ANCESTRY-SPECIFIC GENETIC RISK SCORES

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Application No. 62/772,565, filed November 28, 2018, and U.S. Patent Application No. 16/216,940, filed December 11, 2018, which applications are hereby incorporated by reference herein in their entireties.

SEQUENCE LISTING

[0002] The instant application contains a Sequence Listing which is hereby incorporated by reference in its entirety. Said ASCII copy, created on October 28, 2019, is named 55075-701_601_SL.txt and is 47.9 KB in size.

SUMMARY ~~OF THE INVENTION~~

[0003] Genome Wide Association Studies (GWAS) have enabled scientists to identify genetic variations that are associated with a wide range of phenotypic traits. A genetic risk score (GRS) is used to predict whether an individual will develop a trait based on a presence of certain genetic variants detected in a sample obtained from that individual. However, data show that genetic variation and patterns underlying discrete ancestral populations differ. Thus, whether the detected genetic variants confer a risk that the individual will develop the trait depends in large part on the ancestry of that individual. Current genetic risk prediction methods either do not account for the ancestry of the individual at all, or account for ancestry using consumer surveys leading to imprecise, and often, inaccurate genetic risk predictions.

[0003a] Disclosed herein, in one aspect, is a computer-implemented method of determining a likelihood that an individual has, or will develop, a specific phenotypic trait based on the ancestry of the individual, the method comprising: a) assigning an ancestry of the individual by using a distance-based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants; b) selecting, from a trait-associated variants database comprising ancestry-specific genetic variants derived from subjects with the same ancestry as the individual (subject group), one or more ancestry-specific genetic variants based, at least in part, on the ancestry of the individual, wherein each of the one or more ancestry-specific genetic variants correspond to: i) an individual-specific genetic variant of the one or more individual-specific genetic variants, or ii) a predetermined genetic variant in a linkage disequilibrium (LD) with an individual-specific genetic variant of the one or more individual-specific genetic variants in a subject population with the same ancestry as the individual, wherein the predetermined genetic variant is predetermined by: 1) phasing unphased genotype data from the individual to generate individual-specific phased haplotypes based on the ancestry of the individual; 2) imputing individual-specific genotypes not present in the phased individual-specific phased haplotypes using phased haplotype data from a reference group that has the same ancestry as the individual; and 3) selecting a genetic variant from the imputed individual-specific genotypes that matches with the individual-specific genetic variant associated with a

likelihood that the individual has, or will develop, a specific phenotypic trait and corresponding to the one or more ancestry specific variants, wherein each of the one or more ancestry-specific genetic variants and each of the one or more individual specific genetic variants comprise one or more units of risk; and c) calculating a genetic risk score for the individual based on the selected one or more ancestry-specific genetic variants, wherein the genetic risk score is indicative of the likelihood that the individual has, or will develop the specific phenotypic trait.

[0003b] Disclosed herein, in another aspect, is a wellness reporting system comprising: a computing device comprising at least one processor, a memory, and a software program including instructions executable by at least one processor to assess a likelihood that an individual has, or will develop, a specific phenotypic trait, the instructions comprising the steps of: a. assigning an ancestry of the individual by using a distance-based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants; b. selecting, from a trait-associated variants database comprising ancestry-specific genetic variants derived from subjects with the same ancestry as the individual (subject group), one or more ancestry-specific genetic variants based, at least in part, on the ancestry of the individual, wherein each of the one or more ancestry-specific genetic variants correspond to: i. an individual-specific genetic variant of the one or more individual-specific genetic variants, or ii. a predetermined genetic variant in a linkage disequilibrium (LD) with an individual-specific genetic variant of the one or more individual-specific genetic variants in a subject population with the same ancestry as the individual, wherein the predetermined genetic variant is predetermined by: 1. phasing unphased genotype data from the individual to generate individual-specific phased haplotypes based on the ancestry of the individual; 2. imputing individual-specific genotypes not present in the phased individual-specific phased haplotypes using phased haplotype data from a reference group that has the same ancestry as the individual; and 3. selecting a genetic variant from the imputed individual-specific genotypes that matches with the individual-specific genetic variant associated with a likelihood that the individual has, or will develop, a specific phenotypic trait and corresponding to the one or more ancestry specific variants, wherein each of the one or more ancestry-specific genetic variants and each of the one or more individual specific genetic variants comprise one or more units of risk; and c. calculating a genetic risk score for the individual based on the selected one or more ancestry-specific genetic variants, wherein the genetic risk score is indicative of the likelihood that the individual has, or will develop the specific phenotypic trait; a reporting module configured to generate a report comprising the genetic risk score of the individual for the specific phenotypic trait; and an output module configured to display the report to the individual.

