discoveries have facilitated the development of therapeutics that target the products of those genes.

[0006] However, further elucidation of genetic factors that affect health and disease and the development of targeted therapeutics based on this information are needed to drive the implementation of precision medicine, and to identify more biological targets for pharmacological intervention. One approach for identifying putative biological targets is to statistically associate a variant of interest with a phenotype (or *vice versa*) in a large population of subjects for whom genetic variant and phenotype information is available (for example, Wellcome Trust Case Control Consortium, *Nature* 2007; 447: 661; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium, *Circulation: Cardiovascular Genetics* 2009; 2: 73). However, such efforts generally do not utilize a sufficient number of subjects or sufficiently deep characterization of genetic variation to find rare, high-impact, loss of function variants, which is due at least in part to insufficient statistical power, and insufficient genetic variant-phenotype association data from which clinically relevant putative targets can be nominated.

[0007] Furthermore, despite increased investment in research and development by biopharmaceutical industry, over 90% of molecules that enter Phase I clinical trials fail to demonstrate sufficient safety and efficacy to obtain regulatory approval. Most failures occur in Phase II clinical trials, with about half of the failures due to lack of efficacy, and about a quarter of the failures due to toxicity. Reasons for failure include that pre-clinical models may be poor predictors of clinical benefit.

[0008] Thus, there is a need in the art for an integrated electronic system that supports (1) the scalable storage of genetic variant and phenotype data for hundreds of thousands of subjects, (2) the scalable, automated analysis of genetic variant-phenotype associations, and (3) the automated computational analysis of genetic variant-phenotype association results.

SUMMARY

[0009] It is to be understood that both the following general description and the following detailed description are exemplary and explanatory only and are not restrictive. Methods and systems for generating and analyzing genetic variant-phenotype association results are disclosed.

[0009a] According to a first aspect of the present invention, there is provided a method, comprising:

querying a genetic data component for a variant associated with a gene of interest;

passing the variant to a phenotypic data component as a query for a cohort possessing the variant;

passing the variant and the cohort to a genetic variant-phenotype association data component to determine an association result between the variant and a phenotype of the cohort;

passing the association result to a data analysis component to store and index the association result by at least one of the variant and the phenotype; and

querying the data analysis component by a target variant or a target phenotype, wherein the association result is provided in response.

[0009b] According to a second aspect of the present invention, there is provided a computer-readable medium having computer-executable instructions adapted to cause a computer system to perform the method of:

querying a genetic data component for a variant associated with a gene of interest;

passing the variant to a phenotypic data component as a query for a cohort possessing the variant;

passing the variant and the cohort to a genetic variant-phenotype association data component to determine an association result between the variant and a phenotype of the cohort;

passing the association result to a data analysis component to store and index the association result by at least one of the variant and the phenotype; and

querying the data analysis component by a target variant or a target phenotype, wherein the association result is provided in response.

[0009c] According to a third aspect of the present invention, there is provided an apparatus comprising:

one or more processors; and

a memory storing processor executable instructions that, when executed by the one or more processors, cause the apparatus to:

query a genetic data component for a variant associated with a gene of interest;

pass the variant to a phenotypic data component as a query for a cohort possessing the variant;

pass the variant and the cohort to a genetic variant-phenotype association data component to determine an association result between the variant and a phenotype of the cohort;

pass the association result to a data analysis component to store and index the association result by at least one of the variant and the phenotype; and

query the data analysis component by a target variant or a target phenotype, wherein the association result is provided in response.

[0010] The present methods and systems provide an integrated electronic system comprising a genetic data component, a phenotypic data component, an automated

What is claimed is:

1. A method, comprising:

querying a genetic data component for a variant associated with a gene of interest;

passing the variant to a phenotypic data component as a query for a cohort possessing the variant;

passing the variant and the cohort to a genetic variant-phenotype association data component to determine an association result between the variant and a phenotype of the cohort;

passing the association result to a data analysis component to store and index the association result by at least one of the variant and the phenotype; and

querying the data analysis component by a target variant or a target phenotype, wherein the association result is provided in response.

2. The method of claim 1, wherein querying the genetic data component for the variant associated with the gene of interest comprises:

receiving a plurality of variants from exome sequencing data;

assessing a functional impact of the plurality of variants;

generating an effect prediction element for each of the plurality of variants; and

assembling the effect prediction element into a searchable database comprising the plurality of variants, wherein the query for the variant is applied to the searchable database.

3. The method of claim 2, wherein the effect prediction element is a prediction of an effect of a variant on a biochemical structure and function of an expression product of a variant gene.

4. The method of claim 2, wherein generating the effect prediction element for each of the plurality of variants comprises identifying at least one of the plurality of variants as a potential loss-of-function (pLoF) candidate.

5. The method of claim 4, wherein identifying each of the plurality of variants as a pLoF candidate comprises identifying a level of quality associated with each variant call for each of the plurality of variants and applying a pLoF definition based on the level of quality.

6. The method of claim 1, further comprising

receiving a plurality of Variant Call Format (VCF) files;

determining one or more variant sites in common among the plurality of VCF files;

generating an index identifying presence or absence of the one or more variant sites for each of the plurality of VCF files;

encoding a plurality of attributes as a single value for each of the plurality of VCF files; and

generating a final VCF file comprising the index and the encoded plurality of variables, wherein the query for the variant is applied to the final VCF file.

7. The method of any of claims 1-6, wherein passing the variant to the phenotypic data component as the query for the cohort possessing the variant comprises identifying one or more individuals associated with the variant as the cohort.

8. The method of claim 7, further comprising:

determining one or more de-identified medical records associated with the one or more individuals; and

grouping the one or more de-identified medical records into a data set.

9. The method of claim 8, wherein the one or more de-identified medical records comprises one or more of phenotype data or medical information.

10. The method of any of claims 1-9, wherein passing the variant and the cohort to the genetic variant-phenotype association data component to determine the association result between the variant and the phenotype of the cohort comprises performing one or more statistical tests.

11. The method of claim 10, wherein the one or more statistical tests comprises one or more of, Hardy–Weinberg equilibrium (HWE) analysis, Fisher's exact test, a BOLT-LMM analysis, a logistic regression, and a linear mixed model, for a binary phenotype.

12. The method of claim 10, wherein the one or more statistical tests comprises one or more of, a linear regression, a linear mixed model, ANOVA, for a quantitative phenotype.

13. The method of any of claims 1-12, further comprising providing the association result to a pedigree interface, wherein the pedigree interface constructs a pedigree indicating one or more relationships between one or more individuals in the cohort.

14. A computer-readable medium having computer-executable instructions adapted to cause a computer system to perform the method of:

querying a genetic data component for a variant associated with a gene of interest;

passing the variant to a phenotypic data component as a query for a cohort possessing the variant;

passing the variant and the cohort to a genetic variant-phenotype association data component to determine an association result between the variant and a phenotype of the cohort;

passing the association result to a data analysis component to store and index the association result by at least one of the variant and the phenotype; and

querying the data analysis component by a target variant or a target phenotype, wherein the association result is provided in response.

15. An apparatus comprising:

one or more processors; and

a memory storing processor executable instructions that, when executed by the one or more processors, cause the apparatus to:

query a genetic data component for a variant associated with a gene of interest;

pass the variant to a phenotypic data component as a query for a cohort possessing the variant;

pass the variant and the cohort to a genetic variant-phenotype association data component to determine an association result between the variant and a phenotype of the cohort;

pass the association result to a data analysis component to store and index the association result by at least one of the variant and the phenotype; and

query the data analysis component by a target variant or a target phenotype, wherein the association result is provided in response.