Title: ESTIMATION OF RECENT SHARED ANCESTRY

Figure 1

Abstract: Methods and systems are described for the estimation of recent shared ancestry (ERSA) from the number and lengths of identical-by-descent (IBD) nucleotide segments derived from, e.g., high-density single-nucleotide polymorphism data or whole-genome sequence data. ERSA is accurate to within one degree of relationship for 97% of first-through fifth-degree relatives and 80% of sixth- and seventh-degree relatives. ERSA’s statistical power approaches the maximum theoretical limit imposed by the fact that distant relatives frequently share no DNA through a common ancestor. ERSA greatly expands the range of relationships that can be estimated from genetic data.
UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published: with international search report (Art. 21(3))
ESTIMATION OF RECENT SHARED ANCESTRY

[0001] This invention was made with government support under R01-CA040641, GM-59290, and K99HG005846 awarded by the National Institutes of Health; DK069513 awarded by the National Institute of Diabetes and Digestive and Kidney Diseases; and P01-CA073992 and NOI-PC-35141 awarded by the National Cancer Institute. The Government has certain rights to this invention.

Background

[0002] Knowledge about recent shared ancestry between individuals is fundamental to a wide variety of genetic studies. Detecting cryptic relatedness is a valuable technique for mapping disease-susceptibility loci and for identifying other at-risk individuals (Neklason et al. 2008; Thomas et al. 2008). For case-control association studies and population-based genetic analyses, related individuals should be identified and removed from samples that are intended to be random representatives of their populations (Pemberton et al. 2010; Simonson et al. 2010; Voight and Pritchard 2005; Xing et al. 2010). Using genetic data to correct pedigree errors increases the power of disease mapping in families (Cherny et al. 2001). Genetic identification of relatives has proven invaluable in forensic identification of missing persons, victims of mass disasters, and suspects in criminal investigations (Bieber et al. 2006; Biesecker et al. 2005; Zupanic Pajnic et al. 2010). Studies of conservation biology, quantitative genetics, and evolutionary biology are greatly illuminated when the recent shared ancestry between individuals being observed or sampled can be reconstructed, especially in agricultural and wild populations (DeWoody 2005; Slate et al. 2010).

Summary

[0003] Some embodiments of the disclosure include methods of estimating genetic relatedness between members of a first pair of conspecific organisms comprising receiving, by a processor, a value indicating a number of nonoverlapping polynucleotide segments longer than a threshold length (t) that are identical, by at least about 90 percent sequence identity, between members of the first pair; receiving, by a processor, values indicating lengths of the identical segments; comparing the number of the first pair's identical segments to a number of
nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a second pair of organisms, the members of the second pair having an established degree of genetic relatedness to each other; comparing the lengths of the first pair's identical segments to lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a third pair of organisms, the members of the third pair having an established degree of genetic relatedness to each other; based on the number comparison and the length comparison, estimating, by a processor, a degree of genetic relatedness between the members of the first pair.

In some embodiments, the members of the first pair are human. In certain embodiments, first pair's polynucleotide segments comprise DNA, mitochondrial DNA, sex-linked nucleotide segments, and/or RNA. In certain embodiments, \( t \) is equal to or greater than about 2.5 cM.

[0004] Some embodiments of the disclosure include methods of estimating genetic relatedness between members of a first pair of conspecific organisms further comprising comparing the lengths of the first pair's identical segments to a background distribution of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the background distribution. In some embodiments the identical segments of the background group are no longer than about 10 cM. In certain embodiments members of the background group are selected randomly from a larger population.

[0005] In certain embodiments, the methods further comprise comparing the number of the first pair's identical segments to a first distribution, of numbers of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and wherein the estimating is further based on the comparison of the number of the first pair's identical segments to the numbers in the first distribution.
In some embodiments, the methods further comprise comparing the lengths of the first pair's identical segments to a second distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a second group, the members of each pair in the second group having an established degree of genetic relatedness to each other; wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the second distribution.

In certain embodiments, the methods further comprise comparing the lengths of the first pair's identical segments to a second distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a second group, the members of each pair in the second group having an established degree of genetic relatedness to each other; wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the second distribution. In some embodiments, the methods further comprise comparing the lengths of the first pair's identical segments to a background distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the background distribution.

Some embodiments of the disclosure include methods of estimating genetic relatedness between members of a first pair of conspecific organisms wherein the estimating further comprises estimating a likelihood \( L_p \) that the first pair are no more related than two individuals selected randomly from a population, wherein: \( L_p(n,s \setminus t) = N_p(n \setminus t) \cdot S_p(s \setminus t) \); wherein \( S_p(s \setminus t) = \prod_{i \in s} F_p(i \mid t) \); wherein \( N_p(n \setminus t) \) comprises the likelihood of sharing \( n \) segments, \( S_p(s \setminus t) \) comprises the likelihood of the set of segments \( s \), and \( F_p(i \mid t) \) comprises the likelihood of a segment of size \( i \). In some embodiments, \( F_p(n \setminus t) \) is approximated as \( F_p(i \mid t) = \frac{e^{-(i-t)/\Theta}}{\Theta} \); wherein \( \Theta \)
is equal to a mean shared segment size in the population for all segments of size greater than \( t \) and less than a maximum length. In some embodiments the maximum length is about 10 cM.

[0009] In some aspects, the estimating further comprises estimating a likelihood \( L_R \) that the first pair share one or two ancestors, wherein:

\[
L_R = L_A(n_A, s_A \mathbf{Id}, d, a, t) L_p(s_p, \mathbf{Id});
\]

wherein \( n_p + n_A = n \), where \( n_A \) is equal to the number of shared segments inherited from ancestors, \( n_p \) is the number of segments shared by the population; wherein \( s_p \) and \( s_A \) are two mutually exclusive subsets of \( s \), where \( s_A \) is the subset of segments inherited from ancestor(s) with \( n_A \) elements, and \( s_p \) is the subset of segments shared by the population with \( n_p \) elements; wherein \( a \) represents the number of ancestors shared, and \( d \) represents the combined number of generations separating the individuals from their ancestor(s).

[0010] In some embodiments, the estimating further comprises estimating a likelihood \( L_A \) that the first pair share \( n \) segments from ancestor(s) specified by \( d \) and \( a \), with the segment sizes specified by \( s_A \), wherein:

\[
L_A(n_A, s_A \mathbf{Id}, d, a, t) = N_A(n \mathbf{Id}, d, a, t) S_A(s_A \mathbf{Id}, t);
\]

wherein \( S_A(s \mathbf{Id}, t) = \prod_{i \in s} F_A(l_i t) \); wherein \( N_A(n \mathbf{Id}, d, a, t) \) is the likelihood of sharing \( n \) segments, \( S_A(s_A \mathbf{Id}, t) \) is the likelihood of the set of segments \( s_A \), and \( F_A(l_i t) \) is the likelihood of a segment of size \( t \); wherein \( s_A \) and \( s_p \) are two mutually exclusive subsets of \( s \), where \( s_A \) is the subset of segments inherited from ancestor(s) with \( n_A \) elements, and \( s_p \) is the subset of segments shared by the population with \( n_p \) elements; wherein \( n_p + n_A = n \), where \( n_A \) is equal to the number of shared segments inherited from ancestors, \( n_p \) is the number of segments shared by the population; wherein \( a \) represents the number of ancestors shared, and \( d \) represents the combined number of generations separating the individuals from their ancestor(s). In certain aspects, the estimating further comprises estimating a likelihood \( L_A \) that the first pair share \( n \) segments from ancestor(s) specified by \( d \) and \( a \), with the segment sizes specified by \( s_A \), wherein:

\[
L_A(n_A, s_A \mathbf{Id}, d, a, t) = N_A(n \mathbf{Id}, d, a, t) S_A(s_A \mathbf{Id}, t); \]

wherein \( N_A(n \mathbf{Id}, d, a, t) \) is the likelihood of sharing \( n \) segments, \( S_A(s_A \mathbf{Id}, t) \) is the likelihood of the set of segments \( s_A \), and \( F_A(l_i t) \) is the likelihood of a segment of size \( t \).
[0011] In some embodiments, \( N_A(n, l, a, t) = \frac{e^{-a(l+c)p(t)}}{n!} \); wherein

\( p(t) \) is the probability that a shared segment is longer than \( t \). \( c \) comprises an average number of chromosomes in the organisms, and \( r \) comprises an average number of recombination events per haploid genome in the organisms. In certain embodiments, \( p(t) \) is assumed to be equal to or about

\( e^{-\ln(\infty)} \), In certain embodiments, \( F_A(n, l, a, t) = \frac{e^{-d(l-t)100}}{100/d} \).

[0012] In certain aspects, the estimating further comprises estimating a maximum likelihood of \( L_R \) (ML\(_R\)), wherein:

\[
ML_R(n_p, n_a, s, l, a, t) = N_p(n_p, l, t)N_a(n_a, l, a, t) - \sum_{s_p \in \{s_{n_p}, \ldots, s_{n_p, n_a}\}} S_{x,y} = \{s_{n_p, l, a, t}\};
\]

where \( s_{x,y} \) is equal to the \( X^{th} \) smallest value in \( s \). In certain embodiments, the methods further comprise evaluating, by a processor, a ratio of \( ML_R(n_p, n_a, s, l, a, t) \) and \( L_p(n, s, l, t) \) using a chi-square approximation with two degrees of freedom. In some embodiments, the estimating further comprises estimating a maximum likelihood of \( L_R \) (ML\(_R\)), wherein:

\[
ML_R(n_s, s, l, a, t) = \text{Max}\{\text{MLR}(n_p, n - n_p, s) : n_p \in [0, n]\}.
\]

[0013] In some embodiments, the methods of the invention further comprise receiving, by a processor, values indicating locations of the identical segments; comparing the locations of the first pair’s identical segments to locations of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a fourth pair of organisms, the members of the fourth pair having an established degree of genetic relatedness to each other; and wherein the estimating is further based on the location comparison. In some embodiments, the methods further comprise comparing the locations of the first pair’s identical segments to a background distribution of locations of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins;
and wherein the estimating is further based on the comparison of the locations of the first pair's identical segments to the locations in the background distribution.

[0014] In some embodiments, the methods further comprise comparing the locations of the first pair's identical segments to a first distribution, of locations of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and wherein the estimating is further based on the comparison of the locations of the first pair's identical segments to the locations in the first distribution.

[0015] Some embodiments of the disclosure include a computer-readable medium encoded with a computer program comprising instructions executable by a processor for estimating genetic relatedness between members of a first pair of conspecific organisms, the instructions including instruction code for: receiving, by a processor, a value indicating a number of nonoverlapping polynucleotide segments longer than a threshold length \( (t) \) that are identical, by at least about 90 percent sequence identity, between members of the first pair; receiving, by a processor, values indicating lengths of the identical segments; comparing the number of the first pair's identical segments to a number of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a second pair of organisms, the members of the second pair having an established degree of genetic relatedness to each other; comparing the lengths of the first pair's identical segments to lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a third pair of organisms, the members of the third pair having an established degree of genetic relatedness to each other; based on the number comparison and the length comparison, estimating, by a processor, a degree of genetic relatedness between the members of the first pair. In some embodiments, the members of the first pair are human. In certain embodiments, first pair's polynucleotide segments comprise DNA, mitochondrial DNA, sex-linked nucleotide segments, and/or RNA. In certain embodiments, \( t \) is equal to or greater than about 2.5 cM.

[0016] In certain aspects, the computer-readable medium further comprises comparing the lengths of the first pair's identical segments to a background distribution of
lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the background distribution. In some embodiments, the identical segments of the background group are no longer than about 10 cM. In certain embodiments the members of the background group are selected randomly from a larger population.

[0017] In some aspects, the medium further comprises comparing the number of the first pair's identical segments to a first distribution, of numbers of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and wherein the estimating is further based on the comparison of the number of the first pair's identical segments to the numbers in the first distribution.

[0018] In some embodiments, the medium further comprises comparing the lengths of the first pair's identical segments to a second distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a second group, the members of each pair in the second group having an established degree of genetic relatedness to each other; wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the second distribution.

