical professionals or technicians to be “definite” coronary heart disease for subjects who died during the follow-up period.
chd_death_probable_1	An indicator of whether the cause of death was determined by medical professionals or technicians to be “probable” or “definite” coronary heart disease for subjects who died during the follow-up period.
coronary_angioplasty_incident_1	An indicator of whether a subject had a coronary angioplasty procedure (that was verified by adjudication or by medical professionals) during the follow-up period.
mi_incident_1	An indicator of whether a subject had a myocardial infarction (MI) event (that was verified by adjudication or by medical professionals) during the follow-up period.
pad_incident_1	An indicator of whether a subject had peripheral arterial disease (that was verified by adjudication or by medical professionals) during the follow-up period.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	FHS	WHI	Total
angina_incident_1	15,154	142,539	157,693
cabg_incident_1	11,814	142,539	154,353
cad_followup_start_age_1	15,154	143,213	158,367
chd_death_definite_1	15,154	142,539	157,693
chd_death_probable_1	15,154	142,539	157,693
coronary_angioplasty_incident_1	0	142,539	142,539
mi_incident_1	15,154	142,539	157,693
pad_incident_1	15,154	142,539	157,693

Return to top [1]

Back to top [17]

Atherosclerosis events prior

Phenotype	description
angina_prior_1	An indicator of whether a subject had an angina event prior to the baseline visit.
cabg_prior_1	An indicator of whether a subject had a coronary artery bypass graft (CABG) procedure prior to the start of the baseline visit.
coronary_angioplasty_prior_1	An indicator of whether a subject had a coronary angioplasty procedure prior to the start of the baseline visit.
coronary_revascularization_prior_1	An indicator of whether a subject had a coronary revascularization procedure prior to the start of the baseline visit. This includes angioplasty, CABG, and other coronary revascularization procedures.
mi_prior_1	An indicator of whether a subject had a myocardial infarction (MI) prior to the start of the baseline visit.
pad_prior_1	An indicator of whether a subject had peripheral arterial disease prior to the baseline visit.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	CHS	COPDGene	FHS	GENOA	JHS	MESA	WHI	Total
angina_prior_1	0	0	5,531	10,371	15,154	0	0	6,429	142,250	179,735
cabg_prior_1	0	14,817	5,493	10,370	11,814	0	3,501	6,429	141,106	193,530
coronary_angioplasty_prior_1	0	14,817	5,482	10,369	0	0	3,501	6,429	141,124	181,722
coronary_revascularization_prior_1	0	0	0	0	0	3,431	0	0	0	3,431
mi_prior_1	1,113	14,717	5,531	10,371	15,154	3,426	3,507	6,429	143,136	203,384
pad_prior_1	0	14,388	5,531	10,370	15,154	0	3,126	6,429	142,216	197,214

Return to top [1]

Back to top [17]

