AUTOMATED GENOME INTERPRETATION AS A UTILITY TO PRIORITIZE VARIANTS FOR CLINICAL AND STATISTICAL FOLLOW-UP

Xiaodi Wu

A thesis submitted to the Department of Biology

in partial fulfilment of the requirements for the degree of

Bachelor of Arts

Harvard University

Cambridge, Massachusetts

March 2009

Abstract

Recent developments in genome sequencing technology have moved closer to ushering in an era of unprecedented amounts of genetic information. However, data linking genetic variants to traits are currently quite incomplete, and efforts to interpret the four currently published genomes have not yielded many results of note. In this project, I present an interpretation utility that draws upon several data sources and a substitution matrix-based predictive algorithm to prioritize variants for follow-up. Application of this utility to existing genomes replicates phenotypes claimed by previous authors but also casts doubt on their significance, in addition to suggesting additional phenotypes. While demonstrating weaknesses in the underlying data, these results are a step forward in personal genome interpretation and represent part of an evolving bioinformatic approach to extracting functional predictions from large sets of genetic variants.

Acknowledgments

With gratitude I acknowledge the help and advice of Prof. George Church, whose support has been tremendously helpful in these past years; the data, ideas, and technical support I have received from Shawn Douglas, Alexander Zaranek, Abraham Rosenbaum, Michael Chou, Jason Bobe, John Aach, and others; the clinical expertise provided by Joe Thakuria over many months; the technical support I have received Tom Clegg and Ward Vandewege; and the work done by Kay Aull, Tiffany Chan, Resmi Charalel, Cynthia Chi, Katie Fifer, Hetmann Hsieh, Deniz Kural, Christopher Nabel, Zachary Sun, and Michael Wang as a part of the initial exercise that ultimately initiated this project. Indispensable also were the many conversations about statistics with Robert Sinnott, who has forgotten more about the subject than I shall ever know.

I authored all code produced in association with this specific project, with the exception of those portions drawn from open source implementations as indicated in the text. Any clinical interventions presented were suggested by JT, while I was responsible for other variant analyses. GC provided the utility's name ("Trait-o-matic").

AR, AZ, JA, JB, JT, MC (and GC) collected and analysed data from participants of the Personal Genome Project (PGP). TC and WV implemented several information technology resources in use at the laboratory, including the sequence placement pipeline used for PGP data ("Genomerator").

Contents

Introduction

Alongside the release of individual genomes for J. Craig Venter,[1] James D. Watson,[2] and two anonymous individuals,[3, 4] recent efforts have focused on methods to produce higher-quality genome sequences more cheaply, a movement that will lead eventually to the general availability of personal genomes for a modest cost.[5] For clinicians and scientists, the arrival of these data will usher in an era of unprecedented quantities of information about human genetics.

Yet the availability of ever-increasing amounts of sequence data alone will not be sufficient to bring about greater insight into individuals' genetic constitution or to promote a better understanding of how such information can be used responsibly. In addition to lowering the cost of producing a complete genetic sequence for individuals, efforts are under way to provide insight into their microbiome,[6] VDJome,[7] and other such "-omes." These sources of information hold promise in bridging the gap between genetic and environmental factors by providing quantitative data on the individual's multicellular environment.

Furthermore, studies continue to investigate disease-causing alleles and their mechanisms. The data produced by studies linking genotype and phenotype are essential, of course, for evidence-based interpretation of any individual's genetic sequence; today, these data are significantly incomplete even in the coding regions, the best-studied fraction of the human genome.[8] Traditional studies of genotype–phenotype associations have generally examined a handful of genes and traits due to cost and technical limitations. While such studies have been relatively successful at elucidating the genetic bases of diseases caused by a single locus with classical Mendelian inheritance (so-called "Mendelian diseases"), finding reproducible results has been less successful for more complex traits.[9] More recent studies use haplotype-tagging single nu-

cleotide polymorphisms (htSNPs) as statistical proxies for neighbouring variants, making use of linkage disequilibrium in order to examine a wider array of genotype–phenotype correlations. Since htSNPs have non-trivial minor allele frequencies in the population, the result has been the identification of many somewhat common alleles conferring modest increased risk for various diseases, often difficult to replicate.[10] Hence, a naïve interpretation based on the presence or absence of these genomic variants is guaranteed (statistically) to generate a plethora of "risk alleles" for each individual that have questionable relevance after correcting for multiple hypothesis testing. The possibility of triggering potentially invasive or harmful interventions based on false positives renders the approach undesirable for clinical use, while the inability to ascertain disease-causing loci precisely due to the use of htSNPs as proxies also presents limitations for study in the laboratory setting.

This project builds on these studies of genotype and phenotype to explore our understanding of what individual genomes can reveal about traits. I examine sources of information available for genome interpretation, present a utility that draws on these sources to interpret personal genomes, and analyse the effectiveness of methods employed in such interpretation. The motivations for this project can best be made clear first through a broad survey of five sequencing methods to explain the current state of genome sequencing, then through an overview of existing methods and resources used in genome interpretation.

Sequencing. The classical Sanger method of DNA sequencing is based on the use of labelled chain-terminating dideoxynucleotide triphosphates (ddNTPs). These are randomly incorporated into a growing DNA strand complementary to the template of interest, separated by gel electrophoresis, and visualized.[11] As the accuracy of Sanger sequencing drops after several hundred base pairs (bp), techniques for splitting apart longer sequences of interest for sequencing

and then reassembling these fragments computationally are required. The Human Genome Project elected to address this issue by using hierarchical shotgun sequencing; in this procedure, the genomic DNA was split into large-insert clones (between 100–200 kilobases); each selected clone was then individually sequenced by random fragmentation into overlapping Sanger reads later reassembled computationally (a strategy known as the shotgun method).[12, 13] By contrast, Venter's genome was produced using whole-genome shotgun sequencing, where the entire genome is fragmented for Sanger sequencing without the use of large-insert clones. This process produced approximately 32 million reads for Venter's genome which, when assembled, yielded an average of 7.5-fold coverage.[1]

To reduce costs, modern high-throughput systems attempt to increase output using alternative chemistries that afford massive parallelism and cyclic data capture. Pyrosequencing, as developed by 454 Life Sciences, is a cyclic sequencing-by-synthesis method that does not use chemically bound fluorophores. Rather, the pyrophosphate released upon nucleotide incorporation is detected by the use of a luciferase; to correlate light release with a particular base, the four sets of deoxynucleotide triphosphates (dNTPs) must be added individually between washing steps.[14, 15] Watson's genome was produced in this manner, with over 24 billion bases generated in 100 million reads; 88% of these reads aligned to the reference sequence to yield an average of 7.4-fold coverage across the genome.[2] Note that reads produced by Sanger sequencing, which can exceed 800 bp, are generally longer than 454 reads, which can extend up to 400 bp.

The fluorescent sequencing method available from Illumina uses labelled reversible nucleotide terminators to interrogate DNA. Fragments of DNA tethered to a glass surface undergo rounds of amplification by polymerase chain reaction, after which sequencing primers and reversible dNTP terminators are introduced. When unincorporated dNTPs have been washed away,

the identity of the incorporated base is revealed by laser excitation of the bound fluorophore (unique for each base). The terminating group and fluorophore are then cleaved for additional cycles of dNTP incorporation and laser excitation;[14] according to genome centres, this process can be repeated for as many as 110 cycles. Two subsequently completed whole genomes make use of this next-generation sequencing technology: one of an unidentified Yoruba Nigerian male (HapMap NA18507), in which ~4 billion paired 35-base reads covered 99.9% of the reference sequence at ~40-fold average coverage,[3] and one of an unidentified Han Chinese male, in which 3.3 billion 35-base reads (some paired, some single) covered 99.97% of the reference sequence at ~36-fold average coverage.[4]

Sequencing-by-ligation protocols use the discriminating capacity of ligases rather than polymerases to interrogate DNA. Briefly, synthetic anchor primers are first hybridized to immobilized DNA, after which populations of labelled degenerate oligomers are introduced to be ligated. The identity of a particular position (or positions) in the oligomer is correlated with its attached fluorophore, so the identity of the corresponding position on the immobilized DNA can be queried to the extent that the ligase is sensitive to complementarity; this ligase sensitivity is sufficiently accurate for sequencing purposes only up to a certain distance between the query position and ligation junction. One protocol, used in the Polonator sequencing instrument, repeats rounds of single ligations by stripping the anchor primer and probe, starting anew using probes that have been labelled to query different positions.[16] In the Applied Biosystems (ABI; now Life Technologies) SOLiD protocol, 8-bp oligomers are used that consist of three degenerate bases at one end but three universal bases at the other, which are cleaved after each round of ligation to expose a 5′ phosphate for further rounds of extension by ligation; this permits cyclic querying of every fifth base. Furthermore, these oligomers are labelled according to the combination of bases

at the fourth and fifth positions (16 combinations), increasing error detection possibilities as two adjacent bases are queried simultaneously.[14] ABI has claimed sequencing of a Yoruba Nigerian male (HapMap NA18507) using this technology, though as of this writing no report has been published presenting the genome. Meanwhile, Complete Genomics has announced plans to offer US$5000 human genome sequencing based on proprietary sequencing-by-ligation approaches.[17]

Finally, nanopore sequencing is a technology still under development that eschews the use of either polymerases or ligases during readout. In this method, unlabelled RNA or single-stranded DNA (ssDNA) molecules are driven electrophoretically through a nanoscale pore, which is sufficiently small that characteristic changes in ionic current caused by translocating nucleotides can be used to distinguish between the differently sized bases. Unfortunately, it has been found that the rate at which ssDNA polymers translocate through a nanopore under an electric potential is too high to resolve individual nucleotide identity via ionic current.[18] One attempt to surmount this limitation involves the use of an exonuclease to cleave individual nucleotides from ssDNA polymers in sequence; when the exonuclease is appropriately attached to the nanopore, these liberated deoxynucleotide monophosphates (dNMPs) have a high probability of translocating through the pore in the order in which they are cleaved. Recently, one group reported a successful implementation of single-molecule nanopore sequencing with an average accuracy of 99.8%.[19]

Here, I have given only a very brief treatment of some particular sequencing technologies involved in recent sequencing efforts, and have omitted techniques for DNA amplification prior to these readout chemistries as well as algorithms for image processing that follow; both are areas of active research. Differences in technology have an impact not only on cost, but also on

the kinds of error encountered and consequent strategies for evaluating sequence quality, among other considerations.