[0019] In certain aspects, the medium further comprises comparing the lengths of the first pair's identical segments to a second distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a second group, the members of each pair in the second group having an established degree of genetic relatedness to each other; wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the second distribution.
In certain embodiments, the medium further comprises comparing the lengths of the first pair's identical segments to a background distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the background distribution. In certain aspects, the estimating further comprises estimating a likelihood \( L_p \) that the first pair are no more related than two individuals selected randomly from a population, wherein:

\[
L_p(n, s \setminus t) = N_p(n \setminus t) \cdot S_p(s \setminus t); \\
\text{wherein} S_p(s \setminus t) = \prod_{i \in s} F_p(i \setminus t); \\
\text{wherein} N_p(n \setminus t) \text{ comprises the likelihood of sharing } n \text{ segments,}
\]

\( S_p(s \setminus t) \) comprises the likelihood of the set of segments \( s \), and \( F_p(i \setminus t) \) comprises the likelihood of a segment of size \( i \). In certain aspects, \( F_p(i \setminus t) \) is approximated as:

\[
F_p(i \setminus t) = \frac{e^{-(i-t)/\Theta}}{\Theta};
\]

wherein \( \Theta \) is equal to a mean shared segment size in the population for all segments of size greater than \( t \) and less than a maximum length. In some embodiments, the computer-readable medium of claim 45, wherein the maximum length is about 10 cM.

In some aspects of the computer-readable medium the estimating further comprises estimating a likelihood \( L_R \) that the first pair share one or two ancestors, wherein:

\[
L_R = L_A(n_A, s_A \setminus d, a, t) L_p(s_p \setminus t); \\
\text{wherein} n_p + n_A = n, \text{ where } n_A \text{ is equal to the number of shared segments inherited from ancestors, } n_p \text{ is the number of segments shared by the population;}
\]

wherein \( s_p \) and \( \frac{3}{4} \) are two mutually exclusive subsets of \( s \), where \( \frac{3}{4} \) is the subset of segments inherited from ancestor(s) with \( n_A \) elements, and \( s_p \) is the subset of segments shared by the population with \( n_p \) elements; wherein \( a \) represents the number of ancestors shared, and \( d \) represents the combined number of generations separating the individuals from their ancestor(s).

In some embodiments the estimating further comprises estimating a likelihood \( L_A \) that the first pair share \( n \) segments from ancestor(s) specified by \( d \) and \( a \), with the segment sizes specified by \( s_A \), wherein:

\[
L_A(n_A, s_A \setminus d, a, t) = N_A(n \setminus d, a, t) \cdot S_A(s_A \setminus d, a, t); \\
\text{wherein} S_A(s \setminus d, a, t) = \prod_{i \in s} F_A(i \setminus t);
\]
wherein $N_A(n \setminus d,a,t)$ is the likelihood of sharing $n$ segments, $S_A(s_A \setminus d,t)$ is the likelihood of the set of segments $s_A$, and $F_A(n)$ is the likelihood of a segment of size $i$; wherein $s_p$ and $s_A$ are two mutually exclusive subsets of $s$, where $s_A$ is the subset of segments inherited from ancestor(s) with $n_A$ elements, and $s_p$ is the subset of segments shared by the population with $n_p$ elements; wherein $n_p + n_A = n$, where $n_A$ is equal to the number of shared segments inherited from ancestors, $n_p$ is the number of segments shared by the population; wherein $a$ represents the number of ancestors shared, and $d$ represents the combined number of generations separating the individuals from their ancestor(s).

[0022] In some embodiments of the computer-readable medium, the estimating further comprises estimating a likelihood $L_A$ that the first pair share $n$ segments from ancestor(s) specified by $d$ and $a$, with the segment sizes specified by $s_A$, wherein:

$$L_A(n_A, s_A \setminus d, a, t) = N_A(n \setminus d, a, t) \cdot S_A(s_A \setminus d, t);$$

wherein $N_A(n \setminus d, a, t)$ is the likelihood of sharing $n$ segments, $S_A(s_A \setminus d, t)$ is the likelihood of the set of segments $s_A$, and $F_A(n)$ is the likelihood of a segment of size $i$.

$$N_A(n \setminus d, a, t) = \frac{e^{a(rd + c)p(t)}}{2^{d - 1}} \cdot \frac{(a(rd + c)p(t))^n}{n!};$$

wherein $p(t)$ is the probability that a shared segment is longer than $t$, $c$ comprises an average number of chromosomes in the organisms, and $r$ comprises an average number of recombination events per haploid genome in the organisms. In certain embodiments, $p(t)$ is assumed to be equal to or about $e^{-d/10}$. In certain embodiments, of the computer-readable medium $F_A(i \setminus d, t) = e^{-d(i - t)/10} / 100/d$.

[0023] In certain aspects, the estimating further comprises estimating a maximum likelihood

$$L_R (ML_R).$$

wherein:

$$ML_R(n_p, n_A, s \setminus d, a, t) = N_p(n_p \setminus t) \cdot N_A(n_A \setminus d, a, t) - S_p(\{s_{s_{n_p}}\} \setminus t) S_A(\{s_{s_{n_p+k}}\} \setminus d, a, t);$$

where $s_c$ is equal to the $x^{th}$ smallest value in $s$. In some embodiments of the medium, evaluating further comprises evaluating, by a processor, a ratio of $ML_R(n_p, n_A, s \setminus d, a, t)$ and $L_p(n, s \setminus t)$. Using a chi-square
approximation with two degrees of freedom. In certain aspects, the estimating further comprises estimating a maximum likelihood of $L_R$ ($ML_R$), wherein:

$$ML_R(n,s,l,d,a,t) = \max\{MLR(n_p, n - n_p, s) : n_p \in \{0..n\}\}.$$ 

[0025] In some embodiments, the computer-readable medium further comprises receiving, by a processor, values indicating locations of the identical segments; comparing the locations of the first pair's identical segments to locations of nonoverlapping polynucleotide segments longer than $t$ that are identical, by at least about 90 percent sequence identity, between members of a fourth pair of organisms, the members of the fourth pair having an established degree of genetic relatedness to each other; and wherein the estimating is further based on the location comparison.

[0026] In some aspects, the computer-readable medium further comprises comparing the locations of the first pair's identical segments to a background distribution of locations of nonoverlapping polynucleotide segments longer than $t$ that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and wherein the estimating is further based on the comparison of the locations of the first pair's identical segments to the locations in the background distribution. In certain aspects, the computer-readable medium further comprises comparing the locations of the first pair's identical segments to a first distribution, of locations of nonoverlapping polynucleotide segments longer than $t$ that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and wherein the estimating is further based on the comparison of the locations of the first pair's identical segments to the locations in the first distribution.

[0027] Additional features and advantages of the subject technology will be set forth in the description below, and in part will be apparent from the description, or may be learned by practice of the subject technology. The advantages of the subject technology will be realized and attained by the structure particularly pointed out in the written description and claims hereof as well as the appended drawings.
[0028] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory and are intended to provide further explanation of the subject technology as claimed.

[0029] All publications, patents, and GenBank sequences cited in this disclosure are incorporated by reference in their entirety.

**Brief Description of the Drawings**

[0030] The accompanying drawings, which are included to provide further understanding of the subject technology and are incorporated in and constitute a part of this specification, illustrate aspects of the subject technology and together with the description serve to explain the principles of the subject technology.

[0031] Figure 1. 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). (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) The lengths of the IBD segments are approximately exponentially distributed, with mean length depending on the relationship between individuals (theoretical distributions shown for $d = 2, 4, 6, 8$).

[0032] Figure 2. Characteristics of HapMap CEU (Utah Americans of Northern and Western European descent) parents as a background reference population. (A) Principal components analysis comparing 36 individuals from the three pedigrees set forth in Table 1 (no pair closer than seventh-degree relatives) to 85 unrelated individuals from three European
populations (60 HapMap CEU parent-offspring trios and 25 HapMap TSI (Toscani in Italia) individuals) based on pairwise allele-sharing distances computed from -247,000 single-nucleotide polymorphisms (SNPs) typed on the Affymetrix SNP array (see Xing et al. 2010). The percentage of genetic variation explained by each component is given on the corresponding axis. (B) Distribution of the number of segments with length ≥ 2.5 cM that are inferred to be shared IBD by GERMLINE in pairs of CEU individuals (Observed), with fitted Poisson distribution (Expected). (C) Distribution of the lengths of IBD segments longer than 2.5 cM in CEU pairs (Observed), with fitted exponential distribution (Expected). (D) Scatterplot of the number of IBD segments per pair vs. mean length of segments in the pair.

[0033] Figure 3. Estimated degree of relationship between pairs of individuals vs. known degree of relationship. (A) 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). 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. (B) The number of pairs in each category is indicated by the histogram below.

[0034] Figure 4. (A) Power to detect recent common ancestry between pairs of individuals known to be related at varying degrees. Each pair of individuals has exactly two known ancestors in the pedigree, and both inheritance paths connecting the pair (one through each ancestor) have the same number of meioses in them. Maximum theoretical power is shown by the solid black line (the probability that a pair of individuals with the given relationship are genetically related at all, calculated from Eq. 7 with $a = 2$ and $t = 0$). The power of ERSA using IBD segments estimated by GERMLINE, with $\alpha = 0.05$ and $\alpha = 0.001$ (2 degrees of freedom d.f.), is indicated by the dotted and solid red lines respectively. Using IBD segments estimated by fastlBD of the Beagle 3.3 package available on Sharon Browning’s or Brian Browning’s University of Washington webpages), ERSA achieves the power shown by the green line ($\alpha = 0.001$, 2 d.f.). The power of RELPAIR (Epstein et al. 2000) to detect a relationship is indicated by the dotted blue line (using 9,990 evenly-spaced autosomal markers with minor allele
frequency (MAF) > 0.4, default likelihood ratio (LR) threshold of 10 for reporting a relationship as significant). The power of GBIRP (Stankovich et al. 2005) is shown by the solid blue line (10,028 evenly-spaced autosomal markers with MAF > 0.4, LOD threshold of 2.34 for significance as in Stankovich et al. 2005, corresponding to $\alpha = 0.001$ with 1 d.f.).

[0035] Figure SI: ERSA’s power and accuracy for one-ancestor relationships. Figures 3 and 4 display results for all known two-ancestor relationships in the pedigree where the two inheritance paths are the same length, such as full siblings and full cousins. This figure displays the equivalent results for all relationships with exactly one known one-ancestor relationships, i.e. half siblings and half cousins. (A) Known vs. estimated degree of relationship. (B) Number of pairs in the pedigree with the specified known degree of relationship. (C) Power to detect a significant relationship at the $\alpha = 0.001$ significance level plotted against the maximum theoretical power (calculated from Eq. 7 with $a = 1$ and $t = 0$).

[0036] Figure S2: Known vs. estimated degree of relationship for individuals that share exactly two common ancestors and where both paths connecting the pair have the same length, using (A) ERSA with $\alpha = 0.05$ based on IBD segments estimated by GERMLINE (Gusev et al. 2009) IBD segments; (B) ERSA with $\alpha = 0.001$ and GERMLINE IBD segments (same as Figure 3); (C) ERSA with $\alpha = 0.05$ and Beagle 3.3 fastIBD (available on Sharon Browning’s or Brian Browning’s University of Washington webpages) segments; (D) GBIRP and 10,028 evenly-spaced SNPs with MAF > 0.4, with a LOD threshold of 2.34 for significance (as in Stankovich et al. 2005); and (E) RELPAIR with 9,990 evenly-spaced SNPs and requiring a likelihood ratio > 10 for significance (the default in RELPAIR; Epstein et al. 2000). (F) The number of pairs in each relationship class. For GBIRP analysis, SNP data was thinned (following Berkovic et al. 2008) after phasing and imputation as described in Methods, then written to GBIRP-readable data format files (fdist, ffreq, fhaplos, and fLastMarkers; available on the Walter + Eliza Hall Institute of Medical Research Bioinformatics/GBIRP webpages), with allele frequencies estimated from the entire sample of 169 individuals. GBIRP analyses were performed with various numbers of markers (from 1,000 to 50,000) with different minimum MAF values (from 0.1 to 0.4); the optimal results are shown.
[0037] Figure S3: Performance of ERSA’s nominal (A) 95% and (B) 99% confidence intervals (C.I.) - The proportion of pairs for which the nominal C.I. contains the known value is plotted vs. the known relationship (degree of relationship for a pair of individuals that share two common ancestors, where both paths through those ancestors have the same length, with \(a = 2\)).

[0038] Figure S4: Realized vs. expected sums of shared IBD segment lengths between pairs of related individuals sharing exactly two ancestors. The dotted lines enclose the middle 90% of observed values. The expectation for the sum of IBD segment lengths (dashed line) is adjusted to account for the fact that IBD segments detected by GERMLINE do not distinguish between haploid and diploid sharing and for the expected overlap of IBD segments in siblings.

[0039] Figure S5: Bioinformatic merging of shared segments in full siblings. Two homologous autosomal chromosomes are shown for two parents, each colored differently. Meiosis and recombination occurs and two sibling offspring inherit recombinant chromosomes. Although the siblings share three distinct IBD segments, two of these segments overlap and are thus merged bioinformatically (by GERMLINE or BEAGLE) into a single shared segment (black bar, far right). Eq. S1 and S2 account for this process of bioinformatic merging.

