**Overall study structure**

500k px has subcomponents:

- phenotyping changed over course of study (eg personality only available for a subset)
- - QC datafile contains batch variable for every individual- see if it contains "BiLEVE"
- differences between online/in person data
- two genotypings
- - 50k on one of the chips where half heavy smokers
- - two affy arrays but there are sig difs in call rates for particular SNPs
- - phenotyping confounding with snp arrays and ascn for heavy smoking
- - smoking also confounded with batch
- Phenotype data available as .csv and .Rdata file generated by provided R script; possible for SAS as well

!! rdata file is large and will excede memory allocated to login nodes

- object is `bd`
- - each "project"/request has it's own file as IDs have been randomized; `f.eid` is randomized day linking phen/gene data within requests; can establish bijection between eids across projects via plink sample files (ie eid1 <-> pos <-> eid2)
- - f.50.0.**0 :** 0 is initial visit; 1: reax (-20k indiv); 2: imaging visit;
- - 50^ is var id
- - details on phenotype page on wiki
- - can get ukb_field.tsv from data showcase to id specific vars without loading entire data set
-

**Phenotypes available**

- psychiatric sx data (now available) -- need to submit additional application if interested in using (particularly suicide)
- wiki with list of fields out to email
- data on rc `/work/ibg/^ but some still in kellerlab still waiting on data availability
- for storage, important to use generic bgen files

**Data cleaning - need to ensure consistency across projects**

- genotype data
- - vcf files,
- - ld-pruned relatedness files
- - gargi will send out parameters (HWE, MAF cutoffs, etc) of cleaned files and location on directory (discussed previously by gargi and luke)
- - QC
- - raw data will remain available
- - one set of files that have a bare min of QC (e.g., for imputed data, info score >=.3, removing indels, individs whose self-rep vs genetic sex differs excluded, singleton doubleton excld, two phases of imputation with some error--should use HRC snps, so luke removed uk10k and 1kg only snps)
- - bed files / chrm done
- - saved into plink bin files -->> gzip vcf in progress but will take a long time; will likely die as wall time < compute time

- - luke will just post QCd bgen files instead
- - plink binaries lose uncertainty info present in bgen and vcf
- - gargi has ID'd ethnic subsets: see
- /work/IBG/gada5574/ethnicities
- - relatives identification only done for 350k indiv so far but ukb provides kinship matrices up to 3rd degree for 500k; gargi will post script for IDing unrelated (currently removes both indiv, but will be modified to include only one of each pair; only for 350k currently)
- - need a list of folks to exclude for genetically unrelated sample
- - might recompute PCs only for caucasian subset -
- - need a subset of ld pruned files for caucasian only, then calc PCs; gargi is going to take care of this; but luke will calc PCs
- - HRC SNPs ~36k
- - best practices: all derrived data created in scratch (blanca: rcscratch..; summit ... // can't read between)
- - use globus for LFT between scratches on blanca/summit
- - procedure: init create in scratch; if cp to work/ explain in wiki QC, purpose, etc
-