METHODS AND SYSTEMS FOR DETERMINING ANCESTRAL RELATEDNESS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 62/936,879, filed November 18, 2019, which is incorporated by reference herein in its entirety.

BACKGROUND

[0002] Small differences in DNA within populations of individuals may be responsible for much of the genetic and phenotypic variation observed between individuals. Certain identical-by-descent (IBD) haplotypes (e.g., small regions of DNA that are inherited together) that are shared by individuals may be indicative of a familial relationship between those individuals. Therefore, IBD analysis may be performed on a plurality of individuals to determine their ancestral relatedness.

SUMMARY

[0003] The present disclosure provides methods of estimating a degree of ancestral relatedness between individuals. In an example, a method comprises receiving haplotype data comprising genetic markers shared among a population of individuals; dividing the haplotype data into segments based on the genetic markers; for each of the population of test individuals: (i) based on the genetic markers, matching segments of the haplotype data that are identical-by-descent between two individuals, (ii) for each of the matched segments: dividing the matched segment into discrete genomic intervals, scoring each of the discrete genomic intervals based on a degree of matching within or between the individuals, correcting the scores for consistency, and (iii) calculating a weighted sum over the discrete genomic intervals of the matched segment, based on the corrected scores and assigned weights; and (d) estimating the degree of ancestral relatedness between the individuals based on the weighted sums of the matched segments.

[0004] In an aspect, the present disclosure provides a computer-implemented method for estimating a degree of ancestral relatedness between two individuals of a diploid population, comprising: (a) receiving haplotype data for a population of test individuals, the haplotype data comprising a plurality of genetic markers shared among the population of test individuals; (b) dividing the haplotype data into segments based on the plurality of genetic markers; (c) for each of the population of test individuals: (i) based on the plurality of genetic markers, matching segments of the haplotype data that are identical-by-descent (IBD) between a first individual and a second individual among the population of test individuals, each of the matched segments having a first size that is at least a pre-determined threshold size and comprising at least a pre-determined number of genetic markers; (ii) for each of the matched segments between the first

individual and the second individual: dividing the matched segment into a plurality of discrete genomic intervals; scoring each of the plurality of discrete genomic intervals based on (i) a degree of homozygosity matching of the discrete genomic interval within the first individual or the second individual or (ii) a degree of pairwise matching of the discrete genomic interval between the first individual and the second individual, thereby generating a plurality of scores; correcting each of the plurality of scores based on a consistency of the plurality of scores, thereby producing a plurality of corrected scores; and assigning a plurality of weights to the plurality of discrete genomic intervals, based on the plurality of corrected scores of the discrete genomic intervals; and (iii) calculating a weighted sum over the plurality of discrete genomic intervals of the matched segment, based on the plurality of corrected scores and the plurality of weights; and (d) estimating the degree of ancestral relatedness between the first individual and the second individual based on the weighted sums of the matched segments.

[0005] In some embodiments, the diploid population is a mammal population. In some embodiments, the mammal population is a canine population, a feline population, a sport animal population, or a rodent population. In some embodiments, the mammal population is a canine population. In some embodiments, the canine population is a dog population. In some embodiments, the mammal population is a feline population. In some embodiments, the feline population is a cat population. In some embodiments, the mammal population is a sport animal population. In some embodiments, the sport animal population is a horse population. In some embodiments, the dog population comprises one or more dog breeds selected from the group consisting of: Affenpinscher, Afghan Hound, Africanis, Aidi, Airedale Terrier, Akbash Dog, Akita Inu, Alangu Mastiff, Alano Español, Alapaha Blue Blood Bulldog, Alaskan Klee Kai, Alaskan Malamute, Alaunt, Alopekis, Alpine Dachsbracke, Alsatian Shepalute, American Akita, American Bulldog, American Cocker Spaniel, American Eskimo Dog, American Foxhound, American Hairless Terrier, American Mastiff, American Pit Bull Terrier, American Staffordshire Terrier, American Water Spaniel, Anatolian Shepherd Dog, Anglo-Français de Petite Vénerie, Appenzeller Sennenhund, Argentine Dogo, Ariege Pointer, Ariegeois, Armant, Artois Hound, Australian Bulldog, Australian Cattle Dog, Australian Kelpie, Australian Shepherd, Australian Silky Terrier, Australian Stumpy Tail Cattle Dog, Australian Terrier, Austrian Black and Tan Hound, Austrian Pinscher, Azawakh, Bakharwal Dog, Barbet, Basenji, Basque Shepherd Dog, Basset Artésien Normand, Basset Bleu de Gascogne, Basset Fauve de Bretagne, Grand Basset Griffon Vendéen, Petit Basset Griffon Vendéen, Bavarian Mountain Hound, Beagle, Beagle-Harrier, Bearded Collie, Beauceron, Bedlington Terrier, Belgian Shepherd Dog, Belgian Shepherd Dog (Groenendael), Belgian Shepherd Dog (Laekenois),