[0040] Figure S6: The effect of allowing \(a\) to vary under the null model. The cumulative probability for values of the observed LRT statistic comparing models with \(a\) free to vary or fixed equal to 2 is shown in blue. The cumulative distribution for a \(\chi^2\) distribution with one degree of freedom is shown in red for comparison.

Table 1

Proportions of the total possible number of ancestors of the 169 genotyped individuals up to a given depth (in generations) that are listed in the three pedigrees. For example, for the combined dataset (the 1st column), 99.4% of the second-generation ancestors of the 169 genotyped individuals are included in the pedigree.

<table>
<thead>
<tr>
<th>Generation</th>
<th>Proportion of ancestors in pedigree</th>
<th>Combined (61,569)</th>
<th>Pedigree 1 (169; 115; 58,329)*</th>
<th>Pedigree 2 (30; 2,017)*</th>
<th>Pedigree 3 (24; 1,223)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.994</td>
<td>0.991</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.966</td>
<td>0.972</td>
<td>0.967</td>
<td>0.938</td>
<td></td>
</tr>
</tbody>
</table>

-14-
Table 2

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

<table>
<thead>
<tr>
<th>Nominal false positive rate</th>
<th>Observed false positive rate</th>
<th>Observed false positive counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.044</td>
<td>89/2,025</td>
</tr>
<tr>
<td>0.01</td>
<td>0.0094</td>
<td>19/2,025</td>
</tr>
<tr>
<td>0.001</td>
<td>0.00049</td>
<td>1/2,025</td>
</tr>
</tbody>
</table>

Table SI: Data of Figure S2 and Figure 3.

ERSA + GERMLINE, \( a = 0.05 \)

| Estimated degree | None | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | None
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>None detected</td>
<td></td>
<td>6</td>
<td>14</td>
<td>53</td>
<td>180</td>
<td>263</td>
<td>339</td>
<td>334</td>
<td>103</td>
<td>6</td>
<td>6584</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>10</td>
<td>20</td>
<td>15</td>
<td>63</td>
<td>48</td>
<td>36</td>
<td>10</td>
<td>7</td>
<td>133</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>41</td>
<td>39</td>
<td>94</td>
<td>64</td>
<td>28</td>
<td>16</td>
<td>8</td>
<td>1</td>
<td>184</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>75</td>
<td>65</td>
<td>38</td>
<td>38</td>
<td>15</td>
<td>4</td>
<td>1</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>102</td>
<td>126</td>
<td>28</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>28</td>
<td>164</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>19</td>
<td>85</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>75</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>3</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>12</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ERSA + GERMLINE, \( a = 0.001 \) (data of Figure 3)

<table>
<thead>
<tr>
<th>Estimated degree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>None known</th>
</tr>
</thead>
<tbody>
<tr>
<td>None detected</td>
<td>10</td>
<td>21</td>
<td>57</td>
<td>213</td>
<td>296</td>
<td>360</td>
<td>350</td>
<td>110</td>
<td>7</td>
<td>6829</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>15</td>
<td>14</td>
<td>44</td>
<td>34</td>
<td>23</td>
<td>5</td>
<td>4</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>24</td>
<td>39</td>
<td>36</td>
<td>80</td>
<td>46</td>
<td>20</td>
<td>5</td>
<td>4</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>75</td>
<td>65</td>
<td>38</td>
<td>38</td>
<td>14</td>
<td>4</td>
<td>1</td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>102</td>
<td>126</td>
<td>28</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>164</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>19</td>
<td>85</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>75</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ERSA + BEAGLE, \( a = 0.001 \)

<table>
<thead>
<tr>
<th>Estimated degree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>None known</th>
</tr>
</thead>
<tbody>
<tr>
<td>None detected</td>
<td>2</td>
<td>2</td>
<td>17</td>
<td>64</td>
<td>74</td>
<td>105</td>
<td>323</td>
<td>360</td>
<td>397</td>
<td>361</td>
<td>118</td>
<td>7</td>
<td>6907</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>17</td>
<td>27</td>
<td>18</td>
<td>22</td>
<td>13</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>14</td>
<td>55</td>
<td>39</td>
<td>15</td>
<td>25</td>
<td>11</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>48</td>
<td>87</td>
<td>22</td>
<td>8</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>137</td>
<td>39</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>68</td>
<td>71</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>68</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>28</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GBIRP, LOD > 2.34

<table>
<thead>
<tr>
<th>Estimated degree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>None known</th>
</tr>
</thead>
<tbody>
<tr>
<td>None detected</td>
<td>1</td>
<td>4</td>
<td>63</td>
<td>149</td>
<td>127</td>
<td>123</td>
<td>353</td>
<td>378</td>
<td>405</td>
<td>359</td>
<td>116</td>
<td>7</td>
<td>6905</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>19</td>
<td>10</td>
<td>9</td>
<td>12</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>33</td>
<td>47</td>
<td>23</td>
<td>15</td>
<td>14</td>
<td>7</td>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>14</td>
<td>120</td>
<td>50</td>
<td>8</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>74</td>
<td>68</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>62</td>
<td>24</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>23</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RELPair, likelihood ratio > 10
Table S2: Number of pairs in each relationship degree class (data of lower panel of Figure 3)

<table>
<thead>
<tr>
<th>Known degree detected</th>
<th>None</th>
<th>3+</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
<td>164</td>
<td>150</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>117</td>
<td>250</td>
<td>107</td>
</tr>
</tbody>
</table>

Table S4: Percent detection power for various methods (data of Figure 3)

<table>
<thead>
<tr>
<th>Degree of relationship known</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.98</td>
<td>99.1</td>
<td>92.9</td>
<td>76.8</td>
<td>55.0</td>
<td>35.25</td>
<td>20.9</td>
<td>11.8</td>
<td>6.5</td>
</tr>
<tr>
<td>Theoretical Power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>94</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>7.48</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>ERS+A</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>97.79</td>
<td>91.6</td>
<td>64</td>
<td>52.5</td>
<td>32.7</td>
<td>16.71</td>
<td>7.48</td>
<td>12.7</td>
<td>14.1</td>
</tr>
<tr>
<td>GERMLIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>E, a = 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>62</td>
<td>43.8</td>
<td>24.3</td>
<td>11.55</td>
<td>3.05</td>
<td>6.78</td>
<td>0</td>
</tr>
<tr>
<td>ERS+A</td>
<td>100</td>
<td>93.3</td>
<td>97.89</td>
<td>100</td>
<td>100</td>
<td>96.31</td>
<td>87.5</td>
<td>62</td>
<td>43.8</td>
<td>24.3</td>
<td>11.55</td>
<td>3.05</td>
<td>6.78</td>
<td>0</td>
</tr>
<tr>
<td>BEAGLE, E, a = 0.001</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>46</td>
<td>8</td>
<td>7.93</td>
<td>14.7</td>
<td>7.93</td>
<td>2.46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>GBIRP</td>
<td>100</td>
<td>96.8</td>
<td>96.5</td>
<td>8</td>
<td>8</td>
<td>78.2</td>
<td>45.02</td>
<td>24.4</td>
<td>18</td>
<td>6.86</td>
<td>3.32</td>
<td>0.49</td>
<td>0.55</td>
<td>1.69</td>
</tr>
<tr>
<td>RELPAIR</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>86.2</td>
<td>39.48</td>
<td>10.7</td>
<td>2.6</td>
<td>0.79</td>
<td>0</td>
<td>0.49</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

* For very distant relationships, estimated power sometimes exceeds the maximum expected power. This is likely due to the existence of some undocumented distant relationships, since the pedigrees are not complete at such depths, as well as to false positive results.
Table S5: ERSA + GERMLINE, $a = 0.001$, one-ancestor model and data set (data of Figure SI)

<table>
<thead>
<tr>
<th>Known degree of relationship</th>
<th>Estimate degree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>None known</th>
</tr>
</thead>
<tbody>
<tr>
<td>None detected</td>
<td></td>
<td>14</td>
<td>57</td>
<td>50</td>
<td>38</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6826</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>6</td>
<td>13</td>
<td>13</td>
<td>6</td>
<td>1</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>4</td>
<td>24</td>
<td>27</td>
<td>17</td>
<td>6</td>
<td>1</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>5</td>
<td>29</td>
<td>58</td>
<td>34</td>
<td>12</td>
<td>2</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>16</td>
<td>59</td>
<td>29</td>
<td>4</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>2</td>
<td>44</td>
<td>21</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of Pairs

<table>
<thead>
<tr>
<th>Number of Pairs</th>
<th>Estimate</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>89.39</th>
<th>57.78</th>
<th>45.65</th>
<th>26.92</th>
<th>22.22</th>
<th>6930</th>
</tr>
</thead>
</table>

| d Power | 100 | 100 | 100 | 100 | 100 | 89.39 | 57.78 | 45.65 | 26.92 | 22.22 | 6930 |
**Table S6:** Estimates of significant recent ancestry \((a=0.001)\) among pairs of parent individuals in the HapMap CELT dataset.

<table>
<thead>
<tr>
<th>Individual 1</th>
<th>Individual 2</th>
<th>Number of shared ancestors ((a))</th>
<th>Estimated degree of relationship</th>
<th>99.9% Confidence Interval for the degree of relationship (a=2, \gamma=1)</th>
<th>(\ln(L(\text{Related})))</th>
<th>(\ln(L(\text{Unrelated})))</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA12154</td>
<td>NA12892</td>
<td>2</td>
<td>9</td>
<td>6-21 6-21</td>
<td>12.90</td>
<td>19.98</td>
</tr>
<tr>
<td>NA06985</td>
<td>NA12812</td>
<td>1</td>
<td>7</td>
<td>5-13 5-13</td>
<td>23.86</td>
<td>67.49</td>
</tr>
<tr>
<td>NA06993</td>
<td>NA07022</td>
<td>2</td>
<td>4</td>
<td>3-6 3-6</td>
<td>81.95</td>
<td>499.50</td>
</tr>
<tr>
<td>NA11995</td>
<td>NA12145</td>
<td>2</td>
<td>8</td>
<td>5-16 5-16</td>
<td>16.74</td>
<td>26.85</td>
</tr>
<tr>
<td>NA11840</td>
<td>NA12717</td>
<td>2</td>
<td>8</td>
<td>6-16 5-16</td>
<td>15.70</td>
<td>30.77</td>
</tr>
<tr>
<td>NA12056</td>
<td>NA12872</td>
<td>2</td>
<td>8</td>
<td>5-13 5-13</td>
<td>18.67</td>
<td>27.12</td>
</tr>
<tr>
<td>NA07034</td>
<td>NA12145</td>
<td>1</td>
<td>9</td>
<td>6-19 5-19</td>
<td>16.33</td>
<td>37.98</td>
</tr>
<tr>
<td>NA12146</td>
<td>NA12812</td>
<td>2</td>
<td>8</td>
<td>5-19 5-19</td>
<td>21.11</td>
<td>30.25</td>
</tr>
<tr>
<td>NA11881</td>
<td>NA12762</td>
<td>2</td>
<td>8</td>
<td>5-17 5-17</td>
<td>14.62</td>
<td>23.63</td>
</tr>
<tr>
<td>NA06993</td>
<td>NA07056</td>
<td>2</td>
<td>4</td>
<td>3-6 3-6</td>
<td>85.14</td>
<td>510.44</td>
</tr>
<tr>
<td>NA11993</td>
<td>NA12239</td>
<td>2</td>
<td>8</td>
<td>5-16 5-18</td>
<td>17.78</td>
<td>27.13</td>
</tr>
<tr>
<td>NA11829</td>
<td>NA12815</td>
<td>2</td>
<td>7</td>
<td>5-13 5-13</td>
<td>22.46</td>
<td>32.26</td>
</tr>
<tr>
<td>NA07034</td>
<td>NA11882</td>
<td>2</td>
<td>6</td>
<td>5-8 4-8</td>
<td>33.72</td>
<td>139.83</td>
</tr>
<tr>
<td>NA07000</td>
<td>NA12057</td>
<td>2</td>
<td>8</td>
<td>5-18 5-18</td>
<td>23.27</td>
<td>42.08</td>
</tr>
<tr>
<td>NA12155</td>
<td>NA12264</td>
<td>2</td>
<td>4</td>
<td>3-5 3-5</td>
<td>103.79</td>
<td>631.83</td>
</tr>
<tr>
<td>NA12006</td>
<td>NA12155</td>
<td>2</td>
<td>9</td>
<td>6-20 6-20</td>
<td>10.12</td>
<td>19.43</td>
</tr>
<tr>
<td>NA07034</td>
<td>NA12750</td>
<td>2</td>
<td>8</td>
<td>5-19 5-19</td>
<td>20.75</td>
<td>41.10</td>
</tr>
<tr>
<td>NA12236</td>
<td>NA12716</td>
<td>1</td>
<td>9</td>
<td>5-17 5-17</td>
<td>18.32</td>
<td>60.64</td>
</tr>
<tr>
<td>NA06994</td>
<td>NA07000</td>
<td>1</td>
<td>9</td>
<td>6-17 5-18</td>
<td>13.29</td>
<td>49.92</td>
</tr>
<tr>
<td>NA07022</td>
<td>NA07056</td>
<td>2</td>
<td>8</td>
<td>5-18 5-18</td>
<td>19.80</td>
<td>35.36</td>
</tr>
<tr>
<td>NA12043</td>
<td>NA12760</td>
<td>2</td>
<td>8</td>
<td>6-18 5-18</td>
<td>12.42</td>
<td>19.73</td>
</tr>
<tr>
<td>NA11994</td>
<td>NA12146</td>
<td>2</td>
<td>8</td>
<td>5-19 5-19</td>
<td>15.21</td>
<td>24.71</td>
</tr>
<tr>
<td>NA06994</td>
<td>NA12892</td>
<td>2</td>
<td>5</td>
<td>4-7 4-6</td>
<td>65.19</td>
<td>296.69</td>
</tr>
</tbody>
</table>

