Maximum-likelihood estimation of recent shared ancestry (ERSA)

Chad Huff, David Witherspoon, Tatum Simonson, et al.

Genome Research, published online February 8, 2011

Triinu Kõressaar

Seminar in Bioinformatics

TARTU 2011

Fields where is required to estimate shared ancestry

Case-control association studies and population-based genetic analyses

Disease mapping in families

Forensic identification of missing persons, victims of mass disasters, and suspects in criminal investigations

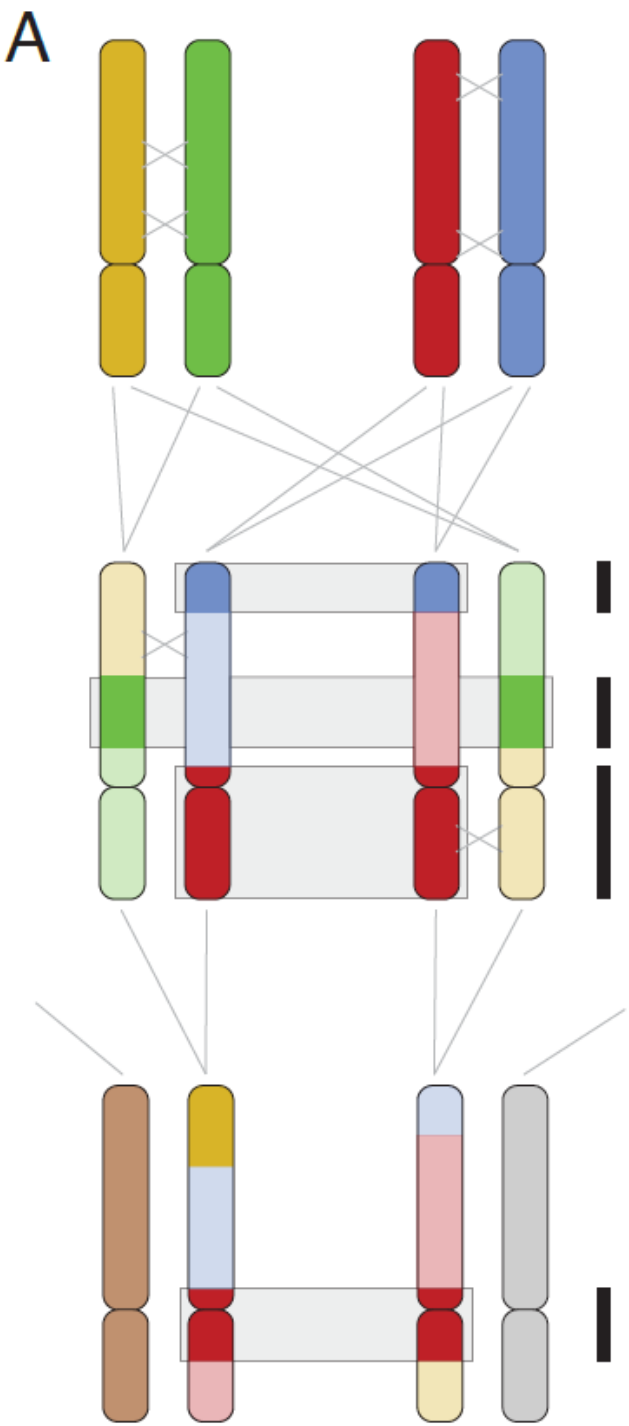
Conservation biology, quantitative genetics, and evolutionary biology (breeding programs, population dynamics, taxonomic issues, wildlife management - cannibalism, immune system and mate choise)

Current methods and their drawbacks

Based on genome-wide averages of the estimated number of alleles shared identity by descent (IBD) between two individuals

Cannot identify more than about third-degree relatives (e.g first cousins) distant relationships

(**IBD** – two or more alleles are identical by descent if they are identical copies of the same ancestral alleles)



Expected distributions of IBD chromosomal segments between pairs of individuals.