Several ambitious projects have been announced to sequence large numbers of individuals using some of these next-generation sequencing technologies. An international consortium announced in 2008 the launch of the 1000 Genomes Project, with the aim of providing more detailed and biomedically relevant information on genetic variants than is available from current sources.[20] With a focus on the human exome (coding exons of the genome), the Personal Genome Project (PGP) has completed a pilot effort with 10 individuals and is developing protocols to sequence 100,000 participants. Analysis of this sequence information in conjunction with personal medical records will be undertaken with the aim of better connecting traits to both genes and environment.[21]

Interpretation. With the impending arrival of a large number of genomes, sequencing projects have also turned to efforts to interpret these data, systematically and—because manual efforts are prohibitively laborious—with automated aids. The first attempts at analysing individual genomes for phenotype have been, in general, unimpressive.

Levy *et al.* described several loci in Venter's genome examined (apparently manually) for phenotype. They determined that the donor is neither affected nor a carrier for Huntington disease (HD), and has alleles associated with tobacco addition, increased risk for heart disease, decrease risk for heart disease, and a small number of other traits.[1] Subsequently, a follow-up report by Ng *et al.* examined systematically Venter's variants in disease genes by consulting a table in dbSNP mapping phenotypes in Online Mendelian Inheritance in Man (OMIM) to dbSNP rs IDs, finding seven of note. However, they found all seven to be common (minor allele frequency (MAF) > 0.05), functionally neutral, or both.[22]

Wheeler *et al.* described 11 single nucleotide polymorphisms (SNPs) in Watson's genome matching disease-causing or other recognizable phenotypes based on consultation with the Human Gene Mutation Database (HGMD), though Watson exhibits none of these phenotypes.[2] (Note that while SNPs have been traditionally defined as benign single base changes, or those with a minimum allele frequency of 1%, current usage with respect to individual genomes has not discriminated on the basis of allele frequency or function.) Ultimately, the Wheeler study showed little information of clinical value, and at least one observer suggested that reliable predictions from genome sequences are several years away.[23]

In their analysis of the Han Chinese male, Wang *et al.* surveyed the same database table used by Ng *et al.* and found one mutation for a recessive deafness disorder for which the subject is claimed to be a carrier. A search of alleles linked to complex diseases from curated data sources additionally revealed several predisposing alleles for tobacco addiction, Alzheimer disease (AD), and diabetes, among others; the authors revealed that their subject is a heavy smoker.[4] By contrast, Bentley *et al.* included no summary of phenotype inferences for their report on the sequence of the Yoruba Nigerian male.[3]

Subsequent to the publication of these four genomes, two groups with as yet unpublished individual genomes have used this project's utility as an interpretative aid for SNPs; we have named this utility "Trait-o-matic." Below, I survey some resources used in genome interpretation—namely, the reference sequence and five databases—and very briefly discuss a handful of predictive algorithms related to the task.

As the Human Genome Project produced the first largely complete sequence of the human genome, all four published genomes have been aligned against this reference sequence, and variants lists are produced from comparisons with it. This reference sequence, now curated by

the Genome Reference Consortium (GRC), is available from the National Center for Biotechnology Information (NCBI) and other international public databases, the most recent release being build 36.3 (March 2008). Point increments of NCBI reference build numbers have indicated annotation updates (protein alignments, repetitive sequence masking, etc.) or addition of alternative assemblies without changes in the reference genome assembly; hence, a particular set of coordinates given for build 36.1 (March 2006) refers to the same sequence in build 36.3.

It is to be noted that the reference sequence itself is haploid and based on a composite of tissue lines so that it is not the sequence of any particular (half-)person.[24] Donors were recruited via newspaper advertisements in 1997, and libraries were first constructed from one male (RPCI-11) and one female (RPCI-13); as the experiment was double-blind, additional samples could not be obtained from RPCI-11 when necessary.[25] Other libraries were used in addition to these to construct clones for sequencing, but it is known that most clones were drawn from the RPCI-11 library;[12] consequently, the reference sequence is largely that of a male from Buffalo, New York. Still, the current sequence is clearly neither a true reference sequence, as it is a patchwork of sequences from different individuals, nor a consensus sequence, as it contains rare alleles.

In the capacity in which we (and others) use this reference, it would be optimal to have annotations that describe loci where the reference contains a rare or potentially deleterious allele. Others have argued that the genomes of James D. Watson and other generally healthy individuals should be used to improve the reference sequence directly, with the aim of producing a consensus sequence representing only major alleles.[22] The use of a sequence that consistently contains alleles most common in the population does necessarily simplify the task of identifying minor alleles when comparing genomes. However, as our ultimate aim is to identify not minor alleles but disease-associated alleles, editing the reference sequence to produce a consensus would not

be the most productive approach. Rather, direct editing would obscure the distinction between minor and disease-associated alleles, and the resultant data would remain unable to (a) represent actual allele frequency data, or indicate whether alleles appear to be in Hardy-Weinberg equilibrium, (b) describe situations where a heterozygous genotype is most common in the population due to balancing selection, or (c) accommodate differences due to population structure without the proliferation of multiple reference sequences representing various ancestries.

Instead, the revised Cambridge reference sequence (rCRS) of human mitochondrial DNA (mtDNA) demonstrates an alternative approach, where sequencing errors but not rare polymorphisms have been corrected, to the extent that correction would not affect compatibility with historical numbering.[26] Seven rare polymorphisms are noted separately from the sequence, allowing the rCRS to remain unchanged indefinitely and without consideration of population structure because it is not a consensus sequence meant to reflect allele frequencies in the population. Those seeking to compare mtDNA from particular populations may still align against a common reference, while data on the alleles most common for those populations need to be consulted separately.

It is likely that, with refinements in sequencing technologies, the use of the reference sequence as a scaffold for genome assembly will become outdated. Currently, genome-wide surveys of large-scale structural changes are typically performed after alignment of sequence data against the reference,[3, 4] and targeted *de novo* assembly of reads in regions with evidence of structural changes is carried out where necessary to investigate these changes in detail. When entire genomes are assembled *de novo*, all structural changes are represented in the consensus sequence without the need for an independent step to survey their presence, offering an advantage over the current workflow and removing one key use for the reference sequence. However,

some standard will still be necessary in order to express differences between genomes. For variants not near coding regions, it is of course necessary to give coordinates and sequences with reference to some genomic contig; for those in or near coding regions, well-developed standards for notation provide options for representing changes with respect to a genomic contig, coding sequence, or (where applicable) RNA or polypeptide sequence.[27] Whatever becomes used as a reference for comparison, it will remain necessary to consider the presence or absence of particular alleles in that reference to understand fully what is implied by a deviation from it. I analyse some properties of the current reference sequence as part of this project; though it may be that future genomes will use a different reference, a similar exercise would be productive for any such sequence.

Authors of published genomes have used several sources in addition to the reference sequence for interpretation. Reports presenting all four genomes have made extensive use of an NCBI database known as dbSNP to classify variants as previously observed or "novel." Launched during sequencing efforts for the Human Genome Project, dbSNP functions as a repository where individual submissions, represented by "ss" IDs, are collated into references for each variant designated by "rs" (reference SNP) IDs.[28] Besides SNPs, dbSNP accepts small indels, microsatellite repeats, and other types of variation regardless of frequency,[29] and it displays genomic locations, population frequencies, and other information (principally in the form of cross-references to federated databases) alongside user submissions. Although it is possible to access dbSNP using a web interface or to query dbSNP programmatically for small amounts of data on particular variants, these options are not practical for use on larger datasets. For millions of SNPs, one option is to retrieve the entire database in a structured XML format to be parsed and stored for local use, and another is to replicate the entire database as it is stored at NCBI for

local use. Both represent high barriers to entry because of the size and complexity of dbSNP, as well as the use of proprietary features from Microsoft in its internal database schemas. The University of California, Santa Cruz (UCSC) provides, among other bioinformatics resources, one method to work around these limitations by offering parsed tables that summarize genomic location and inferred function for each dbSNP rs ID.[30] Corresponding to the most recent dbSNP build (129; April 2008) is the UCSC table "snp129."

Ng *et al.* have reported in their analysis of Venter's genome that most changes likely to affect protein function tend to be heterozygous, rare, or novel.[22] While zygosity is determined based on each individual's data, and novelty from presence or absence in dbSNP, other resources need to be consulted for data on allele frequency in the population. In 2002, the International HapMap Consortium began cataloguing a collection of SNPs across the genome in individuals from four geographically distinct populations; the purpose was to construct a haplotype map (HapMap) in order to enable the use of subsets of these SNPs as proxies.[31] In addition to linkage disequilibrium (LD) data, the HapMap is a valuable source for allele and genotype population frequencies. Data from phase I of the project, published in 2005, included over 1 million SNPs; additional data comprising over 2 million additional SNPs were published in 2007 as part of phase II.[32] The most recent HapMap draft as of writing is release 27 (February 2009), which contains information collected from 1115 individuals in 11 populations when pre-release data from phase III genotyping is included. HapMap frequency data are also integrated into dbSNP, but are more difficult to parse from dbSNP than in the format retrievable directly from the International HapMap Consortium.

It has been mentioned that some authors interpreting individual genomes have used a database table in dbSNP linking phenotypes in OMIM to dbSNP rs IDs. Prior to this project, this

table was the most complete freely available and machine-readable catalogue of allele–phenotype correlations, known as "OmimVarLocusIdSNP" and (as of dbSNP build 127, retrieved August 2007) mapping 1391 rs IDs to 695 OMIM articles. OMIM itself, the online counterpart to a print publication known as *Mendelian Inheritance in Man* (MIM), is a free-text catalogue of genotype–phenotype information which as of 2007 included over 11,000 gene entries and 6000 phenotype entries. The resource includes both Mendelian and complex phenotypes, and lists allelic variants for each gene deemed to be of interest but not all variants that have been associated with a phenotype.[33] Hence, OMIM represents an interesting and extensive (though intentionally incomplete) collection of allele–phenotype data, of which only a small subset has been parsed into a computer-readable format. As part of this project, I construct a database table by parsing all nsSNP annotations in OMIM.