Demographic

Phenotype	description
annotated_sex_1	Subject sex, as recorded by the study.
geographic_site_1	Recruitment/field center, baseline clinic, or geographic region.
hispanic_or_latino_1	Indicator of reported Hispanic or Latino ethnicity.
hispanic_subgroup_1	classification of Hispanic/Latino background for Hispanic/Latino subjects where country or region of origin information is available
race_us_1	Reported race of participant according to the United States administrative definition of race.
subcohort_1	A distinct subgroup within a study, generally indicating subjects who share similar characteristics due to study design. Subjects may belong to only one subcohort.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	BAGS	CARDIA	CCAF	CFS	CHS	COPDGene	CRA	DHS	FHS	GALAII	GeneSTAR	GENOA	GOLDN	HCHS_SOL	HVH	JHS	Mayo_VTE	MESA	MGH_AF	Partners	SAGE	Samoan	VAFAR	VU_AF	WGHS	WHI	Total
annotated_sex_1	1,123	14,940	1,335	3,622	363	1,469	5,531	10,371	1,533	405	15,154	4,458	1,787	3,434	968	12,520	1,204	3,536	2,935	8,296	1,025	128	2,104	3,501	173	1,134	118	143,213	246,380
geographic_site_1	0	14,940	0	3,622	0	0	5,531	10,371	0	0	0	0	0	3,434	968	12,520	0	3,536	0	8,296	0	0	0	3,501	0	0	0	143,213	209,932
hispanic_or_latino_1	0	0	1,527	0	363	1,469	5,511	10,371	1,527	0	6,665	4,458	0	1,577	0	12,895	1,182	0	1,959	3,096	999	121	0	0	173	1,134	0	142,865	197,892
hispanic_subgroup_1	0	0	0	0	0	0	0	0	1,527	0	0	0	0	0	0	12,100	0	0	0	2,156	0	0	0	0	0	0	0	2,829	18,612
race_us_1	1,123	14,940	0	3,622	363	1,469	5,531	10,371	0	405	12,848	4,458	1,787	3,434	968	12,895	1,204	3,602	2,864	8,296	1,025	127	2,106	0	173	1,134	118	143,127	237,990
subcohort_1	1,123	15,678	1,527	3,622	363	1,473	5,531	10,371	1,533	405	15,154	4,458	1,787	3,462	968	12,895	1,204	3,602	2,935	8,296	1,025	128	2,106	3,501	173	1,134	118	143,213	247,785

Return to top [1]

Back to top [17]

Baseline Common Covariates

Phenotype	description
bmi_baseline_1	Body mass index calculated at baseline.
current_smoker_baseline_1	Indicates whether subject currently smokes cigarettes.
ever_smoker_baseline_1	Indicates whether subject ever regularly smoked cigarettes.
height_baseline_1	Body height at baseline.
weight_baseline_1	Body weight at baseline.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	BAGS	CARDIA	CCAF	CFS	CHS	COPDGene	CRA	DHS	FHS	GALAII	GeneSTAR	GENOA	GOLDN	HCHS_SOL	HVH	JHS	Mayo_VTE	MESA	MGH_AF	Partners	SAGE	Samoan	VAFAR	VU_AF	WGHS	WHI	Total
bmi_baseline_1	1,120	14,915	385	3,612	362	1,452	5,513	10,371	881	405	15,134	2,904	1,779	3,432	968	12,486	1,194	3,528	2,809	8,262	990	127	1,701	3,477	173	1,101	113	142,083	241,277
current_smoker_baseline_1	1,079	14,926	816	3,560	0	1,203	5,497	10,371	592	0	15,100	4,458	1,786	3,432	0	12,508	1,195	3,505	2,841	8,259	0	0	1,707	3,494	0	0	118	141,382	237,829
ever_smoker_baseline_1	0	14,930	861	3,578	0	1,203	5,519	10,371	592	0	14,905	0	1,783	3,433	0	12,514	1,195	3,530	2,841	8,238	0	0	0	3,482	0	0	118	142,060	231,153
height_baseline_1	1,122	14,921	385	3,614	362	1,454	5,521	10,371	881	405	15,141	0	1,780	3,432	0	12,504	1,194	3,530	2,812	8,262	990	127	0	3,479	173	1,101	116	142,368	236,045
weight_baseline_1	1,120	14,915	385	3,613	362	1,453	5,514	10,371	881	405	15,143	0	1,779	3,433	0	12,495	1,195	3,530	2,827	8,262	990	128	0	3,480	173	1,128	115	142,745	236,442

Return to top [1]

Back to top [17]

Sleep

Phenotype	description
sleep_duration_1	Usual amount of time slept per day.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	ARIC	CARDIA	CFS	CHS	FHS	HCHS_SOL	JHS	MESA	WHI	Total
sleep_duration_1	5,976	3,269	1,354	1,167	11,985	11,912	3,509	5,432	142,504	187,108

Return to top [1]

Back to top [17]