[0003c] Disclosed herein, in another aspect, is a non-transitory computer readable storage medium, comprising computer-executable code configured to cause at least one processor to perform steps comprising: a) assigning an ancestry of the individual by using a distance-based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants; b) selecting, from a trait-associated variants database comprising ancestry-specific genetic variants derived from subjects with the same ancestry as the individual (subject group), one or more ancestry-specific

genetic variants based, at least in part, on the ancestry of the individual, wherein each of the one or more ancestry-specific genetic variants correspond to: i) an individual-specific genetic variant of the one or more individual-specific genetic variants, or ii) a predetermined genetic variant in a linkage disequilibrium (LD) with an individual-specific genetic variant of the one or more individual-specific genetic variants in a subject population with the same ancestry as the individual, wherein the predetermined genetic variant is predetermined by: 1) providing unphased genotype data from the individual; 2) phasing the unphased genotype data to generate individual-specific phased haplotypes based on the ancestry of the individual; 3) imputing individual-specific genotypes not present in the phased individual-specific phased haplotypes using phased haplotype data from a reference group that has the same ancestry as the individual; and 4) selecting a genetic variant from the imputed individual-specific genotypes that matches with the individual-specific genetic variant associated with a likelihood that the individual has, or will develop, a specific phenotypic trait; and c) calculating a genetic risk score for the individual based on the selected one or more ancestry-specific genetic variants, wherein the genetic risk score is indicative of the likelihood that the individual has, or will develop the specific phenotypic trait.

[0004] Disclosed herein, in certain embodiments, are methods, media, and systems for calculating a GRS by analyzing the genotype of the individual to determine an ancestry of the individual and calculating a GRS based on the ancestry-specific genetic risk variants derived from GWAS of subjects of the same ancestry as the individual. In some embodiments, genetic variant(s) accounted for in a GRS may include single nucleotide variants (SNVs), insertions or deletions of nucleotide bases (indels), or copy number variants (CNVs). In some embodiments, if a genetic variant detected in a sample obtained from the individual does not correspond to genetic variant reported in the GWAS of the ancestry-specific subject group (unknown genetic variant), a proxy genetic variant is selected based on the non-random association, known as linkage disequilibrium (LD), with the unknown genetic variant within the particular ancestral population, which serves as the basis for risk prediction. Studies show that patterns in LD in the human genome differ across different ancestral populations.

[0005] Disclosed herein, in some embodiments, are computer-implemented methods comprising: (a) assigning an ancestry of an individual using a distance-based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants; and (b) detecting in the genotype of the individual, an ancestry-specific variant associated with a specific phenotypic trait, the ancestry-specific variant corresponding to: (i) an individual-specific genetic variant detectable in the genotype of the individual; or (ii) a genetic variant in linkage disequilibrium (LD) with the

CLAIMS

What is claimed is:

1. A computer-implemented method comprising:
    a) assigning an ancestry of an individual using a distance based or a models based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants; and
    b) detecting in the genotype of the individual, an ancestry-specific variant associated with a specific phenotypic trait, the ancestry-specific variant corresponding to:
        i) an individual-specific genetic variant detectable in the genotype of the individual; or
        ii) a genetic variant in linkage disequilibrium (LD) with the individual-specific genetic variant as determined by imputing the individual-specific variant missing from ancestry-specific phased haplotypes determined using a reference group of individuals that has the same ancestry as the individual; and
    c) calculating a genetic risk score (GRS) for the individual based on the ancestry-specific variant detected in (b), wherein the GRS is indicative of a likelihood that the individual has, or will develop, the specific phenotypic trait.
2. The computer-implemented method of claim 1, wherein the ancestry-specific genetic variant and the individual-specific genetic variant is selected from the group consisting of a single nucleotide variant (SNV), a copy number variant (CNV), and an indel.
3. The computer-implemented method of claim 1, wherein imputing in step (ii) comprises:
    a) phasing unphased genotype data from the individual to generate ancestry-specific phased haplotypes based on the ancestry of the individual; and
    b) imputing individual-specific genotypes not present in the ancestry-specific phased haplotypes using phased haplotype data from the reference group that has the same ancestry as the individual to select the genetic variant in LD with the individual-specific genetic variant.
4. The computer-implemented method of claim 1, wherein the LD is defined by a D' value comprising at least about 0.80 or a r2 value comprising at least 0.80.
5. The computer-implemented method of claim 1, wherein the specific trait comprises a nutritional trait, a clinical trait, a subclinical trait, a physical exercise trait, a skin trait, an allergy trait, or a mental trait, or combination thereof.
6. The computer-implemented method of claim 1, wherein the genotype of the individual is obtained by subjecting, or having subjected, genetic material obtained from the individual to a genotyping assay.
7. The computer-implemented method of claim 6, wherein the genotyping assay comprises a deoxyribonucleic acid (DNA) array, ribonucleic acid (RNA) array, sequencing assay, or a combination thereof.