It has also been mentioned that Wheeler *et al.* use HGMD as a data source for mutations causing disease. This database, maintained at Cardiff University, extracts information from hundreds of journals and locus-specific databases (LSDBs), with one reference to the literature provided for each entry.[34] All published reports of heritable, nuclear lesions responsible for inherited disease or associated with disease are eligible for inclusion,[35] and the database is likely to be the most complete collection of machine-readable allele–phenotype associations currently available. Others have in fact attempted to compare the completeness and accuracy of OMIM and HGMD, finding that both databases contain useful mutations but also inconsistencies;[36] these conclusions, however, were not well accepted by HGMD authors.[37] Complicating such comparisons, HGMD maintains two simultaneous releases: one set of up-to-date data (HGMD Professional) is available for purchase and can be downloaded for local use, while a subset of these data (excluding the most recent additions and a fraction of older entries) is freely available for academic use but

can only be accessed through a web interface. In this project, I use a purchased instance of HGMD Professional 7.1 (March 2007), which contains 44,776 nsSNPs; the most recent version as of writing (Professional 2008.4, December 2008) contains 48,343 nsSNPs.

Furthermore, an unpublished resource known as "SNPedia" presents SNPs and disease associations in a "wiki" format. These data can be edited by users, but are chiefly maintained by two individuals. Many SNPs available on SNPedia are found on microarrays used by direct-to-consumer services such as 23andMe, deCODEme, and Navigenics. Like reports offered by these services, SNPedia describes relative risk estimates for many genotypes listed. These estimates are of limited reliability, however, partly because interactions between an arbitrary number of variants is not well characterized. It is conceivable, for example, that a person found with two SNPs associated with increased risk for a disease may actually have reduced risk for that disease. Only at a time when the sample size is sufficiently large (i.e. the number of genomes analysed) will it be possible to calculate meaningful relative risk values for a large array of phenotypes.

Because SNPedia covers many loci analysed by direct-to-consumer services, this resource also offers an interpretation utility known as "Promethease," which accepts genotype data retrieved from these services and queries SNPedia for information on each genotype. Despite the appearance of a desktop application, Promethease transmits users' genotype data to SNPedia and requires approximately two hours to complete unless the user provides an optional monetary payment to reduce runtime. Although information from SNPedia can be useful, particularly when comparing complete genome data to microarray data, we are unable to use Promethease due to pay-per-run limitations; instead, I retrieve and parse SNPedia data into a table as part of this project.

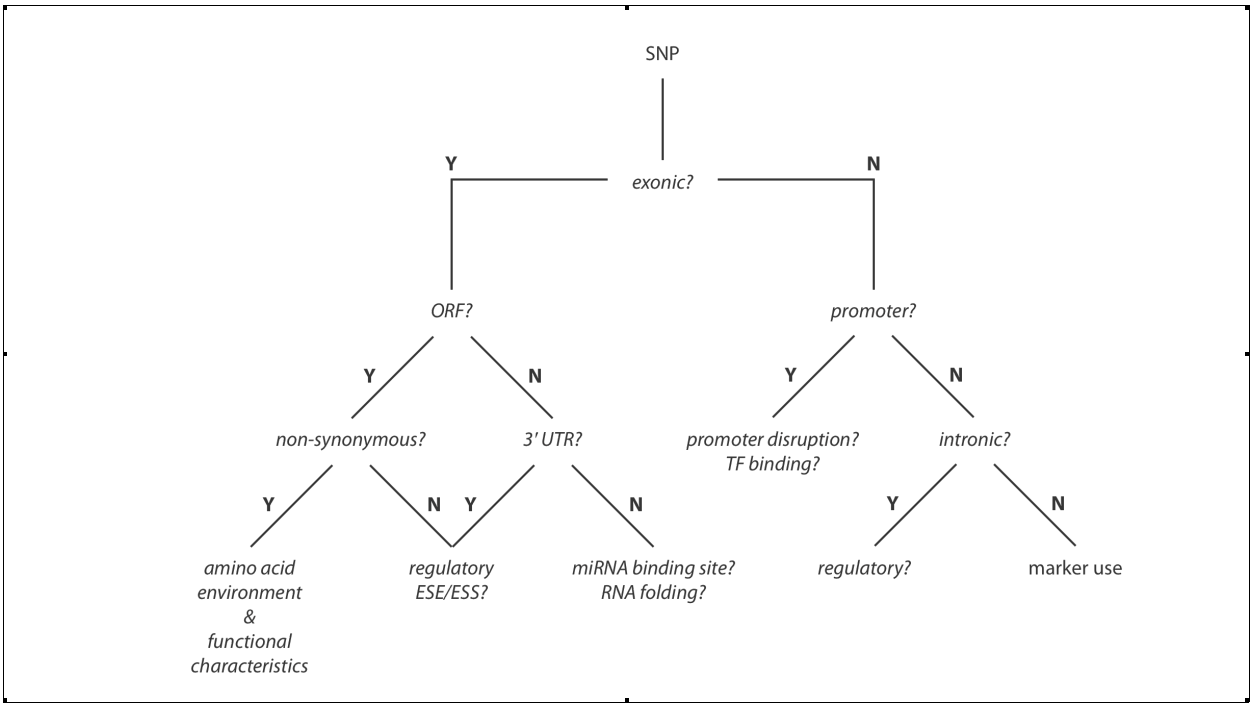

Figure 1. Prediction decision tree for SNPs. Adapted from Plumpton and Barnes.[38]

Finally, as not all alleles have been sufficiently studied or even observed, it is important that interpretation tools are capable of predicting to some extent the effect of a novel mutation. Several considerations can be used to determine the potential effects of a SNP, beginning with whether or not it exists in an exon. Non-exonic SNPs and SNPs in untranslated regions may disrupt promoter or transcription factor (TF) binding, or have effects on RNA folding, splicing, or microRNA (miRNA) binding (**Figure 1**).[38] Several algorithms for detecting regulatory elements[39, 40] and several databases that catalogue these elements[41, 42] are available.

For a first implementation, however, I have chosen to focus prediction on nsSNPs. Several existing amino acid substitution prediction tools are widely used to infer nsSNP impact on protein function. SIFT (Sorting Intolerant From Tolerant) attempts to distinguish deleterious and benign alleles using protein sequence homology, assigning scores that take into account the amino acid change and the degree to which the amino acid position is conserved. Scores given are between 0 and 1, representing the normalized probability that a change is benign; by default,

scores less than 0.05 are considered deleterious, although this threshold can be adjusted.[43] PolyPhen is another tool that attempts to distinguish deleterious and benign alleles; in addition to sequence conservation, this algorithm considers the three-dimensional structure of the affected protein and consults annotations of important residues in the protein database Swiss-Prot.[44] Unlike SIFT, output from PolyPhen simply places each amino acid into one of four categories (benign, possibly damaging, probably damaging, unknown). Besides SIFT and PolyPhen, several other prediction tools using similar methods have also been published.[45]

A chief disadvantage of SIFT, PolyPhen, and related tools concerns the computational resources required for their operation. While precomputed tables for dbSNP entries overcome this issue for previously known SNPs, predictions for novel mutations still need to be calculated using multiple alignments and, in the case of PolyPhen, three-dimensional modelling. Since it would be difficult to evaluate all novel SNPs in this manner within our computational limitations, I attempt to use a simpler algorithm in this project to predict the effect of nsSNPs, and I evaluate the loss of accuracy incurred as compared to SIFT and PolyPhen.

For this project, I have pulled together sequences, databases, and algorithms to create Trait-o-matic as a utility to prioritize variants for manual follow-up. I have also participated in a study of methods to limit Trait-o-matic results to genes available for diagnostic sequencing, and in the application of Trait-o-matic interpretation to PGP partial exomes; reports on both efforts are forthcoming. Here, I present the creation of Trait-o-matic itself and attempt to evaluate the effectiveness of this utility in comparison with existing methods.

Methods

Sequence alignment and variant identification. Input data for early software prototypes were retrieved from publicly released data available from the NCBI Trace Archive. Sequencing reads of approximately 100–300 bp each were available for James D. Watson and J. Craig Venter, and were aligned using BLAT.[46] Subsequently, additional data emerged listing all variants identified in these genomes, obviating the need to align sequences. The genome sequence for an unnamed Han Chinese individual was also published with variant data, while variant data corresponding to the publicly released genome of an unnamed Yoruba Nigerian individual were obtained via correspondence with the authors of that study. Variant data generated as part of the Personal Genome Project were produced by Solexa (now Illumina) sequencing and were aligned using MAQ.[47]

The algorithm used to identify variants is left to each data source. For example, Venter's genome includes short insertions and deletions, while Watson's genome lacks this data. Due to short reads, Solexa reads mapped by MAQ generally lack insertions and deletions. Data provided for Watson, Venter, and the unnamed Han Chinese male were in a common format known as GFF, though each file had its own idiosyncrasies. Using Python, I wrote scripts to convert all data formats into a common GFF format, and subsequent analysis steps all accept this common format as input.

Information extraction. As previously discussed, a table from dbSNP incompletely maps OMIM alleles to dbSNP rs IDs. To extract a more complete set of information, I implemented a script in Python capable of extracting non-synonymous single amino acid changes from the OMIM full text. This led to the creation of a database table with over 11,000 entries, representing a much larger set of OMIM allelic variants. These data are not strictly a superset of

senting a much larger set of OMIM allelic variants. These data are not strictly a superset of the dbSNP table, however, as certain variants either have incorrect information in the text or are compound mutations.

I implemented a script to extract the reference allele for each nsSNP in dbSNP (build 129) to examine reference alleles that appear in OMIM. An attempt using dbSNP tables failed to produce results because a SQL query joining separate database tables to retrieve all relevant fields could not be evaluated within the computational resource limits encountered. A second script made use of dbSNP data retrieved from UCSC which contained all relevant fields in the same table ("snp129").