**Detailed Description**

At line 106, a phrase such as "an aspect" does not imply that such aspect is essential to the subject technology or that such aspect applies to all configurations of the subject technology. A disclosure relating to an aspect may apply to all configurations, or one or more configurations.
An aspect may provide one or more examples of the disclosure. A phrase such as "an aspect" may refer to one or more aspects and vice versa. A phrase such as "an embodiment" does not imply that such embodiment is essential to the subject technology or that such embodiment applies to all configurations of the subject technology. A disclosure relating to an embodiment may apply to all embodiments, or one or more embodiments. An embodiment may provide one or more examples of the disclosure. A phrase such "an embodiment" may refer to one or more embodiments and vice versa. A phrase such as "a configuration" does not imply that such configuration is essential to the subject technology or that such configuration applies to all configurations of the subject technology. A disclosure relating to a configuration may apply to all configurations, or one or more configurations. A configuration may provide one or more examples of the disclosure. A phrase such as "a configuration" may refer to one or more configurations and vice versa.

1. OVERVIEW

[0044] Most established methods for detecting and estimating genetic relationships are based on genome-wide averages of the estimated number of alleles shared that are identical by descent (IBD) between two individuals (Weir et al. 2006). These methods are accurate and efficient for relationships as distant as third-degree relatives (e.g., first cousins) but cannot identify more distant relationships. In contrast, aspects of the instant disclosure provide novel methods and apparatus of estimation of recent shared ancestry (ERSA) that accurately estimate the degree of relationship for up to eighth-degree relatives (e.g., third cousins once removed), and detect relationships as distant as twelfth-degree relatives (e.g., fifth cousins once removed).

[0045] Some methods of detecting relatedness (for example, the method implemented in PLINK; Purcell et al. 2007) rely on genome-wide averages of genetic identity coefficient estimates. These statistics incompletely summarize the information contained in the IBD segment data: genetic identity coefficients can be calculated from IBD segment data, but the reverse is not true. To illustrate the importance of this difference, the typical amount of genetic sharing between a pair of fourth cousins is considered. The probability that fourth cousins share at least one IBD segment is 77%, and the expected length of this segment is 10 centiMorgans (cM) (Donnelly 1983). Because a 10 cM segment represents less than 0.3% of the genome, this excess
of IBD has very little effect on estimates of relatedness averaged over the genome. However, because unrelated individuals are unlikely to share a 10 cM segment in most populations, the novel ERSA methods and apparatus disclosed herein are capable of detecting many fourth-cousin relationships.

Another family of methods for detecting relationships models the IBD states between haplotypes as a Markov process along a chromosome, with different transition probability matrices corresponding to different hypothesized relationships. The likelihoods of various relationship models are then estimated from the data. Examples of these methods include RELPAIR (Boehnke and Cox 1997; Epstein et al. 2000), PREST (extending the methods in Boehnke and Cox, 1997; McPeek and Sun 2000; Sun et al. 2002), and GBIRP (extending PREST to the problem of general relationship estimation; Stankovich et al. 2005). These tools were initially designed for use with hundreds of microsatellite loci spaced at intervals of several cM, but they have also been applied to high-density single-nucleotide polymorphism (SNP) data (e.g., Berkovic et al. 2008; Pemberton et al. 2010). However, they do not model the patterns of linkage disequilibrium (LD) that exist between very closely spaced SNP markers and instead assume that markers are not in strong LD. High-density SNP data sets must be thinned to approximately 10,000 markers before they can be used (see, e.g., Berkovic et al. 2008; Pemberton et al. 2010). The key information used by such Markov-process methods is the match between the hypothesized transition probability matrix and the pattern of IBD state transitions induced by the genotype data.

In contrast, some embodiments of the instant ERSA methods and apparatus use explicit IBD segment information to estimate the relationships between pairs of individuals in a maximum-likelihood framework. This makes better use of the information present in high-density SNP genotyping data. The power of ERSA disclosed herein to detect relationships between second cousins or closer relatives is essentially perfect and exceeds 85% for third cousins even at the a = 0.001 level. ERSA is also more accurate than RELPAIR or GBIRP.

The number, lengths, and locations of chromosomal segments that are shared IBD by a pair of individuals essentially constitute the genetic information that bears on their recent shared genetic ancestry. Figure 1 illustrates the process that generates IBD segments and
shows how the expected distributions of segment number and length depend on the relationship between two individuals.

[0049] Algorithms can be used to detect the number, lengths, and locations of chromosomal segments IBD between two individuals. (Browning and Browning 2010; Gusev et al. 2009; Thomas et al. 2008) In some embodiments, ERSA uses a likelihood ratio test to compare the null hypothesis that the two individuals are unrelated with the alternative hypothesis that the individuals share recent ancestry. Because of the qualitative difference between genome-wide averages of relatedness and the information contained in IBD segments, aspects of the present disclosure greatly expand the range of relationships that can be detected from genetic data.

[0050] ERSA is immediately applicable to a number of problems. It can be used to identify cryptic relatedness between individuals with the same rare genetic disorder. In analyzing large pedigrees, ERSA can verify distant relationships without genotyping intervening family members. This can sharply reduce sample collection and genotyping requirements.

[0051] In the forensic field, a common DNA-based method for identifying the remains of missing persons is based on comparisons of kinship statistics computed from a modest number (13-17) of STR loci, with useful comparisons generally limited to second-degree relationships (Alonso et al. 2005; e.g., MDKAP, Leclair et al. 2007; M-FISys, Budimlija et al. 2003; Cash et al. 2003). The International Commission on Missing Persons (ICMP) has generated matches for more than 18,000 persons missing from armed conflicts or mass disasters at a significance level exceeding 99.95% (personal communication from TJ Parsons, ICMP). However, this level of certainty requires typing multiple first- or second-degree relatives. Such close relatives are often unavailable, due either to disasters and conflicts that disperse entire families or to the passage of time (Brenner 2006; Leclair 2004). For example, DNA profiles exist for over 2,000 individuals killed in the armed conflict in Bosnia for which identifications cannot be made due to insufficient family reference samples (TJ Parsons, ICMP). ERSA allows the use of a much larger pool of distant relatives (Bieber et al. 2006) and enables definitive conclusions to be drawn based on single closer relatives. For the first time, with ERSA, even a single individual searching for a family member would be able to provide a definitive reference.
The methods described here are computationally efficient, make near-optimal use of the genetic signal of relatedness between individuals, achieve a statistical power very close to the theoretical maximum and have multiple applications. These methods can be implemented by machine-readable code, e.g., in software or hardware, and over computer networks such as the Internet.

2. IDENTICAL BY DESCENT

As used herein, "IBD-segments" are nonoverlapping polynucleotide segments longer than a threshold length \( t \) that are identical, in certain embodiments, by at least about 90% identical; in certain embodiments about 95% identical; in certain embodiments about 98% identical; in certain embodiments about 99% identical; and in certain embodiments about 100% identical.

Any IBD segment number and length data can be used in aspects of the present disclosure. Likewise, any IBD segment detection method can be used. Examples of software programs for IBD segment detection are GERMLINE (Gusev et al. 2009); fastIBD in Beagle 3.3 (Browning and Browning 2010), MERLIN (via -extended, Abecasis et al.) and Thompson (tech report, U Wash). IBD segments are determined using, for example, SNP data, whole-genome sequencing data, and/or higher-density microarray data.

3. POLYNUCLEOTIDES

As used herein, "polynucleotides" are in certain embodiments deoxyribonucleic acids (DNA), in certain embodiments ribonucleic acids (RNA), in certain embodiments mitochondrial DNA (mtDNA), in certain embodiments sex-linked nucleotide segments, such as those found on the Y or X chromosomes.

In certain embodiments, autosomal segments is a source of the polynucleotides used in estimating recent shared ancestry. In certain embodiments, RNA is a source of the polynucleotides used in estimating recent shared ancestry.

In certain embodiments, mtDNA or the Y chromosome(s) is a source of the polynucleotides used in estimating recent shared ancestry. For a hypothesized alternative relationship with \( a \) ancestors on a path \( d \) meioses long, the likelihood of the observed mtDNA or
Y chromosome data is computed by integrating over all possible pedigrees with \(a\) ancestors and \(d\) meioses, specifying the sex of each individual in the inheritance path so that the probabilities can be calculated. The likelihood of the null hypothesis (no relationship) is calculated based on the frequencies of the observed mtDNA or Y chromosome haplotypes in the background population. In both calculations, an allowance is made for an appropriate genotyping or sequencing error rate. The log-likelihoods based on the mtDNA and Y chromosome data are then added to the log-likelihoods computed from the autosomal data (for the corresponding null and alternative hypotheses), and the relationship is estimated using standard likelihood theory as before.

[0058] In certain embodiments, the X chromosome(s) is a source of the polynucleotides used in estimating recent shared ancestry. IBD segment data from the X chromosome is used in a similar way as Y chromosome and mtDNA data. To calculate the likelihood of the null hypothesis given observed X chromosome SNP genotype or sequence data, the observed IBD segments are compared to distributions estimated from unrelated individuals in the source population. For each alternative hypothesis, likelihoods are calculated by integrating over all possible sex-specified pedigrees in the class of relationships with \(a\) ancestors on a path \(d\) meioses long. This allows the method to account for the number of meioses in the path in which recombination occurred (only in females), which determines the IBD segments length distribution, and for the probability that the ancestral X chromosome is lost altogether (due to two consecutive male parents in the inheritance path.) The log-likelihoods for null and alternative hypotheses based on X chromosome data are added to the log-likelihoods for the autosomal data, and the final likelihood ratio test is carried out as before.

4. DEFINITIONS

[0059] As used herein, the term "ancestor" is a parent or, recursively, the parent of an ancestor, e.g., a grandparent, great-grandparent, or great-great-grandparent.

[0060] As used herein, the term "random selection" is a broad term that includes, without limitation, selections that are any combination of (a) truly random, such as a random number generated by a random physical process, e.g., radioactive decay; (b) pseudo-random, such as a computer-generated random selection; (c) semi-random, including constraints in a selection process such as database size, and (d) quasi-random, such as a selection of \(n\) items that
fills n-space more uniformly than uncorrelated random items, sometimes also called a low-discrepancy sequence. (The outputs of quasi-random sequences are generally constrained by a low-discrepancy requirement that has a net effect of points being generated in a highly correlated manner, i.e., the next point "knows" where the previous points are).

[0061] As used herein, the word "module" refers to logic embodied in hardware or firmware, or to a collection of software instructions, possibly having entry and exit points, written in a programming language, such as, for example C++. A software module may be compiled and linked into an executable program, installed in a dynamic link library, or may be written in an interpretive language such as BASIC. It will be appreciated that software modules may be callable from other modules or from themselves, and/or may be invoked in response to detected events or interrupts. Software instructions may be embedded in firmware, such as an EPROM or EEPROM. It will be further appreciated that hardware modules may be comprised of connected logic units, such as gates and flip-flops, and/or may be comprised of programmable units, such as programmable gate arrays or processors. The modules described herein are preferably implemented as software modules, but may be represented in hardware or firmware. It is contemplated that the modules may be integrated into a fewer number of modules. One module may also be separated into multiple modules. The described modules may be implemented as hardware, software, firmware or any combination thereof. Additionally, the described modules may reside at different locations connected through a wired or wireless network, or the Internet.