8. The computer-implemented method of claim 1, wherein the distance-based computer program is principle component analysis, and wherein the models-based computer program is a maximum likelihood or a Bayesian method.

9. The computer-implemented method of claim 1, wherein the GRS for the individual based on the ancestry-specific variant is more accurate than a corresponding GRS of the individual based on a variant that is not ancestry-specific.

10. The computer-implemented method of claim 1, further comprising providing a notification comprising the GRS for the specific phenotypic trait of the individual.

11. The computer-implemented method of claim 10, wherein the notification further comprises a behavior recommendation to the individual based on the GRS for the specific phenotypic trait.

12. The computer-implemented method of claim 11, wherein the behavioral modification related to the specific phenotypic trait comprises increasing, reducing, or avoiding an activity comprising performance of a physical exercise, ingestion of a drug, vitamin, or supplement, exposure to a product, usage of a product, a diet modification, sleep modification, alcohol consumption, or caffeine consumption.

13. A system comprising:

a computing device comprising at least one processor, a memory, and a software program including instructions executable by at least one processor to assess a likelihood that an individual has, or will develop, a specific phenotypic trait, the instructions comprising the steps of:

a) assigning an ancestry of an individual using a distance based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants; and

b) detecting in the genotype of the individual, an ancestry-specific variant associated with a specific phenotypic trait, the ancestry-specific variant corresponding to:

i) an individual-specific genetic variant detectable in the genotype of the individual; or

ii) a genetic variant in linkage disequilibrium (LD) with the individual-specific genetic variant as determined by imputing the individual-specific variant missing from ancestry-specific phased haplotypes determined using a reference group of individuals that has the same ancestry as the individual; and

c) calculating a genetic risk score (GRS) for the individual based on the ancestry-specific variant detected in (b), wherein the GRS is indicative of a likelihood that the individual has, or will develop, the specific phenotypic trait.

14. The system of claim 13, wherein the ancestry-specific genetic variant and the individual-specific genetic variant is selected from the group consisting of a single nucleotide variant (SNV), a copy number variant (CNV), and an indel.

15. The system of claim 13, wherein imputing in step (2) comprises:

a) phasing unphased genotype data from the individual to generate ancestry-specific phased haplotypes based on the ancestry of the individual; and

b) imputing individual-specific genotypes not present in the ancestry-specific phased haplotypes using phased haplotype data from the reference group that has the same ancestry as the individual to select the genetic variant in LD with the individual-specific genetic variant.

16. The system of claim 13, wherein the LD is defined by a D' value comprising at least about 0.80 or a r2 value comprising at least 0.80.

17. The system of claim 13, wherein the specific trait comprises a nutritional trait, a clinical trait, a subclinical trait, a physical exercise trait, a skin trait, an allergy trait, or a mental trait.

18. The system of claim 13, further comprising a genotyping assay.

19. The system of claim 18, wherein the genotyping assay comprises a deoxyribonucleic acid (DNA) array, ribonucleic acid (RNA) array, sequencing assay, or a combination thereof.

20. The system of claim 13, wherein the distance-based computer program is principle component analysis, and wherein the models-based computer program is a maximum likelihood or a Bayesian method.

21. The system of claim 13, wherein the GRS for the individual based on the ancestry-specific variant is more accurate than a corresponding GRS of the individual based on a variant that is not ancestry-specific.

22. The system of claim 13, further comprising a reporting module configured to generate a report comprising the GRS of the individual for the specific phenotypic trait.

23. The system of claim 13, further comprising an output module configured to display the report to the individual.

24. The system of claim 23, wherein the report comprises the risk that the individual has, or will develop, the specific trait.

25. The system of claim 23, wherein the report further comprises a recommendation of a behavior recommendation to the individual based on the GRS for the specific phenotypic trait.

26. The system of claim 25, wherein the behavioral modification related to the specific phenotypic trait comprises increasing, reducing, or avoiding an activity comprising performance of a physical exercise, ingestion of a drug, vitamin, or supplement, exposure to a product, usage of a product, a diet modification, sleep modification, alcohol consumption, or caffeine consumption.

27. Use of the system of claim 13, for recommending a behavior modification or a product to the individual, based on the GRS calculated in (c).