Inflammation

Phenotype	description
cd40_1	Cluster of differentiation 40 ligand (CD40) concentration in blood.
crp_1	C-reactive protein (CRP) concentration in blood.
eselectin_1	E-selectin concentration in blood.
icam1_1	Intercellular adhesion molecule 1 (ICAM1) concentration in blood.
il1_beta_1	Interleukin 1 beta (IL1b) concentration in blood.
il10_1	Interleukin 10 (IL10) concentration in blood.
il18_1	Interleukin 18 (IL18) concentration in blood.
il6_1	Interleukin 6 (IL6) concentration in blood.
isoprostane_8_epi_pgf2a_1	Isoprostane 8-epi-prostaglandin F2 alpha (8-epi-PGF2a) concentration in urine.
lppla2_act_1	Activity of lipoprotein-associated phospholipase A2 (LP-PLA2), also known as platelet-activating factor acetylhydrolase, measured in blood.

Phenotype	description
lppla2_mass_1	Mass of lipoprotein-associated phospholipase A2 (LP-PLA2), also known as platelet-activating factor acetylhydrolase, measured in blood.
mcp1_1	Monocyte chemoattractant protein-1 (MCP1), also known as C-C motif chemokine ligand 2, concentration in blood.
mmp9_1	Matrix metalloproteinase 9 (MMP9) concentration in blood.
mpo_1	Myeloperoxidase (MPO) concentration in blood.
opg_1	Osteoprotegerin (OPG) concentration in blood.
pselectin_1	P-selectin concentration in blood.
tnfa_1	Tumor necrosis factor alpha (TNFa) concentration in blood.
tnfa_r1_1	Tumor necrosis factor alpha receptor 1 (TNFa-R1) concentration in blood.
tnfr2_1	Tumor necrosis factor receptor 2 (TNFR2) concentration in blood.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	CARDIA	CFS	CHS	FHS	GENOA	HCHS_SOL	JHS	MESA	Total
cd40_1	0	0	0	0	0	3,274	0	0	0	964	4,238
crp_1	781	5,512	3,170	707	5,455	7,980	2,693	12,509	3,478	7,251	49,536
eselectin_1	0	0	0	0	0	0	0	0	0	1,215	1,215
icam1_1	0	0	2,532	706	2,132	7,691	0	0	0	2,815	15,876
il1_beta_1	0	0	0	708	0	0	0	0	0	0	708
il10_1	0	0	0	708	0	0	0	0	0	2,747	3,455
il18_1	0	0	0	0	0	3,159	0	0	0	0	3,159
il6_1	0	0	695	708	5,063	7,646	0	0	0	6,278	20,390
isoprostane_8_epi_pgf2a_1	0	0	0	0	0	7,523	0	0	0	0	7,523
lppla2_act_1	0	0	0	0	5,379	7,616	0	0	0	5,122	18,117
lppla2_mass_1	0	0	0	0	5,392	7,615	0	0	0	5,042	18,049
mcp1_1	0	0	0	0	0	7,557	0	0	0	0	7,557
mmp9_1	0	0	0	0	0	0	0	0	0	964	964
mpo_1	0	0	0	0	0	3,162	0	0	0	0	3,162
opg_1	0	0	0	0	0	7,648	0	0	0	0	7,648
pselectin_1	0	0	0	0	0	8,037	0	0	0	0	8,037
tnfa_1	0	0	0	708	0	2,516	0	0	0	1,851	5,075
tnfa_r1_1	0	0	0	0	0	0	0	0	0	2,802	2,802
tnfr2_1	0	0	0	0	0	7,962	0	0	0	0	7,962

Return to top [1]

Back to top [17]

Lipids

Phenotype	description
fasting_lipids_1	Indicates whether participant fasted for at least eight hours prior to blood draw to measure lipids phenotypes.
hdl_1	Blood mass concentration of high-density lipoprotein cholesterol
ldl_1	Blood mass concentration of low-density lipoprotein cholesterol

