Module 17: Computational Pipeline for WGS Data

Relatedness Inference from Genetic Data

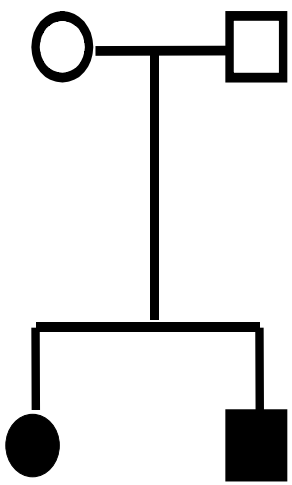
Timothy A. Thornton

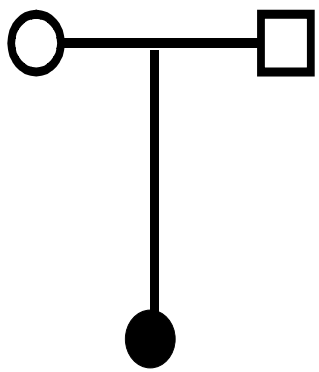
University of Washington, TOPMed Data Coordinating Center

Summer Institute in Statistical Genetics
July 2019

Incomplete Genealogy

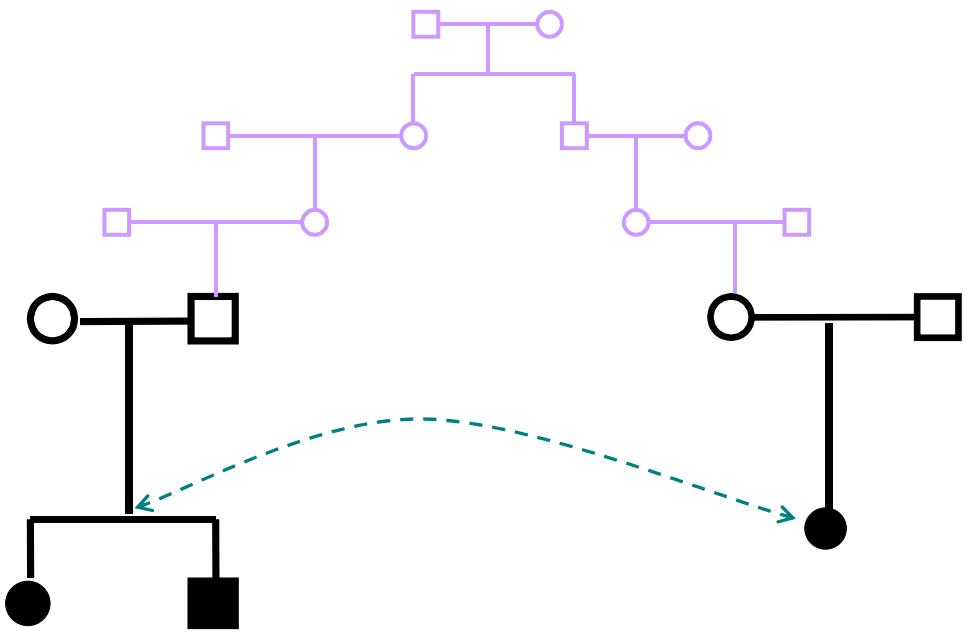
- Widely used statistical methods for the analysis of large-scale genetic data often assume independent samples or samples with known pedigree relationships; e.g., standard linkage analysis and association analysis methods





Incomplete Genealogy

- Misspecified and cryptic relationships can invalidate many of these methods if correlated genotypes among relatives are not properly accounted for in the analysis



Identifying Relative Pairs

- In principle, we could determine the relationship between two individuals by simply looking at the percentage of the genome that are **identical by descent (IBD)** for a pair where:
  - parent-offspring sharing: 50% of genome IBD
  - sibs: 50% of genome (on average) IBD
  - avuncular: 25% of genome (on average) IBD
- However, we do not directly observe IBD sharing.
- With SNP genotyping data or DNA sequencing data, we can estimate IBD sharing.