[0062] In general, it will be appreciated that the processors can include, by way of example, computers, program logic, or other substrate configurations representing data and instructions, which operate as described herein. In other embodiments, the processors can include controller circuitry, processor circuitry, processors, general purpose single-chip or multi-chip microprocessors, digital signal processors, embedded microprocessors, microcontrollers and the like.

[0063] Furthermore, it will be appreciated that in one embodiment, the program logic may advantageously be implemented as one or more components. The components may advantageously be configured to execute on one or more processors. The components include, but are not limited to, software or hardware components, modules such as software modules,
object-oriented software components, class components and task components, processes
methods, functions, attributes, procedures, subroutines, segments of program code, drivers,
firmware, microcode, circuitry, data, databases, data structures, tables, arrays, and variables.

[0064] The foregoing description is provided to enable a person skilled in the art to
practice the various configurations described herein. While the subject technology has been
particularly described with reference to the various figures and configurations, it should be
understood that these are for illustration purposes only and should not be taken as limiting the
scope of the subject technology.

[0065] There may be many other ways to implement the subject technology. Various
functions and elements described herein may be partitioned differently from those shown without
departing from the scope of the subject technology. Various modifications to these
configurations will be readily apparent to those skilled in the art, and generic principles defined
herein may be applied to other configurations. Thus, many changes and modifications may be
made to the subject technology, by one having ordinary skill in the art, without departing from
the scope of the subject technology.

[0066] It is understood that the specific order or hierarchy of steps in the processes
disclosed is an illustration of exemplary approaches. Based upon design preferences, it is
understood that the specific order or hierarchy of steps in the processes may be rearranged. Some
of the steps may be performed simultaneously. The accompanying method claims present
elements of the various steps in a sample order, and are not meant to be limited to the specific
order or hierarchy presented.

[0067] As used herein, the singular forms "a," "an" and "the" include plural
references unless the content clearly dictates otherwise.

[0068] The term "about," as used herein, can refer to +/- 10% of a value.

[0069] Furthermore, to the extent that the term "include," "have," or the like is used
in the description or the claims, such term is intended to be inclusive in a manner similar to the
term "comprise" as "comprise" is interpreted when employed as a transitional word in a claim.

[0070] The word "exemplary" is used herein to mean "serving as an example,
instance, or illustration." Any embodiment described herein as "exemplary" is not necessarily to
be construed as preferred or advantageous over other embodiments.
[0071] A reference to an element in the singular is not intended to mean "one and only one" unless specifically stated, but rather "one or more." Pronouns in the masculine (e.g., his) include the feminine and neuter gender (e.g., her and its) and vice versa. The term "some" refers to one or more. Underlined and/or italicized headings and subheadings are used for convenience only, do not limit the subject technology, and are not referred to in connection with the interpretation of the description of the subject technology. All structural and functional equivalents to the elements of the various configurations described throughout this disclosure that are known or later come to be known to those of ordinary skill in the art are expressly incorporated herein by reference and intended to be encompassed by the subject technology. Moreover, nothing disclosed herein is intended to be dedicated to the public regardless of whether such disclosure is explicitly recited in the above description.

Examples

[0072] Aspects of the invention now being generally described, it will be more readily understood by reference to the following examples, which are included merely for purposes of illustration of certain aspects and embodiments of the present disclosure, and are not intended to limit the invention.

EXAMPLE 1: Genotyping and inference of IBD segments

[0073] Some aspects of the present disclosure employ a likelihood ratio test for which the data are the number and lengths of autosomal genomic segments shared between two individuals, with segment length measured in centiMorgans (cM). The null hypothesis is that the individuals are no more related than two persons picked at random from the population; the alternative hypothesis is that the two individuals share recent ancestry. When the alternative model is not significantly more likely than the null model, it is concluded that there is no evidence for recent shared ancestry. Otherwise, the maximum-likelihood estimate for the degree of relationship between two individuals by maximizing the likelihood over all possible relationships is obtained in the alternative model. Significance levels and confidence intervals are determined from standard chi-square approximations for the likelihood ratio test.

[0074] An embodiment of ERSA according to the present disclosure was applied to three well-defined pedigrees with predominantly Northern European ancestry (Table 1). Informed
consent was obtained from all study subjects, and all procedures were approved by the Western Institutional Review Board. DNA samples were collected and purified from blood as described in Xing et al. (2010). Affymetrix 6.0 SNP arrays were used to genotype 169 individuals selected from these pedigrees (Table 1), per the manufacturer's instructions (see Xing et al. 2010). Beagle 3.2 (Browning and Browning 2010) was used to phase and impute missing genotypes, using the Affymetrix 6.0 SNP genotypes of the 30 HapMap CEU trios as a reference (CEL files provided by Affymetrix). Of 868,155 autosomal SNP loci with unique positions on the array (not including controls, whose probe set IDs begin with 'AFFX-SNP'), 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. On the basis of the pedigree genotypes, GERMLINE 1.4.1 (Gusev et al. 2009; software available on Columbia University's Computer Science webpage (Gusev; GERMLINE)) inferred the locations and extents of IBD segments for all pairs of individuals (parameters err_het = 2, err_hom = 1, and min_m = lCM, with marker positions given on the HapMap r22 genetic map). GERMLINE identifies short regions of exact matches between haplotypes using a library of short seeds, then extends and merges those regions using an efficient hashing and matching algorithm. ERSA was applied to the output of GERMLINE. The program fastIBD in Beagle vers. 3.3 (Browning, University of Washington website) was also used to generate IBD segments for analysis by ERSA (default options). Although principal component analysis (Figure 2A) can distinguish the closely-related HapMap CEU and TSI sample sets, the pedigree and HapMap CEU samples are indistinguishable.

Methods
A. Null Hypothesis

[0075] The likelihood of the null hypothesis is estimated from the empirical distribution of autosomal shared segments in the population. Only shared segments longer than a given threshold, \( t \), are considered because shorter segments are more difficult to detect and provide little information about recent ancestry. Let \( s \) equal the set of segments shared between two individuals and \( n \) equal the number of elements in \( s \). For this calculation, it is assumed that the number of segments shared and the length of each segment are independent, which is
approximately true for the HapMap CEU population (see Figure 2D). The likelihood of the null hypothesis is:

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

where

2. \[ S_p(s \mid t) = \prod_{i} F_p(i \mid t). \]

[0076] \( N_p(n \mid t) \) is the likelihood of sharing \( n \) segments, \( S_p(s \mid t) \) is the likelihood of the set of segments \( s \), and \( F_p(i \mid t) \) is the likelihood of a segment of length \( i \). \( N_p(n \mid t) \) is approximated from a Poisson distribution with mean equal to the sample mean of the number of segments shared in the population (Figure 2B). Under a model of random mating and complete ascertainment of shared segments, \( F_p(i \mid t) \) specifies a geometric distribution, for which an exponential approximation is substituted.

[0077] The variable \( t \) is set to the smallest value that can achieve a false-negative rate of 1% or lower. This setting maximizes the use of available data while ensuring that the exponential approximation to the distribution of segment lengths in the population holds. Here, the choice of \( t = 2.5 \) cM was based on GERMLINE’s previously reported false-negative rate of 1% for segments 2.5 cM and longer (Gusev et al. 2009). In the HapMap CEU population, the distribution of segments detected by GERMLINE that are longer than 2.5 cM is approximately exponential, with the exception of a few significant outliers (Figure 2C). These outlying segments (those longer than \( h = 10 \) cM) are excluded when estimating the population distribution of shared segment lengths for two reasons. First, the outliers are inconsistent with the assumption of random mating used in the approximation. Second, the outliers are examples of shared recent ancestry, and including them in the population distribution would decrease the power to detect recent ancestry. Therefore, \( F_p(i \mid t) \) is approximated from the maximum likelihood estimate of the mean of a truncated exponential distribution:

3. \[ F_p(i \mid t) = \frac{e^{-(i-t)\theta}}{\theta}. \]
where Θ is equal to the mean shared segment length in the population for all segments of size greater than t and less than h. For HapMap CELT with t = 2.5 cM and h = 10 cM, the estimate of Θ is 3.12 cM.

B. Alternative Hypothesis

The alternative hypothesis is that the pair of individuals share either one or two recent ancestors. Let a represent the number of ancestors shared, and let d equal the combined number of generations separating the individuals from their ancestors(s), e.g., d = 6 and a = 1 for half-second cousins. Under the alternative hypothesis, segments shared by two individuals come from two sources: recent ancestry and the population background (denoted by subscripts A and P, respectively). Let \( n_p + n_A = n \), where \( n_A \) is equal to the number of shared segments inherited from recent ancestors, and \( rip \) is the number of segments shared due to the population background. \( sp \) and \( s_A \) are two mutually exclusive subsets of s, with \( s_A \) equal to the subset of segments inherited from recent ancestor(s) with \( n_A \) elements and \( sp \) equal to the subset of segments shared due to the background with \( rip \) elements. The likelihood of the alternative hypothesis of recent ancestry, \( L_R \), is then:

\[
L_R = L_A(n_A, s_A \setminus d, a, t) L_P(n_s, s_p \setminus t).
\]

Because \( sp \) is distributed according to the population distribution, \( L_P \) follows the description in Eq. 1. \( L_A \) is the likelihood that two individuals share \( n \) autosomal segments from recent ancestor(s) specified by \( d \) and \( a \), with the segment lengths specified by \( s_A \). \( L_A \) can be expressed as the product of likelihoods of the number of shared segments and the length of each segment, which parallels Eqs. 1 and 2:

\[
L_A(n_A, s_A \setminus d, a, t) = N_A(n \setminus d, a, t) \cdot s_A(s_A \setminus d, t).
\]

\[
S_A(s \setminus d, t) = \prod_{i \in 2} F_A(i \setminus t).
\]

Eq. 6 assumes that, for a given value of \( d \), the lengths of segments are independent. This assumption is not strictly true. One might imagine that the presence of a particularly long segment would reduce the genomic space available for additional segments. However because the length of any one segment is small relative to the length of the genome, and
because the genome is physically divided into chromosomes, the segment lengths are approximately independent (Thomas et al. 1994).

For two individuals who are related by an inheritance path that is $d$ meioses long, the probability that they will inherit any particular autosomal segment from a common ancestor on that path is equal to $1/2^{d-1}$. The expected number of shared autosomal segments that could potentially be inherited from a common ancestor is equal to $rd+c$, where $c$ is the number of autosomes and $r$ is the expected number of recombination events per haploid genome per generation. Therefore, the expected number of shared segments is equal to $a(rd+c)/2^{d-1}$ (Thomas et al. 1994). In humans, $c$ is equal to 22 and $r$ is approximately 35.3 (McVean et al. 2004). Given $d$, the expected value of $i$ is $100d$. Without conditioning on $t$, the distribution of segment length is exponential with mean $100/d$. Conditioning on $t$,

$$F_s(i \mid d,t) = \frac{e^{-d(i-1/100)}}{100/a}.$$  

The probability that a shared segment is longer than $t$, $p(t)$, is equal to $e^{-dt/100}$ (Thomas et al. 1994). Because the distribution of the number of shared segments is approximately Poisson (Thomas et al. 1994),

$$N_A(n \mid d,a,t) = e^{-\frac{a(rd+c)p(t)}{2^{d-1}}} \left(\frac{a(rd+c)p(t)}{2^{d-1}}\right)^n.$$  

Given $n_A$ and $np$, the maximum value of the likelihood function (Eq. 4) is equal to:

$$ML_R(n_p,n_A,s \mid d,a,t) = N_A(n_A,t)N_A(n_A \mid d,a,t) \cdot$$  

$$S_p(\{s_{n_A} \cdot \ s_{n_A} \cdot \ s_{n_A} \cdot \ s_{n_A} \cdot \}) \cdot t \cdot \left(\left\{s_{n_A} \cdot s_{n_A} \cdot s_{n_A} \cdot s_{n_A} \cdot \right\}\right)$$  

where $s_{x:n}$ is equal to the $x$th smallest value in $s$. Eq. 9 asserts that the likelihood is maximized when the set of segments resulting from recent ancestry is equal to the longest $n_A$ segments in $s$, with the remaining $np$ segments being due to the population background.

