Integrated Measurements of Causal Variations in Human Biology

The Personal Genome Project (PGP) has developed an "open consent" protocol so that volunteers may responsibly share their genomic, epigenomic, and associated phenotype data without restriction. The study is IRB-approved at Harvard Medical School for genome sequencing, as well as data and sample collection from 100,000 individuals. Nearly all participants (95%) have indicated availability for on-going, minimally invasive collections (saliva, hair, blood, skin). More than 85% are interested in donating surgical samples and more than 90% in donating samples post mortem. The PGP has begun collecting whole genomes, but an individual's sequence is only one part of the etiology of health phenotypes; synthesis of genomic data with other external measurements is a fundamental barrier for personalized medicine research. Yet no framework exists for systematically prioritizing which of many possible measurements one might undertake with these samples.

Our Specific Aim is to develop technologies and protocols for integrated measurements of causal variations in human biology. The resulting open, standardized data-sets will enable collaborative integration of biological data, including time-series "omics profiling" (see below), wearable physiological measures, and imaging. Broad sharing of genome, environment, and trait (GET) data will enable diverse research communities to cross-check novel combinations of these data for consistency, quality, and informativeness. This resource will serve as a springboard for general research in human biology, and eventually, personalized therapies.

The availability of surgical samples from the PGP cohort presents unique opportunities for technology development. We will select manageable, representative sets of samples spanning important organ systems for integrated "omics profiling." This profiling will involve integrated haplotype-phased measurements of genomic, epigenomic, transcriptomic, proteomic, and metabolomic information. We will use technologies already under development by our groups, or novel combinations of such technologies, to achieve a target cost of $300 per sample. Simultaneously, we will also advance the quality of omics profiling, specifically reducing genome sequencing error rates to 1×10^{-10} to accurately detect somatic variation or de novo mutations. We are already tracking more than 2,000 samples in a public "virtual" biobank. As resources permit, we will expand the range of protocols for donating samples and measuring them once they are part of the collection.

Integration of omics profiling and other measurements of clinical or scientific interest–such as microbiome profiling, immune profiling, text-mining of health records and/or automated processing of other biomedical modalities–will require technological advances that will fundamentally alter the applicability of these technologies to human biology. The resulting open datasets will also serve as a kernel for nascent data standards aimed at integrating the plethora of measurements envisioned. It is difficult to say with certainty which measurements will be most useful for any particular future question. Thus, we are focused on making measurements that are simultaneously more integrated, less expensive and of higher quality.

The more comprehensive the individual datasets become, the more researchers will want to augment and use them and, therefore, the more identifiable they will become. Consequently, privacy preservation and anonymization of such integrated datasets will always be in competition with new ways to re-identify them. Through the use of open consent, in which privacy is neither assumed nor implied, the PGP provides a uniquely safe environment for ongoing exploration of what is possible.

By weaving together data from all participants, we will build open, scalable tools for the integration and interpretation of the resulting datasets. These tools will help illuminate causal variations and their relationships across different data types. Any organization will be able to securely use our computational and storage resources to demonstrate their own tools on our public datasets. To benefit from existing and ongoing omics profiling studies, we will develop a system for recording and organizing variation data from the literature. This system will provide a platform to drive consensus interpretation of such variation.

Our unprecedented, public research effort will pilot the generation of extremely comprehensive data: omics and other profiling time-series, surgical samples, and post mortem "body maps" (final pilots of 4-6 individuals each). The ongoing and final pilot data as well as analysis methods (see Data and Resource Sharing Plan) will enable the usage of these technologies and facilitate sharing of such data among research groups worldwide.

Research Strategy

1 Challenge, Innovation and Impact Statement

As genomic data generation advances, its complexity creates tremendous challenges in analysis, validation, and standardized data sharing. New tools are required for integrated molecular characterization of diverse tissue samples as well as standardized combination of these data with other measurements of scientific or clinical interest. Currently these data are expensive, error-prone, isolated and difficult to integrate. Using novel combinations of new and existing technologies we will demonstrate integrated "omics profiling"–including genomic, epigenomic, transcriptomic, proteomic, and, metabolomic information–at a cost of $300 per sample. Simultaneously we will advance sequencing techniques to generate haplotype-specific information to detect allele-specific associations, and reduce error rates to 1×10^{-10} for detection of *de novo* and somatic mutations. Finally we will use advancing capabilities from our CEGS to "edit" human cell-lines and demonstrate the causal consequences of prioritized variations. The resulting cost, quality and integration advances can amplify the effectiveness of numerous NIH efforts. The methods developed as well as ongoing and final pilot data – omics and other profiling time-series, surgical samples, and post mortem "body maps" (final pilots of 4-6 individuals each) – will enable researchers to collaboratively explore the vast landscape of human genome, environment and trait (GET) data.

2 Rationale

If we could read every nucleic acid sequence in the human body, the uncompressed information would exceed all stored digital information on Earth by more than ten orders of magnitude (10^{32} bits versus 10^{21} bits). Considering the activity of retrotransposons, somatic genetic variation, epigenetic modifications, RNA editing, post-translational modifications, the variety of metabolites present in each human cell, immune system antibody diversity and human associated microbes, measuring the information content of an individual human represents a staggering challenge. Yet, only a small fraction of this information varies from cell to cell or person to person and an even smaller fraction has measurable causal consequences. Identifying the most important causal variations is a difficult problem spanning numerous areas of research.

To tame the vast complexity of human biology we must dedicate resources to integration. We will tackle these challenges by combining and improving existing measurements–spanning single cells, tissues or individual humans and over time–in a standardized, integrated fashion. The Personal Genome Project (PGP) aims to be an ideal testbed for this integration (Figure 1.)

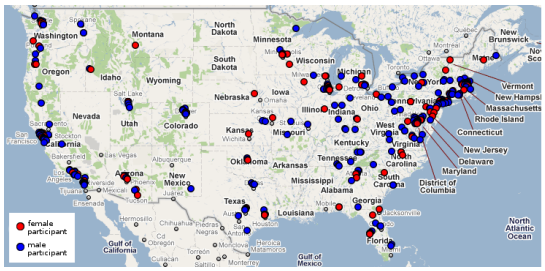


Figure 1: Participants in the PGP represent a diverse, active population.

Working closely with numerous collaborators, we are rapidly advancing toward a $100 genome. Advances in Long Fragment Read (LFR) technology have achieved error rates of 1×10^{-7} in nearly complete, phased whole genome sequences (Peters, submitted). In addition to DNA sequence, there is growing recognition that integration of other types of genomic datasets (eg, transcriptomes, epigenomes) and associated phenome datasets is necessary to advance genomic research, but these data only compound existing research challenges. Raw data alone, from a combined omics profile based on LFR technology, consumes more than 1TB for a single sample. Interpretation of this data is extremely difficult due to the unprecedented number of implied hypotheses.

A focus on improved technologies and protocols for integrating diverse omic data as well as environmental and trait measurements, unbiased by the study of any particular disease, is an important new paradigm for human biology research. Since it will be impossible to ascertain which measurements are most informative (and cost-effective) at the outset, our central motivation is to lower the cost as well as increase the quality and informativeness of integrated data-sets. We will pilot the generation of extremely comprehensive data: omics and other profiling time-series, surgical samples, and *post mortem* "body maps" (final pilots of 4-6 individuals each). These technologies can then be used in future disease specific studies.

3 Approach

3.1 Participant enrollment and scale of project

All volunteers will be enrolled under an open consent process pioneered by the Harvard PGP, which our Institutional Review Board (IRB) has approved for the enrollment of up to 100,000 individuals (Lunshof et al., 2008). Our automatic enrollment system currently has over 1,000 health records uploaded by consented PGP volunteers from a diverse population of individuals, illustrating broad interest in participation (Figure 1, previous page). Participants can be recontacted so that data in these records can be systematically confirmed either through additional testing or additional self-reported survey instruments. Collection of outcomes data at regular three monthly intervals is also facilitated by our participants' consent for recontact and automated infrastructure.