I implemented another script to extraction data from SNPedia via the MediaWiki API. Structured markup was parsed for information about genotypes and their corresponding effects, and links to PubMed literature references were isolated from the free text and recorded as accompanying references. Where descriptions of effect specified degree but not condition (i.e. "increased risk" instead of "increased risk for [disease]"), free text was parsed for links preceded by the text "associated with" or "association with"; where found, the accompanying text was appended to the effect description. Where descriptions of effect suggested that a genotype was associated only with an "average," "common," or "normal" phenotype, that particular genotype–phenotype pair was discarded. Output data were then manually edited for spelling and consistency. I implemented an accompanying script to format the edited data (~1500 entries) for insertion into a database table, with an optional flag to output only those entries with genotypes homozygous for the reference allele.

Subsequently, I implemented a script to insert HapMap allele frequency data into a database table, summing allele counts for populations of related geographic origin. For example, al-

lele frequencies for East Asians were aggregated from HapMap data for two Chinese populations (CHB, CHD) and one Japanese population (JPT). Aggregation was intended to maximize the HapMap frequency data available for each aggregated region.

Phenotype inferences. Trait-o-matic queries several databases to retrieve information for each variant. I implemented this functionality in a set of scripts written in Python.

A fundamental set of functionality was first constructed as a "utils" library. Portions of this library drew from pre-existing open source code; most notably, a selection of code written in C by W. James Kent, then partially ported to Python as part of the Pennsylvania State University "Galaxy" project, implements functionality to read and write efficiently from a compressed sequence format known as "2bit," by which the entire reference genome can be represented in approximately 700 MB on disk.

Above this foundation, I created Python scripts that query the necessary databases and perform scoring and filtering on the data provided to round out the utility's "core" functionality. These scripts can be invoked via the command line, and are also exposed by a Python script that implements an XML-RPC server, which responds to XML-formatted requests transmitted over HTTP. In addition to core functionality, I implemented a web interface to permit users to view sample data, upload data to the server, and retrieve results in sortable tabular format. This interface passes data to the core via XML-RPC calls; separation of functionality allows the outward-facing web components to be located separately from the core utility, thus making any future necessity to run many interface or core instances in parallel much easier to implement.

Results

Data flow. I have constructed Trait-o-matic as a utility that finds, for each SNP: (a) the reference allele at that locus, and whether it is a previously known SNP in dbSNP (build 127); (b) whether it is a SNP associated with disease, as listed in OMIM, HGMD (Professional 7.1), or SNPedia; and (c) if the SNP is contained within a coding region, what amino acid change is produced if it is non-synonymous, and whether the change is conservative (**Figure 2**). This information is presented in sortable tables separated by data source, with highlighting for rare alleles (frequency < 0.05) and alleles of unknown frequency. Total processing time using one process for ~3 million SNPs is approximately six hours.

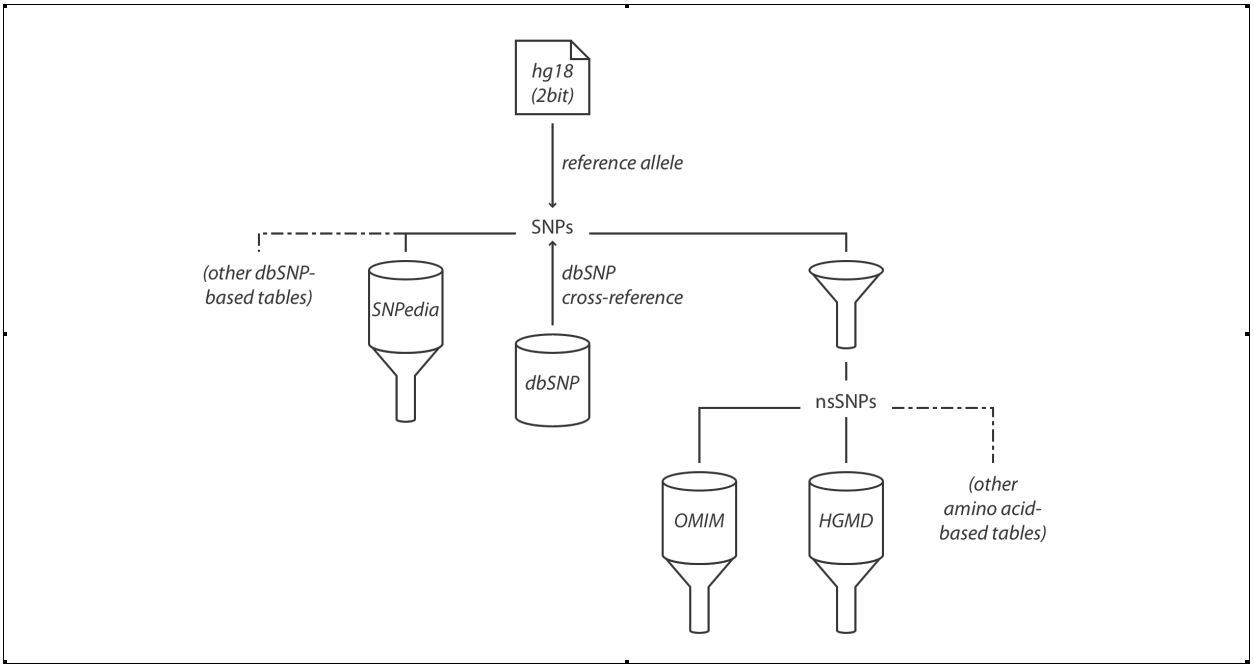

Figure 2. Trait-o-matic data flow. SNPs are checked against the reference genome and dbSNP for additional annotations before being filtered through SNP-based sources (*left*) and through a nsSNP filter, followed by amino acid-based sources (*right*). Matrix-based scoring not shown; dashed lines are not yet implemented.

As OMIM and HGMD data are stored in gene and amino acid coordinates, only nsSNPs are queried against these datasets, and the amino acids corresponding to a given SNP are calculated by a script that translates DNA sequences in the appropriate frame after consulting a database table that describes coding sequence locations (UCSC table "refFlat"). Comparison of the Trait-o-matic function inference algorithm against dbSNP revealed an error in the dbSNP algorithm whereby amino acids encoded across a splice junction may be incorrectly reckoned. In rs9370096, for example, a C→T mutation in coding strand of the polycystic kidney and hepatic disease 1 (*PKHD1*) gene produces a cysteine residue. While the third letter of the corresponding codon is located at the beginning of exon 22, the dbSNP algorithm apparently ignores the splice junction and proceeds incorrectly into the intron, reading [C/T]GG instead of [C/T]GC. Genetic code redundancy shields the effect of this error in the case of the C allele but not the T allele, with the result that dbSNP gives the amino acid change as R760W instead of R760C (**Figure 3**). There is likely to be a small but appreciable number of SNPs in dbSNP systematically affected by this error. As these data also happen to be difficult to parse from dbSNP sources, Trait-o-matic makes no use of dbSNP amino acid calculations and instead relies entirely on its own algorithm.

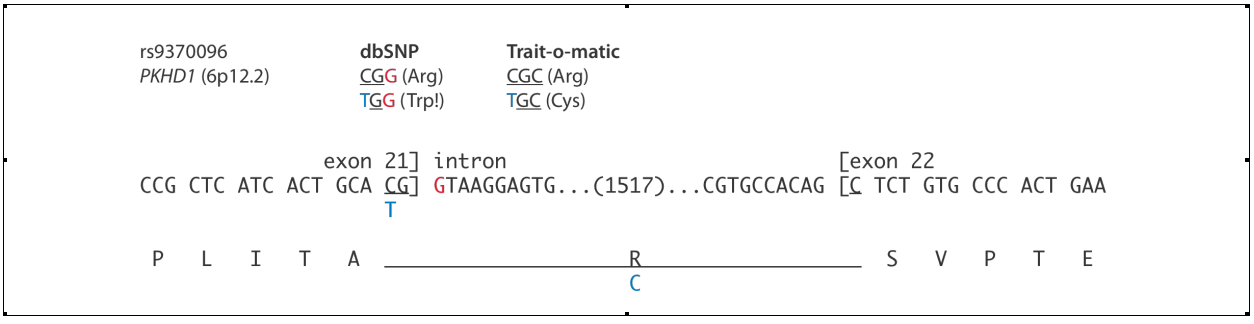

Figure 3. dbSNP incorrectly calculates amino acid changes for codons split by splice junctions. Differences arise between dbSNP and Trait-o-matic function inferences due to erroneous use of intron sequences by dbSNP. *Top*, dbSNP and Trait-o-matic function inferences for rs9370096; *middle*, excerpt of coding strand nucleotide sequence for *PKHD1* near rs9370096 (blue); *bottom*, corresponding amino acid sequence.

Proof-of-concept genomes. Trait-o-matic interpretation was applied to the genomes of James D. Watson, J. Craig Venter, the Han Chinese male ("YH"), and the Yoruba Nigerian male to evaluate the effectiveness of Trait-o-matic procedures and to find variants of interest.

Variant information retrieved for some genomes show discrepancies with published claims. Of Watson's 3,322,093 claimed SNPs, 2,060,544 (62%) were present in the GFF data file retrieved from ‹http://jimwatsonsequence.cshl.edu/› in December 2007. Of these, 6602 were determined to be nsSNPs by Trait-o-matic, 62% of the published claim of 10,569, including 9 of 11 SNPs claimed to match mutations in HGMD that cause disease or other phenotypes. Approximately 11% of these 6602 nsSNPs were not found in dbSNP ("novel"). Of Venter's 3,213,401 claimed SNPs, 3,074,686 (96%) were present in the GFF data file retrieved from ‹http://huref.jcvi.org/› in September 2007; 761,148 more SNPs were retrieved from the same source one year later. In 2007, 6428 nsSNPs were designated by Trait-o-matic, 105% of the published claim of 6114; subsequently, 10,690 nsSNPs were designated from 2008 data, 103% of the published claim of 10,389. Of these 10,690 nsSNPs, ~10% were novel.