[0018] In some embodiments, the method further comprises identifying, for each of one or more individuals of the population of test individuals, one or more dog breeds of the individual. In some embodiments, the method further comprises identifying, for each of one or more individuals of the population of test individuals, one or more proportions of the one or more dog breeds of the individual. In some embodiments, the method further comprises generating a family tree of a plurality of individuals of the population of test individuals. In some embodiments, the method further comprises identifying, for each of one or more individuals of the population of test individuals, a presence, absence, or risk of a phenotype or trait in the individual. In some embodiments, the phenotype or trait is selected from the group consisting of: base coat color (e.g., dark or light fur, color of pigment, and color dilution), color coat modifiers (e.g., hidden patterning, body pattern, and facial pattern), coat traits (e.g., furnishings, coat length, shedding, coat texture, hairlessness (Xolo type), hairlessness (Terrier type), and albinism), body features (e.g., muzzle length, tail length, hind dew claws, back muscling and bulk, and eye color), body size (e.g., smaller, intermediate, and larger), performance (e.g., altitude adaptation), genetic diversity (e.g., degree of inbreeding, and diversity in immune response).

[0019] In some embodiments, the method further comprises generating a report indicative of one or more of: the degree of ancestral relatedness; the familial relationship; the presence, the absence, the risk, or the carrier status of the genetic or health condition; the one or more dog breeds; the one or more proportions of the one or more dog breeds; the family tree; the presence, the absence, the risk, or the carrier status of the phenotype or trait; and any combination thereof. In some embodiments, the method further comprises transmitting the report to a veterinarian.

[0020] In another aspect, the present disclosure provides a computer system for estimating a degree of ancestral relatedness between two individuals of a diploid population, comprising: a database that is configured to store haplotype data for a population of test individuals, the haplotype data comprising a plurality of genetic markers shared among the population of test individuals; and one or more computer processors operatively coupled to the database, wherein the one or more computer processors are individually or collectively programmed to: (a) divide the haplotype data into segments based on the plurality of genetic markers; (b) for each of the population of test individuals: (i) based on the plurality of genetic markers, match segments of the haplotype data that are identical-by-descent (IBD) between a first individual and a second individual among the population of test individuals, each of the matched segments having a first size that is at least a pre-determined threshold size and comprising at least a pre-determined number of genetic markers; (ii) for each of the matched segments between the first individual

and the second individual: divide the matched segment into a plurality of discrete genomic intervals; score each of the plurality of discrete genomic intervals based on (i) a degree of homozygosity matching of the discrete genomic interval within the first individual or the second individual or (ii) a degree of pairwise matching of the discrete genomic interval between the first individual and the second individual, thereby generating a plurality of scores; and correct the each of the plurality of scores based on a consistency of the plurality of scores, thereby producing a plurality of corrected scores; and assign a plurality of weights to the plurality of discrete genomic intervals, based on the plurality of corrected scores of the discrete genomic intervals; and (iii) calculate a weighted sum over the plurality of discrete genomic intervals of the matched segment, based on the plurality of corrected scores and the plurality of weights; and (c) estimate the degree of ancestral relatedness between the first individual and the second individual based on the weighted sums of the matched segments.