IBD Sharing Probabilities and Kinship coefficients

- IBD sharing probabilities and kinship coefficients are commonly used measures of relatedness for pairs of individuals
- For any pair of outbred individuals i and j, let δ_k^{ij} be the probability that i and j share k alleles IBD at a locus where k is 0, 1, or 2.
- Let ϕ_{ij} to be the kinship coefficient for i and j. The kinship coefficient is the probability that a random allele selected from individual i and a random allele from individual j are IBD.
  - Note that in outbred populations, $\phi = \frac{1}{2}\delta_2^{ij} + \frac{1}{4}\delta_1^{ij}$

Relatedness Measures for a Few Relationships

Relationship	ϕ_{ij}	δ_2^{ij}	δ_1^{ij}	δ_0^{ij}
Parent-Offspring	$\frac{1}{4}$	0	1	0
Full Siblings	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Half Siblings	$\frac{1}{8}$	0	$\frac{1}{2}$	$\frac{1}{2}$
Uncle-Nephew	$\frac{1}{8}$	0	$\frac{1}{2}$	$\frac{1}{2}$
First Cousins	$\frac{1}{16}$	0	$\frac{1}{4}$	$\frac{3}{4}$
First-Cousins Once Removed	$\frac{1}{32}$	0	$\frac{1}{8}$	$\frac{7}{8}$
Unrelated	0	0	0	1

Genome Screen Data to Identify Relative Pairs

- High-throughput genotyping data facilitated new opportunities for the detection of pedigree errors as well assessing the degree of relatedness among sampled individuals in genetic studies.
- A number of methods have been proposed for identifying relatives using genome-screen data for samples from a single homogenous population

Inference for Close Relatives from a Homogenous Population

- McPeek and Sun (2000) developed an approximate likelihood method (using HMM) to identify relative pairs for close relationships
- Purcell et al. (2007) proposed a method of moments estimator for IBD sharing probabilities using genome-screen data (implemented in the PLINK software)
- Choi, Wijsman, and Weir (2009) proposed using an EM algorithm to estimate the IBD probabilities and kinship coefficients
- Thornton and McPeek (2010) proposed a method of moments estimator for kinship coefficients by estimating genotypic correlations across the genome

Identification of More Distant Relatives from a Homogenous Population

- A variety of methods have been developed for identifying long shared IBD segments for inference on more distant relatives
- Stankovich et al. (2005) extended HMM method of McPeek and Sun for more distant relative pairs
- Huff et al. (2011)
- Browning and Browning (2013)
- and others...

Estimating Kinship Coefficients via Genotypic Correlations

- Thornton and McPeek (2010) proposed a method of moments approach for estimating kinship coefficients from SNP genotyping data in samples from homogenous populations based on genetic correlations.
- Consider two individuals i and j the sample. Assume genome screen data is available for i and j at M autosomal markers, indexed by $m = 1, 2, \ldots, M$.
- Let g_{im} be the genotype value at marker m for individual i, where g_{im} takes values 0, 1, or 2, corresponding to the number of reference alleles individual i has.
- Let p_m be the frequency of allelic type 1, where $0 < p_m < 1$.

Estimating Kinship Coefficients in Homogenous Populations

- It can be shown that the covariance of g_{im} and g_{jm} at marker m is $Cov(g_{im}, g_{jm}) = 4p_m(1-p_m)\phi_{ij}$, where ϕ_{ij} is the kinship coefficient for i and j.
- Rearranging terms, we see that $\phi_{ij} = \frac{Cov(g_{im},g_{jm})}{4p_m(1-p_m)}$
- This relationship holds for markers across the genome (but with the allele frequency distribution changing for each marker).
- It follows that the kinship coefficients can be estimated for pairs of individuals using genotype data from a genome-screen.