Private funding has already been provided for 250 combined whole genomes and medical records; as sequencing costs fall and the size of our existing database grows we anticipate continued private funding support for additional genome and medical data-sets. In addition, each participant will have cell lines established and made available through the Coriell Institute. Such cell lines can be used for functional studies that provide additional evidence for the causality of specific measurements.

We will select a few of the 250 individuals with genome and medical data for more in-depth measurement with technologies contemplated in this proposal. Our aim is to utilize this existing resource to integrate and enhance numerous new whole genome sequencing, other omics profiling technologies and other measurements of scientific and clinical interest on ~50 of these individuals (see budget justifications for Harvard and Stanford). Furthermore, informatics tools developed to prioritize participants for additional detailed measurements will be systematically applied to the entire cohort which will grow to between 10,000 and 100,000 participants during the funding period.

Our initial pilot sequencing project has provided valuable information regarding the process of nation-wide enrollment of participants, sample-collection and data interpretation. We can now draw on this resource to integrate numerous omics profiling technologies – initially via time-series analysis of individual participants, subsequently on analysis of (healthy and diseased) surgical samples from representative body sites, and finally by developing a *post mortem*, molecular "body-map" of deceased participants (final pilots of 4-6 individuals each).

In addition, our experience is that clinical analysis of whole genome sequencing currently results in a high rate of incidental findings (Kohane et al., 2006). Openly consented individuals are critical to improving tools that can ameliorate such findings because they allow us to recontact individuals and perform additional measurements and follow-up.

3.2 Advances in omics profiling of arbitrary tissues

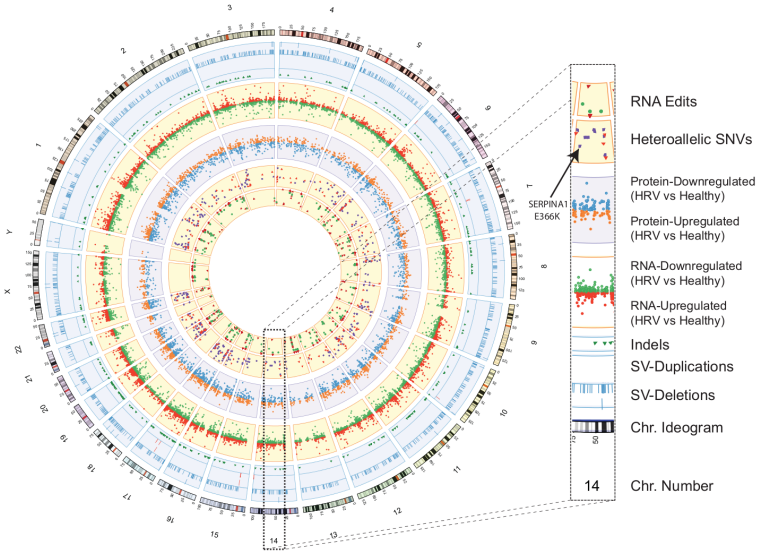


Figure 2: Integratd personal omics profile

We have made considerable progress developing integrated omics profiling applicable to many tissues. In Figure 2, previous page, the Circos plot summarizes an integrated personal omics profile. From outer to inner rings: chromosome ideogram; genomic data (pale blue ring) - structural variants > 50 bp (deletions (blue tiles), duplications (red tiles)), indels (green triangles); transcriptomic data (yellow ring) – FPKM expression ratio of HRV infection to healthy states; proteomic data (light purple ring) - ratio of protein levels during HRV infection to healthy states; transcriptomic data (yellow ring) – differential heteroallelic expression ratio of alternative allele to reference allele for missense and synonymous variants (purple dots) and candidate RNA missense and synonymous edits (red triangles, purple dots, orange triangles and green triangles, respectively).

We will continue to simultaneously lower the cost and improve quality of these profiles. Omics profiling will be applied to both minimally invasive participant samples (i.e. blood, saliva, skin, and hair) as well as surgical samples and, in years 3-5, samples collected *post mortem*. Where possible samples from the same tissue of origin will be analyzed in a time-series so variation can be measured both between tissues and over time.

3.2.1 Sequencing technologies

Current low-cost sequencing technologies, which have short read lengths, fail to distinguish which chromosome heterozygous variations are associated with. Such phased data is critical for analyses which seek to detect causal variations that create allele-specific differences in gene transcription and splicing. While phased data can be inferred using parental genomic data, it could also be determined directly with appropriate technological improvements. To that end, we are continuing our collaboration with Complete Genomics and their development of “Long Fragment Read” (LFR) technology. For only $100 in additional sample-preparation and informatics, this technique phases whole genome data and also improves sequencing accuracy by more than an order of magnitude (1×10^{-7}) due to bioinformatic improvements taking advantage of subsetted data. Cost reduction of this technology will include reduced use of reagents through smaller features & volume sizes ($750nm \times 750nm \times 50um \rightarrow 200nm \times 200nm \times 10\mu m$), increased instrument throughput (more cameras: 4 vs 2, higher pixel density: $5\times$ increase, higher frames per second: 100 vs 30), computational improvements (hardware $10\times$ in 5 years, software $4\times$ in 5 years).

Briefly, LFR whole genome data (Figure 3) is produced through dilution and tagging of genomic DNA to segregate subsets of the genome, facilitating later bioinformatic analysis and capturing phase information. Genomic DNA (10-20 cells) is physically separated into 384 wells and separately amplified, fragmented, and ligated to unique barcode adapters (A-C). All 384 wells are then combined and sequenced using Complete Genomics’ DNA nanoarray sequencing platform (D). Mate-paired reads are mapped to the genome using a custom alignment program that uses the barcode sequences to group tags and create haplotype contigs (E). This process successfully phases 90-97% of heterozygous single nucleatide polymorphisms (SNPs) into assembled into contigs with N50 lengths of ~500 kb for samples of European ethnicity and ~1 Mb for an African sample; this difference is due to more regions of low heterozygosity in European samples introduced by a population bottleneck ~50,000 years ago.

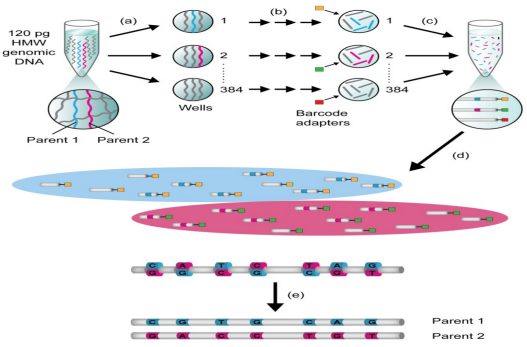


Figure 3: LFR Overview

Due to short read lengths, high throughput sequencing is currenlty limited in its ability to analyze some types of variation critical for genetic analysis. These include trinucleotide repeats (e.g. Huntington’s disease and Fragile X) and chromosomal rearrangements (extensive insertions and deletions, inversions, and translocations). While these can sometimes be profiled through the measurement of copy number variations and/or breakpoint detection, when detected these data often lack basepair resolution.

There are several different cutting edge sequencing technologies which promise to provide greatly increased read lengths – these technological improvements would address almost all current issues with unreported variation in current technology. We plan to continue working with these sequencing companies, including: Halcyon Molecular (which uses electron microscope imaging of DNA molecules), NABSys (nanopore sequencing), Pacific Biosystems (real time single molecule sequencing) and Oxford Nanopore (nanopore sequencing). Combining

a long read technology with LFR sequencing will allow us to increase the N50s of LFR assemblies as well as resolve many remaining limitations. Simultaneously, each of these technologies have their own cost reduction and quality improvement trajectories. We will integrate the best of these throughout our omics profiling effort.