(A) The process underlying the pattern of IBD segments. Two homologous autosomal chromosomes are shown for two parents, each colored differently. Meiosis and recombination occurs and two sibling offspring inherit recombinant chromosomes (just one crossover per homologous pair for each meiosis event is depicted, marked by an 'X'). For some segments of the chromosome in question, the siblings share a stretch that was inherited from one of the four parental chromosomes. The three IBD segments are identifiable as regions that share the same color (boxed and marked at right by black bars). The siblings mate with unrelated individuals and the offspring each inherit an unrelated chromosome (tan or gray) and one that is a recombinant patchwork of the grandparental chromosomes. These first cousins share one segment IBD at this chromosome (red, boxed).

The ***degree*** (first, second, third cousin, etc.) indicates one less than the minimum number of generations between both cousins and the nearest common ancestor.

For example, a person with whom one shares a grandparent (but not a parent) is a first cousin; someone with whom one shares a great-grandparent (but not a grandparent or a parent) is a second cousin; and someone with whom one shares a great-great-grandparent (but not a great-grandparent or grandparent or parent) is a third cousin; and so on..

The ***removal*** (once removed, twice removed, etc.) indicates the number of generations, if any, separating the two cousins from each other.

The child of one's first cousin is one's "first cousin once removed" because the one generation separation represents one "removal". Oneself and the child are still considered first cousins, as one's grandparent (this child's great-grandparent), as the most recent common ancestor, represents one "degree“.

ERSA uses a likelihood ratio test

Null hypothesis - the two individuals are unrelated
Alternative hypothesis - the individuals share recent ancestry

The data are the number and lengths of autosomal genomic segments shared between two individuals, with segment length measured in centiMorgans (cM)

One centimorgan corresponds to about 1 million base pairs in humans on average

Two markers on a chromosome are one centimorgan apart if they have a 1% chance of being separated from each other by a crossing over in a single generation

ERSA accurately estimates the degree of relationship for up to eighth-degree relatives (*e.g. third* cousins once removed), and detects relationships as distant as twelfth-degree relatives (*e.g. fifth cousins once removed).*

Null hypothesis

Segments longer than a given threshold *t*

Set of segments shared between two individuals *s*
Number of elements in *s* is *n*

The number of segments shared and the length of each segment are independent

The likelihood of the null hypothesis

1. $$L_p(n,s \mid t) = N_p(n \mid t) \cdot S_p(s \mid t),$$

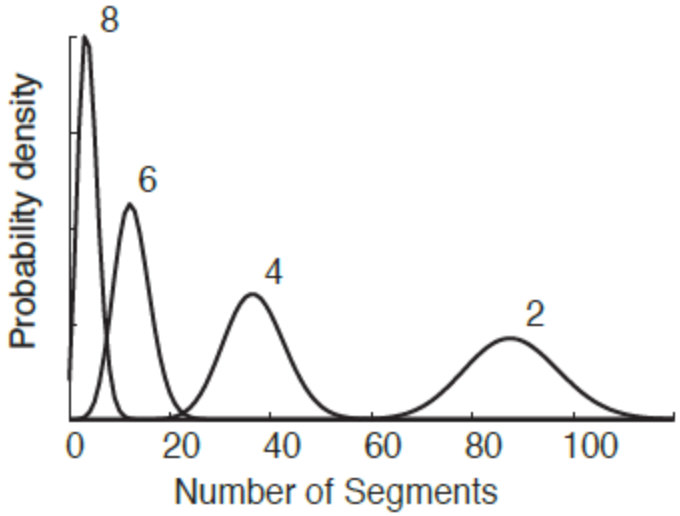
where

2. $$S_p(s \mid t) = \prod_{i \in s} F_p(i \mid t).$$

Recommended is setting *t* to the smallest value that can achieve a false-negative rate of 1% or lower