Estimating Kinship Coefficients in Homogenous Populations

- For any pair of individuals i and j, we can estimate ϕ_{ij} using method of moments where

$$\hat{\phi}_{ij} = \frac{1}{M} \sum_{m=1}^{M} \frac{(g_{im} - 2\hat{p}_m)(g_{jm} - 2\hat{p}_m)}{4\hat{p}_m(1 - \hat{p}_m)}$$

  where $\hat{p}_m$ is an allele frequency estimate for the reference allele at marker m
- Note that this estimator is essentially the same (up to a constant factor) as the previously discussed GRM estimator used for population structure inference with PCA! It is also the same estimator that is used to construct a GRM for association testing with linear/logistic mixed models that will be discussed later!

Relatedness Inference in Structured Populations

- The aforementioned algorithms for relatedness inference assume population homogeneity.
- This assumption is often untenable. Many genetic studies, (such as TOPMed) have samples from populations with different ancestries.
- relationship estimation methods that assume homogeneity can give extremely biased results in the presence of population structure.
- The degree of relatedness among related and unrelated sample individuals with similar ancestry can be systematically inflated

KING: Relatedness inference with Distinct Ancestral Subpopulations

- The KING estimator (Manichaikul A et al., 2010) discussed in the previous lecture was developed for estimating kinship coefficients for pairs of individuals from ancestrally distinct subpopulations
- KING-robust estimates kinship coefficients for a pair of individuals by using the shared genotype counts as a measure of the genetic distance between the pair.
- The method does not require allele frequency estimates at the marker: is based on allele sharing counts for individuals
- A limitation of the method is that it gives biased kinship estimates for individuals with different ancestry, including close relatives who are admixed.

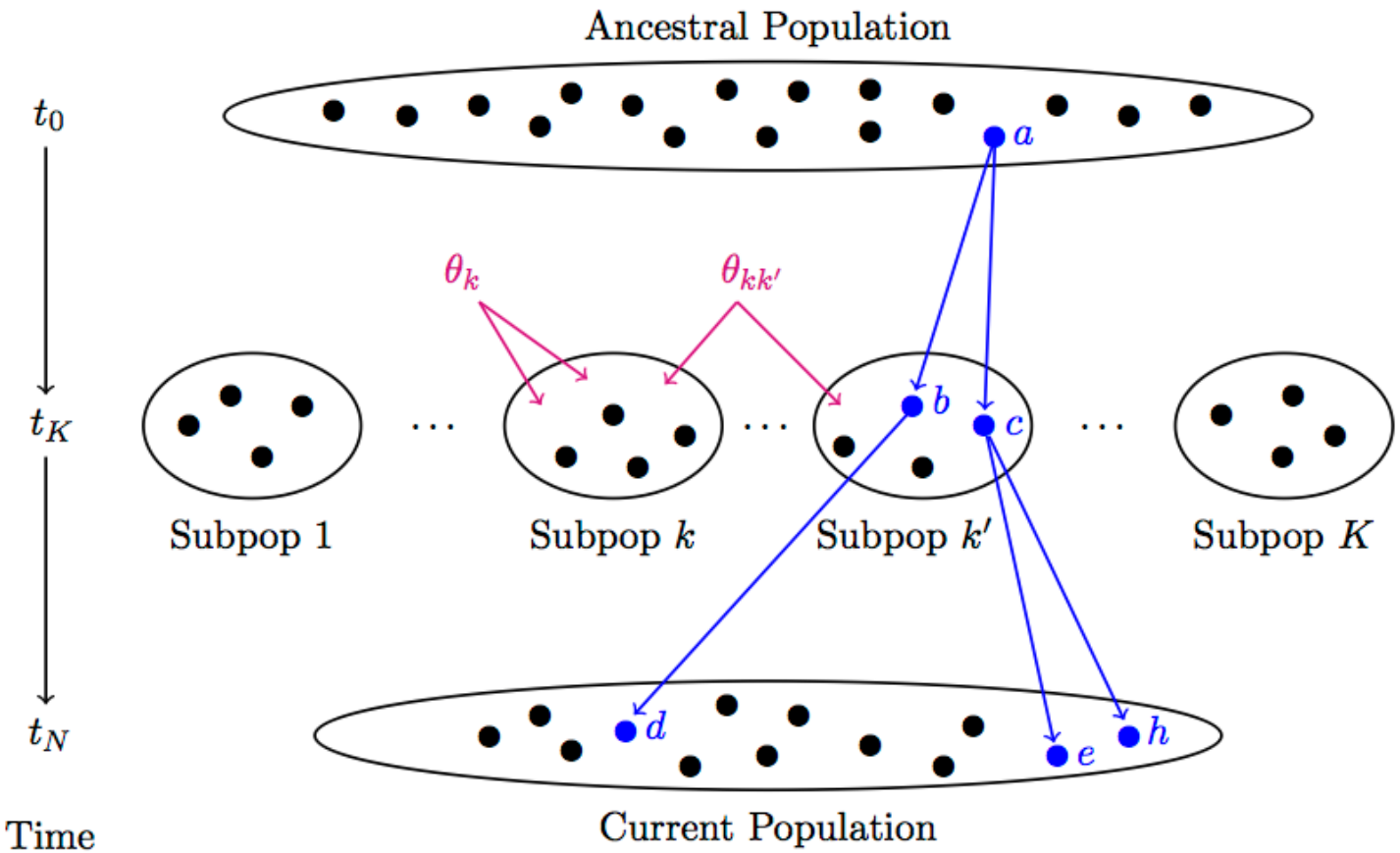
Relatedness Inference in Admixed Samples