3.2.2 Epigenomics

We will be focused on providing whole genome allele-specific quantitative measurement of cytosine methylation. Initially we will apply our existing technologies to quantitively assess percent methylation at subsets of genomic sites using bisulfite padlock probes (BSPP) and/or methyl-sensitive cut-site counting (MSCC) (Ball et al. 2009). Both technologies use high-throughput sequencing and thereby take advantage of anticipated dramatic cost reductions. BSPP uses padlock probes to isolate 10,000 short regions from bisulfite-treated DNA (BS-DNA). MSCC quantitatively assays a subset of CpGs in specific sequence contexts, assessing the fraction that have been digested by a methylation-sensitive restriction enzyme (when used with HpaII this assesses “CCGG” sequence context, 5% of all genomic CpGs).

We will improve methylation profiling by applying Complete Genomics’ LFR technology to BS-DNA. Currently the bioinformatic challenge of placing short reads from BS-DNA, which has reduced sequence uniqueness, has been the major obstacle to whole genome bisulfite sequencing. By subsetting the genome regions LFR greatly reduces read placement complexity. In addition, LFR sequencing will capture haplotype-specific methylation information. Separating and independently assessing the methylation patterns on each chromosome provides valuable information for linking specific genotypes to associated methylation patterns.

Hydroxymethyl modifications are an additional cytosine modification of interest. We will develop haplotype-specific hydroxymethyl profiling technology, taking advantage of molecular labelling combined with single molecule sequencing technology as developed by Pacific Biosciences (Song et al. 2011).

3.2.3 Transcriptomics

Allele-specific expression represents an important source of causal variation leading to numerous physiological effects since (potentially) one sees the effect of sequence variations on expression with a perfectly matched control. We have been developing technologies for complete, phased genomes that when combined with RNA-sequencing data can measure allele specific expression with extremely high accuracy. Reconstruction of a diploid personal genome sequence and using it instead of the reference genome is a critical step preceding allele-specific transcript analysis (Rozowsky et al 2011). First, using a generic reference genome introduces biases in read mapping—reads originating from the non-reference allele are more susceptible to mismapping since, when aligned to the reference allele, they contain at least one mismatch (in case of SNPs) or gap (in case of indels)—the reference bias effect, that is, both alleles are not treated equally by default. Second, expression or binding in regions of genome Structural Variation (SV) could be misinterpreted as allele specific expression (ASE). For example, duplication of an allele in the studied genome will double binding signal for the allele while signal for the allele on another haplotype will be unchanged. Furthermore, SNP calling in the regions of SV is likely to be less precise and contain more false positives compared with non-SV regions (“compression regions” in Dewey et al., 2011). To investigate causal variants influencing splicing, known to be important in genetic disease, we will align reads to personal phased splice-junction libraries to determine splice-junction ASE SNPs.

Read counts for phased alleles are generated at each heterozygous SNP nucleotide positions, and ASE events are reported by applying a binomial test followed by correction for multiple hypothesis testing. We correct for multiple hypothesis testing by estimating the false-discovery rate (FDR) by explicit simulation of the number of false positives given an even null background (i.e., no allele-specific events).

Cell populations are often heterogeneous in their expression, and this information is lost when profiling transcripts from pooled cell. Extending our current work developing single-cell in-depth transcriptome assays (Center for Causal Consequences of Variation), we will apply this technology to generate single cell expression data as part of integrated omics profiling.

In parallel, continuing our existing collaborations with sequencing companies, we anticipate working to extend sequencing improvements that are currently limited to whole genome data (e.g. Complete Genomics sequencing technology) to also realize dramatic cost improvements in transcript profiling.

3.2.4 Proteomics

Proteins will be analyzed using liquid chromatography and mass spectrometry after peptide fragmentation. Briefly, proteins lysates are prepared, digested with trypsin and peptides separated using a reverse phase HPLC directly coupled to a Velos OrbiTrap mass spectrometer. Peaks are further fragmented and subjected to another round of mass spectrometry (MS/MS). When samples are analyzed in triplicate ~7000 proteins can be identified at high confidence (two peptide match 1% FDR).

To quantify protein levels, two strategies can be employed. First, to obtain relative levels, a large reference sample can be prepared from the sample tissue and the labeling of both the reference sample and sample of interest using isobaric tag (TMT from ThermoFinnegan) can be used to identify experimental/realtive levels for each peak. Second, absolute protein levels can be quantified using isotoically labeled “detectable” peptide for each protein which are run in parallel. Several groups (notably Aebersold/Agilent) are working to generate such peptides for all human proteins. It is likely that such peptide libraries will become universally available for all human proteins during the course of this study.

3.2.5 Metabolomics

Metabolites will be analyzed using an Agilent Q-TOF 6538. Metabolites can be extracted from serum or relevant tissues using a mixture of methanol, acetonitrile and acetone. Proteins are precipitated at -20℃ and the extracted metabolites separated using Agilent 1260 liquid chromatography that is directly coupled in-line with an Agilent 6538 accurate mass Q-TOF MS with electrospray ionization (ESI) operated at positive and negative mode. To assure the mass accuracy of the recorded ions, continuous internal calibration ions will be infused. Each sample is run in triplicate at MS mode first at both positive and negative modes. Metabolites of interest are further subjected to MS/MS. Over 5,000 diverse metaoblites can be sepearated from serum using this procedure. Additional solvent fractionation conditions can be explored to further maximize the number of metabolite peaks that can be detected. Moreover, we will attempt fractionations using sizing columns and solid phase extraction conditions to selective extract compounds from complex mixtures.

A significant challenge in metabolomics is the identification of the metabolite peaks. Presently approximately 1000 peaks (20%) can be tentatively assigned a chemical composition or compound name based on retention time and molecular mass using MassHunter Workstation software (Agilent Technologies); the majority can not be assigned. Analysis of purified reference metabolites can allow the identification of additional metabolites and we will attempt to prepare mixtures of purified metabolites and separate them under similar conditions to further identify the metabolites that can be anlayzed. Our goal will be to increase the number of metabolites that can be monitored to >2000 peaks representing >1000 distinct compounds. The use of isotopically labeled reference compounds will allow quantification of metabolite levels.

3.3 Other measurements of scientific or clinical interest

To build integrated data associated with an individual that explores the relationship of genomes and environment to traits, we will also apply and improve high-throughput methods of profiling microbiome and immune system information.

3.3.1 Microbiome profiling

Our microbiome profiling builds upon our experience with metagenomic studies (Dantas et al., 2008, Sommer et al., 2009). The gut microbiome is implicated in playing a role in obesity, crohn’s disease, and irritable bowel syndrome. An individual’s environment in the form of diet and antibiotic drugs also interacts with their microbiome. To study how genetic and environmental factors interact to affect microbial diversity in the human digestive system, we plan to recruit subjects who volunteer to collect the required samples, as well as record dietary and drug information.

Our ongoing microbiome research extends our earlier work on molecular inversion probe (MIP) capture technology (Porreca et al., 2007, Li et al., 2009, Ball et al., 2009) to develop low cost methods of capturing specific subsets of microbial genes from complex microbiomes. To broadly capture phylogenetic biomarker information

we will design MIPs specific for the 16S ribosomal DNA (rDNA). Microbiome characterization will include the capture of clinically relevant biomarkers such as pathogenicity islands, mobilization elements and antibiotic resistance genes. Barcode DNA sequences on the probes will allow sample parallelization, allowing us to realize very low cost for microbiome profiling (less than $100 per sample). Given acceptable cost, we will then focus on extending *in-situ* sequencing techniques (see Figure 4a-d below), currently under development, to catalog individiual microbial cells.