[0021] In some embodiments, the diploid population is a mammal population. In some embodiments, the mammal population is a canine population, a feline population, a sport animal population, or a rodent population. In some embodiments, the mammal population is a canine population. In some embodiments, the canine population is a dog population. In some embodiments, the mammal population is a feline population. In some embodiments, the feline population is a cat population. In some embodiments, the mammal population is a sport animal population. In some embodiments, the sport animal population is a horse population. In some embodiments, the dog population comprises one or more dog breeds selected from the group consisting of: Affenpinscher, Afghan Hound, Africanis, Aidi, Airedale Terrier, Akbash Dog, Akita Inu, Alangu Mastiff, Alano Español, Alapaha Blue Blood Bulldog, Alaskan Klee Kai, Alaskan Malamute, Alaunt, Alopekis, Alpine Dachsbracke, Alsatian Shepalute, American Akita, American Bulldog, American Cocker Spaniel, American Eskimo Dog, American Foxhound, American Hairless Terrier, American Mastiff, American Pit Bull Terrier, American Staffordshire Terrier, American Water Spaniel, Anatolian Shepherd Dog, Anglo-Français de Petite Vénerie, Appenzeller Sennenhund, Argentine Dogo, Ariege Pointer, Ariegeois, Armant, Artois Hound, Australian Bulldog, Australian Cattle Dog, Australian Kelpie, Australian Shepherd, Australian Silky Terrier, Australian Stumpy Tail Cattle Dog, Australian Terrier, Austrian Black and Tan Hound, Austrian Pinscher, Azawakh, Bakharwal Dog, Barbet, Basenji, Basque Shepherd Dog, Basset Artésien Normand, Basset Bleu de Gascogne, Basset Fauve de Bretagne, Grand Basset Griffon Vendéen, Petit Basset Griffon Vendéen, Bavarian Mountain Hound, Beagle, Beagle-Harrier, Bearded Collie, Beauceron, Bedlington Terrier, Belgian Shepherd Dog, Belgian Shepherd Dog (Groenendael), Belgian Shepherd Dog (Laekenois),

the one or more computer processors are individually or collectively programmed to further transmit the report to a veterinarian.

[0036] In another aspect, the present disclosure provides a non-transitory computer readable medium comprising machine-executable code that, upon execution by one or more computer processors, implements a method for estimating a degree of ancestral relatedness between two individuals of a diploid population, the method comprising: (a) receiving haplotype data for a population of test individuals, the haplotype data comprising a plurality of genetic markers shared among the population of test individuals; (b) dividing the haplotype data into segments based on the plurality of genetic markers; (c) for each of the population of test individuals: (i) based on the plurality of genetic markers, matching segments of the haplotype data that are identical-by-descent (IBD) between a first individual and a second individual among the population of test individuals, each of the matched segments having a first size that is at least a pre-determined threshold size and comprising at least a pre-determined number of genetic markers; (ii) for each of the matched segments between the first individual and the second individual: dividing the matched segment into a plurality of discrete genomic intervals; scoring each of the plurality of discrete genomic intervals based on (i) a degree of homozygosity matching of the discrete genomic interval within the first individual or the second individual or (ii) a degree of pairwise matching of the discrete genomic interval between the first individual and the second individual, thereby generating a plurality of scores; correcting each of the plurality of scores based on a consistency of the plurality of scores, thereby producing a plurality of corrected scores; and assigning a plurality of weights to the plurality of discrete genomic intervals, based on the plurality of corrected scores of the discrete genomic intervals; and (iii) calculating a weighted sum over the plurality of discrete genomic intervals of the matched segment, based on the plurality of corrected scores and the plurality of weights; and (d) estimating the degree of ancestral relatedness between the first individual and the second individual based on the weighted sums of the matched segments.

[0037] In some embodiments, the diploid population is a mammal population. In some embodiments, the mammal population is a canine population, a feline population, a sport animal population, or a rodent population. In some embodiments, the mammal population is a canine population. In some embodiments, the canine population is a dog population. In some embodiments, the mammal population is a feline population. In some embodiments, the feline population is a cat population. In some embodiments, the mammal population is a sport animal population. In some embodiments, the sport animal population is a horse population. In some embodiments, the dog population comprises one or more dog breeds selected from the group

CLAIMS

WHAT IS CLAIMED IS:

1. A computer-implemented method for estimating a degree of ancestral relatedness between two individuals of a diploid population, comprising:

(a) receiving haplotype data for a population of test individuals, the haplotype data comprising a plurality of genetic markers shared among the population of test individuals;

(b) dividing the haplotype data into segments based on the plurality of genetic markers;

(c) for each of the population of test individuals:

(i) based on the plurality of genetic markers, matching segments of the haplotype data that are identical-by-descent (IBD) between a first individual and a second individual among the population of test individuals, each of the matched segments having a first size that is at least a pre-determined threshold size and comprising at least a pre-determined number of genetic markers;

(ii) for each of the matched segments between the first individual and the second individual:

dividing the matched segment into a plurality of discrete genomic intervals;

scoring each of the plurality of discrete genomic intervals based on (i) a degree of homozygosity matching of the discrete genomic interval within the first individual or the second individual or (ii) a degree of pairwise matching of the discrete genomic interval between the first individual and the second individual, thereby generating a plurality of scores;

correcting each of the plurality of scores based on a consistency of the plurality of scores, thereby producing a plurality of corrected scores; and

assigning a plurality of weights to the plurality of discrete genomic intervals, based on the plurality of corrected scores of the discrete genomic intervals; and

(iii) calculating a weighted sum over the plurality of discrete genomic intervals of the matched segment, based on the plurality of corrected scores and the plurality of weights; and

(d) estimating the degree of ancestral relatedness between the first individual and the second individual based on the weighted sums of the matched segments.