- Genetic models used to identify related individuals from large scale genetic data often make simplifying assumptions about population structure – either random mating or simple structures.
- In reality, human populations do not mate at random nor are there simple endogamous subgroups.
- While large-scale genetic studies have primarily examined populations of European ancestry, more recent studies, including TOPMed, involve multi-ethnic cohorts with samples from admixed populations.

Recent versus Distant Genetic Relatedness

- Distinguishing familial relatedness from population structure using genotype data is difficult, as both manifest as genetic similarity through the sharing of alleles.
- It is important to note that relatedness and ancestry are a continuum
- Two alleles that are considered to be identical copies (e.g., IBD) of an ancestral allele is relative to some choice of previous reference point in time, with the implication being that more distant allele sharing prior to that time is not considered in the determination of IBD

Recent versus Distant Genetic Relatedness



Deconvolution of Recent and Distant Genetic Relatedness

- Conomos et al. [Am J Hum Genet, 2016]

ARTICLE

Model-free Estimation of Recent Genetic Relatedness

Matthew P. Conomos,[1,*] Alexander P. Reiner,[2,3] Bruce S. Weir,[1] and Timothy A. Thornton[1,*]

- Conomos et al. [Genet Epidemiol, 2015]

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness

Matthew P. Conomos,[1] Michael B. Miller,[2] and Timothy A. Thornton[1]*

PC-Relate: Relatedness Inference in Diverse Samples

- The PC-Relate method of Conomos et al. (2016) estimates IBD sharing probabilities and kinship coefficients in the presence of unknown population structure
- Let g_{im} and g_{jm} be the previously defined genotype value at marker m for individuals i and j respectively.
- For all individuals in the sample, PC-Relate uses a regression model to estimate the expected genotypic values for each marker m conditional on i's inferred ancestry using principal components from PC-AiR.
- A regression model is used to estimated the expected genotypic count, where the top principal components (PCs) from PC-AiR are included as predictors.