For both anonymous genomes, 100% of claimed SNPs were available from supplemental data (YH data retrieved from ‹http://yh.genomics.org.cn/›, Yoruba data retrieved via correspondence with authors). However, Trait-o-matic retrieved 8742 nsSNPs from YH data, 24% more than the published claim of 7062. Subsequent examination revealed that Wang *et al.*, authors of the original data, had already labelled 8166 SNPs as nonsynonymous in the retrieved file, 16% more than their own published claim. 8128 of these 8166 nsSNPs (99.5%) were among the 8742 designated as such by Trait-o-matic. Manual follow-up on a random subset of the remaining 38 shows a mixture of SNPs where: (a) the coding sequence claimed by Wang *et al.* has been permanently suppressed in NCBI databases due to insufficient evidence; (b) dbSNP and Trait-o-

matic agree that the SNP is synonymous; or (c) the claimed exon is absent in the table of coding sequences ("refFlat") consulted by Trait-o-matic, presumably because it belongs to a suppressed isoform. Conversely, 11 of the 8742 nsSNPs designated by Trait-o-matic were erroneously labelled as synonymous by Wang *et al.* because these SNPs are silent in at least one coding sequence but nonsynonymous in at least another (data not shown). The remaining 603 nsSNPs designated by Trait-o-matic but not Wang *et al.* were annotated as non-coding by the authors, likely reflecting the use of an older dataset for coding sequence locations. Again, ~11% of nsSNPs retrieved by Trait-o-matic were novel for YH. For the Yoruba genome, Trait-o-matic identified 9650 nsSNPs, ~20% of which were novel, significantly higher than the remaining genomes and consistent with expectations of greater genetic diversity among African populations.

Table 1. nsSNPs in four published individual genomes, analysed by Trait-o-matic. All genomes show 45–48% homozygous nsSNPs in dbSNP, with the exception of Watson, for which only 62% of the claimed quantity of data were retrieved; of homozygous SNPs, C and G consistently outnumber A and T.

Watson	dbSNP	Novel	*Total*
A/A or T/T	469	48	517
C/C or G/G	719	27	746
Heterozygous	4661	678	5339
Total	5849	753	6602

YH	dbSNP	Novel	*Total*
A/A or T/T	1487	41	1528
C/C or G/G	2261	30	2291
Heterozygous	4035	888	4923
Total	7783	959	8742

Venter	dbSNP	Novel	*Total*
A/A or T/T	1446	63	1509
C/C or G/G	2107	113	2220
Heterozygous	6046	915	6961
Total	9599	1091	10690

Yoruba	dbSNP	Novel	*Total*
A/A or T/T	1312	73	1385
C/C or G/G	2152	75	2227
Heterozygous	4265	1773	6038
Total	7729	1921	9650

Consistent with previous reports,[2, 22] nearly half of known nsSNPs are homozygous in each of these four genomes, with the exception of Watson, for which only 62% of the claimed variant data was available (**Table 1**). In the case of Watson, it can be inferred both from published claims and available data that the majority of missing nsSNPs are homozygous. Also consistent with these previous reports, novel nsSNPs are predominantly heterozygous; these results confirm that novel SNPs tend to be rare, since rare alleles are much more likely to be heterozygous. It can also be observed that the number of homozygous C/C and G/G variants is consis-

tently 60–70% more than the number of homozygous A/A and T/T variants. As these genomes have been sequenced using different methods, these results are unlikely to be a product of systematic bias in any particular sequencing protocol, and could be another characteristic that remains consistent across genomes aligned against the reference sequence.

Trait-o-matic interpretation of Watson's genome (**Supplemental Interactive Figure S1**; **Supplemental Figure S1**) replicated 9 of 11 variants claimed to be associated with disease or other phenotypes; the remaining two were absent in the original data file retrieved. However, 8 of 9 trait-associated alleles were either major alleles or had population frequencies greater than 0.05 among European HapMap samples; the remaining variant (associated with retinitis pigmentosa) had no population frequency information and may be rare. It is notable that Wheeler *et al.* have claimed, based on the retrieval of these variants from the HGMD subset of the genome, that previous estimates predicting fewer than ten lethal equivalents in each person must be too low; however, these supposedly highly penetrant disease-causing alleles are questionable. Also of interest within the Trait-o-matic interpretation of Watson's genome is one variant in the peroxisome proliferator-activated receptor alpha (*PPARA*) gene associated with increased serum levels of total and LDL cholesterol in men, apolipoprotein B (apoB) in men and women, and apolipoprotein C3 (apoC3) in men,[48] but not associated with diabetes.[49] This particular trait-associated allele is rare according to HapMap data and, although not lethal, appears to be meaningfully associated with lipid metabolism.

Again in the case of Venter's genome (**Supplemental Interactive Figure S2**; **Supplemental Figure S2**), the authors' claimed variants of interest were replicated by Trait-o-matic; the chief difference between Trait-o-matic interpretation and results from Ng *et al.* concerned the A171T and D444H alleles encoded in the biotinidase (*BTD*) gene, associated with biotinidase

deficiency.[50] In both cases, HapMap data claim that the trait-associated allele is extremely rare, contradicted by studies cited in Ng *et al.*; this discrepancy is likely due to small sample size in HapMap. Furthermore, variant data claim Venter as homozygous for H444, which would suggest that he is affected, while Ng *et al.* remark that Venter is in fact heterozygous, presumably on the basis of resequencing results. This example demonstrates how low coverage in a sequence creates issues for interpretation and requires targeted resequencing for confirmation. Also of interest in Venter's genome is a homozygous variant in the protoporphyrinogen oxidase (*PPOX*) gene associated with recessive porphyria variegata in OMIM. The trait-associated allele is somewhat rare in the European population, according to HapMap data, and has been erroneously associated with porphyria because it was found in affected individuals with a chain-termination or splice mutation in *cis*, but is itself located in a region poorly conserved between species.[51] Hence, manual review of the literature confirms that the subject should not be affected.

In the YH genome, Trait-o-matic replicated the variant associated with recessive deafness (**Supplemental Interactive Figure S3**; **Supplemental Figure S3**), the allele frequency of which was not available from HapMap for Asian populations. Of interest among variants shortlisted by Trait-o-matic are two that are associated with recessive congenital insensitivity to pain with anhidrosis (CIPA) according to one source,[52] but are rare neutral polymorphisms according to another.[53] In any case, it is not expected that YH should be affected as a heterozygote, although the possibility of this allele being a lethal variant carried by YH is of interest. Meanwhile, in Trait-o-matic interpretation of the Yoruba genome (**Supplemental Interactive Figure S4**; **Supplemental Figure S4**), we noted a variant in the potassium voltage-gated channel, Isk-related family, member 2 (*KCNE2*) gene associated with susceptibility to acquired long QT syndrome. This variant had no associated information in either HapMap or dbSNP, though at least one study has

identified its presence in 3% of healthy black subjects.[54] The degree to which arrhythmia susceptibility may be affected by this variant is hence somewhat unclear, although our review of the literature suggests that clinical follow-up would be appropriate.

On the basis of these Trait-o-matic results, we are compiling additional information and observations on these four genomes, including clinical recommendations, as part of a forthcoming report on all public individual genomes.

To score nsSNPs not found in database sources, Trait-o-matic uses a simple scoring system based on a substitution matrix to minimize computational demands, inferring that a serious mutation—a non-conservative or nonsense mutation—in a gene associated with a particular disease is more likely to be deleterious. Of the two commonly used sets of substitution matrices, I have chosen to use a BLOSUM (block substitution matrix) over a PAM (point accepted mutation) matrix because the latter is asymmetric. The presence of symmetry obviates consideration of which allele at a polymorphic locus is ancestral; this is not difficult for novel mutations (the reference allele can reasonably be taken as ancestral) but requires additional database queries for common polymorphisms. Of the block substitution matrices, I have used the highest threshold of sequence similarity (BLOSUM100), since we are only examining variants within the same species. I have also taken the negative of the matrix in answer to the intuitive notion that less conservative amino acid changes ought to have a higher score, since these scores are used as a measure of potential deleterious effect. Though this matrix-based procedure assigns integer values normalized so that stop codons have a score of 10, I make no assumption about any linear progression of phenotypic effect between integer increments, and hence use non-parametric methods to assess the effectiveness of this scoring method in discriminating benign and deleterious mutations.

As expected, pairwise Mann-Whitney *U* tests give that score distributions do not differ significantly for nsSNPs in the four complete genomes. Two pairwise tests showed $P < 0.05$ (between the Venter and YH genomes ($U = 4.75 \times 10^7$, $n_1 = 10{,}690$, $n_2 = 8742$, $P = 0.0437$ with continuity correction, two-tailed) and between the Venter and Yoruba genomes ($U = 5.26 \times 10^7$, $n_1 = 10{,}690$, $n_2 = 9650$, $P = 0.0118$ with continuity correction, two-tailed)), but these results are not significant after Bonferroni correction for multiple hypotheses. However, the score distribution of any complete genome does differ significantly with that of aggregated PGP exome data (the least significant *P*-value arising between the Venter genome and the PGP, $U = 3.62 \times 10^7$, $n_1 = 10{,}690$, $n_2 = 7189$, $P = 3.36 \times 10^{-11}$ with continuity correction, two-tailed). Also as expected, the OMIM dataset is shifted significantly towards higher matrix-based scores as compared with any genome; a representative test between OMIM and Watson's genome gives $U = 2.10 \times 10^7$, $P < 2.2 \times 10^{-16}$ with continuity correction, one-tailed (**Figure 4a–f**).

Since nearly all OMIM alleles are deleterious and ~80% of nsSNPs in an individual's genome are estimated to be benign,[44] I use OMIM alleles as positive controls and Watson's alleles negative controls to estimate true and false positive rates for the use of matrix-based scores in predicting deleterious alleles. In this case, plotting these data yields an area under the receiver operating characteristic (ROC) curve of 0.695, and setting a decision boundary at $\geq +3$ yields a maximal difference of 29.1% between true and false positive rates (**Figure 4g**). As deleterious alleles are correlated with higher scores, one expects that a true set of negative controls would show lower classification error.