The alternative model contains three additional parameters relative to the null model, $d$, $a$, and $n_A$ ($np = n - n_A$). However, when the behavior of $d$ and $a$ was evaluated
empirically, it was found that they effectively act as a single parameter (Figure S6). Therefore, the ratio of Eq. 1 and Eq. 9 was evaluated using a $\chi^2$ approximation with two degrees of freedom ($-2\ln[L_R/L_N] \sim \chi^2$). For closely related individuals, the distribution of $N_R(n, p, t)$ should theoretically be adjusted to account for segments shared from the population background that could not be observed because they occur within longer segments shared due to recent ancestry. Although ERSA optionally includes this adjustment, the algorithm performs slightly better without the adjustment due to the occasional imprecise definition of very long IBD segments in GERMLINE. To identify the maximum value of the likelihood function (Eq. 4) given $d$, $a$, and $t$, all possible values of $n_p$ and $n_A$ are evaluated in Eq. 9:

10. $ML_R(n,s \mid d,a,t) = \max \{ML^\wedge_R(n_p,n - n_p,s); n_p \in \{0,1,\ldots\} \}$

a. Individuals ascertained based on a shared genetic variant

[0087] If the two individuals have been ascertained because they both share the same genetic variant, as in the case of a shared disease-causing variant, then the likelihood calculation must be conditioned on this ascertainment. In the case of such ascertainment, the shared segment that contains the variant is equivalent to two shared segments, with the segment boundaries defined by the original boundaries and the location of the ascertained variant. (Thomas et al. 2008; Thomas et al. 1994) Thomas et al. have shown that the lengths of these segments, $g_1$ and $g_2$, are exponentially distributed, with the mean equal to the unconditional length of a segment. Excluding the ascertained segment from $n$ and $s$, the maximum value of the likelihood function is equal to:

$$AML_R(n,s,g_1,g_2 \mid d,a,t) = ML_R(n,s \mid d,a,t) - \max \{ \lambda^\wedge S_A (\{g_1,g_2\} \setminus d,a,t^\wedge) \}$$

11. Proof of Equation 9

[0088] Equation 9 holds as long as $\theta \leq a(rd+c)$, which is true whenever $a$ and $d$ specify shared ancestry that is recent relative to pairs of individuals selected at random from the population. Given a set of shared segment lengths between two individuals, $s$, the objective is to identify the subset of these segments, $m$, containing the $n_A$ elements that are most likely to have been inherited from recent ancestor(s). Eq. 9 assumes that $m$ is equal to the largest $n_A$ elements in $s$. Here, it is shown why this assumption holds: Let $\theta t = 100/d$, which is the expected length of a
shared segment inherited from a recent ancestor. Let \( \Theta = \Theta \) which is the expected length of a shared segment if it is not inherited from a recent ancestor. If the average time to the most recent common ancestor between individuals in the population is greater than \( d/2 \), then \( \Theta > \Theta \). If \( \Theta < \Theta \), then individuals selected at random from the population are more closely related than the relationship being analyzed, and therefore there is no power to detect a relationship.

To demonstrate that \( m \) is equal to the set containing the largest \( m \) elements of \( s \), consider two mutually exclusive subsets of \( s \), \( z_p \) and \( z_A \), with \( z_A \) containing \( m \) elements. Let \( x_i \) equal the largest element in \( z_p \) and \( x_2 \) equal the smallest element in \( z_A \). Let \( y_p \) and \( y_A \) respectively equal the sets \( z_p \) and \( z_A \) with the exception that \( x_i \) and \( x_2 \) are swapped. As long as \( x_i > x_2 \), the likelihood of \( z_p \) and \( z_A \) is less than the likelihood of \( y_p \) and \( y_A \):

\[
L_p(n_p, n_A, z_A, z_p) < L_p(n_p, n_A, z_A, z_p).
\]

The components of \( L_R \) are \( N_A, N_p, S_A, \) and \( S_p \). Because \( N_A \) and \( N_p \) depend only on \( n_p \) and \( n_A \), the above condition simplifies to:

\[
S_p(y_p \setminus z_A) S_A(y_A \setminus z_A, z_p) < S_p(y_p \setminus z_A) S_A(z_A \setminus z_p).
\]

The elements in both \( z_p \) and \( z_A \), and \( y_p \) and \( y_A \), are equal, with the exception of \( x_i \) and \( x_2 \). Therefore, by Eq. 6, the inequality becomes

\[
F_p(x_2 I_r) F_A(x_1 I_d, a, t) < F_p(x_1 I_r) F_A(x_2 I_d, a, t).
\]

which (by Eqs. 3 and 7) is equal to

\[
\frac{1}{e^\frac{\theta_1}{\theta_2}} - \frac{1}{e^\frac{\theta_2}{\theta_1}} \cdot \frac{x_1}{x_2} \cdot \frac{\theta_1}{\theta_2}.
\]

This simplifies to

\[
\frac{x_2 - x_1}{\theta_2} < \frac{x_1 - x_2}{\theta_1}.
\]

Q.E.D.

D. **Parameters \( d \) and \( a \) in the likelihood ratio test**

Although \( d \) and \( a \) are specified as two separate parameters in the likelihood ratio test, analyses indicated that allowing \( a \) to vary has almost no effect on the distribution of likelihood scores under the null hypothesis. To demonstrate this behavior, the likelihood scores for pairs of individuals from two closely-related populations, the CHB (45 Han Chinese in Beijing) and JPT (45 Japanese in Tokyo) samples, were evaluated using the HapMap phase 2 SNP genotype data (HapMap Consortium 2005). For each pair of individuals, the maximum
likelihood for two alternative models \((Li \text{ and } Li)\) was calculated. In model 1, \(a\) is allowed to vary, and in model 2, \(a\) is fixed equal to 2 (\(d\) is estimated in both). To evaluate the effect of allowing \(a\) to vary, a likelihood ratio test (LRT) statistic for the two models \((-2 \ln[L_i/L_2]\) was calculated; Figure S6, blue ("Observed" line). For comparison, the expected cumulative distribution of a \(\chi^2\) with one degree of freedom was calculated (red). As the cumulative distribution illustrates, all of the observed LRT values are less than \(10^{-4}\), indicating that there is very little difference between the likelihoods of the two models. Thus \(d\) and \(a\) can be treated as a single parameter when applying the \(\chi^2\) approximation to the likelihood ratio test statistic.

**Results**

[0096] The performance of ERSA was assessed by analyzing high-density SNP microarray data on three deep, well-defined pedigrees composed of 24, 30, and 115 individuals (Table 1). The output from this analysis was a maximum-likelihood estimate and confidence interval (C.I.) for the degree of relationship of each pair of individuals in the sample. The computation time taken by ERSA to analyze all 14,196 pairs of individuals in this sample was approximately 9 minutes running on one core of a 2.3 GHz AMD Opteron processor. In Figure 3 presents results for all 2,677 known pairs of first- through twelfth-degree relatives with exactly two known common ancestors in the pedigree and for which the two inheritance paths between the individuals have the same length (e.g., full sibs, full cousins). Results for relatives with exactly one common ancestor (e.g., half cousins) were qualitatively similar (see Figure SI).

[0097] For pairs of individuals as distantly related as eighth-degree relatives, ERSA's estimates are generally accurate to within one degree of the known relationship. ERSA predicted the exact degree of relationship for 66% of the 549 pairs of first- through fifth-degree relative and was accurate to within one degree of relationship for 97% of those pairs (Figure 3 and Table SI). Point estimates were accurate to within one degree of relationship for more than 80% of sixth- and seventh-degree relatives, and 60% of eighth-degree relatives (Figure 3), but accuracy drops off rapidly beyond this point (Figure 3).

[0098] ERSA has nearly 100% power to detect first- through fifth-degree relatives and substantial power to detect ancestry as distant as eleventh-degree relatives. A significant relationship was detected among all 549 pairs of first- through fifth-degree relatives in the
sample $a = 0.001$, where the null hypothesis is no relationship (Figure 4). Although the power to detect more distant ancestry is constrained by the fact that distant relatives often share no genetic material (Donnelly 1983), ERSA retains relatively high power for these relationships. Eighty-eight percent of seventh-degree relatives, 44% of ninth-degree relatives, and 12% of eleventh-degree relatives were detected at a significance level of 0.001 (red line in Figure 4), which closely approaches the maximum theoretical power (black line in Figure 4).

For comparison, the same relationships were analyzed by applying RELPAIR (Epstein et al. 2000) and GBIRP (Stankovich et al. 2005) to a subset of the SNP loci (see Figures 4 and S2). Both methods had high power to detect third- and fourth-degree relatives (dotted and solid blue lines in Figure 4), although RELPAIR reports all relationships beyond second degree as simply "cousins" (i.e., more distant than second degree). The power of RELPAIR and GBIRP drops off rapidly beyond fourth-degree relationships, approximately three degrees before ERSA's power begins to decline (Figure 4).

As shown in Table 2, ERSA's probability of detecting a significant relationship between unrelated individuals (the empirical false positive rate) is approximately equal to the nominal significance level ($a$). To estimate the empirical false positive rate, high-density SNP data on a set of individuals with no recent shared ancestry was needed. Given the sensitivity of ERSA to distant relationships, acquiring an appropriate dataset from pedigree data would require complete ancestry information for each individual in the sample extending back at least seven generations. Because such pedigrees are extremely rare, the false positive rate from two closely related populations, the CHB (45 Han Chinese in Beijing) and JPT (45 Japanese in Tokyo) samples, using the HapMap phase 2 SNP genotype data was estimated (HapMap Consortium 2005). Because these populations can be distinguished genetically (HapMap Consortium 2005), estimating the false positive rate from the CHB-JPT comparison is not ideal. However, the allele frequency and haplotype distributions of these populations are very similar (HapMap Consortium 2005), and pairs of CHB and JPT individuals are unlikely to have shared an ancestor in the past 200 years. Therefore, false-positive rates from the proportions of CHB-JPT pairs in which significant recent ancestry was detected was estimated. The estimated false positive rates closely matched the nominal rates (Table 2). For the significance level of $a = 0.001$
used in Figures 3 and 4, the estimated false positive rate was 0.0005 (95% C.I. 1.3 x 10\(^{-5}\) to 0.0028).

ERSA can also accurately detect relationships between individuals who share a disease-causing mutation transmitted from a common founder. The process of ascertaining individuals based on a shared mutation introduces biases in the estimation of recent ancestry, but this bias can be taken into account (see Methods). The test case was composed of seven previously described individuals who are affected with attenuated familial adenomatous polyposis (AFAP) due to a single disease-causing mutation (c.426_427delAT in the APC gene; Neklason et al. 2008). The available pedigree information identified four pairs of these individuals as sixth-degree relatives and one pair as eighth-degree relatives. The point estimates from ERSA were accurate to within one degree of relationship for all five of these pairs.

Discussion

ERSA uses explicit IBD segment information to estimate the relationships between pairs of individuals in a maximum-likelihood framework. This makes better use of the information present in high-density SNP genotyping data, as shown by the power curves in Figure 4. The power of aspects of the instant invention to detect relationships between second cousins or closer relatives is essentially perfect and exceeds 85% for third cousins even at the \(\alpha = 0.001\) level. ERSA is also more accurate than RELPAIR or GBIRP (Figure S2 and Table S1). Beyond third cousins, genetic methods inherently become more limited by the fact that two individuals with a common genealogical ancestor frequently do not share any genetic material inherited from that ancestor: such genealogical links cannot be directly detected by genetic methods. This limitation is illustrated in Figure 4, which demonstrates that ERSA's power decreases in lockstep with the maximum theoretical power as the degree of relationship increases.

Because denser and more accurate genetic data will improve the ability to detect and delineate IBD segments, it is expected that the accuracy of IBD segment inference will improve as whole-genome sequencing becomes more affordable and as higher-density microarrays become available. In addition, while the IBD segment detection methods used here
(GERMLINE; Gusev et al. 2009; fastIBD in Beagle 3.3) perform well, further improvements are expected as phasing and imputation methods advance (e.g., Genovese et al. 2010).

ERSA detects recent shared ancestry by identifying an excess of IBD segment-sharing relative to the population background. Therefore, the power to detect shared ancestry between individuals depends on the demographic history of the population to which those individuals belong. If the population size is small, or if the population has experienced a founder effect or recent bottleneck, then the level of IBD segment-sharing among unrelated individuals will increase. In such populations, ERSA's power to detect distant relationships will be diminished.

EXAMPLE 2: Estimating recent ancestry in admixed populations

The pedigree samples analyzed in Example 1 are from a homogeneous population. As shown here, it is predicted that ERSA will retain its high detection power in admixed populations.

Analysis of the European samples of Example 1 demonstrates that ERSA performs well in a homogeneous population with no history of recent admixture from a more distantly related population. Because pedigree data for an admixed population was not available, ERSA's performance in the presence of admixture could not be directly analyzed. Impacts of admixture on ERSA's performance would most likely be mediated through effects on the expected distributions of the number and lengths of IBD segments shared between unrelated individuals. Admixture should reduce the number and lengths of such segments. The reasoning for this expected reduction is as follows. The detection of IBD segments is based largely on long runs of consecutive loci at which the genotypes are consistent with identity-by-state (IBS). Admixture will introduce alleles that are frequently IBS among pairs of individuals in the population due to shared ancestry. However, in the absence of founder effect, given that two admixed individuals are of identical ancestry at a particular genomic segment, they are no more likely to share long runs of IBS than individuals chosen at random from the appropriate reference population. When individuals are not required to share ancestry at any particular genomic segment (as would be the case for ascertainment for a shared genetic disease), it results in an
expectation of fewer and smaller shared segments among unrelated individuals relative to at least one of the reference populations.