PC-Relate: Relatedness Inference in Diverse Samples

- We denote $\mu_{im} = \frac{1}{2}E[g_{im}|PCs]$ to be the **individual specific allele frequency** for individual i at marker m based on the PCs. Note that $2\mu_{im}$ is the expected value of g_{im} conditional on i's ancestry that is represented by the PCs from the regression model.
- The PC-Relate estimator of ϕ_{ij} for i and j is obtained via method of moments:

$$\hat{\phi}^A_{ij} = \frac{\sum_{m=1}^{M}(g_{im} - 2\hat{\mu}_{im})(g_{jm} - 2\hat{\mu}_{jm})}{\sum_{m=1}^{M} 4\sqrt{\hat{\mu}_{im}(1-\hat{\mu}_{im})\hat{\mu}_{jm}(1-\hat{\mu}_{jm})}}$$

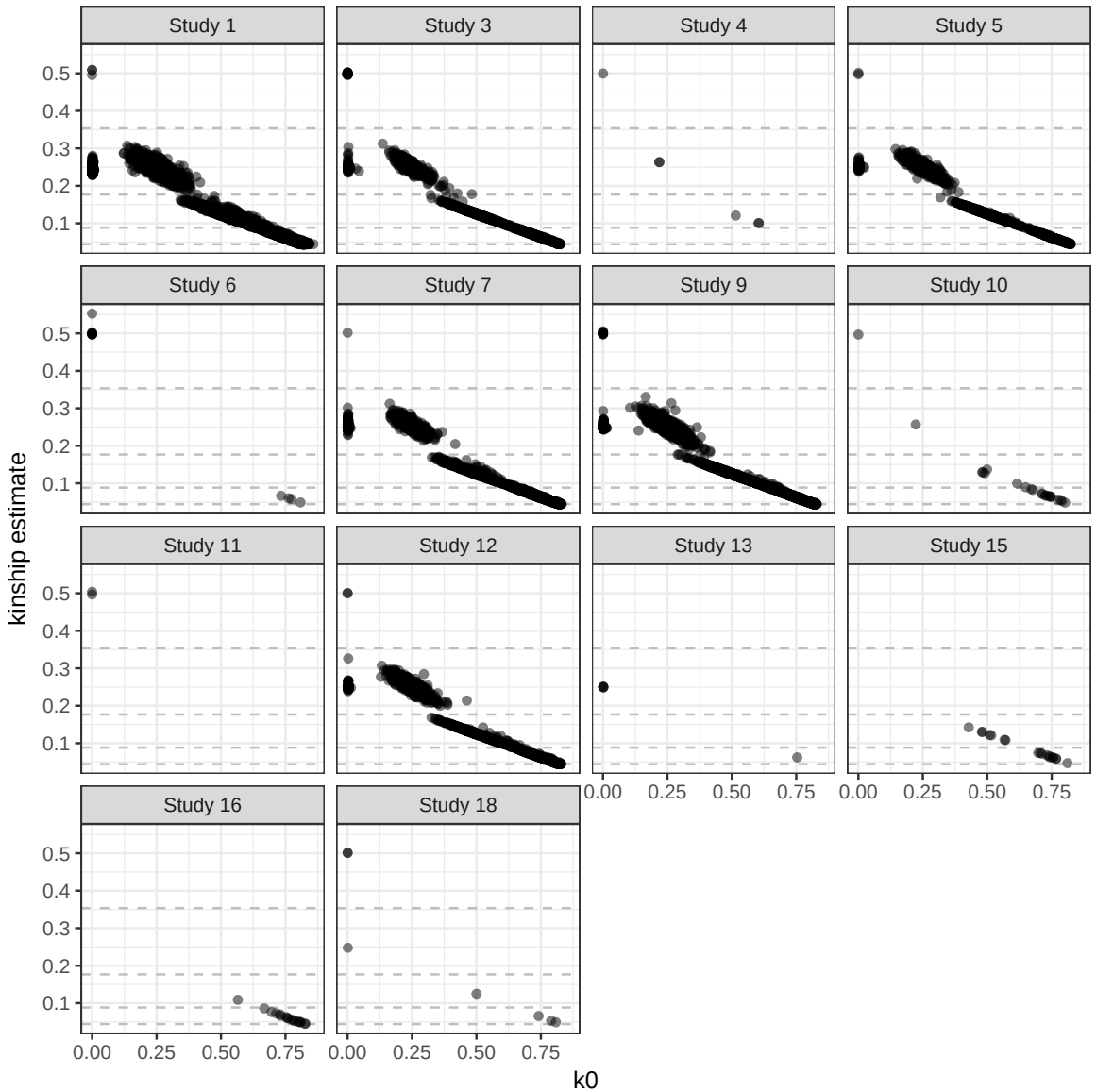
PC-Relate: Relatedness Inference in Diverse Populations

