TOPMed freeze.4 QC

Stephanie Gogarten

August 7, 2017

Genotype data on the exchange area

Freeze	Samples	Variants
freeze.1c	2,643	112,275,224
freeze.2a	9,109	140,980,783
freeze.3a	16,558	185,970,832
freeze.3a.phased	18,258	200,750,986
freeze.4	18,526	219,154,455

For recent freezes, IRC has distributed BCF (Binary VCF) files.

passgt.minDP0	no missing data
passgt.minDP10	genotype calls with depth < 10 set to missing

QC done by IRC - Variant quality

Filtering has evolved over the different freezes. The current filtering scheme (freeze.4+), as described by Hyun Min Kang:

- Primary filter is based on support vector machine (SVM)
  - Known array-polymorphic variants as positive labels
  - Variants with many Mendelian inconsistencies as negative labels
  - Classifier trained using site-level features in the full VCF
  - HWE statistics are adjusted for population heterogeneity
- Additional hard filters applied
  - DISC : Variants with excessive Mendelian discordances
  - EXHET : Variants with excessive heterozygosity

QC done by IRC - Sample quality

Sequence data deemed high quality sufficient for joint variant discovery and genotyping when:

- estimated DNA sample contamination[1] below 3%
- fraction of the genome covered at least 10x 95% or above

[1] Goo Jun, et al. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array based genotype data. American Journal of Human Genetics, v.91, n.5, pp.839-848.

QC done by DCC

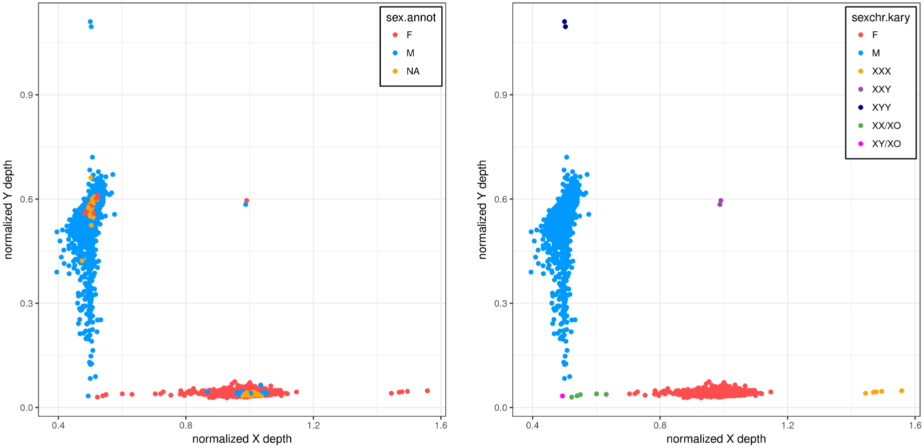
The goal of the DCC's QC process is to verify sample identity. Errors are corrected whenever possible. If the identity of a sample cannot be established, the sequencing data for that sample is dropped.

Steps:

- Genetic vs. annotated sex
- Comparison of genotypes (het/hom) with prior array data
- Comparison of observed kinship with pedigrees

QC is complete for freeze 4.

Genetic vs. annotated sex



Concordance with prior array data

study	n_unique	n_PASS	mean_PASS	n_FAIL	mean_FAIL
Amish	1032	1029	99.99	3	58.06
ARIC	80	232	99.14	1	56.30
CCAF*	345	345	99.52	0	
CFS*	961	1057	99.84	6	57.72
COPDGene**	1886	4566	99.79	10	54.60
CRA	783	782	99.96	1	72.15
EOCOPD	73	73	99.57	0	
FHS*	4056	11449	99.65	16	65.14
GALAII	967	967	99.44	0	
HVH*	74	73	99.74	1	54.83
JHS*	3215	3206	99.80	9	57.26
SAGE	496	496	99.41	0	
SAS	383	383	99.36	0	
WGHS	116	112	99.59	4	57.14