3.3.2 Immune system profiling

An person's history of viral and bacterial exposures is recorded in the antibodies present in their immune system. Our groups are interested in developing biotechnology for surveying an individual's antibody repertoire. Understanding a cell's antibody structure requires isolating and sequencing both heavy and light chain components, which consist of separate transcripts on different chromosomes. To that end, we are exploring different methods for isolating polyacrylamideencapsulated lymphocytes combined with multiplex PCR to produce libraries linking heavy and light chain sequences from individual cells.

Single cell isolation may be achieved through one of four methods, described in Figure : (a) Adapting emulsion PCR protocols currently in use for polony bead PCR in the Church Lab to include cells so that a large number of compartments are formed that contain single cells, (b) performing PCR in fixed, solublized cells, (c) performing a variant of gel polony PCR developed in the Church Lab in which PCR products amplified from cells sparsely dispersed in the gel remain localized in the gel and do not mix, and (d) in collaboration with the Weitz Lab at Harvard University, we are developing a microfluidic device for high-fidelity emulsion formation in which single cells and polony microbeads can be placed together in emulsion compartment. Second, the PCR must be performed in a way that physically co-localizes the products of the heavy chain VDJ and light chain VJ segments of the individual cells. Figure (bottom) illustrates two possible methods: (e) Use distinct primer extensions for the two PCR reactions along with polony beads bearing sequences complementary to these extensions, so that the same beads may be sequenced in turn for the VDJ and VJ segments starting from different sequencing primers. (f) Use overlap extension PCR to generate a single product in which VJ and VDJ sequences are concatenated.

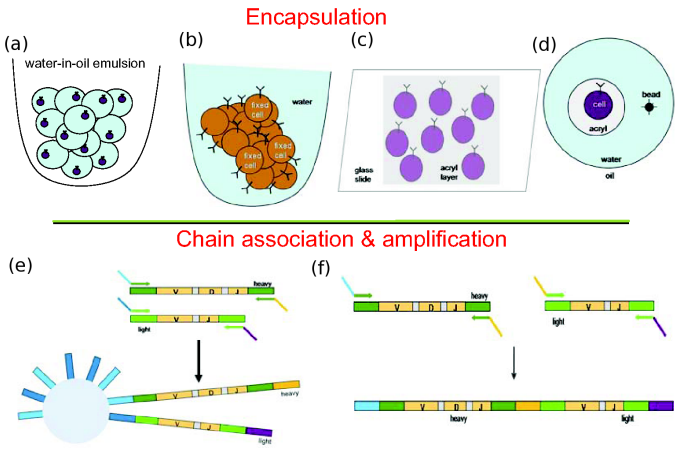


Figure 4: Top: Possible methods for massively parallel intra-cellular PCR. Bottom: Proposed methods for associating heavy chain VDJ and light chain VJ DNA sequences originating in the same immune cells.

Before this technology matures, we will determine the autoantibody profile using an established method. Serum from each individual will be used to profile an array of ~9,000 unique proteins from Invitrogen. Reactive antigens will be scored using ProCat; we expect 30-60 antigens that exhibit greater than three standards deviations over background for each antigen that is spotted in duplicate on the array. We will correlate the antigen activity with that of the VDJ sequences and other omics as described below.

3.4 Data integration and sharing

3.4.1 Open software platform for Genome-Environment-Trait (GET) Evidence Integration

Peer production recording of genetic variant interpretation Whole genome analysis is currently hampered by a lack of genome-scale methods for prioritizing genetic variants with potential associated phenotypes. These issues have become clear in our own work with genome review (Ashley et al., 2010, Kim et al., 2009, Drmanac et al., 2010, Dewey et al, 2011). Our "Trait-o-matic" tool (Drmanac et al., 2010, Kim et al., 2009) combines data from several genome wide databases to give lists of all variants found within them, including: Online Mendelian

Inheritance in Man (OMIM), the Pharmacogenomics Knowledge Base (PharmGKB), and the Human Genome Epidemiology Network (HuGENet). Integrating variant databases (general and disease-specific) is an important first step in establishing a central tool for understanding genetic variation. To this end, we will be working together with the MutaDATABASE project and PharmGKB to associate genetic variants with diverse reported associations.

We find that comparing whole genomes against these variant databases produces a large number of variants with published data (Figure 5), and existing databases are severely limited in providing additional data needed to automatically sort and prioritize variants. Interpretation of a whole genome thus involves a large amount of labor, and this work is often redundant as many variants will be seen again in other genomes. The number of new variants in each additional genome declines dramatically – we find that the number of new variants in an additional genome drops 10-fold after ~20 genomes have been reviewed (Figure 5). We believe this redundancy is best addressed with a shared system where genome reviewers may record analyses of relevant literature and form a consensus interpretation of genetic variants.

Our GET-Evidence system, attempts to address these issues and we plan extensive improvements in the course of this research. In particular, we will record the following variant and gene specific data to allow automatic genome interpretations: (a) association of the variant with coded phenotype categories, (b) strength of evidence supporting the proposed effect, (c) severity or strength of the proposed effect. Our first advances will be to facilitate manual entry of these data in a user-editable database, allowing re-use, correction and rapid updating under a "peer production" model similar to that used by Linux and Wikipedia. In keeping with the public nature of PGP participant data, the associated software for analysis and user contributions will be released under free software and creative commons licenses (see Resource Sharing Plan).

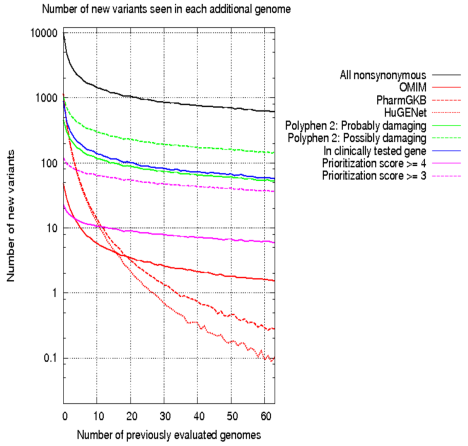


Figure 5: The number of new variants in each additional genome drops as more genomes are reviewed.

Integration of coded phenotypes from health records As we advance GET-Evidence to utilize coded phenotypes in peer production variant interpretations, we will integrate personal health records from PGP participants to create coded phenotype information associated with each individual. The open source IndivoX platform (Adida et al., 2010) will be used as the basis for a system that allows PGP participants to import and extend their health records. Initially phenotypes will be defined as corresponding to ICD codes. As we may find phenotypes not defined within this coding system an internal coding system may be built which maintains connections to ICD when available.

Many of the records within the existing health system are in documents or unformatted records, lacking standardized codings. For scanned documents, we plan to automatically convert these using existing software (eg. *Ocropus*) and use semantic detection of phenotypes to semi-automate the extraction of coded data from these documents by PGP participants and other volunteers (see below).

Automated extraction of coded phenotypes from literature and health records Because computational linguistics is not sufficiently advanced to provide for completely automated semantic analysis of biomedical texts, GET-Evidence currently uses a semi-automated approach to import structured knowledge from literature using the Bionotate platform (Cano, et al. 2009). As papers are added to variant pages using a PubMed identifier, a link is provided to BioNotate. Named entities such as gene and variant mentions are automatically identified and high-lighted. The curator's task is to qualify the complete text segment as supporting (or not) of a variant-phenotype relation, as well as identify, highlight specific verbal support and attribute it's type (i.e. as familial, quantitative, etc). The annotation is stored in the GET-Evidence database as well as presented as a marked-up text to facilitate subsequent validation and comprehension.