Phenotype	description
lipid_lowering_medication_1	Indicates whether participant was taking any lipid-lowering medication at blood draw to measure lipids phenotypes
total_cholesterol_1	Blood mass concentration of total cholesterol
triglycerides_1	Blood mass concentration of triglycerides

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	CARDIA	CFS	CHS	FHS	GENOA	HCHS_SOL	JHS	MESA	Samoan	Total
fasting_lipids_1	1,123	14,872	3,608	712	4,639	9,467	3,433	11,759	3,519	8,262	3,501	64,895
hdl_1	1,110	14,706	3,592	708	5,471	9,488	3,429	12,510	3,471	8,240	2,951	65,676
ldl_1	1,110	14,484	3,580	696	5,405	9,381	3,331	12,250	3,433	8,132	2,913	64,715
lipid_lowering_medication_1	1,123	14,827	0	712	5,526	9,573	3,433	12,280	3,234	8,254	0	58,962
total_cholesterol_1	1,110	14,705	3,592	709	5,479	9,507	3,429	12,511	3,471	8,243	2,951	65,707
triglycerides_1	1,110	14,707	3,591	709	5,479	9,505	3,429	12,511	3,471	8,243	2,951	65,706

Return to top [1]

Back to top [17]

VTE

Phenotype	description
vte_case_status_1	An indicator of whether a subject experienced a venous thromboembolism event (VTE) that was verified by adjudication or by medical professionals.
vte_followup_start_age_1	Age of subject at the start of the follow up period during which venous thromboembolism (VTE) events were reviewed and adjudicated.
vte_prior_history_1	An indicator of whether a subject had a venous thromboembolism (VTE) event prior to the start of the medical review process (including self-reported events).

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	ARIC	CHS	FHS	HVH	Mayo_VTE	WHI	Total
vte_case_status_1	14,562	5,199	8,620	987	2,925	30,799	63,092
vte_followup_start_age_1	14,562	5,531	10,021	0	0	31,578	61,692
vte_prior_history_1	14,562	5,291	10,028	990	0	31,574	62,445

Return to top [1]

Back to top [17]

Blood Cell Count

Phenotype	description
basophil_ncnc_bld_1	Count by volume, or number concentration (ncnc), of basophils in the blood (bld).
eosinophil_ncnc_bld_1	Count by volume, or number concentration (ncnc), of eosinophils in the blood (bld).
hematocrit_vfr_bld_1	Measurement of hematocrit, the fraction of volume (vfr) of blood (bld) that is composed of red blood cells.
hemoglobin_mcnc_bld_1	Measurement of mass per volume, or mass concentration (mcnc), of hemoglobin in the blood (bld).

Phenotype	description
lymphocyte_ncnc_bld_1	Count by volume, or number concentration (ncnc), of lymphocytes in the blood (bld).
mch_entmass_rbc_1	Measurement of the average mass (entmass) of hemoglobin per red blood cell(rbc), known as mean corpuscular hemoglobin (MCH).
mchc_mcnc_rbc_1	Measurement of the mass concentration (mcnc) of hemoglobin in a given volume of packed red blood cells (rbc), known as mean corpuscular hemoglobin concentration (MCHC).
mcv_entvol_rbc_1	Measurement of the average volume (entvol) of red blood cells (rbc), known as mean corpuscular volume (MCV).
monocyte_ncnc_bld_1	Count by volume, or number concentration (ncnc), of monocytes in the blood (bld).
neutrophil_ncnc_bld_1	Count by volume, or number concentration (ncnc), of neutrophils in the blood (bld).
platelet_ncnc_bld_1	Count by volume, or number concentration (ncnc), of platelets in the blood (bld).
pmv_entvol_bld_1	Measurement of the mean volume (entvol) of platelets in the blood (bld), known as mean platelet volume (MPV or PMV).
rbc_ncnc_bld_1	Count by volume, or number concentration (ncnc), of red blood cells in the blood (bld).
rdw_ratio_rbc_1	Measurement of the ratio of variation in width to the mean width of the red blood cell (rbc) volume distribution curve taken at +/- 1 CV, known as red cell distribution width (RDW).
wbc_ncnc_bld_1	Count by volume, or number concentration (ncnc), of white blood cells in the blood (bld).