* = array data from dbGaP fingerprints

** = multiple sources of array data including dbGaP fingerprints

n_unique = number of unique TOPMed samples checked

n_PASS = number of sample pairs with concordance > 90%

mean_PASS = mean concordance (percent) of passing samples

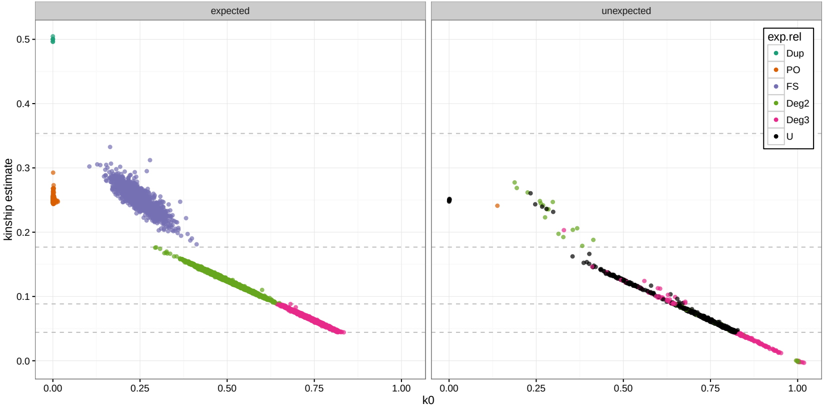
n_FAIL = number of sample pairs with concordance < 90%

mean_FAIL = mean concordance (percent) of failing samples

Correcting sample identity

- 44 of 14,467 samples checked had one or more array discordances. Of these, all but 10 were assigned to the correct subject ID, or their identity was confirmed by pedigree relationships. In several cases, a sample that did not match to its expected array counterpart instead matched to a different subject ID.
- Always use study subject ID (and not NWD ID) when assigning phenotypes to subjects, since sample-subject mapping may change after QC.

Comparison with pedigree (example study)



Many of the pedigree errors were resolved by the studies; some samples excluded.

Sample numbers

	n
Total in file	18,526
Control samples (HapMap and FHS trios)	31
Samples after QC	18,446
Pairs of identical twins	13
Unique samples	18,389
Unrelated (less than degree 3)	11,939

Files on exchange area

exchange/Combined_Study_Data/Genotypes/freeze.4/sample_sex/

- freeze4_sex.txt : Genetic sex of samples including sex chromosome karyotype

exchange/Combined_Study_Data/Genotypes/freeze.4/relatedness/

- freeze4_round2_pcrelate.gds : GDS file with kinship estimates and IBD sharing probability (k0, k1, k2) from PC-Relate
- freeze4_round2_pcrelate_kinship.txt.gz : Kinship estimates only
- freeze4_round2_pcair.RData : Principal components, eigenvalues, and variance proportions from PC-AiR
- freeze4_round2_pcair_pcs.txt : Principal components only
- freeze4_duplicates.txt : Duplicate samples (by NWD_ID) including monozygotic twins
- freeze4_duplicate_subjects.txt : Duplicate subjects (by study_subject_id) including monozygotic twins

Sample annotation on exchange area

exchange/Combined_Study_Data/Genotypes/freeze.4/sample_sets_2017-06-13/

- freeze4_samples_postQC_2017-06-13.txt : annotation of samples passing QC
- Consent group coming soon!

variable	description
sample.id	NWD ID
CENTER	sequencing center
topmed_project	TOPMed project name
PI	principal investigator for TOPMed project
phs	dbGaP phs number
study	study short name (1:1 with phs)
study_subject_id	subject ID to match with phenotypes (unique within study)
sex	genetic sex (inferred from X and Y chromosome depth)
sexchr.kary	inferred sex chromosome karyotype (e.g., XXX, XXY)
MZtwinID	monozygotic twin indicator
TRIO.dups	indicator for Framingham trio sequenced at all centers as controls
keep	samples eligible for analysis (no controls or QC failures)
unique	unique samples across all of TOPMed (excludes duplicates and twins)
unrel.deg3	samples unrelated at degree 3 level (less than first cousins)