This prediction was tested by comparing individuals from a sample of 25 Bolivian individuals genotyped on Affymetrix SNP 6.0 arrays (Xing et al. 2010). Substantial European admixture (19-41%; data not shown) in 9 Bolivians was identified using the Admixture software (Alexander et al. 2009). The Bolivian population was divided into groups with and without admixture. All non-admixed Bolivians were estimated to have < 0.1% admixture. The same process was then applied to identify shared segments in the European sample, i.e., using Beagle (Browning and Browning 2010) to phase and impute the data and GERMLINE (Gusev et al. 2009) to identify all shared segments longer than 2.5 cM. Consistent with predictions, on average, the admixed Bolivians shared 43 segments (95% C.I. 41-45 segments) with an average size of 3.5 cM (95% C.I. 3.4-3.7 cM), compared to 88 segments (95% C.I. 86-92 segments) with an average size of 4.2 cM (95% C.I. 4.1-4.3 cM) in non-admixed Bolivians.

In comparisons of distantly-related admixed individuals, the smaller expected number and size of background segments could slightly improve ERSA’s detection power: short but meaningful shared IBD segments could become statistically significant when compared to a shorter background size distribution. In comparisons of distantly-related individuals with ancestries mostly confined to one of the reference populations, however, the admixed population background distributions would be incorrect. Using them might cause ERSA to suffer a slightly increased false positive rate or a bias towards overestimating the degree of relationship due to the misattribution of some short background segments to a distant relationship.

EXAMPLE 3: Inferring first-degree relationships

Many existing methods for detecting IBD segments do not distinguish segments that overlap on homologous chromosomes, and rather than consider them to be separate, merge them into one (see Figure S5). For two or more degrees of relationship, Eqs. 7 and 8 provide close approximations to the results of this procedure (Thomas et al. 2008). However, in the case of full siblings, Eq. 7 systematically overestimates the number of detected shared segments, and Eq. 8 systematically underestimates the length of the merged segment.
Therefore, for \( d = 2 \) and \( a = 2 \), the calculation for \( N_A \) and \( F_A \) was adjusted to account for shared segments that have been bioinformatically merged:

\[
N_A(n \mid d = 2, a = 2) = \frac{e^{\frac{3}{4} + 2d\frac{3}{4}}}{n!} \left[ \frac{3}{4} - 2d - \frac{1}{4} \right]^n.
\]

\[
S1.
F_A(i \mid d = 2, a = 2, t) = \frac{(i-t)^{k-1} e^{-d(i-t)/\hat{k} \hat{d}}}{(100/d)^t \hat{k} \hat{k}^{-1}}.
\]

S2. 

where \( \hat{k} \) is the maximum likelihood estimate for the number of merged segments. Because Eq. S2 introduces additional estimated parameters into the full-sibling model, ERSA only reports the full-sibling model as the maximum likelihood estimate if it is significantly more likely than all other models at the 0.05 level.

ERSA is designed to detect ancestry at a single node in a pedigree; incorporating information about human biodiversity (HBD) would result in a near-perfect detection of full sibling relationships, but would have little to no effect on estimates of other relationships. HBD information will be incorporated into future evaluation of full-sibling models as the tools for IBD and HBD segment detection improve.

Many existing IBD methods are also unable to detect the recombination breakpoints between parent-offspring pairs and usually report the length of each entire chromosome as a shared segment (Gusev et al. 2009; Thomas et al. 2008). With this detection scheme, a probabilistic description of the number and size of shared segments is no longer appropriate. Therefore, to identify parent-offspring relationships, a different statistic, the total proportion of the genome shared between the two individuals, was considered. A sibling relationship is rejected in favor of a parent-offspring relationship when the proportion of the genome shared exceeds a specified significance level for siblings (default is 0.01). ERSA includes options to bypass Eqs. S1, S2, and/or the parent-offspring option for situations where the overlapping segments can be accurately identified.
REFERENCES


WHAT IS CLAIMED IS:

1. A method of estimating genetic relatedness between members of a first pair of conspecific organisms, the method comprising:

   receiving, by a processor, a value indicating a number of nonoverlapping polynucleotide segments longer than a threshold length \( t \) that are identical, by at least about 90 percent sequence identity, between members of the first pair;

   receiving, by a processor, values indicating lengths of the identical segments;

   comparing the number of the first pair's identical segments to a number of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a second pair of organisms, the members of the second pair having an established degree of genetic relatedness to each other;

   comparing the lengths of the first pair's identical segments to lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a third pair of organisms, the members of the third pair having an established degree of genetic relatedness to each other;

   based on the number comparison and the length comparison, estimating, by a processor, a degree of genetic relatedness between the members of the first pair.

2. The method of claim 1, wherein the members of the first pair are human.

3. The method of claim 1, wherein the first pair's polynucleotide segments comprise DNA.

4. The method of claim 1, wherein the first pair's polynucleotide segments comprise mitochondrial DNA.

5. The method of claim 1, wherein the first pair's polynucleotide segments comprise sex-linked nucleotide segments.
6. The method of claim 1, wherein the first pair's polynucleotide segments comprise RNA.

7. The method of claim 1, wherein \( t \) is equal to or greater than about 2.5 cM.

8. The method of claim 1, further comprising:
   comparing the lengths of the first pair's identical segments to a background distribution of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and

   wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the background distribution.

9. The method of claim 8, wherein the identical segments of the background group are no longer than about 10 cM.

10. The method of claim 8, wherein members of the background group are selected randomly from a larger population.

11. The method of claim 1, further comprising:
   comparing the number of the first pair's identical segments to a first distribution, of numbers of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and

   wherein the estimating is further based on the comparison of the number of the first pair's identical segments to the numbers in the first distribution.
12. The method of claim 1, further comprising:
   comparing the lengths of the first pair's identical segments to a second distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a second group, the members of each pair in the second group having an established degree of genetic relatedness to each other;
   wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the second distribution.

13. The method of claim 11, further comprising:
   comparing the lengths of the first pair's identical segments to a second distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a second group, the members of each pair in the second group having an established degree of genetic relatedness to each other;
   wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the second distribution.

14. The method of claim 13, further comprising:
   comparing the lengths of the first pair's identical segments to a background distribution, of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and
   wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the background distribution.

15. The method of claim 1, 13, or 14, wherein the estimating further comprises estimating a likelihood \( L_p \) that the first pair are no more related than two individuals selected randomly from a population, wherein:
\[ L_p(n, s \mid t) = N_p(n \mid t) \cdot S_p(s \mid t); \]

wherein
\[ S_p(s \mid t) = \prod_{i \in s} F_p(i \mid t); \]

wherein \( N_p(n \mid t) \) comprises the likelihood of sharing \( n \) segments, \( S_p(s \mid t) \) comprises the likelihood of the set of segments \( s \), and \( F_p(i \mid t) \) comprises the likelihood of a segment of size \( i \).

16. The method of claim 15, wherein \( F_p(i \mid t) \) is approximated as:
\[ F_p(i \mid t) = e^{-(i-t)\theta} / \Theta; \]

wherein \( \Theta \) is equal to a mean shared segment size in the population for all segments of size greater than \( t \) and less than a maximum length.

17. The method of claim 16, wherein the maximum length is about 10 cM.

18. The method of claim 15, wherein the estimating further comprises estimating a likelihood \( L_R \) that the first pair share one or two ancestors, wherein:
\[ L_R = L_A(n_A, s_A \mid d, a, t) L_p(s_p \mid t); \]

wherein \( n_P + n_A = n \), where \( n_A \) is equal to the number of shared segments inherited from ancestors, \( n_P \) is the number of segments shared by the population;

wherein \( s_P \) and \( s_A \) are two mutually exclusive subsets of \( s \), where \( s_A \) is the subset of segments inherited from ancestor(s) with \( n_A \) elements, and \( s_P \) is the subset of segments shared by the population with \( n_P \) elements;

wherein \( a \) represents the number of ancestors shared, and \( d \) represents the combined number of generations separating the individuals from their ancestor(s).
19. The method of claim 15, wherein the estimating further comprises estimating a likelihood $L_A$ that the first pair share $n$ segments from ancestor(s) specified by $d$ and $a$, with the segment sizes specified by $\frac{3}{4}$, wherein:

$$L_A(n_A, s_A \setminus d, a, t) = N_A(n \setminus d, a, t) - S_A(s_A \setminus d, t);$$

wherein

$$S_A(s \setminus d, t) = \prod_{i \in t} F_A(i \mid t);$$

wherein $N_A(n \setminus d, a, t)$ is the likelihood of sharing $n$ segments, $S_A(s_A \setminus d, t)$ is the likelihood of the set of segments $s_A$, and $E_A(i \mid t)$ is the likelihood of a segment of size $i$;

wherein $s_p$ and $s_A$ are two mutually exclusive subsets of $s$, where $s_A$ is the subset of segments inherited from ancestor(s) with $n_A$ elements, and $s_p$ is the subset of segments shared by the population with $n_p$ elements;

wherein $n_p + n_A = n$, where $n_A$ is equal to the number of shared segments inherited from ancestors, $n_p$ is the number of segments shared by the population;

wherein $a$ represents the number of ancestors shared, and $d$ represents the combined number of generations separating the individuals from their ancestor(s).

20. The method of claim 18, wherein the estimating further comprises estimating a likelihood $L_A$ that the first pair share $n$ segments from ancestor(s) specified by $d$ and $a$, with the segment sizes specified by $\frac{3}{4}$, wherein:

$$L_A(n_A, s_A \setminus d, a, t) = N_A(n \setminus d, a, t) - S_A(s_A \setminus d, t);$$

wherein

$$S_A(s \setminus d, t) = \prod_{i \in t} F_A(i \mid t);$$

wherein $N_A(n \setminus d, a, t)$ is the likelihood of sharing $n$ segments, $S_A(s_A \setminus d, t)$ is the likelihood of the set of segments $s_A$, and $E_A(i \mid t)$ is the likelihood of a segment of size $i$.

21. The method of claim 19 or 20, wherein:
wherein \( p(t) \) is the probability that a shared segment is longer than \( t \), \( c \) comprises an average number of chromosomes in the organisms, and \( r \) comprises an average number of recombination events per haploid genome in the organisms.

22. The method of claim 21, wherein \( p(t) \) is assumed to be equal to or about \( e^{-d/t^{d_0}} \).

23. The method of claim 19 or 20, wherein:

\[
F_A(i \mid d, t) = \frac{e^{-d(i - n^{d_0})}}{100/d}.
\]

24. The method of claim 18, wherein the estimating further comprises estimating a maximum likelihood of \( L_R(ML_R) \), wherein:

\[
ML_R(n_p, n_A, s \mid d, a, t) = N_p(n_p \mid t)N_A(n_A \mid d, a, t) - S_p(\{s_{n_p} \ldots s_{n_A} \mid t\})S_A(\{s_{n_p} \ldots s_{n_A} \mid d, a, t\});
\]

where \( s_{x,n} \) is equal to the \( x \)th smallest value in \( s \).

25. The method of claim 24, further comprising evaluating, by a processor, a ratio of \( ML_R(n_p, n_A, s \mid d, a, t) \) and \( L_p(n, s \mid t) \) using a chi-square approximation with two degrees of freedom.

26. The method of claim 18, wherein the estimating further comprises estimating a maximum likelihood of \( L_R(ML_R) \), wherein:

\[
ML_R(n, s \mid d, a, t) = \text{Max}\{MLR(n_p \mid n - n_p, s) : n_p \in \{0 \ldots \}}\}.
\]
27. The method of claim 1, further comprising:

   receiving, by a processor, values indicating locations of the identical segments;

   comparing the locations of the first pair's identical segments to locations of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a fourth pair of organisms, the members of the fourth pair having an established degree of genetic relatedness to each other; and

   wherein the estimating is further based on the location comparison.

28. The method of claim 28, further comprising:

   comparing the locations of the first pair's identical segments to a background distribution of locations of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and

   wherein the estimating is further based on the comparison of the locations of the first pair's identical segments to the locations in the background distribution.