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	CARDIA	CHS	FHS	HCHS_SOL	JHS	MESA	WHI	Total
basophil_ncnc_bld_1	787	10,911	2,672	0	5,348	11,698	2,832	2,338	0	36,586
eosinophil_ncnc_bld_1	787	10,956	3,287	0	5,348	11,718	2,992	2,338	0	37,426
hematocrit_vfr_bld_1	1,116	14,907	3,582	5,447	8,065	12,420	3,410	2,756	141,766	193,469
hemoglobin_mcnc_bld_1	1,116	14,907	3,582	5,447	8,010	12,420	3,410	2,756	141,719	193,367
lymphocyte_ncnc_bld_1	787	12,889	3,582	0	5,348	11,717	3,041	2,338	0	39,702
mch_entmass_rbc_1	1,116	8,710	3,582	0	8,010	12,420	3,055	2,756	0	39,649
mchc_mcnc_rbc_1	1,116	14,907	3,582	5,447	8,010	12,420	3,055	2,756	0	51,293
mcv_entvol_rbc_1	1,116	13,654	3,582	0	8,010	12,420	3,055	2,756	0	44,593
monocyte_ncnc_bld_1	787	12,861	3,555	0	5,348	11,721	3,037	2,338	0	39,647
neutrophil_ncnc_bld_1	787	11,472	3,582	0	5,348	11,717	3,041	2,338	0	38,285
platelet_ncnc_bld_1	1,109	14,815	3,581	5,417	5,254	12,413	3,413	2,750	141,425	190,177
pmv_entvol_bld_1	0	5,413	0	0	5,349	0	3,054	0	0	13,816
rbc_ncnc_bld_1	1,116	8,776	3,583	0	8,004	12,420	3,055	2,756	0	39,710
rdw_ratio_rbc_1	0	7,209	0	0	5,352	12,419	3,054	0	0	28,034
wbc_ncnc_bld_1	1,116	14,907	3,583	5,447	8,007	11,722	3,055	2,756	141,753	192,346

Return to top [1]

Back to top [17]

Blood Pressure

Phenotype	description
antihypertensive_meds_1	Indicator for use of antihypertensive medication at the time of blood pressure measurement.
bp_diastolic_1	Resting diastolic blood pressure from the upper arm in a clinical setting.
bp_systolic_1	Resting systolic blood pressure from the upper arm in a clinical setting.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	CARDIA	CFS	CHS	COPDGene	FHS	GENOA	GOLDN	HCHS_SOL	JHS	MESA	Samoan	WHI	Total
antihypertensive_meds_1	1,123	14,854	3,618	712	5,526	0	14,377	3,433	0	12,280	3,335	8,254	886	138,732	207,130
bp_diastolic_1	1,123	14,926	3,622	712	5,515	10,366	14,501	3,432	968	12,507	3,526	8,258	3,443	143,035	225,934
bp_systolic_1	1,123	14,926	3,622	712	5,515	10,366	14,501	3,432	968	12,507	3,526	8,258	3,443	143,035	225,934

Return to top [1]

Back to top [17]

Atherosclerosis

Phenotype	description
cac_score_1	Coronary artery calcification (CAC) score using Agatston scoring of CT scan(s) of coronary arteries
cac_volume_1	Coronary artery calcium volume using CT scan(s) of coronary arteries
carotid_plaque_1	Presence or absence of carotid plaque.
carotid_stenosis_1	Extent of narrowing of the carotid artery.
cimt_1	Common carotid intima-media thickness, calculated as the mean of two values: mean of multiple thickness estimates from the left far wall and from the right far wall.
cimt_2	Common carotid intima-media thickness, calculated as the mean of four values: maximum of multiple thickness estimates from the left far wall, left near wall, right far wall, and right near wall.