28. A non-transitory computer readable storage medium, comprising computer executable code configured to cause at least one processor to perform steps provided in claim 1.

29.1. A computer-implemented method of determining a likelihood that an individual has, or will develop, a specific phenotypic trait based on the ancestry of the individual, the method comprising:

a) assigning an ancestry of the individual by using a distance-based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants;

b) selecting, from a trait-associated variants database comprising ancestry-specific genetic variants derived from subjects with the same ancestry as the individual (subject group), one or more ancestry-specific genetic variants based, at least in part, on the ancestry of the individual, wherein each of the one or more ancestry-specific genetic variants correspond to:

i) an individual-specific genetic variant of the one or more individual-specific genetic variants, or

ii) a predetermined genetic variant in a linkage disequilibrium (LD) with an individual-specific genetic variant of the one or more individual-specific genetic variants in a subject population with the same ancestry as the individual, wherein the predetermined genetic variant is predetermined by:

1) phasing unphased genotype data from the individual to generate individual-specific phased haplotypes based on the ancestry of the individual;

2) imputing individual-specific genotypes not present in the phased individual-specific phased haplotypes using phased haplotype data from a reference group that has the same ancestry as the individual; and

3) selecting a genetic variant from the imputed individual-specific genotypes that matches with the individual-specific genetic variant associated with a likelihood that the individual has, or will develop, a specific phenotypic trait and corresponding to the one or more ancestry specific variants,

wherein each of the one or more ancestry-specific genetic variants and each of the one or more individual specific genetic variants comprise one or more units of risk; and

~~4)~~c) calculating a genetic risk score for the individual based on the selected one or more ancestry-specific genetic variants, wherein the genetic risk score is indicative of the likelihood that the individual has, or will develop the specific phenotypic trait.

~~30.~~2. The method of claim ~~29~~1, wherein the one or more ancestry-specific genetic variants, the one or more individual-specific genetic variants, and the genetic variants in LD with the one or more individual-specific genetic variants comprise a Single Nucleotide Variant (SNV), an indel, and/or a Copy Number Variant (CNV).

~~31.~~3. The method of claim ~~29~~2, wherein the one or more units of risk of the SNV comprises a risk allele; the one or more units of risk of the indel comprises a presence (I) or an absence (D) of the nucleotide; and the one or more units of risk of the CNV comprises an insertion or a deletion of a nucleic acid sequence.

32.4. The method of claim 291, further comprising providing a notification to the individual comprising the risk that the individual has, or will develop, the specific phenotypic trait.

33.5. The method of claim 291, wherein the specific phenotypic trait comprises a nutritional trait, a clinical trait, a subclinical trait, a physical exercise trait, a skin trait, a hair trait, an allergy trait, or a mental trait.

34.6. The method of claim 334, wherein the notification further comprises a recommendation for a behavior modification related to the specific phenotypic trait.

35.7. The method of claim 346, wherein the behavior modification related to the specific phenotypic trait comprises increasing, reducing, or avoiding an activity comprising performance of a physical exercise, ingestion of a drug, vitamin, or supplement, exposure to a product, usage of a product, a diet modification, sleep modification, alcohol consumption, or caffeine consumption.

36.8. The method of claim 291, wherein the distance-based computer program is principle component analysis, and wherein the models-based computer program is a maximum likelihood or a Bayesian method.

37.9. A wellness reporting system comprising:

a) a computing device comprising at least one processor, a memory, and a software program including instructions executable by at least one processor to assess a likelihood that an individual has, or will develop, a specific phenotypic trait, the instructions comprising the steps of:

i)a. assigning an ancestry of the individual by using a distance-based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants;

ii)b. selecting, from a trait-associated variants database comprising ancestry-specific genetic variants derived from subjects with the same ancestry as the individual (subject group), one or more ancestry-specific genetic variants based, at least in part, on the ancestry of the individual, wherein each of the one or more ancestry-specific genetic variants correspond to:

1)i. an individual-specific genetic variant of the one or more individual-specific genetic variants, or

2)ii. a predetermined genetic variant in a linkage disequilibrium (LD) with an individual-specific genetic variant of the one or more individual-specific genetic variants in a subject population with the same ancestry as the individual, wherein the predetermined genetic variant is predetermined by:

a)1. phasing unphased genotype data from the individual to generate individual-specific phased haplotypes based on the ancestry of the individual;

~~b)~~2. imputing individual-specific genotypes not present in the phased individual-specific phased haplotypes using phased haplotype data from a reference group that has the same ancestry as the individual; and

~~c)~~3. selecting a genetic variant from the imputed individual-specific genotypes that matches with the individual-specific genetic variant associated with a likelihood that the individual has, or will develop, a specific phenotypic trait and corresponding to the one or more ancestry specific variants,

wherein each of the one or more ancestry-specific genetic variants and each of the one or more individual specific genetic variants comprise one or more units of risk; and

~~d)~~c. calculating a genetic risk score for the individual based on the selected one or more ancestry-specific genetic variants, wherein the genetic risk score is indicative of the likelihood that the individual has, or will develop the specific phenotypic trait;

~~b)~~ a reporting module configured to generate a report comprising the genetic risk score of the individual for the specific phenotypic trait; and

~~c)~~ an output module configured to display the report to the individual.