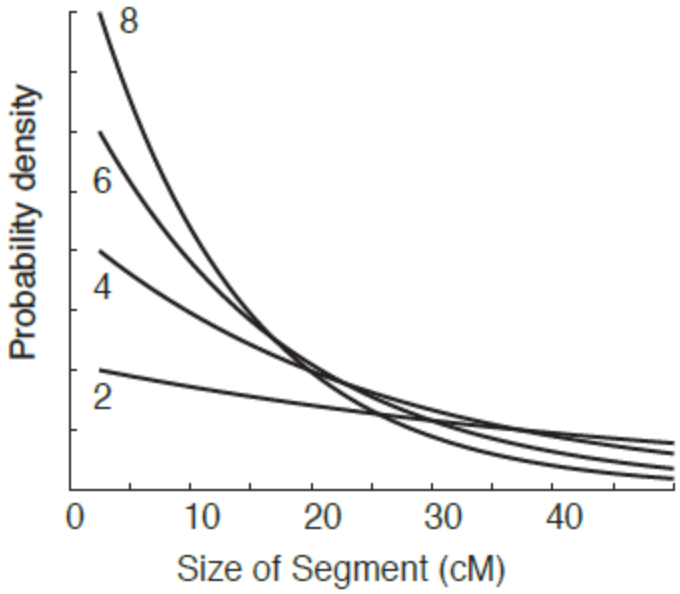
t = 2.5 cM based on GERMLINE's previously reported false negative rate of 1% for segments 2.5 cM and longer

B



(B) The number of segments that a pair of individuals shares IBD, across all chromosomes, is approximately Poisson distributed with a mean that depends on the degree of relationship *d between the individuals (d = 2, 4, 6, 8,* corresponding to siblings through third cousins).

C



(C) The lengths of the IBD segments are approximately exponentially distributed, with mean length depending on the relationship between individuals (theoretical distributions shown for *d = 2, 4, 6, 8).*

Alternative hypothesis

Pair of individuals share one or two recent ancestors.

The number of ancestors shared *a*
Combined number of generations separating the individuals from their ancestors *d*

Segments shared by two individuals come from two sources:
1. recent ancestry A
2. population background P

For a given value of *d*, the lengths of segments are independent

The likelihood of the alternative hypothesis of the recent ancestry:

$$L_R = L_A(n_A, s_A \mid d, a, t) L_P(n_P, s_P \mid t)$$

The expected number of shared segments (Thomas *et al. 1994*)

$$a(rd+c)/2^{d-1},$$

where c is the number of autosomes

r is the expected number of recombination events per haploid genome per generation

rd+c – the expected number of shared autosomal segments that could potentially be inherited from a common ancestor

$1/2^{d-1}$ is the probability that two individuals will inherit any particular autosomal segment from a common ancestor on that path

In humans r ~ 35.3 (McVean et al. 2004)

Data used

169 individuals
three well-defined pedigrees (24,30,115 ind per pedigree)
predominantly northern European ancestry
DNA from blood
Affymetrix 6.0 SNP arrays

868,155 autosomal SNP loci
18,610 were excluded from the final data set because they exhibited more than three Mendelian inheritance errors in the CEU trios or more than 10% missing data in either the CEU or pedigree individuals

GERMLINE and fastIBD in Beagle vers. 3.3- locations and extents of IBD segments for all pairs of individuals