2. The method of claim 1, wherein the diploid population is a mammal population.

3. The method of claim 1, wherein the haplotype data is generated by processing genotype data of the population of test individuals using a haplotype phasing algorithm.

4. The method of claim 3, wherein the haplotype phasing algorithm comprises a reference-based haplotype phasing algorithm comprising a Hidden Markov Model (HMM)-based search or a cohort-based haplotype phasing algorithm.

5. The method of claim 3, wherein the genotype data is obtained by assaying biological samples obtained from the population of test individuals.

6. The method of claim 5, wherein the assaying comprises sequencing the biological samples or derivatives thereof.

7. The method of claim 1, wherein the plurality of genetic markers comprises at least about 500 distinct genetic markers.

8. The method of claim 1, wherein matching the segments of the haplotype data that are identical-by-descent comprises using a GERMLINE algorithm, a PLINK algorithm, a PREST algorithm, a Random Projection for IBD Detection (RaPID) algorithm, a Find IBD Shared Haplotypes Rapidly (FISHR) algorithm, a refined identical-by-descent (IBD) algorithm, a fastIBD algorithm, a KING algorithm, a HaploScore algorithm, a TRUFFLE algorithm, or a combination thereof.

9. The method of claim 1, wherein the pre-determined threshold size is at least about 100 kilobase pairs (kbp).

10. The method of claim 1, wherein the pre-determined number of genetic markers is at least about 30 genetic markers.

11. The method of claim 1, further comprising dividing the matched segments such that the discrete genomic intervals of the plurality of discrete genomic intervals have an equal size.

12. The method of claim 1, further comprising dividing the matched segments such that the discrete genomic intervals of the plurality of discrete genomic intervals have a variable size.

13. The method of claim 12, wherein the variable size of a given discrete genomic interval of the plurality of discrete genomic intervals is determined based at least in part on a start position and an end position of IBD matches proximal to the given discrete genomic interval, a density of genetic markers in the given discrete genomic interval, a maximum number of markers for the given discrete genomic interval, a maximum length of the given discrete genomic interval, or a combination thereof.

14. The method of claim 1, further comprising scoring each of the plurality of discrete genomic intervals based on (i) the degree of homozygosity matching of the discrete genomic interval within the first individual or the second individual and (ii) the degree of pairwise matching of the discrete genomic interval between the first individual and the second individual, thereby generating a plurality of homozygosity matching scores and a plurality of pairwise matching scores.

15. The method of claim 14, further comprising correcting the plurality of pairwise matching scores based on a consistency of a given pairwise matching score with a corresponding homozygosity matching score, thereby producing a plurality of corrected pairwise matching scores.

16. The method of claim 14, further comprising assigning the plurality of weights to the plurality of discrete genomic intervals, based on a plurality of identity states for two alleles in two diploid individuals.

17. The method of claim 16, wherein the plurality of identity states comprises identity states selected from **Table 1**, and wherein the plurality of weights are assigned based on a plurality of contributions to relatedness r_{xy} as listed in **Table 1**.

18. The method of claim 17, wherein the degree of ancestral relatedness comprises a coefficient of relatedness.

19. The method of claim 18, further comprising calculating the weighted sum over the plurality of discrete genomic intervals of the matched segment, wherein the weighted sum is expressed by:

$r_{xy} = \frac{\Delta 1 + \Delta 7 + (0.75\ x\ \Delta 3) + (0.5\ x\ \Delta 8)}{L}$, wherein Δi represent the sum total of the genome length assigned to one of the four match count states, *i*, and *L* is the total length of the genome considered.

20. The method of claim 17, wherein the degree of ancestral relatedness comprises a coefficient of kinship.

21. The method of claim 26, further comprising calculating the weighted sum over the plurality of discrete genomic intervals of the matched segment, wherein the weighted sum is expressed by:

$k_{xy} = \frac{\Delta 1 + (0.5\ x\ (\Delta 3 + \Delta 7)) + (0.25\ x\ \Delta 8)}{L}$, wherein Δi represent the sum total of the genome length assigned to one of the four match count states, *i*, and *L* is the total length of the genome considered.