~~38.~~10. The system of claim ~~37~~9, wherein the one or more ancestry-specific genetic variants, the one or more individual-specific genetic variants, and the genetic variants in LD with the one or more individual-specific genetic variants comprise a Single Nucleotide Variant (SNV), an indel, and/or a Copy Number Variant (CNV).

~~39.~~11. The system of claim ~~38~~10, wherein the one or more units of risk of the SNV comprises a risk allele; the one or more units of risk of the indel comprises an insertion (I) or a deletion (D) of the nucleotide; and the one or more units of risk of the CNV comprises an insertion or a deletion of a nucleic acid sequence.

~~40.~~12. The system of claim ~~37~~9, wherein the report further comprises a recommendation for a behavior modification related to the specific phenotypic trait.

~~41.~~13. The system of claim ~~37~~9, wherein the specific phenotypic trait comprises a nutritional trait, a clinical trait, a subclinical trait, a physical exercise trait, a skin trait, a hair trait, an allergy trait, or a mental trait.

~~42.~~14. The system of claim ~~37~~9, further comprising a personal electronic device with an application configured to communicate with the output module via a computer network to access the report.

43.15. The system of claim 379, wherein the distance-based computer program is principle component analysis, and wherein the models-based computer program is a maximum likelihood or a Bayesian method.

44.16. A non-transitory computer readable storage medium, comprising computer-executable code configured to cause at least one processor to perform steps comprising:

a) assigning an ancestry of the individual by using a distance-based or a models-based computer program to analyze a genotype of the individual, the genotype comprising one or more individual-specific genetic variants;

b) selecting, from a trait-associated variants database comprising ancestry-specific genetic variants derived from subjects with the same ancestry as the individual (subject group), one or more ancestry-specific genetic variants based, at least in part, on the ancestry of the individual, wherein each of the one or more ancestry-specific genetic variants correspond to:

i) an individual-specific genetic variant of the one or more individual-specific genetic variants, or

ii) a predetermined genetic variant in a linkage disequilibrium (LD) with an individual-specific genetic variant of the one or more individual-specific genetic variants in a subject population with the same ancestry as the individual, wherein the predetermined genetic variant is predetermined by:

1) providing unphased genotype data from the individual;

2) phasing the unphased genotype data to generate individual-specific phased haplotypes based on the ancestry of the individual;

3) imputing individual-specific genotypes not present in the phased individual-specific phased haplotypes using phased haplotype data from a reference group that has the same ancestry as the individual; and

4) selecting a genetic variant from the imputed individual-specific genotypes that matches with the individual-specific genetic variant associated with a likelihood that the individual has, or will develop, a specific phenotypic trait; and

c) calculating a genetic risk score for the individual based on the selected one or more ancestry-specific genetic variants, wherein the genetic risk score is indicative of the likelihood that the individual has, or will develop the specific phenotypic trait.

45.17. The medium of claim 4416, wherein the one or more ancestry-specific genetic variants, the one or more individual-specific genetic variants, and the genetic variants in LD with the one or more individual-specific genetic variants comprise a Single Nucleotide Variant (SNV), an indel, and/or a Copy Number Variant (CNV).

46.18. The medium of claim 4517, wherein each of the one or more ancestry-specific genetic variants and each of the individual specific genetic variants comprises one or more units of risk, and wherein the one or more units of risk of the SNV comprises a risk allele; the one or more units of risk of the indel

comprises an insertion (I) or a deletion (D) of a nucleotide; and the one or more units of risk of the CNV comprises an insertion or a deletion of a nucleic acid sequence.

~~47.~~19. The medium of claim ~~46~~16, wherein the steps further comprise providing a notification to the individual comprising the likelihood that the individual has, or will develop, the specific phenotypic trait.

~~48.~~20. The medium of claim ~~47~~16, wherein the specific phenotypic trait comprises a nutritional trait, a clinical trait, a subclinical trait, a physical exercise trait, a skin trait, a hair trait, an allergy trait, or a mental trait.