This semi-automated extraction of knowledge from literature will be extended to add coded phenotype information for automatic highlighting and manual review. A hypothesis regarding the publication's hypothesized relationship between the genetic variant and the phenotype will also be recorded (e.g. "causal", "protective", or "unknown"). Literature review will consequently provide the following linked information for a given publication:

(a) Pubmed identifier, (b) Gene name, (c) variant(s) identifier by amino acid change and/or dbSNP identifier, (d) coded phenotype information, (e) the published hypothesis of how the genetic variant is connected to the phenotype.

These methods for automatic highlighting and extraction of knowledge will also be adapted for the semi-automated extraction of phenotypes from uploaded health record text by PGP participants.

Integrated omics profiling and phenotype data to allow structured queries In total, many diverse data may be linked with PGP participants spanning many aspects of how genetic inheritance and environmental history interact to produce traits. We will integrate the quantitative data produced with omics profiling so that the standardized result permits structured queries. Such queries may be used to extract selected data from the database for testing hypothesis and discovering relationships. Examples of queries we wish support include (a) report metabolite information for all PGP participants taking a given drug at time of sampling, along with any coding variants in genes reported to interact with this drug, (b) report all PGP participants who carry an unevaluated coding variant in a metabolism gene and exceed some threshold value for the metabolite, (c) report genetic variants in a given genomic region and the associated transcript levels, separated by a splicing variant.

The coded phenotype information combined with genome sequences (more than 250 WGS are already privately funded outside this project) will be linked to create a large database allowing queries of linked genetic and phenotype information. Integration of omics profiling pilot data with other NIH resources, will similarly allow variation within profiles to be querried.

For example, we will create expression profiles specific to each of the human tissues for which public RNA-seq data are available. Examples of public data sources include ENCODE, which contains RNA-seq derived genome-wide expression levels for common cell types including both cell lines and primary cell types; the Illumina BodyMap 2.0 project contains RNA-seq data derived from 16 tissues across a variety of human subjects at different ages; and the Cancer Genome Atlas (TGCA) will produce expression profiles from over 3,000 samples derived from 26 different cancer tissues. Such expression profiles are required by a variety of existing single sample prediction methods and can be based on, for example, correlating variance and mean-standardized whole-genome mRNA abundance values (Perou et al., 2000), binary membership of clusters of similarly expressing genes, or use of a unique 'barcode' of genes expressing at specific levels specific to each tissue (Zilliox and Irizarry, 2007). Expression data from the 'random' tissue of interest would be summarized by the same method and classification performed using the closest expression profile match to the combined public reference.

Automated prioritization of candidate causal variations In the absence of existing reviews within the system, the multitude of potentially interesting variants (SNVs, indels, SVs, etc.) in a whole genome needs to be prioritized to enable interpretation. There are two different goals in doing this: (1) prioritize variants reported on in published literature and/or present in other databases and (2) prioritize variants which computational methods predict as likely to have a pathogenic effect. The first goal should become less important as the system becomes populated with existing reviews, while the second goal will become more important for interpretation of unpublished variants.

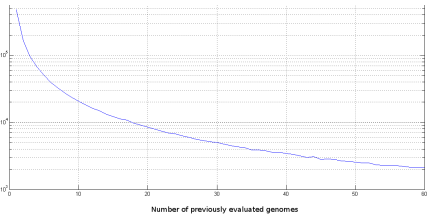


Figure 6: The number of discovered SNPs in transcription factor binding sites (or peaks) per each additional sequenced genome. Thousands of new SNPs are found after 20 genomes suggesting manual curation of non-coding variants will be challenging.

We are currently developing a pipeline for annotating noncoding variations called ncVAR (Mu et al., 2011). Noncoding variants are annotated using two sources of information: 1. Noncoding features annotated in the GENCODE annotation which include long noncoding RNA, pseudogenes, UTR, snoRNA, miRNA and snRNA (`http://www.gencodegenes.org/`) 2. Integrating variations based on annotations derived from functional genomics data from ENCODE consortium. We have developed PeakSeq, an approach to identify peak regions in ChIP-seq data sets that correspond to sites of transcription factor binding (Rozowsky et al, 2009). PeakSeq scores the results of ChIP-seq experiments by compensating for the mapability map and comparing these results against a normalized matching control data set. Using this method, we define genomic regions associated with transcription-factor binding and identify variants within these regions. To better represent the

DNA–protein interaction sites, we scan the TF peaks with consensus sequences of corresponding motifs to obtain sites representing TF-binding motifs. Therefore, we also annotate variants within specific transcription factor motifs. We plan to extend noncoding annotations to predicted enhancer regions.

After the variants are annotated, the effect of variations can be assessed by a variety of in silico methods. Nonsynonymous coding variations can be assessed by established tools SIFT and Polyphen2. Both these tools primarily predict the effect of nonsynonymous variations based on how well conserved a position is across different species. In addition, Polyphen also uses information derived from the 3D structure of protein to assess the effect of nonsynonymous variations. We have studied Loss of Function (LOF) variations in detail in personal genomes as well as in the Pilot phase of 1000 genomes (Balasubramanian et al, 2011). We are currently developing a LOF pipeline, `vat.gersteinlab.org`, for analysing the effect of LOF events such as premature STOP codons due to SNPs and indels, and SNPs affecting splice sites, in terms of their effect on function. Specifically, this LOF module will provide NMD predictions and assess the loss of functional or structural domains due to premature truncations. In general, effect of noncoding variations are harder to predict as conservation is restricted to evolutionarily closely–related species and many regulatory variations are also species-specific. Nonetheless, a variety of conservation-based tools such as GERP scores, Phastcons and PhyloP will be used to assess noncoding variants. We will also develop a LOF module for non-coding variations. This will focus on identifying variations that affect the core part of the transcription-factor binding motif in conjunction with sequence-conservation-based analyses.

Variations in an individual genome will be classified as common and rare based on the 1000genomes data. Variants that are present in the 1000G catalog at allele frequencies greater than 1% frequency will be identified as common variants. The remaining variants will be rare variations. Coding variations at less than 1% frequency can be identified from the 1000 genomes exome data as well as the NHLBI Exome Sequencing Project (ESP). This will also enable us to further classify an individual's rare coding variant as either a rare variant seen in the general population or a family-specific/de novo change private to the individual.

Automated prioritization of new measurements and functional testing Functional information supporting the impact of genetic variants is critical to genetic interpretation, and genetic linkage often means that it is difficult to tell which genetic variant is causal when a particular genomic region is implicated in an association study. This is particularly difficult for noncoding variants, which do not directly impact protein structure but may have a functional effect through modification of transcription factor binding sites, thereby affecting gene expression. Current technology has been limited in its ability to test these variants.

As one of the Centers for Excellence in Genomic Science, the Church laboratory has been pioneering technologies to enable study of these variants in a high throughput manner. We plan to pilot our technology by investigating noncoding variants from PGP participants.

In brief, our currently-funded ongoing technology development has been attacking this problem from multiple angles to develop a high-throughput system. We are using allele-specific expression profiling to associate individual haplotypes with expression differences (Lee et al., 2009). Working with engineered DNA-targeting nucleases (zinc finger nucleases, TALNs) and multiplex oligonucleotide-based targeted genome alteration (Wang et al., 2009) we are developing methods to introduce designed changes in a high throughput manner. Finally, pioneering work with high throughput single cell sequencing promises high throughput association of single-cell transcriptional differences with various introduced DNA changes in a diverse population of modified cells.

Expression Quantitative Trait Loci (eQTL) has been used to uncover the genetic variations underlying expression variations in both human and model organisms (Schadt et al., 2003, Yvert et al., 2003, Stranger et al., 2007). Some eQTL approaches have even been used to identify causal genetic variations of diseases (Schadt et al., 2005, Chen et al, 2008). We will use all variants (SNPs, indels, SVs) for these analyses.