22. The method of claim 1, wherein estimating the degree of ancestral relatedness between the first individual and the second individual comprises determining a degree of inbreeding of the first individual and/or the second individual.

23. The method of claim 22, further comprising determining a familial relationship between the first individual and the second individual based at least in part on the degree of inbreeding of the first individual and the second individual.

24. The method of claim 23, wherein the familial relationship is a parent-child relationship, a sibling relationship, an aunt/uncle-nephew/niece relationship, a cousin relationship, or a grandparent-grandchild relationship.

25. The method of claim 1, further comprising generating a social connection between a first person associated with the first individual and a second person associated with the

second individual, based at least in part on the estimated degree of ancestral relatedness between the first individual and the second individual.

26. The method of claim 25, wherein the social connection is generated when the estimated degree of ancestral relatedness between the first individual and the second individual exceeds a pre-determined threshold value, wherein the pre-determined threshold value is at least about 0.001.

27. The method of claim 25, wherein the first person is a pet owner of the first individual, and wherein the second person is a pet owner of the second individual.

28. The method of claim 1, further comprising identifying a familial relationship between the first individual and the second individual based at least in part on the degree of ancestral relatedness, wherein the familial relationship is a parent-child relationship, a sibling relationship, an aunt/uncle-nephew/niece relationship, a cousin relationship, or a grandparent-grandchild relationship.

29. A non-transitory computer readable medium comprising machine-executable code that, upon execution by one or more computer processors, implements a method for estimating a degree of ancestral relatedness between two individuals of a diploid population, the method comprising:

(a) receiving haplotype data for a population of test individuals, the haplotype data comprising a plurality of genetic markers shared among the population of test individuals;

(b) dividing the haplotype data into segments based on the plurality of genetic markers;

(c) for each of the population of test individuals:

(i) based on the plurality of genetic markers, matching segments of the haplotype data that are identical-by-descent between a first individual and a second individual among the population of test individuals, each of the matched segments having a first size that is at least a pre-determined threshold size and comprising at least a pre-determined number of genetic markers;

(ii) for each of the matched segments between the first individual and the second individual:

dividing the matched segment into a plurality of discrete genomic intervals;

scoring each of the plurality of discrete genomic intervals based on (i) a degree of homozygosity matching of the discrete genomic interval within the first individual or the second individual or (ii) a degree of pairwise matching of the discrete genomic interval between the first individual and the second individual, thereby generating a plurality of scores;

correcting each of the plurality of scores based on a consistency of the plurality of scores, thereby producing a plurality of corrected scores; and

assigning a plurality of weights to the plurality of discrete genomic intervals, based on the plurality of corrected scores of the discrete genomic intervals; and

(iii) calculating a weighted sum over the plurality of discrete genomic intervals of the matched segment, based on the plurality of corrected scores and the plurality of weights; and

(d) estimating the degree of ancestral relatedness between the first individual and the second individual based on the weighted sums of the matched segments.

30. A computer system for estimating a degree of ancestral relatedness between two individuals of a diploid population, comprising:

a database that is configured to store haplotype data for a population of test individuals, the haplotype data comprising a plurality of genetic markers shared among the population of test individuals; and

one or more computer processors operatively coupled to the database, wherein the one or more computer processors are individually or collectively programmed to:

(a) divide the haplotype data into segments based on the plurality of genetic markers;

(b) for each of the population of test individuals:

(i) based on the plurality of genetic markers, match segments of the haplotype data that are identical-by-descent (IBD) between a first individual and a second individual among the population of test individuals, each of the matched segments having a first size that is at least a pre-determined threshold size and comprising at least a pre-determined number of genetic markers;

(ii) for each of the matched segments between the first individual and the second individual:

divide the matched segment into a plurality of discrete genomic intervals;

score each of the plurality of discrete genomic intervals based on (i) a degree of homozygosity matching of the discrete genomic interval within the first individual or the second individual or (ii) a degree of pairwise matching of the discrete genomic interval between the first individual and the second individual, thereby generating a plurality of scores; and

correct the each of the plurality of scores based on a consistency of the plurality of scores, thereby producing a plurality of corrected scores; and

assign a plurality of weights to the plurality of discrete genomic intervals, based on the plurality of corrected scores of the discrete genomic intervals; and

(iii) calculate a weighted sum over the plurality of discrete genomic intervals of the matched segment, based on the plurality of corrected scores and the plurality of weights; and

(c) estimate the degree of ancestral relatedness between the first individual and the second individual based on the weighted sums of the matched segments.