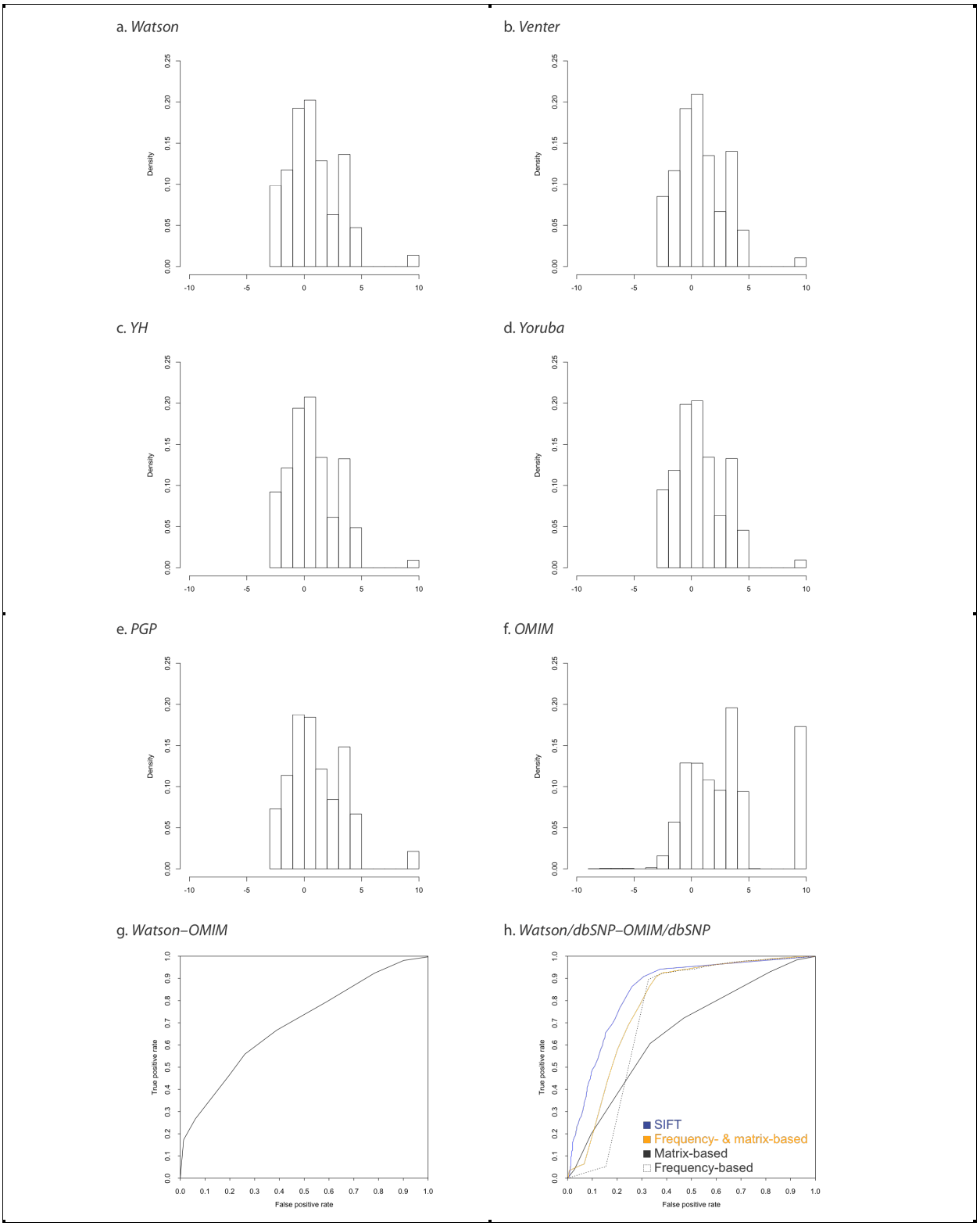

Figure 4. BLOSUM-based scoring of nsSNPs. *(a–f)* Score histograms for individual genomes, aggregated PGP data, and OMIM. *(g)* ROC curve, taking Watson's nsSNPs as benign variants and OMIM as deleterious (area = 0.695). *(h)* ROC curves using only variants mapped to dbSNP.

To compare the effectiveness of BLOSUM100-based scoring with that of SIFT, I retrieved precomputed SIFT predictions for Watson and OMIM nsSNPs. As the most efficient method of accessing these predictions is through submission of dbSNP rs IDs via a web interface, I limited positive controls to OMIM alleles mapped directly to dbSNP rs IDs (in OmimVarLocusIdSNP) and negative controls to Watson alleles found in dbSNP. For additional comparison, ROC curves were plotted for SIFT predictions, BLOSUM100-based scores, MAF-based scores, and summed BLOSUM100/MAF-based scores based on this set of positive and negative controls. To generate MAF-based scores, I consider the lowest MAF across aggregated HapMap populations and apply a logarithmic function $f(\text{MAF})$ such that $f(0.5) = -10$, $f(0.05) = 0$, $f(0.005) = 10$, to a maximum score of 15. I presume that polymorphisms at loci with no frequency data are rare, and assign them a score of 15. This function was somewhat arbitrarily chosen so that I could assign approximately equal weight to the BLOSUM100-based score and MAF-based score by addition, and was devised without the use of a training set for parameter optimization. Examination of ROC curves shows that SIFT (area under curve = 0.848) had the highest predictive accuracy and BLOSUM100-based scores had the lowest (area = 0.659), but summed BLOSUM100/MAF-based scoring dramatically increased accuracy (area = 0.787) and was more effective than MAF-based scoring alone (**Figure 4h**).

PolyPhen claims still higher accuracy than SIFT,[45] but because PolyPhen results are qualitative only (benign, possibly damaging, probably damaging, unknown), an ROC curve cannot be generated. Still, if we accept PolyPhen's claimed true positive rate of 82% and false positive rate of 8%, the BLOSUM100-based scoring method clearly provides inferior results for predicting deleterious alleles, but in exchange for significant increases in computational speed. It remains to be seen if adjusting MAF-based scoring or its combination with BLOSUM100-based

scoring using training sets can push the accuracy of this simple algorithm beyond that of SIFT and PolyPhen.

Reference sequence. I determined phenotypes associated with alleles in the reference sequence largely through data queries to OMIM and SNPedia, but could not rely on automated Trait-o-matic interpretation because the utility compares genome data to the reference. An attempt to interpret the reference sequence via Trait-o-matic correctly yielded no results.

Tabulation of UCSC data revealed that there exist 89,467 claimed nonsynonymous SNPs listed in dbSNP (build 129, retrieved February 2009). Of these, 69,770 (78.0%) were confirmed to be single base-pair alleles within 26,662 coding sequences listed in UCSC data ("refFlat" table) as of June 2008; 22 entries were malformed because the start and end coordinates provided for these alleles were not separated by 1 bp as claimed (one allele had start and end coordinates 100 bp apart, and others had zero or negative length); 19,675 (22.0%) were discarded because they were in a putative or recently characterized coding region not listed in "refFlat."

The majority of OMIM entries implicated by reference alleles are apparently benign. In general, the term "polymorphism" is used to describe an allelic variant in OMIM only when the variant is apparently benign and/or very common. While only 103 of 10,482 parsed OMIM allelic variants (less than 1%) include the term "polymorphism," 17 of 88 (19%) nsSNPs, corresponding to 16 of 77 (21%) unique OMIM allelic variants, emerge from this OMIM analysis of the reference sequence described explicitly as polymorphisms. An additional 28 nsSNPs correspond to haemoglobin polymorphisms listed by name, while two more are associated with skin/hair/eye pigmentation and one with phenylthiocarbamide (PTC) tasting ability (**Supplemental Table S1**).

Table 2. Homozygous genotypes for the reference allele extracted from SNPedia. *(a)* SNPedia entries near *ABCB1*. *(b)* SNPedia entries near *OCA2*.

a.

Genotype	dbSNP rs ID	Phenotype
chr7:86987858(C;C)	rs2235067	7x more likely to respond to certain antidepressants
chr7:86990039(A;A)	rs4148740	7x more likely to respond to certain antidepressants
chr7:86998497(A;A)	rs2032583	7x less likely to respond to certain antidepressants
chr7:86998554(A;A)	rs2032582	6.7x risk (Crohn's disease)
chr7:86998985(T;T)	rs4148739	7x less likely to respond to certain antidepressants
chr7:86999456(T;T)	rs11983225	7x more likely to respond to certain antidepressants
chr7:87002922(A;A)	rs10248420	7x less likely to respond to certain antidepressants
chr7:87003686(C;C)	rs2235040	7x more likely to respond to certain antidepressants
chr7:87007292(C;C)	rs12720067	7x more likely to respond to certain antidepressants
chr7:87017537(A;A)	rs1128503	likely to require more methadone during heroin withdrawal
chr7:87037500(C;C)	rs2235015	7x less likely to respond to certain antidepressants

b.

Genotype	dbSNP rs ID	Phenotype
chr15:25903913(C;C)	rs1800407	blue/grey eyes more likely
chr15:25933648(G;G)	rs1800401	blue/grey eyes possible
chr15:26017833(A;A)	rs7495174	blue/grey eyes more likely
chr15:26030454(C;C)	rs1129038	brown eye colour
chr15:26032853(G;G)	rs12593929	brown eye colour
chr15:26039213(A;A)	rs12913832	brown eye colour, 80% of the time
chr15:26039328(C;C)	rs7183877	blue eye colour if part of blue eye colour haplotype
chr15:26101581(T;T)	rs7170852	usually brown eye colour
chr15:26126810(A;A)	rs2238289	blue eye colour if part of blue eye colour haplotype
chr15:26167797(T;T)	rs2240203	blue eye colour if part of blue eye colour haplotype
chr15:26175874(A;A)	rs11631797	usually brown eye colour
chr15:26186959(T;T)	rs916977	usually brown eye colour
chr15:30782048(T;T)	rs4779584	1.70x risk for colorectal cancer
chr15:46213776(A;A)	rs1426654	probably light-skinned, European ancestry

Of 1575 genotype entries at 787 unique loci derived from semi-automated curation of SNPedia data, 315 concerned homozygous genotypes for the reference allele (**Supplemental Table S2**), representing 40.0% of loci in the data table. Of interest among these homozygous genotypes (**Table 2**), 12 located in or near the oculocutaneous albinism II (*OCA2*) gene form part of a haplotype associated with eye colour; six of these claim blue eye colour, while six—including those apparently most strongly associated with eye colour[55, 56]—claim brown. Another nearby genotype in the solute carrier family 24, member 5 (*SLC24A5*) gene claims light-skinned,

European ancestry. Additionally, nine genotypes are located in the ATP-binding cassette, subfamily B (MDR/TAP), member 1 (*ABCB1*) gene which have been associated with response to antidepressant treatment;[57] of these, four claim a sevenfold reduced likelihood of responding to certain treatments, while five claim a sevenfold increased likelihood. These data are to be interpreted with the caveat that the reference sequence itself is haploid; hence, a hypothetical individual bearing this sequence could be either affected or a carrier for any recessive phenotype. In addition, these data are of limited conclusiveness because they are drawn from a selective list of genotype–phenotype correlations based largely off of GWAS results.