When performing eQTL analysis, linear regression models (with possible L1 and L2 regularization constraints) are among the most popular approaches due to their simplicity and interpretability (Stranger et al., 2007, Lee et al., 2009). However, due to linkage disequilibrium and population structures, identifying causal genetic variations underlying complicated high-level phenotypes remains a challenging task, especially in humans, which poses great challenges for the clinical applications of personal genomics. Previous researchers often performed simple case controls or simple population corrections and tended to ignore detailed population structures and genetic structures (Stranger et al., 2007). We plan to develop more advanced regression models by incorporating prior biological expert domain knowledge to tackle these problems.

With the availability of more and more clinical record data, we can apply similar eQTL or eQTL-like approaches to combine gene expression and splicing data, genetic variation data, other omics measurements, and clinical record data. Thereby, we can identify disease-related genes and genetic variations, and pilot the selection of the most effective, personalized treatment based on patients' expression data and personal omic variation data.

3.4.2 Open shared hardware resource for community analysis

The high volumes of data produced by this project make sharing difficult. We have created infrastructure which allows this data to be shared freely. Our "compute and storage clouds" (see Figure for example user interaction) provide hosting, data storage, and batch processing services for a number of web applications (Zaranek et al., 2008, 2010). These applications involve data-intensive computation: they are conducive to asynchronous parallel processing, but their performance is limited by the available disk I/O bandwidth. Their demands for CPU time are highly variable, so it is sensible for them to share a pool of CPU resources by submitting batch jobs. They also tend to share data sets with one another, so it is sensible to share a large data storage system. The application developers have common goals rather than being in competition, so it is beneficial to let them see the source code and results of one another's batch processing jobs. The applications themselves may be maintained by different development teams, so each should run in its own independent virtual machine.

We will permit developers using our infrastructure to write "Freegols" – web services that utilize cluster computing and storage resources. The term Freegol, or Free Golem, emphasizes the idea that the web services are developed and maintained independently of the cluster infrastructure, and independently of one another.

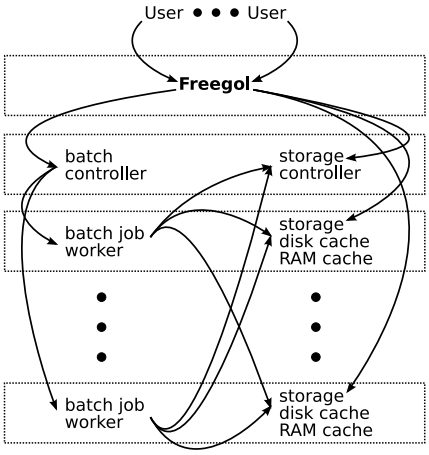


Figure 7: Dotted lines denote virtual machines. Batch processing workers in warehouse instances are dispatched by the batch controller on behalf of Freegols (see text). RAM cache, disk cache, and long term storage services are accessed by Freegols and batch jobs using the same client library.

3.4.3 Anonymization method development

New data reduction file formats will allow us to store data from NGS experiments while attempting to protect individual privacy, and allowing a broad range of data analysis. Variant Call Format (VCF), initially proposed by the 1000 Genomes Project (Danecek et al.), is now widely used for efficiently storing variant calls, including single nucleotide polymorphisms (SNPs), short deletions and insertions (indels), and structural variations (SVs), from NGS datasets. However, conventional VCF provides little protection to personal genomic privacy; an individual in the set can be uniquely identified based on only a few dozen independent variants. Instead, we aim to develop a software tool to encrypt VCF files with a key that is known only to the researcher, while still allowing for a wide variety of calculations – that, for instance, can be used to test pipelines and do demographic analyses. The encrypted file, an "eVCF", would still appear to be a typical human genome from a particular population – albeit de-personalized. Action of the key would uniquely restore the original non-anonymizable information. In addition, given the current migration trend of large-scale data sets to the cloud, we will develop a software workflow to optimize our VCF encryption for cloud computing. The workflow would also enable researchers to retrieve the output from their analyses and map it back to the correct personal genome safely in a local environment. This would add further security for NGS data analyses in the context of cloud computin

Different experiments may require retention of different genomic features, depending on the nature of downstream analyses of the variant calls. To this end, multiple anonymization options will be provided. One possibility in development is to blur the variations within a dataset, as with non-military global positioning system devices, building 'synthetic' personal genomes from a pool of individual genomes in a group. Specifically, we would permute the variants or variant blocks between individual genomes, such that the representative variations of the entire group are readily seen, but not those of any particular individual. While the exact manipulation of the eVCF file would be reversible and uniquely determined by the encryption key, persons without the key would never have adequate information to recover the data, and genomic privacy could remain protected.

Further, we aim to develop a Mapped Read Format (MRF) that can protect the sequences data—as well as reduce the file size, while retaining the ability to perform data analyses. The format is being developed within RSEQtools (Habegger et al., 2011) - a modular framework to analyze RNA-Seq data using compact and anonymized data summaries. Much of the relevant information in gene expression data resides in the localization of the reads rather than the sequence. Our MRF decouples the genomic coordinates of reads from the actual sequences. By using only the genomic coordinates, researchers and clinicians can determine which genes are expressed, quantify the expression of genes, exons, isoforms, identify novel regions of active transcription, and detect chimeric transcripts. Anonymized and small MRF files could potentially enable centralized repositories to make these files available to the research community, similar to what exists for gene expression microarray data. The clear separation of “private” information (sequences) from “public” information (alignment location) information would allow better control of functional genomic data. However, we still need to carefully test whether significant amounts of variant information could potentially "leak" back into the alignment file (e.g. a significant amount of personal deletions being exposed via obviously zero gene expression values).

3.5 Profiling time series, surgical samples, and body maps

We will demonstrate the proposed integrated technologies by generating in-depth, public time-series, omics profiling of surgical samples and post mortem “body maps” as follows:

- Time series with omics profiling and other profiling methods
  - 5 omics time-points per year in years 1-4; and 10 omics time-points in year 5; 10 other profiling time-points years 1-5 to include immune and microbiome profiling;
  - 2 individuals in years 1-4; 4-6 individuals in year 5.
- Surgical samples
  - 1-2 surgical samples per year in years 1-4; final pilot data of 10-20 samples in year 5
  - samples will be accepted, as received subject only to the availability of a standardized collection protocol; the repetoire of standardized protocols will grow throughout the project
- “Body map” of omics profiling from *post mortem* samples
  - 5 tissues with omics profiling with 4-6 participants starting in year 3-5
  - tissue samples will be selected to most closely match other resources such as Genotype-Tissue Expression (GTEx) and BrainSpan