Built upon recently developed algorithms

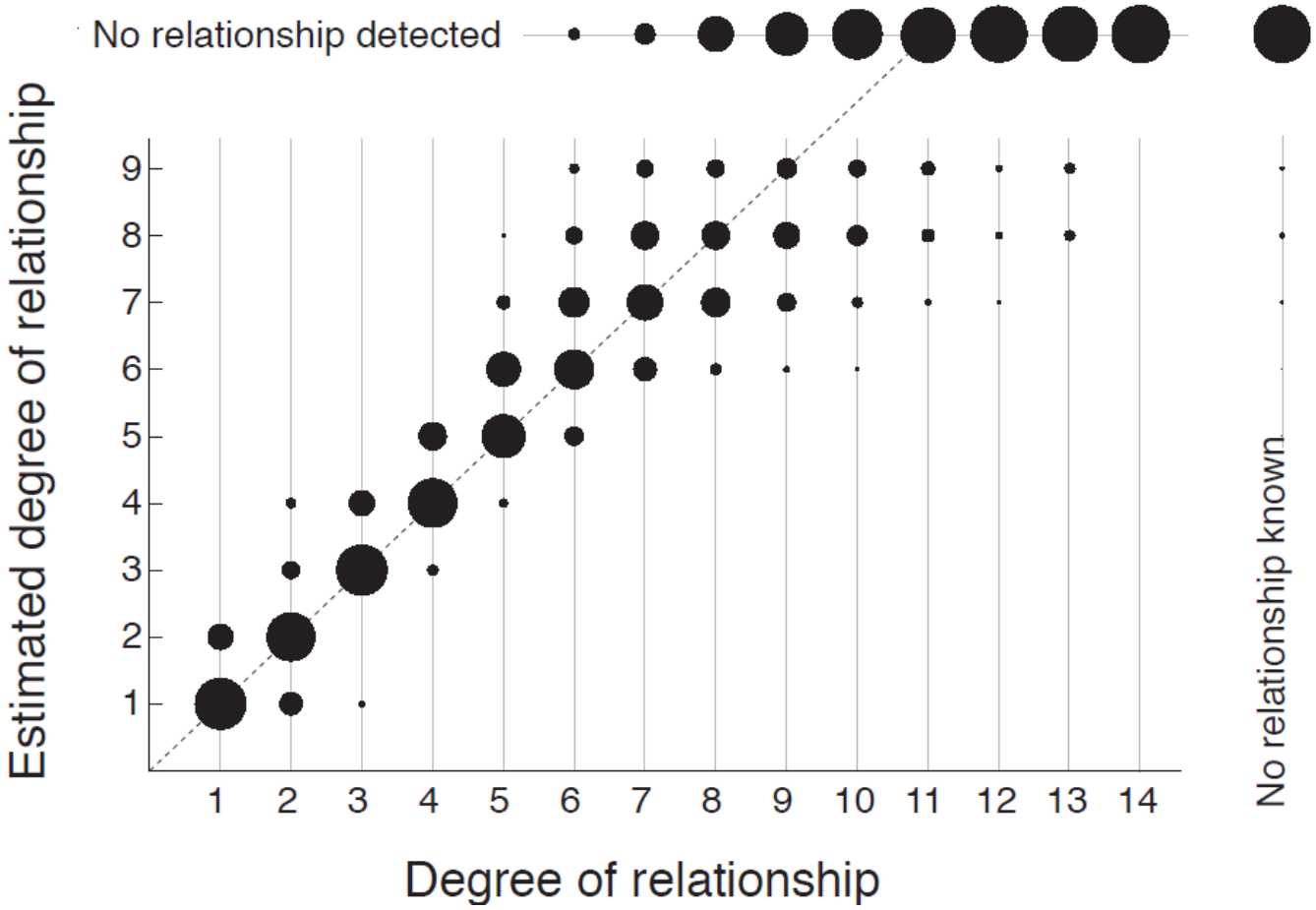
Use high density SNP data to detect the number, lengths, and locations of chromosomal segments identical by descent (IBD) between two individuals

GERMLINE (Gusev et al., 2009) – considers identity by haplotypes rather than genotypes (adv,disadv), searching for IBD is done with linear computing time (generally quadratic in the number of individuals)