Returning to more serious phenotypes from OMIM (**Table 3**), examination of OMIM descriptive text demonstrated two variants for which errors present in the source material flagged benign alleles. One missense mutation implicated in glaucoma was listed with incorrect summary information in OMIM, with benign (Glu) and deleterious (Lys) residues reversed, while summary information for a silent mutation contained notation indicating a Gly→Gly "missense," as a result of which every allele corresponding to glycine at that position would be incorrectly flagged as deleterious. Of the remaining reference alleles with potentially deleterious effects, several suggest that this hypothetical individual could be a carrier for autosomal recessive neurosensory deafness[58] and severe combined immunodeficiency disorder (SCID),[59] although population frequencies for homozygous genotypes associated with the latter suggest that either penetrance is very low or the genotype–phenotype association presented in the literature is spurious. In addition, five potentially deleterious alleles were parts of compound mutations for which other loci were not flagged in the reference sequence. Two other alleles were associated with thyroid carcinoma, though these associations were both reported by the same source,[60] lacked corroborating reports, and implicated the overwhelmingly major alleles at both loci; similarly, an asso-

ciation with hawkinsinuria[61] (an autosomal dominant disease) implicated an allele also found in unaffected controls and present with high frequency in the general population.

Table 3. Susceptibilities and diseases in OMIM associated with the reference allele. Disease susceptibilities are summarized as such in OMIM, with the exception of two that are listed as diseases but do not appear to be causative. Abbreviations: IDDM, insulin-dependent diabetes mellitus; AD, Alzheimer disease; ARMD, age-related macular degeneration; HBV, hepatitis B virus (persistence); SLE, systemic lupus erythematosus; NIDDM, non-insulin-dependent diabetes mellitus; RA, rheumatoid arthritis; SCID, severe combined immunodeficiency disorder.

Susceptibilities:	**Diseases:**
Asthma (4)	Arthropathy (compound mutation)
IDDM (3)	Bardet-Biedl syndrome (compound mutation)
AD (2)	Deafness (AR)
ARMD (2)	DPYD deficiency (compound mutation)
Atopy (2)	~~Glaucoma~~ (incorrect; entered into OMIM with benign and affected amino acids exchanged)
HBV (2)	Hawkinsinuria (high population frequency, found also in controls)
Obesity (2)	~~Progeria~~ (incorrect; entered into OMIM as amino acid change, but in fact silent mutation)
Schizophrenia (2)	Sandhoff disease (compound mutation)
SLE (2)	SCID (2; AR)
Autism	Thrombotic thrombocytopaenic purpura (compound mutation)
Coeliac disease	*Thyroid carcinoma [susceptibility?] (2, same source, lacking corroborating reports; overwhelmingly the major allele)*
Congestive heart failure	
Coronary spasm	
Diabetes	
Glomerulopathy	
Hepatic adenoma	
Hypertension	
Ischaemic heart disease	
Ischaemic stroke	
Myocardial infarction	
Nephrolithiasis	
NIDDM	
Parkinson	
RA	
Thyroiditis	

Among disease susceptibilities present in OMIM, several were flagged for the reference sequence, including asthma (4 variants); insulin-dependent diabetes mellitus (IDDM, 3 variants); age-related macular degeneration (ARMD), Alzheimer disease, atopy, hepatitis B virus (HBV) persistence, and obesity (2 variants each). While these data in no way translate meaningfully to a quantitative measure of risk for these complex traits, they may be useful at least as a rough baseline or guide to the kinds and quantities of complex disease susceptibilities a typical individual genome may reveal. Note also, for example, that the alleles associated with HBV persistence are major alleles in the population; in such cases, it may be more accurate instead to consider the minor alleles as conferring protection.

Discussion

Interpretation. Examination of four genomes has suggested that certain properties are common across human genomes, including ratios of novel to total, homozygous to heterozygous, and homozygous C/G to homozygous A/T nsSNPs. Comparison of these genomes has also shown a characteristic distribution when nsSNPs are scored based on BLOSUM100 values for corresponding amino acid changes. Hence, when aggregated partial exome data from the PGP were scored in the same manner, the distribution of these scores was clearly different from that of nsSNPs from any complete set of coding sequences.

Furthermore, the use of a variety of data sources, including several formatted specifically for this project, has replicated phenotypes previously found by the authors of these genomes and has helped in critically evaluating these claims through HapMap frequency data and cross-references to literature. In addition, interpretation of each genome has yielded results not discussed by the genome authors themselves, including several that may be of clinical interest. These interpretations have also made clear, however, that data available in any curated database, regardless of how recently updated, will be subject to the limitations of the literature source material. As in the case of CIPA-associated alleles in YH, where there is unclear evidence or ongoing debate about a genotype–phenotype association, neither inclusion nor exclusion of such information correctly expresses its status. In these situations, additional annotation, manual inspection, and a human interpreter's clinical or scientific judgment are required beyond what automated processes can accomplish.

Finally, Trait-o-matic includes a simple algorithm intended to form the beginnings of a utility to generate hypotheses for statistical follow-up in association studies. Drawing upon

BLOSUM100-based scoring, this algorithm attempts to identify variants that are likely to have the largest effect on protein function, along with a description of one or more potential phenotypes associated with altered protein function drawn from OMIM. The use of a noisy set of positive and negative controls to evaluate the effectiveness of such a simple algorithm roughly shows, as expected, that simplicity and computational speed come at the cost of accuracy. Somewhat remarkable, however, is the demonstration that accuracy can be dramatically improved by considering allele frequency alongside functional characteristics, which suggests an avenue for further exploration.

Future directions. It goes without saying that many additional areas of improvement for Trait-o-matic can be contemplated. For the purposes of clinical interpretation, the use of machine-readable data more accurate and evidence-based than what is available today is critical for increasing the informative content of these genomes. The currently available OMIM associations between genes and traits are only a selected portion of the associations available in the literature. While parsing OMIM has already increased fivefold the amount of freely available data for this purpose, a still richer database would be useful for our utility. One feasible method to begin the generation of this database would entail automated parsing of the literature; others have already devoted effort to this task, using natural language processing (NLP) and other methods to reconstruct semantic meaning from free text.[62-65]

In light of work presented here using OMIM and SNPedia, however, a literature-based corpus may not bring as radical an improvement as expected. SNPedia represents one approach to compiling a knowledgebase, focusing on GWAS results that largely emphasizes the "common disease, common variants" approach. As a result, a relatively small collection of several thousand variants yields hundreds of results for each genome, but their association with traits will

often be of low penetrance and/or of very small effect—the cumulative effect of which for any one genome may be potentially contradicting effects on phenotype with unknown interactions among them. By contrast, OMIM represents a larger but more selective knowledgebase of alleles of interest. While some of these are common variants, often from GWAS data, the rarest alleles have only been reported once or a handful of times; these variants would be found infrequently in genomes submitted for interpretation, and very little corroborating evidence supports any claim presented about them even when they are found in a submitted genome. As a result, this fairly large database of tens of thousands of variants often yields fewer than 100 results for each genome, not all of which are informative. As we delve into the literature to find less well disseminated variants, we can expect that many will have still weaker associations with their corresponding phenotypes, or still smaller population frequency. Hence, a severalfold increase in underlying database size may increase only slightly the informative content retrievable for any particular individual's genome.

Second, the use of more advanced logic to evaluate SNPs would improve the quality of phenotype inferences even without large increases in the amount of underlying data. Although Trait-o-matic calculations are fairly rudimentary at the present time due to considerations of scale, the presentation of a general list of phenotype inferences is not always the most salient answer to questions that can be asked about an individual genome. Instead, more specific queries, tailored to examine a limited array of mutations or mutations in a limited set of genes, could be useful in answering questions about particular phenotypes. One might then envision the use of Boolean expressions (as are implemented in search engines) to explore in detail such questions as an individual's genetic predisposition to Alzheimer disease. Recently, SNPedia has begun cataloguing Boolean expressions called "genosets" for just this purpose; for example, one ex-

pression claimed to be associated with a roughly sixfold increase in Alzheimer disease risk involves two loci and is represented with the notation *and(rs2071746(T;T),rs242557(A;A))*. One might also use Bayesian statistics instead of simply Boolean logic. For example, Trait-o-matic could automatically examine ancestry-informative markers (AIMs), some of which are already included among SNPedia entries, to arrive at a Bayesian estimate of biogeographical ancestry;[66] anticipating this possibility, I have suggested the inclusion of a small number of AIMs in upcoming exome subsets to be sequenced for PGP participants.

Improvement to the underlying literature is, of course, still another method of improving clinical utility, and it has not escaped our attention that data submitted to Trait-o-matic itself can be used for such purposes. The matrix-based method of scoring is only a first effort at constructing a tool for hypothesis generation with the aim of aiding statistical analysis of genotype–phenotype associations. One might envision the inclusion of analysis toolsets for GWAS[67, 68] to test subsets of variants that are most predicted to be deleterious across all genomes for which consent is given.

Additionally, the ability to deposit any results obtained via NLP or association testing in a semantic, openly accessible location such as SNPedia would both enhance community participation in these efforts and promote independent verification of claims, and would overcome the limitations of small sample size and ascertainment bias found in current datasets as personal genome technologies become more widespread. Ideally, then, Trait-o-matic would be situated in such a way that it can communicate bidirectionally with the data sources from which it draws inferences (**Figure 5**).