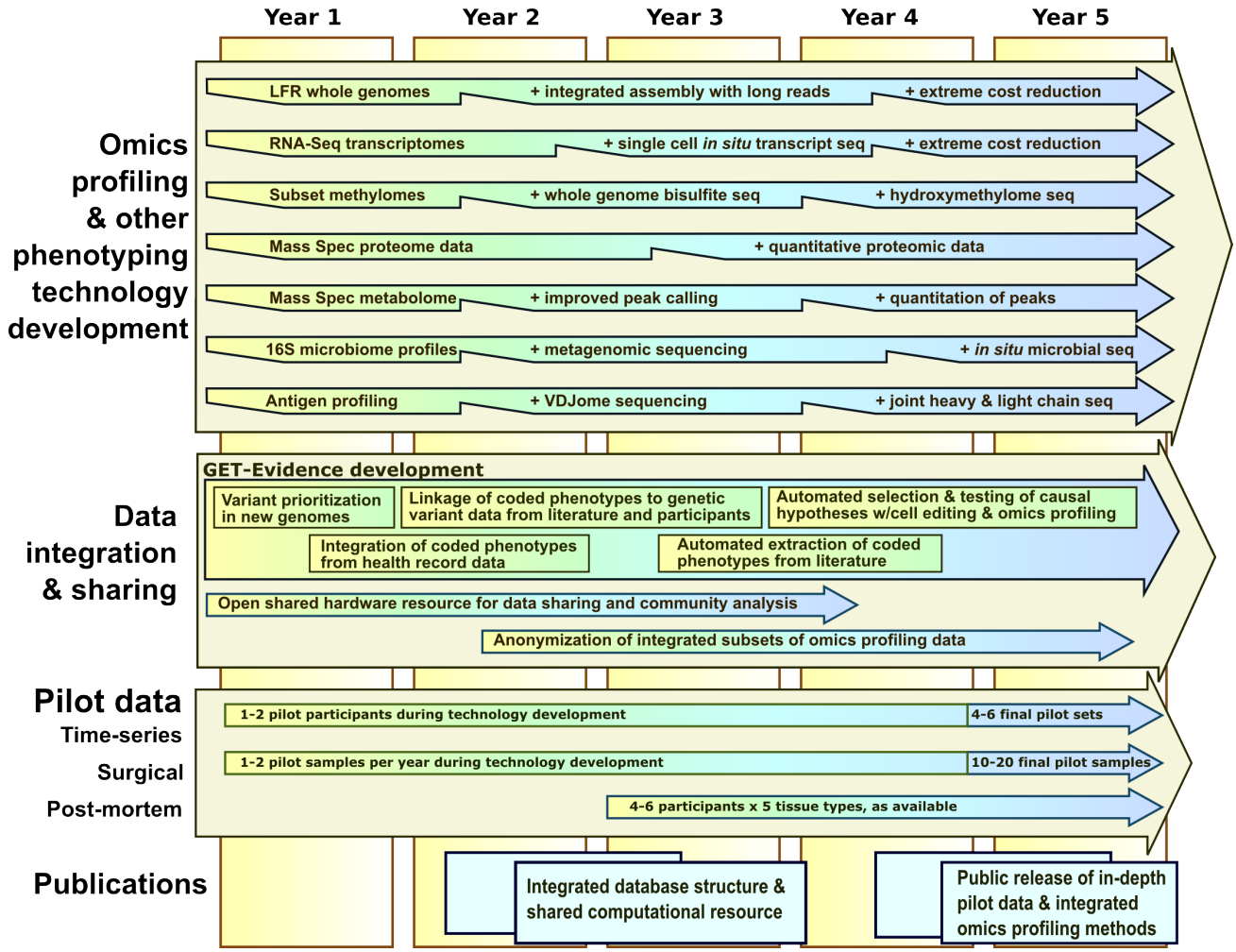
Additional detail and cost projections are provided in Harvard and Stanford budget justifications.

4 Appropriateness

A focus on improved technologies and protocols for integrating diverse omic profiling data as well as environmental and trait measurements, unbiased by the study of any particular disease, is an important new paradigm for human biology research. Demonstrating the power of this new paradigm involves more risk and, to a lesser extent, more funding support than usual for traditional funding mechanisms.

To keep the demonstration manageable we will profile ~40 participants in depth that are selected from a larger group of 10,000-100,000 total participants. Adoption by other groups in South Korea, Israel, Canada and elsewhere represents a growing, world-wide movement of participatory human research. Integration of data across these disparate studies and more traditional “anonymized” data-sets is a tremendous challenge. Explicit support for this integration, under the Tranformational R01 program, will accelerate the study of causal variations in human biology and could allow future population centric studies to focus on the most informative measurements as evidence accumulates in our pilot study.

5 Timeline



The Personal Genome Project anticipates enrollment of 10,000 individuals in 2012 and is IRB approved for enrollment of 100,000. We will use the PGP as a test-bed for technology development. The methods developed will demonstrate integrated omics profiling of arbitrary tissues from ~40 participants with a final post project completion cost of $300 / sample, including haplotype-phased genome sequencing with an error rate of 1×10^{-10}. Other investigators may use the developed technologies for their own studies or test their own measurements and analyses in a growing number of projects using “open-consent” worldwide.

Protection of Human Subjects

The proposed research will include both exempt and non-exempt human subjects research as described below. Our proposed research does not involve a Clinical Trial.

Harvard

We will investigate consented human "Personal Genome Project" samples under existing IRB approvals. The IRB approval date was 2011/02/22; the assurance numbers are M11665 and M15464.

Currently more than 1500 people are fully enrolled and hundreds of participants have provided samples for sequencing. The study is approved for enrollment of 100,000 people. We will select ~50 individuals (including women and minorities) for “in-depth” time-series analysis, analysis of surgical samples and/or, the *post mortem* molecular “body map” in years 3 through 5.

Stanford, Yale

This Human Subjects Research falls under Exemption 4.

Justification for Claimed Exemption:

Human subjects research will be conducted on existing publicly available data, and human cell samples (e.g., HapMap or PGP samples after they have been consented for public release and deposited at the Corriell Institute for Medical Research or other similar bio-repositories).

Inclusion of Women and Minorities

Harvard

We will investigate consented human "Personal Genome Project" samples under existing IRB approvals. The IRB approval date was 2011/02/22; the assurance numbers are M11665 and M15464.

Currently more than 1500 people are fully enrolled and hundreds of participants have provided samples for sequencing. Of currently enrolled participants, more than 500 are women, and more than 100 are non-white. We will select a few individuals (including women and minorities) for “in-depth” time-series analysis, analysis of surgical samples and, eventually, the *post mortem* molecular body map in years 3 through 5.

We anticipate enrollment of 10,000 participants by the start of this project and are currently approved for enrollment of 100,000. These large numbers will allow us to prioritize in-depth analysis of women and minorities.

Stanford, Yale

This Human Subjects Research for these sub-awards falls under Exemption 4.

Justification for Claimed Exemption:

Human subjects research will be conducted on existing publicly available data, and human cell samples (e.g., HapMap or PGP samples after they have been consented for public release and deposited at the Corriell Institute for Medical Research).

Targeted/Planned enrollment

In-depth Time Series

Ethnic Categories	Female	Male	Total
Hispanic or Latino	0-1	0-1	0-2
Not Hispanic or Latino	2-3	2-3	4-6
Ethnic Category: Total	2-4	2-4	4-6
Racial Categories			
American Indian / Alaska Native	0-1	0-1	0-2
Asian	0-1	0-1	0-2
Native Hawaiian or Other Pacific Islander	0-1	0-1	0-2
Black or African American	1	1	2
White	1	1	2
Racial Categories: Total of All Subjects	2-4	2-4	4-6

Post-mortem, molecular body-maps

Ethnic Categories	Female	Male	Total
Hispanic or Latino	0-1	0-1	0-2
Not Hispanic or Latino	2-3	2-3	4-6
Ethnic Category: Total	2-4	2-4	4-6
Racial Categories			
American Indian / Alaska Native	0-1	0-1	0-2
Asian	0-1	0-1	0-2
Native Hawaiian or Other Pacific Islander	0-1	0-1	0-2
Black or African American	1	1	2
White	1	1	2
Racial Categories: Total of All Subjects	2-4	2-4	4-6

In addition, surgical samples will be profiled from 10-20 individuals in the final data-set and 1-2 per year during ongoing technology development.

Inclusion of Children

All institutions

No enrollment of children under the age of 21 years old is permitted in this study. A legal/regulatory bar to inclusion of children as subjects exists as this research requires informed consent to any risks associated with the public release of that individual's combined personal genetic and health data.

Bibliography and References Cited

B. Adida, A. Sanyal, S Zabak, I.S. Kohane, K.D. Mandl. (2010) Indivo X: Developing a Fully Substitutable Personally Controlled Health Record Platform. AMIA Annu Symp Proc. PMC3041305.

E.A. Ashley, E.A. et al. (2010) Clinical assessment incorporating a personal genome. Lancet. PMC2937184.