Number of non-missing measurements by study

Note that NOT all of these participants have been sequenced in TOPMed.

Phenotype	Amish	ARIC	CHS	FHS	GENOA	JHS	MESA	Total
cac_score_1	263	0	551	3,686	657	1,664	8,221	15,042
cac_volume_1	0	0	0	2,877	0	0	8,221	11,098
carotid_plaque_1	936	11,233	5,459	0	0	3,376	6,340	27,344
carotid_stenosis_1	0	0	5,473	3,287	0	0	6,338	15,098
cimt_1	1,008	14,151	5,502	3,279	0	3,358	8,122	35,420
cimt_2	0	10,173	5,502	3,283	0	3,364	8,151	30,473

Return to top [1]

Back to top [17]

Study abbreviations

Abbreviation	Name
Amish	NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish
ARIC	Atherosclerosis Risk in Communities (ARIC) Cohort
BAGS	Barbados Genetics of Asthma Study
CARDIA	CARDIA Cohort
CCAF	Cleveland Clinic Atrial Fibrillation Study
CFS	NHLBI Cleveland Family Study (CFS) Candidate Gene Association Resource (CARe)
CHS	Cardiovascular Health Study (CHS) Cohort

Abbreviation	Name
COPDGene	Genetic Epidemiology of COPD (COPDGene)
CRA	NHLBI TOPMed: The Genetic Epidemiology of Asthma in Costa Rica
DHS	Diabetes Heart Study (DHS)
FHS	Framingham Cohort
GALAII	Genes-Environments and Admixture in Latino Asthmatics (GALA II) Study
GeneSTAR	Genetic Study of Atherosclerosis Risk (GeneSTAR)
GENOA	Genetic Epidemiology Network of Arteriopathy (GENOA)
GenSALT	Genetic Epidemiology Network of Salt Sensitivity (GenSalt)
GOLDN	Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) Lipidomics Study
HCHS_SOL	Hispanic Community Health Study /Study of Latinos (HCHS/SOL)
HVH	Heart and Vascular Health Study (HVH)
HyperGEN	Hypertension Genetic Epidemiology Network Study
JHS	Jackson Heart Study (JHS) Cohort
Mayo_VTE	NHGRI Genome-Wide Association Study of Venous Thromboembolism (GWAS of VTE)
MESA	Multi-Ethnic Study of Atherosclerosis (MESA) Cohort
MGH_AF	Massachusetts General Hospital Atrial Fibrillation Study
Partners	Partners HealthCare Biobank
SAFS	San Antonio Family Heart Study (SAFHS)
SAGE	Study of African Americans, Asthma, Genes and Environment Study
Samoan	Genome-wide Association Study of Adiposity in Samoans
THRV	Taiwan Study of Hypertension using Rare Variants
VAFAR	The Vanderbilt AF Ablation Registry
VU_AF	The Vanderbilt Atrial Fibrillation Registry
WGHS	Women’s Genome Health Study
WHI	Women’s Health Initiative

Back to top [17]

DCC harmonization strategy

The TOPMed DCC is conducting phenotype harmonization to enable cross-study analyses. The main goals of this process are to provide harmonized phenotypes that are well-documented, reproducible, and as homogeneous across studies as possible. In harmonized datasets and documents, the DCC typically uses “phenotype” to refer to the general concept of a measurement and “variable” to refer to the specific data vector values of a given phenotype. The underlying database assigns a “trait_id” to uniquely identify a given variable, which appears in some of the documentation.

Collaboration between working groups, studies, and DCC analysts is essential for rigorous phenotype harmonization. Working group members provide domain expertise in their phenotype area, and liaisons from each study provide guidance about which study variables are appropriate to use. Harmonized phenotypes are constructed from “observed” study variables whenever possible, as opposed to using “derived” variables, unless otherwise specified. The DCC relies on the working groups to provide both the initial harmonization algorithm and the component variables to use, with the study liaisons assisting if necessary.