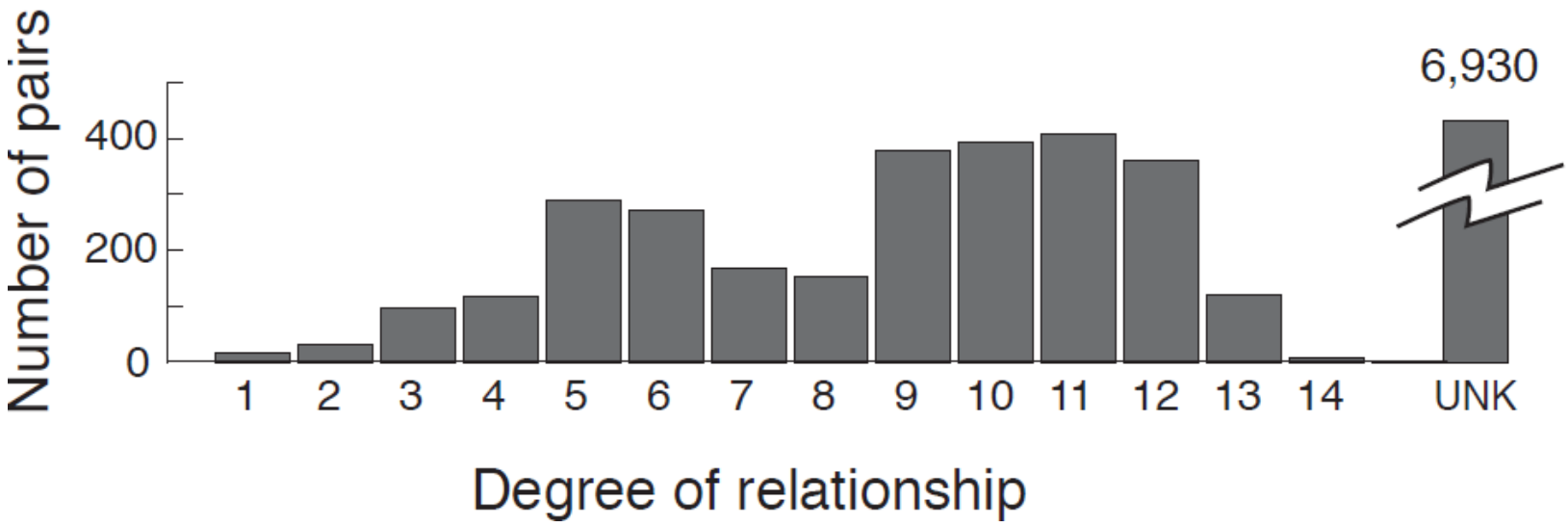
fastIBD (Brownings, 2011) – based on shared haplotype frequency rather than shared haplotype length (referring to GERMLINE). Computation time comparable to GERMLINE.

RELPAIR (Epstein et al., 2000), **GBIRP** (Stankovich et al., 2005)– model the states between haplotypes as a Markov process along a chromosome. Do not model the patterns on LD that exist between very closely spaced SNPs.