S.Balasubramanian, et al. (2011). Gene inactivation and its implications for annotation in the era of personal genomics. Genes Dev. PMC3012931.

M.P. Ball, et al. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnology. PMID19329998.

C. Cano, et al. (2009) Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. J Biomed Inform. PMC2757509.

G. Dantas, et al. (2008) Bacteria subsisting on antibiotics. Science. PMID18388292.

P. Danecek, et al. (2011) The variant call format and VCFtools. Bioinformatics. PMC3137218.

F.E. Dewey, et al. (2011) Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genetics. PMC3174201.

R. Drmanac, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. PMID19892942.

L. Habegger, et al. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics. PMC3018817.

J.-I.Kim, et al. (2009) A highly annotated whole-genome sequence of a Korean individual. Nature. PMC2860965.

I.S. Kohane, et al. (2006) The incidentalome: a threat to genomic medicine. JAMA. PMID16835427.

J.-H. Lee, et al. (2009) A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. PLoS Genet. PMC2766639.

J. B. Li, et al. (2009) Multiplex padlock capture and sequencing reveal human hypermutable CpG variations. Genome Res. PMC2752131.

J. E. Lunshof, et al. (2008) From genetic privacy to open consent. Nat Rev Genet. PMID18379574.

X.J.Mu, et al. (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. PMC3167619.

C.M. Perou, T. Sørlie, M.B. Eisen, et al. (2000) Nature. PMID: 10963602.

G.J.Porreca, et al. (2007) Multiplex amplification of large sets of human exons. Nat Meth. PMID:17934468.

J. Rozowsky, et al. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. PMC3208341.

M.O. Sommer, et al. (2011) Functional characterization of the antibiotic resistance reservoir in the human microflora. Science. PMID19713526.

C. X. Song, et al. (2011) Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods. PMID22101853.

B.E.Stranger, et al. (2007) Population genomics of human gene expression. Nat Genet. PMC2683249.

H.H. Wang, et al. (2009) Programming cells by multiplex genome engineering and accelerated evolution. Nature. PMID19633652.

A. W. Zaranek, et al. (2008) Free factories: Unified infrastructure for data intensive web services. Proc USENIX Annu Tech Conf. PMC2877279.

A. W. Zaranek, et al. (2010) A survey of genomic traces reveals a common sequencing error, RNA editing and DNA editing. PLoS Genetics. PMC2873906.

M.J.Zilliox, R.A.Irizarry. (2007) A gene expression bar code for microarray data. Nat Methods. PMC3154617.

Consortium/Contractual Arrangements

Institution: Yale University
Co-I: Mark Gerstein

The group will develop methods to standardize Omics data and other types of data. They will be instrumental in generating the time-series profiles, omics profiles of surgical samples and omics profiles from *post mortem* "body maps". They will also reconcile the data-sets generated with other NIH resources such as 1K Genomes, Encode and other projects. Finally, they will explore the possibility of of data-anonymization of informative Omics data subsets. All data used will be publicly available.

Yale will provide regular written progress reports documenting the work and progress to date. These reports will be provided at least quarterly and will be accompanied by frequent conference calls, Internet communication and video-conferencing between the teams.

These status meetings and collection of reports will be organized by Yveta Masarova, administrator to the PI.

Institution: Stanford University
Co-I: Mike Snyder

The group will improve methods for Omics profiling of arbitrary tissues focused primarily on Transcriptomics, Proteomics and Metabolomics. They will also improve methods for immune profiling of blood samples. These methods will be applied to generate the time-series profiles, omics profiles of surgical samples and omics profiles of post-mortem "body maps". All samples used will be publicly available.

Stanford will provide regular written progress reports documenting the work and progress to date. These reports will be provided at least quarterly and will be accompanied by frequent conference calls, Internet communication and video-conferencing between the teams.

These status meetings and collection of reports will be organized by Yveta Masarova, administrator to the PI.

cbmi

Center for Biomedical Informatics
10 Shattuck Street • Boston, Massachusetts 02115-6089
617.432.2144 (ph) • 617.432.0693 (fax) • http://cbmi.med.harvard.edu

Professor George M. Church
Department of Genetics
Harvard Medical School
77 Avenue Louis Pasteur
Boston, MA 02115

January 5, 2012

Dear George,

I am excited to support your proposal Integrated Measurements of Causal Variations in Human Biology". The project clearly addresses a critical need in the field which is to enable the effective use of whole genome sequencing in a patient centric context. However, there are technical hurdles that we must overcome to offer such testing as well as significant challenges in understanding human genetic variation. I am particularly interested in establishing standards around which the broader community can more effectively communicate and share data. Such collaborative efforts are the only way in which we as a community will attempt to make use of the complex data sets that represent human genomes.

As you know, I have been involved in spearheading major efforts in using bioinformatics tools to advance healthcare. There is no doubt that the inclusion of genomic data in these efforts will continue to enhance the most personalized and effective healthcare.

In support of your proposal, I would be happy to collaborate with you and examine ways in which our common activities can be interfaced to enable quicker development of public resources for whole genome sequencing and interpretation. I look forward to such opportunities.

Sincerely,

Isaac S. Kohane, MD, PhD
Henderson Professor of Health Sciences and Technology
Children's Hospital and Harvard Medical School
Director, Countway Library of Medicine
Director, i2b2 National Center for Biomedical Computing
Co-Director, HMS Center for Biomedical Informatics

DATA AND RESOURCE SHARING PLAN
The PI and co-investigators on this project are committed to sharing data with other investigators. All samples in this study have been consented for release of molecular and clinical data in an open manner.

DATA SHARING: The investigators will abide by the current NIH data release policy for the sharing of final research data as published at http://grants.nih.gov/grants/policy/data_sharing/ Sequence and phenotype data, in particular, will be deposited in an appropriate NIH repository.

RESOURCE SHARING: The investigators will provide documentation of all data standards used to generate the data and all programs and methods used to analyze the data. All protocols will be published and openly shared. Cell-lines, furthermore, will be made freely available through the Coriell Institute for Medical Research or similar repository.

Harvard, Stanford, and Yale will endeavor to use a set of compatible legal tools to share information with each other, the scientific community and the general public. All participants in this study will be enrolled under an IRB approved “open-consent” protocol and the resulting data will be shared publicly under CC0. These standard legal tools are documented publicly at the Personal Genome Project site and the content is reproduced below.

We are committed to sharing resources with the scientific community and the general public, so that people may collaborate in ways that have not been possible before. We think this will promote discovery and accelerate our ability to understand how genomes and the environment combine to form human traits. To this end, these are some of the legal tools we use in our research

"Open-Consent" protocol
Our consent protocol is available for use and modification by PGP affiliates or other similar "public genomics" research studies.
● http://www.personalgenomes.org/consent/

Data
We are creating a public repository of integrated genomic, environmental, and trait datasets.
● "CC0" - http://creativecommons.org/publicdomain/zero/1.0/legalcode

Software Licenses
Our data-processing methods are publicly available as Free & Open Source Software (FOSS).
● "GNU AGPLv3" - http://www.gnu.org/licenses/agpl-3.0.html
● "GNU GPLv3" - http://www.gnu.org/licenses/gpl-3.0.html
● "CC0" - http://creativecommons.org/publicdomain/zero/1.0/legalcode

Text and Other Media Licenses
Educational and other materials that do not fit in the above categories are also publicly shared.
● "CC-BY-SA" - http://creativecommons.org/licenses/by-sa/3.0/legalcode
● “CC-BY" - http://creativecommons.org/licenses/by/3.0/legalcode
● "CC0" - http://creativecommons.org/publicdomain/zero/1.0/legalcode
The latest version of the above policies is at: http://www.personalgenomes.org/sharing.html