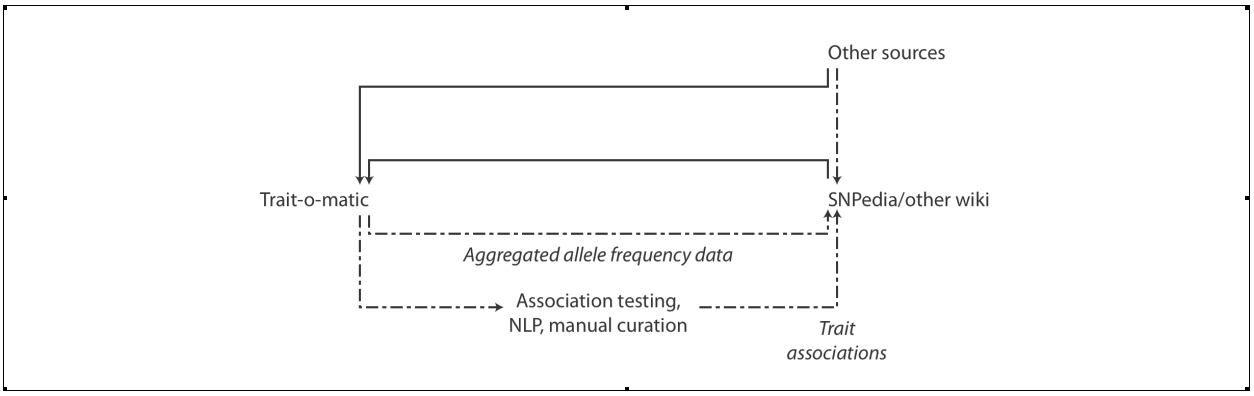

Figure 5. Proposed data flow for information exchange between Trait-o-matic, an interpreter, and a wiki-based source, an evidence-based repository for curated information. Dashed lines are not yet implemented.

Ethics and society. Finally, there is a need to confront, not simply from a scientific perspective, the question of what variants should or should not be presented to stakeholders in the genome sequencing process. A general expectation—and indeed fear—regarding personal genomes is that they will be useful in revealing accurate or deeply intimate information about individuals, voluntarily or not. Some concerns regarding this emerging technology, then, include whether better genome sequencing could lead to new forms of discrimination against those found to have disease or behavioural predispositions, or whether a more comprehensive catalogue of risks could lead to unwary consumers rushing to obtain a barrage of tests and preventive treatments for illnesses they will never face.[69, 70] For now, high costs and limitations on what we are capable of interpreting mean that the full social impact of genome sequencing has yet to be realized, but hints of potential can already be discerned in a project such as this one.

For instance, results presented here demonstrate that a utility such as Trait-o-matic is capable of identifying Mendelian conditions for which an individual is merely a carrier, as well as complex traits for which a heterozygous individual is expected to experience a much milder phenotype than a homozygote. While previously some prospective parents have had to face difficult choices because of their status as carriers for certain Mendelian and well-defined diseases, it

is possible that nearly every prospective parent in future generations could be confronted with a list of alleles they possess that have been correlated with minor increases in risk for serious complex diseases if homozygous or in *trans* with certain other alleles. Hence, negative effects of genome interpretation may include complicating already complex considerations surrounding childbearing decisions, as well as the potential of increasing social stigma for those who choose to forgo—or simply cannot afford—the genetic screening required.

In short, the question of how personal genomes are made available to consumers is a topic in need of consideration by policymakers and society at large. We should hope that, with the evolution of social and scientific approaches, our improved understanding of how to interpret our genetic inheritance can be applied positively to the advancement of human health and welfare.

Supplemental information for this thesis is available at ‹http://thesis.diploid.ca/›.

References

1. Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007).
2. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2008).
3. Bentley, D. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
4. Wang, J. et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65 (2008).
5. Schloss, J. How to get genomes at one ten-thousandth the cost. *Nat Biotechnol* **26**, 1113-1115 (2008).
6. Turnbaugh, P. et al. The human microbiome project. *Nature* **449**, 804-810 (2007).
7. Arnaout, R. Specificity and overlap in gene segment-defined antibody repertoires. *BMC Genomics* **6**, 148 (2005).
8. Guttmacher, A.E. & Collins, F.S. Genomic medicine—a primer. *N Engl J Med* **347**, 1512-1520 (2002).
9. Hirschhorn, J., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet Med* **4**, 45-61 (2002).
10. Janssens, A.C. et al. A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am J Hum Genet* **82**, 593-599 (2008).

11. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463-5467 (1977).

12. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

13. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**, 2601-2610 (1979).

14. Hert, D.G., Fredlake, C.P. & Barron, A.E. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29**, 4618-4626 (2008).

15. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**, 335-344 (2004).

16. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732 (2005).

17. Check Hayden, E. Genome sequencing: the third generation. *Nature* **457**, 768-769 (2009).

18. Branton, D. et al. The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26**, 1146-1153 (2008).

19. Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*, epub ahead of print (2009).

20. McGuire, A. 1000 Genomes: on the road to personalized medicine. *Personalized Med* **5**, 195-197 (2008).

21. Blow, N. Genomics: the personal side of genomics. *Nature* **449**, 627-630 (2007).

22. Ng, P.C. et al. Genetic variation in an individual human exome. *PLoS Genet* **4**, e1000160 (2008).

23. Olson, M.V. Human genetics: Dr Watson's base pairs. *Nature* **452**, 819-820 (2008).

24. Macer, D. Whose genome project? *Bioethics* (1991).

25. Osoegawa, K. et al. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res* **11**, 483-496 (2001).

26. Andrews, R.M. et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147 (1999).

27. den Dunnen, J.T. & Antonarakis, S.E. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* **15**, 7-12 (2000).

28. Sherry, S.T., Ward, M. & Sirotkin, K. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **9**, 677-679 (1999).

29. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).

30. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496 (2004).

31. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-796 (2003).

32. International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).

33. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**, 588-604 (2007).

34. Stenson, P.D. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577-581 (2003).

35. George, R.A. et al. Response to Stenson et al on the review of general mutation databases. *J Med Genet* **45**, 319-320 (2008).

36. George, R.A. et al. General mutation databases: analysis and review. *J Med Genet* **45**, 65-70 (2008).

37. Stenson, P.D. et al. Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet* **45**, 124-126 (2008).

38. Plumpton, M. & Barnes, M.R. in Bioinformatics for geneticists: a bioinformatics primer for the analysis of genetic data, edn. 2. (ed. M.R. Barnes), 249-280 (John Wiley & Sons, Hoboken, NJ; 2007).

39. Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. *Nat Genet* **27**, 167-171 (2001).

40. Jegga, A.G. et al. Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res* **12**, 1408-1417 (2002).

41. Robertson, G. et al. cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* **34**, D68-73 (2006).

42. Griffith, O.L. et al. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* **36**, D107-113 (2008).

43. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).

44. Sunyaev, S. et al. Prediction of deleterious human alleles. *Hum Mol Genet* **10**, 591-597 (2001).

45. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7**, 61-80 (2006).

46. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

47. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858 (2008).

48. Tai, E.S. et al. Association between the PPARA L162V polymorphism and plasma lipid levels: the Framingham Offspring Study. *Arterioscler Thromb Vasc Biol* **22**, 805-810 (2002).

49. Vohl, M.C. et al. Molecular scanning of the human PPARα gene: association of the L162V mutation with hyperapobetalipoproteinemia. *J Lipid Res* **41**, 945-952 (2000).

50. Norrgard, K.J. et al. Double mutation (A171T and D444H) is a common cause of profound biotinidase deficiency in children ascertained by newborn screening the the United States. *Hum Mutat* **11**, 410 (1998).

51. Whatley, S.D. et al. Variegate porphyria in Western Europe: identification of PPOX gene mutations in 104 families, extent of allelic heterogeneity, and absence of correlation between phenotype and type of mutation. *Am J Hum Genet* **65**, 984-994 (1999).

52. Mardy, S. et al. Congenital insensitivity to pain with anhidrosis: novel mutations in the TRKA (NTRK1) gene encoding a high-affinity receptor for nerve growth factor. *Am J Hum Genet* **64**, 1570-1579 (1999).

53. Miranda, C. et al. Novel pathogenic mechanisms of congenital insensitivity to pain with anhidrosis genetic disorder unveiled by functional analysis of neurotrophic tyrosine

receptor kinase type 1/nerve growth factor receptor mutations. *J Biol Chem* **277**, 6455-6462 (2002).

54. Ackerman, M.J. et al. Ethnic differences in cardiac potassium channel variants: implications for genetic susceptibility to sudden cardiac death and genetic testing for congenital long QT syndrome. *Mayo Clin Proc* **78**, 1479-1487 (2003).

55. Eiberg, H. et al. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet* **123**, 177-187 (2008).

56. Sturm, R.A. et al. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am J Hum Genet* **82**, 424-431 (2008).

57. Uhr, M. et al. Polymorphisms in the drug transporter gene ABCB1 predict antidepressant treatment response in depression. *Neuron* **57**, 203-209 (2008).

58. Ouyang, X.M. et al. Mutations in the alternatively spliced exons of USH1C cause non-syndromic recessive deafness. *Hum Genet* **111**, 26-30 (2002).

59. Puel, A., Ziegler, S.F., Buckley, R.H. & Leonard, W.J. Defective IL7R expression in $T^{-}B^{+}NK^{+}$ severe combined immunodeficiency. *Nat Genet* **20**, 394-397 (1998).

60. Gimm, O. et al. Mutation analysis reveals novel sequence variants in NTRK1 in sporadic human medullary thyroid carcinoma. *J Clin Endocrinol Metab* **84**, 2784-2787 (1999).

61. Tomoeda, K. et al. Mutations in the 4-hydroxyphenylpyruvic acid dioxygenase gene are responsible for tyrosinemia type III and hawkinsinuria. *Mol Genet Metab* **71**, 506-510 (2000).

62. Chen, L. & Friedman, C. Extracting phenotypic information from the literature via natural language processing. *Stud Health Technol Inform* **107**, 758-762 (2004).

63. Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M.A. G2D: a tool for mining genes associated with disease. *BMC Genet* **6**, 45 (2005).

64. Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y. & Friedman, C. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pacific Symposium on Biocomputing*, 64-75 (2006).

65. Lee, L.C., Horn, F. & Cohen, F.E. Automatic extraction of protein point mutations using a graph bigram association. *PLoS Comput Biol* **3**, e16 (2007).

66. Halder, I., Shriver, M., Thomas, M., Fernandez, J.R. & Frudakis, T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Hum Mutat* **29**, 648-658 (2008).

67. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

68. Browning, B.L. & Browning, S.R. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol* **31**, 365-375 (2007).

69. Robertson, J.A. The $1000 genome: ethical and legal issues in whole genome sequencing of individuals. *Am J Bioethics* **3**, W-IF1 (2004).

70. My genome. So what? *Nature* **456**, 1 (2008).