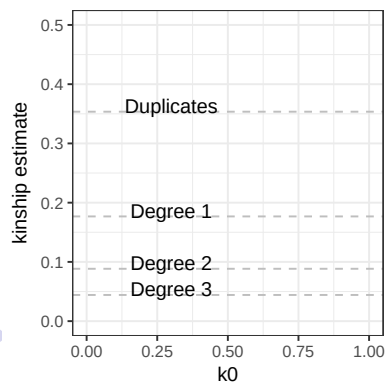
- PC-Relate also estimates IBD sharing probabilities
- In all calculations, baseline differences in genotypic values that are due to ancestry differences (i.e., the PCs) are regressed out (or adjusted for)
- PC-Relate kinship coefficients and IBD sharing probabilities are robust to population structure, admixture, and HWE departures
- See Conomos et al. (2016) for more details.

TOPMed Phase I: Relatedness Inference

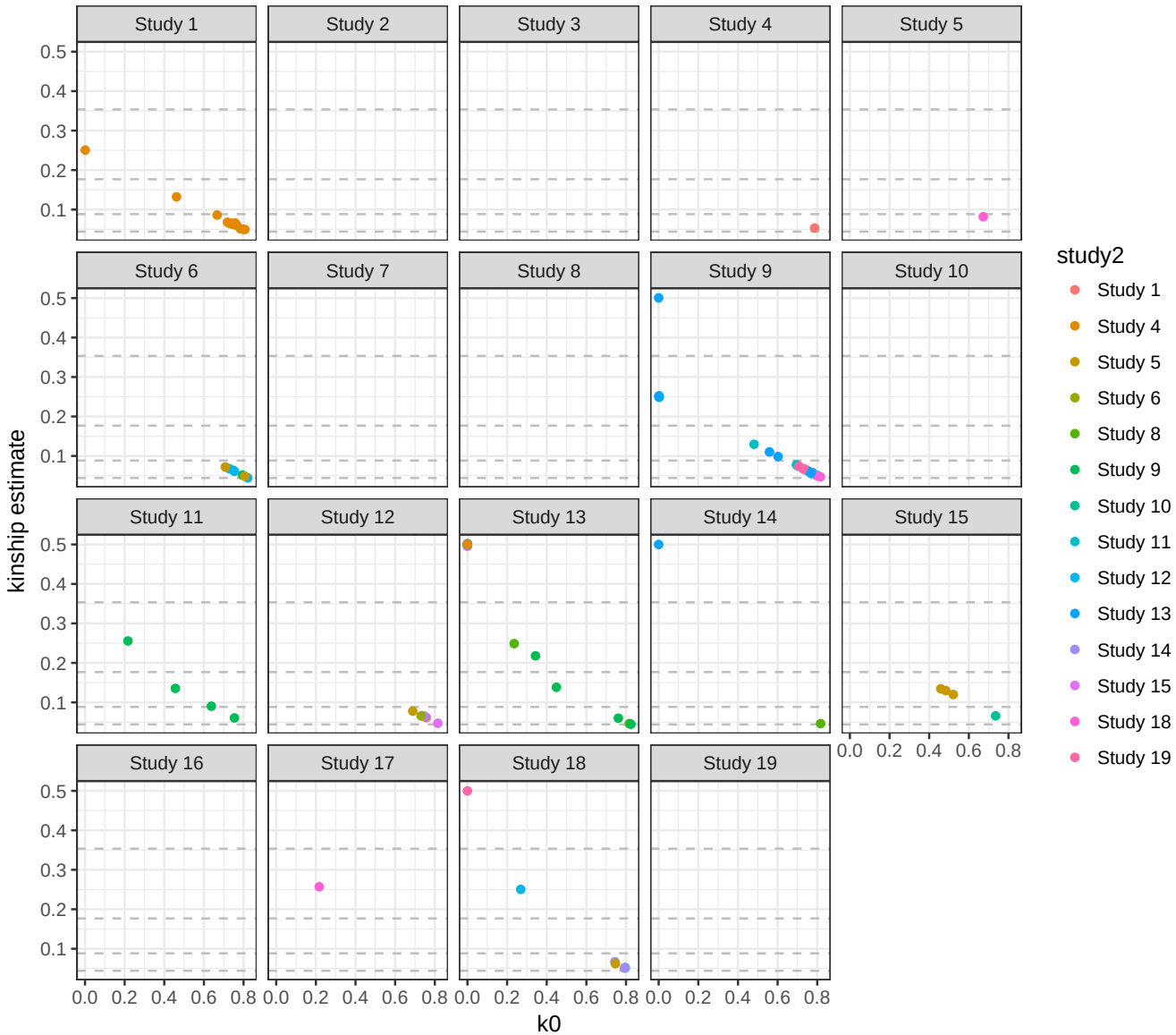
- TOPMed cohorts are multi-ethnic
- There is also extensive relatedness in TOPMed cohorts due to family-based-sampling and samples from founder populations (Amish).
- Also likely a number of cryptic relationships within and among the cohorts.
- The PC-AiR algorithm was first applied to TOPMed Phase I data for inference on population structure in the presence of relatedness
- PC-Relate was used to estimate relatedness in the TOPMed Phase I samples where the top PCs from PC-AiR were adjusted for in the analysis.