29. The method of claim 28, further comprising:

   comparing the locations of the first pair's identical segments to a first distribution, of locations of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and

   wherein the estimating is further based on the comparison of the locations of the first pair's identical segments to the locations in the first distribution.
30. A computer-readable medium encoded with a computer program comprising instructions executable by a processor for estimating genetic relatedness between members of a first pair of conspecific organisms, the instructions including instruction code for:

receiving, by a processor, a value indicating a number of nonoverlapping polynucleotide segments longer than a threshold length \( t \) that are identical, by at least about 90 percent sequence identity, between members of the first pair;

receiving, by a processor, values indicating lengths of the identical segments;

comparing the number of the first pair's identical segments to a number of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a second pair of organisms, the members of the second pair having an established degree of genetic relatedness to each other;

comparing the lengths of the first pair's identical segments to lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of a third pair of organisms, the members of the third pair having an established degree of genetic relatedness to each other;

based on the number comparison and the length comparison, estimating, by a processor, a degree of genetic relatedness between the members of the first pair.

31. The computer-readable medium of claim 30, wherein the members of the first pair are human.

32. The computer-readable medium of claim 30, wherein the first pair's polynucleotide segments comprise DNA.

33. The computer-readable medium of claim 30, wherein the first pair's polynucleotide segments comprise mitochondrial DNA.

34. The computer-readable medium of claim 30, wherein the first pair's polynucleotide segments comprise sex-linked nucleotide segments.
35. The computer-readable medium of claim 30, wherein the first pair's polynucleotide segments comprise RNA.

36. The computer-readable medium of claim 30, wherein \( t \) is equal to or greater than about 2.5 cM.

37. The computer-readable medium of claim 30, further comprising:
   comparing the lengths of the first pair's identical segments to a background distribution of lengths of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and
   wherein the estimating is further based on the comparison of the lengths of the first pair's identical segments to the lengths in the background distribution.

38. The computer-readable medium of claim 37, wherein the identical segments of the background group are no longer than about 10 cM.

39. The computer-readable medium of claim 37, wherein members of the background group are selected randomly from a larger population.

40. The computer-readable medium of claim 30, further comprising:
   comparing the number of the first pair's identical segments to a first distribution, of numbers of nonoverlapping polynucleotide segments longer than \( t \) that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and
   wherein the estimating is further based on the comparison of the number of the first pair's identical segments to the numbers in the first distribution.
41. The computer-readable medium of claim 30, further comprising:
comparing the lengths of the first pair's identical segments to a second
distribution, of lengths of nonoverlapping polynucleotide segments longer than τ that are
identical, by at least about 90 percent sequence identity, between members of pairs of
organisms in a second group, the members of each pair in the second group having an
established degree of genetic relatedness to each other;
wherein the estimating is further based on the comparison of the lengths of the
first pair's identical segments to the lengths in the second distribution.

42. The computer-readable medium of claim 40, further comprising:
comparing the lengths of the first pair's identical segments to a second
distribution, of lengths of nonoverlapping polynucleotide segments longer than τ that are
identical, by at least about 90 percent sequence identity, between members of pairs of
organisms in a second group, the members of each pair in the second group having an
established degree of genetic relatedness to each other;
wherein the estimating is further based on the comparison of the lengths of the
first pair's identical segments to the lengths in the second distribution.

43. The computer-readable medium of claim 42, further comprising:
comparing the lengths of the first pair's identical segments to a background
distribution, of lengths of nonoverlapping polynucleotide segments longer than τ that are
identical, by at least about 90 percent sequence identity, between members of pairs of
organisms in a background group, the members of most pairs in the background group
being more distantly related than fourth cousins; and
wherein the estimating is further based on the comparison of the lengths of the
first pair's identical segments to the lengths in the background distribution.
44. The computer-readable medium of claim 30, 42, or 43, wherein the estimating further comprises estimating a likelihood $L_p$ that the first pair are no more related than two individuals selected randomly from a population, wherein:

$$L_p(n, s \setminus t) = N_p(n \setminus t) - S_p(s \setminus t);$$

wherein

$$S_p(s \setminus t) = \prod_{i \in s} F_p(i \setminus t);$$

wherein $N_p(n \setminus t)$ comprises the likelihood of sharing $n$ segments, $S_p(s \setminus t)$ comprises the likelihood of the set of segments $s$, and $F_p(i \setminus t)$ comprises the likelihood of a segment of size $i$.

45. The computer-readable medium of claim 44, wherein $F_p(n \setminus t)$ is approximated as:

$$F_p(i \setminus t) = \frac{e^{-(i-t)/\Theta}}{\Theta};$$

wherein $\Theta$ is equal to a mean shared segment size in the population for all segments of size greater than $t$ and less than a maximum length.

46. The computer-readable medium of claim 45, wherein the maximum length is about 10 cM.

47. The computer-readable medium of claim 44, wherein the estimating further comprises estimating a likelihood $L_R$ that the first pair share one or two ancestors, wherein:

$$L_R = L_A(n_A, s_A \setminus d, a, t) L_p(s_p \setminus t);$$

wherein $n_{p+riA} = n$, where $n_A$ is equal to the number of shared segments inherited from ancestors, $n_{p}$ is the number of segments shared by the population;

wherein $sp$ and $s_A$ are two mutually exclusive subsets of $s$, where $s_A$ is the subset of segments inherited from ancestor(s) with $n_A$ elements, and $sp$ is the subset of segments shared by the population with $n_p$ elements;
wherein \(a\) represents the number of ancestors shared, and \(d\) represents the combined number of generations separating the individuals from their ancestor(s).

48. The computer-readable medium of claim 44, wherein the estimating further comprises estimating a likelihood \(L_A\) that the first pair share \(n\) segments from ancestor(s) specified by \(d\) and \(a\), with the segment sizes specified by \(\frac{3}{4}\), wherein:

\[
L_A(n_A, s_A \setminus d, a, t) = N_A(n \setminus d, a, t) - S_A(s_A \setminus d, t);
\]

wherein

\[
S_A(s \mid d, t) = \prod_{i \in s} F_A(i \mid t);
\]

wherein \(N_A(n \setminus d, a, t)\) is the likelihood of sharing \(n\) segments, \(S_A(s_A \setminus d, t)\) is the likelihood of the set of segments \(s_A\), and \(E_A(\mid il)\) is the likelihood of a segment of size \(i\);

wherein \(sp\) and \(\frac{3}{4}\) are two mutually exclusive subsets of \(s\), where \(\frac{3}{4}\) is the subset of segments inherited from ancestor(s) with \(n_A\) elements, and \(sp\) is the subset of segments shared by the population with \(np\) elements;

wherein \(n_{p} + n_A = n\), where \(n_A\) is equal to the number of shared segments inherited from ancestors, \(n_{p}\) is the number of segments shared by the population;

wherein \(a\) represents the number of ancestors shared, and \(d\) represents the combined number of generations separating the individuals from their ancestor(s).

49. The computer-readable medium of claim 47, wherein the estimating further comprises estimating a likelihood \(L_A\) that the first pair share \(n\) segments from ancestor(s) specified by \(d\) and \(a\), with the segment sizes specified by \(\frac{3}{4}\), wherein:

\[
L_A(n_A, s_A \setminus d, a, t) = N_A(n \setminus d, a, t) - S_A(s_A \setminus d, t);
\]

wherein

\[
S_A(s \mid d, t) = \prod_{i \in s} F_A(i \mid t);
\]
wherein \( N_A(n | d, a, t) \) is the likelihood of sharing \( n \) segments, \( S_A(s | A, d, a, t) \) is the likelihood of the set of segments \( \frac{n}{4} \), and \( F_A(n) \) is the likelihood of a segment of size \( i \).

50. The computer-readable medium of claim 48 or 49, wherein:

\[
N_A(n | d, a, t) = \frac{e^{\frac{a(rd + c)p(t)}{2^{d-1}}} \left[ \frac{a(rd + c)p(t)}{2^{d-1}} \right]^n}{n!};
\]

wherein \( p(t) \) is the probability that a shared segment is longer than \( t \), \( c \) comprises an average number of chromosomes in the organisms, and \( r \) comprises an average number of recombination events per haploid genome in the organisms.

51. The computer-readable medium of claim 50, wherein \( p(t) \) is assumed to be equal to or about \( e^{-dt/100} \).

52. The computer-readable medium of claim 48 or 49, wherein:

\[
F_A(i | d, t) = \frac{e^{-d(i-1)/100}}{100/d}.
\]

53. The computer-readable medium of claim 47, wherein the estimating further comprises estimating a maximum likelihood of \( L_R(ML_R) \), wherein:

\[
ML_R(n_p, n_A, s | d, a, t) = N_p(n_p, t) N_A(n_A, d, a, t) - \frac{S_p(\{s_{x, a} \ldots s_{x, a}\} | d, a, t)}{S_A(\{s_{x+1, a} \ldots s_{x, a}\} | d, a, t)};
\]

where \( s_{x, a} \) is equal to the \( x \)th smallest value in \( s \).

54. The computer-readable medium of claim 53, further comprising evaluating, by a processor, a ratio of \( ML_R(n_p, n_A, s | d, a, t) \) and \( L_p(n, s \setminus t) \) using a chi-square approximation with two degrees of freedom.
55. The computer-readable medium of claim 47, wherein the estimating further comprises estimating a maximum likelihood of $L_R(ML_R)$, wherein:

$$ML_R(n,s, L, d, a, t) = \max \{ML(n_p, n - n_p, s) : n_p \in \{0..n\}\}.$$  

56. The computer-readable medium of claim 30, further comprising:

receiving, by a processor, values indicating locations of the identical segments;

comparing the locations of the first pair’s identical segments to locations of nonoverlapping polynucleotide segments longer than $t$ that are identical, by at least about 90 percent sequence identity, between members of a fourth pair of organisms, the members of the fourth pair having an established degree of genetic relatedness to each other; and

wherein the estimating is further based on the location comparison.

57. The computer-readable medium of claim 56, further comprising:

comparing the locations of the first pair’s identical segments to a background distribution of locations of nonoverlapping polynucleotide segments longer than $t$ that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a background group, the members of most pairs in the background group being more distantly related than fourth cousins; and

wherein the estimating is further based on the comparison of the locations of the first pair’s identical segments to the locations in the background distribution.

58. The computer-readable medium of claim 56, further comprising:

comparing the locations of the first pair’s identical segments to a first distribution, of locations of nonoverlapping polynucleotide segments longer than $t$ that are identical, by at least about 90 percent sequence identity, between members of pairs of organisms in a first group, the members of each pair in the first group having an established degree of genetic relatedness to each other; and

-59-
wherein the estimating is further based on the comparison of the locations of the first pair's identical segments to the locations in the first distribution.
Figure 2

A

Principal Component 2 (1.02%)

Principal Component 1 (1.59%)

B

Proportion of pairs

Number of segments shared between a pair

C

Proportion of segments

Size of shared segment (cM)

D

Mean segment length

Number of segments shared between a pair
Figure 4

Estimated power to detect a relationship vs. Estimated degree of relationship.

- Maximum Theoretical Limit
- ERSA + GERMLINE $\alpha = 0.001$
- ERSA + GERMLINE $\alpha = 0.05$
- ERSA + BEAGLE $\alpha = 0.001$
- GBIRP LOD > 2.34
- RELPAIR $LR > 10$
Figure S1
A. ERSA + GERMLINE, $\alpha = 0.05$

B. ERSA + GERMLINE, $\alpha = 0.001$

C. ERSA + BEAGLE, $\alpha = 0.05$

D. GBIRP, LOD > 2.34

E. RELPAIR, likelihood ratio > 10

F. Number of pairs in each relationship class

Figure S2
Figure S3

Figure S4

Figure S5

Figure S6
**INTERNATIONAL SEARCH REPORT**

**A. CLASSIFICATION OF SUBJECT MATTER**

- **IPC(8)** - G06F 17/30 (2012.01)
- **USPC** - 707/769

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

- **IPC(8)** - C07H 21/00, 21/04; C12Q 1/68, 1/70; G01N 33/48, 33/50; G06F 17/30, 19/00; G06G 7/48; G06Q 50/00 (2012.01)
- **USPC** - 435/6, 6.1, 6.1.1, 6.12, 6.13, 6.16, 6P, 91.2, 287.2; 536/23.1, 24.3, 24.31; 702/19, 20; 703/1; 705/2; 707/769; E17.014

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>

Further documents are listed in the continuation of Box C.

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier application or patent but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed
- "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
- "Z" document member of the same patent family

**Date of the actual completion of the international search**

10 April 2012

**Date of mailing of the international search report**

02 MAY 2012

**Name and mailing address of the ISA/US**

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450
Facsimile No. 571-273-3201

**Authorized officer:**

Blaine R. Copenhaver
PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774
This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
   because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
   because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 21-23, 50-52
   because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant’s protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant’s protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>