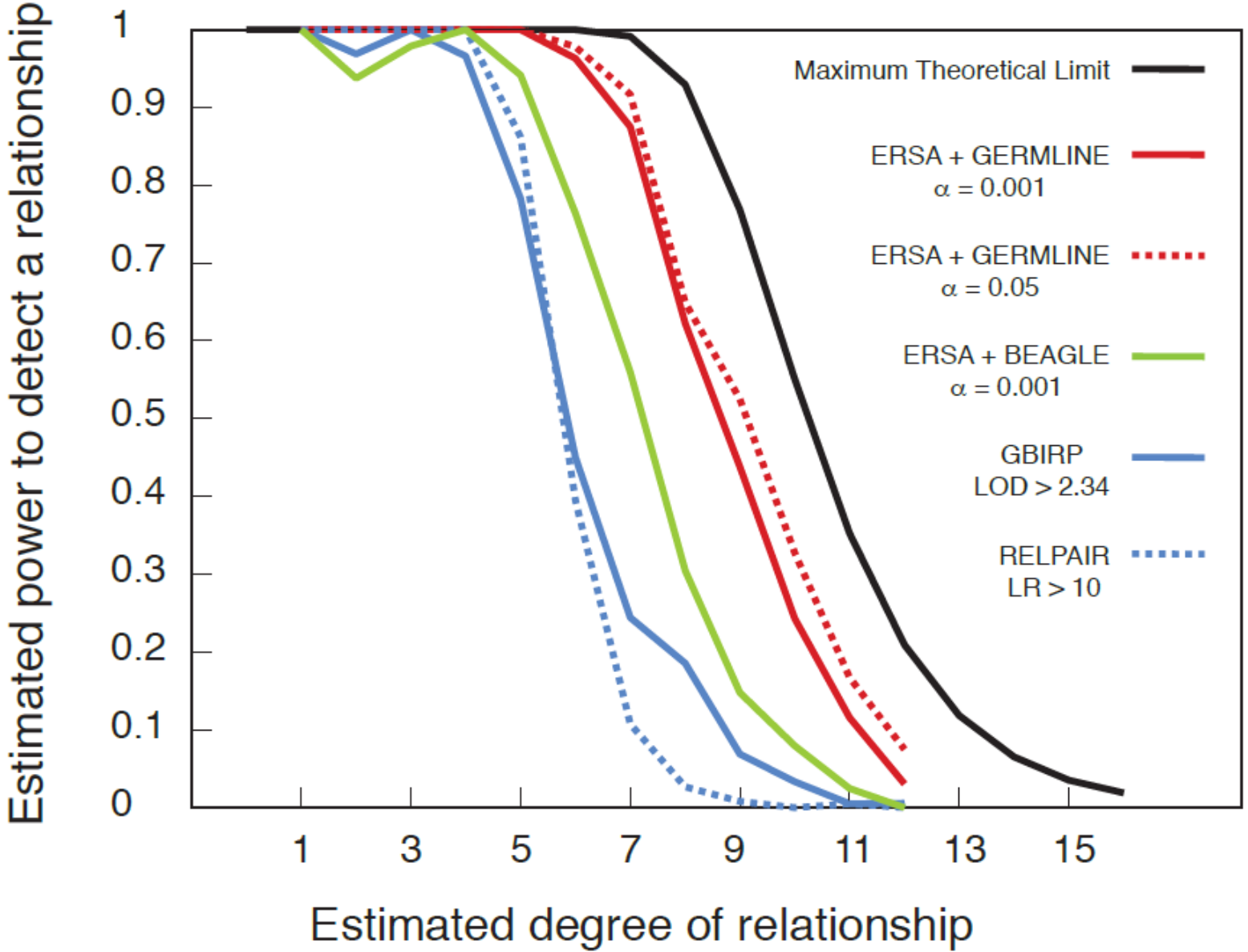
Estimated degree of relationship between pairs of individuals vs. Known degree of relationship. Pedigree information was used to identify **2,802** pairs of genotyped individuals that share **exactly two** common ancestors (a mated pair) and classify them according to the degree of their relationship (horizontal axis). The number of pairs in each category is indicated by the histogram below. Within each category, the areas of the filled circles indicate the proportion of those pairs with various estimated degrees of relationship between a pair (vertical axis; two ancestors, two degrees of freedom, $\alpha = 0.001$). The total area within a category is a constant across categories. Pairs with a known but undetected relationship are represented across the top. Pairs with no known relationship are represented on the right.



The number of pairs in each category is indicated by the histogram below



Power to detect recent common ancestry between pairs of individuals known to be related at varying degrees



False positive rate of detecting recent ancestry among HapMap JPT-CHB pairs

CHB – 45 Han Chinese in Beijing
JPT – 45 Japanese in Tokyo
HapMap phase 2 SNP genotype data

Nominal false positive rate	Observed false positive rate	Observed false positive counts
0.05	0.044	89/2,025
0.01	0.0094	19/2,025
0.001	0.00049	1/2,025

Conclusions

ERSA can be applied to number of problems

Can verify distant relationship without genotyping intervening family members

Computationally efficient

Achieves a statistical power very close to the theoretical maximum

Free for academic usage http://jorde-lab.genetics.utah.edu/ersa