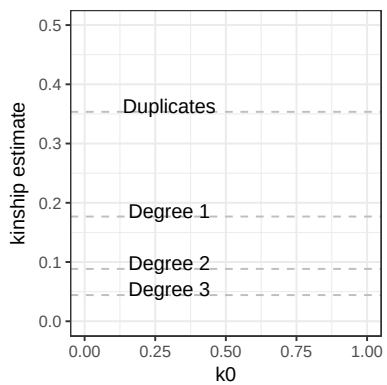
TOPMed Phase I: PC-Relate Kinship by Study





TOPMed Phase I: PC-Relate Kinship Across Studies





References

- Browning BL and SR Browning (2013). Improving the Accuracy and Efficiency of Identity by Descent Detection in Population Data. *Genetics* **194**: 459-471
- Choi Y, Wijsman EM, Weir BS (2009). Case-control association testing in the presence of unknown relationships. *Genet. Epi.* **33**, 668-678.
- Conomos MP, Miller M, Thornton T (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology* **39**, 276-93
- Conomos MP, Reiner AP, Weir BS, and Thornton TA (2016), Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics* **98**: 127-148.

References

- Huff CD, Witherspoon DJ, Simonson, TS, Xing J, et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*, **21**, 768-774.
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies.*Bioinformatics* **26**, 2867-2873.
- McPeek MS and Sun L (2000). Statistical Tests for Detection of Misspecified Relationships by Use of Genome-Screen Data, *Am. J. Hum. Genet.* **66**, 1076-1094.

References

- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559-575.
- Stankovich J, Bahlo M, Rubio JP, Wilkinson CR, Thomson R, Banks A, Ring M, Foote SJ, Speed TP (2005). Identifying nineteenth century genealogical links from genotypes. *Hum. Genet.* **117**, 188-199
- Thornton T, McPeek MS (2010). ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum. Genet.* **86**, 172-184.