The DCC is acquiring phenotype data for all studies from dbGaP. In addition to being a stable repository, dbGaP has already curated and processed the data into a consistent format, which allows for automated processing, and all phenotypic variables have been assigned accession numbers, for tracking provenance. Using the data on dbGaP, the DCC provides harmonized phenotypes for all available study participants instead of just those being sequenced in TOPMed, which allows for non-TOPMed participants to be included in analyses after imputation or future sequencing.

Both the original study phenotypes and the final, harmonized phenotypes are stored in a relational database at the DCC. This setup allows DCC analysts to work with the study data in a consistent format instead of referencing a large number of files. It also provides a mechanism for tracking the provenance of each harmonized phenotype. In addition to storing metadata, the database also tracks the definition of the algorithms used to calculate each harmonized phenotype as well as the exact study phenotypes used in the calculation. Using this information, harmonized phenotypes can automatically be recomputed when updated study data are acquired from dbGaP. It also creates a lasting resource for the broader scientific community, as this detailed information will allow external investigators to augment the phenotypes harmonized in TOPMed with additional, non-TOPMed studies and will likely facilitate additional harmonization in future cross-study projects.

DCC analysts perform QC of component phenotypes, and consult with the relevant Working Group and study liaisons to resolve issues. This work is focused on finding inconsistencies and large batch effects in the study data. The harmonized traits are also QCed to detect possible errors in the harmonization process; this consists of checking whether most values are within the expected range and evaluating differences among studies and sample sets within study that may have been handled differently during harmonization. For each phenotype, the DCC analysts provide comments on the harmonization and QC process. The information in these comments should be considered before including a harmonized phenotype in any analysis. The analysts note outliers, but they are not removed from the harmonized data set for two reasons: (1) they may represent extreme effects of rare loss-of-function variants and (2) the definition of an outlier may vary according to the intended use, so users of the data should be able to make their own decisions about exclusionary criteria. Unless otherwise specified, the precision of phenotypic measurements is not harmonized, and they are not rounded to significant digits because the necessary information is generally not available.

For each phenotypic value for a given subject, an associated age at measurement is provided. These age values may have been winsorized in some studies and, if so, that winsorization carries through to the harmonized phenotype. For example, a study may give an age value as “>89” or “90+” instead of specific ages for subjects greater than 89 years of age; in this case, that text string was converted to a numeric value of 90. Otherwise, if the age measurements have not been winsorized by the study, we provide those age measurements with no winsorization. Analysts should also take care when working with multiple phenotypic variables at the same time, as variables across datasets or even within the same dataset are not necessarily measured at the same time for each subject.

The DCC also provides detailed documentation about which study phenotypes were used and the code that was run to produce a harmonized phenotype. The data values for each subject can be linked to the documentation using harmonization “unit” variables in each dataset. For each harmonized variable, a paired “unit_at_variable” is provided, whose value indicates where in the documentation to look to find the set of component variables and the algorithm used to harmonize those variables.

The DCC recommends that any phenotype variable be carefully inspected before use in analysis. We recommend caution in use of categorical variables as covariates in genetic association tests without first

checking for categories with low counts, which might cause model-fitting problems. Analysts are advised to view plots of the data distribution to identify potential outliers they might want to exclude from analysis.

Back to top [17]

Links

[1] https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#available-datasets [2]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#authorship-guidelines [3]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#available-phenotypes-by-dataset [4]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#atherosclerosis-events-incident [5]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#atherosclerosis-events-prior [6]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#demographic [7]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#baseline-common-covariates [8]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#sleep [9]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#inflammation [10]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#lipids [11]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#vte [12]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#blood-cell-count [13]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#blood-pressure [14]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#atherosclerosis [15]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#study-abbreviations [16]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#dcc-harmonization-strategy [17]
https://topmed.nhlbi.nih.gov/dcc-harmonized-phenotypes#top [18]
https://topmed.nhlbi.nih.gov/authorship-guidelines-dcc-harmonized-phenotypes