Special Article

Human Mutation

Official Journal HGVS Human Genome Variation Society www.hgvs.org

Back to the Future: From Genome to Metabolome

Joseph V. Thakuria,[1,2]* Alexander W. Zaranek,[1] George M. Church,[1] and Gerard T. Berry[3]

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts; [2]Division of Genetics, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; [3]Division of Genetics, Department of Medicine, Children's Hospital Boston, Boston, Massachusetts

For the Deep Phenotyping Special Issue

Received 20 February 2012; accepted revised manuscript 28 February 2012.

Published online 16 March 2012 in Wiley Online Library (www.wiley.com/humanmutation).DOI: 10.1002/humu.22073

ABSTRACT: In the traditional medical genetics setting, metabolic disorders, identified either clinically or through biochemical screening, undergo subsequent single gene testing to molecularly confirm diagnosis, provide further insight on natural disease history, and inform on disease management, treatment, familial testing, and reproductive options. For decades now, this process has been responsible for saving many lives worldwide. Only recently, though, has it become possible to move in the opposite direction by starting with an individual's whole genome or exome, and, guided by this data, study more minor perturbations in the absolute values and substrate ratios of clinically important biochemical analytes. Genomic individuality can also be used to guide more detailed phenotyping aimed at uncovering milder manifestations of known metabolic diseases. Metabolomic phenotyping in the Personal Genome Project for our first 200+ participants—all of whom are scheduled to have full genome sequence at more than 40× coverage available by May 2012—is aimed at uncovering potential subclinical and preclinical disease states in carriers of known pathogenic mutations and in lesser known rare variants that are protein predicted to be pathogenic. Our initial focus targets 88 genes involved in 68 metabolic disturbances with established evidence-based nutritional and/or pharmacological therapy as part of standard medical care.

Hum Mutat 33:809–812, 2012. © 2012 Wiley Periodicals, Inc.

KEY WORDS: genomics; metabolomics; nutritional therapy; pharmacological therapy

Background

In the 1985 American film, "Back to the Future," Marty McFly is accidentally sent back in time to the 1950s by a plutonium powered "flux capacitor" in a modified DeLorean upon reaching 88 mph. Throughout the film, the impact the future has on the past is explored. For decades now, mass spectrometric analysis typically utilizing a cylindrical capacitor ionization source to generate singly charged ions has been the backbone of diagnosis, management, and/or treatment for hundreds of inherited metabolic disorders. Because of proven clinical benefit, a subset of these disorders has made their way into formal newborn screening recommendations [ACMG, 2006]. Used for second-tier biochemical confirmation in conjunction with newborn screening programs, this technology has saved the lives of many newborns, children, and adults the world over. Starting with phenylketonuria in 1953, nutritional therapeutics guided by metabolic screening and serial testing has been conclusively shown to have medical benefit in a wide variety of enzyme deficiencies and other biochemical disorders.

As we enter the genomics era, our most diagnostically challenging cases in a medical genetics clinic are rapidly moving from a state of having no causal molecular candidates to having many candidates that need further evaluation and vetting. Nongenomic axes supporting causality from imaging, biochemical assay, functional cellular work, and other lines of evidence are increasingly important to help verify pathogenicity. Of these, biochemical assays have historically been the axis most frequently correlated with genetic data in a medical genetics practice.

Additionally, although much progress has been made in the screening, prevention, and treatment of inherited and primarily autosomal-recessive biochemical disorders, limited resources have been devoted to studying potential subclinical and preclinical disease states in carriers of known pathogenic mutations as well as in those harboring one or more less well-defined variants in known disease-causing genes. In large part, this is due to newborn screening and other testing modalities reliance on biochemical analytes for screening and diagnosis. In clinical practice, the higher sensitivity, specificity, and cost-effectiveness of screening biochemically are well justified.

Large-scale genomic research studies utilizing next-generation sequencing, however, provides opportunity for researchers to start with comprehensive genomic sequence data and, secondarily, study the resulting phenotype and biochemical profile. If consistent abnormal trends (even trends within the normal range) are found associated with carrier states and/or lesser known mutations in genes causing metabolic disorders, it is intriguing to think of what effect a modified diet specific to the defect will have on the health and well-being of such individuals. In order to explore this possibility, an important first step is identifying whether such trends exist and identifying in which disorders subclinical or preclinical biochemical phenotypes are prevalent. In some disorders, such as galactosemia, the biochemical and phenotypic effect of carrier status, and rarer Duarte allele 1 (GALT N314D + L218L) gain of function mutations have been studied and characterized [Scriver et al., 2012]. In many other metabolic disorders, however, phenotypically, little may be known beyond the scope of classically affected patients on the extreme end of a disease severity spectrum.

In 1908, Archibald Garrod introduced the idea of biochemical individuality and described four of the first known autosomal-recessive disorders: alkaptonuria, cystinuria, albinism,

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Joseph V. Thakuria, Division of Genetics, Massachusetts General Hospital, Boston, MA 02114. E-mail: jthakuria@genetics.med.harvard.edu and pentosuria. Since then, over 300 metabolic disorders with known diagnostic metabolic and genetic alteration have been discovered. And although Norwegian physician, Ivar Asbjorn Folling discovered phenylketonuria in 1934, it was not until approximately 20 years later that dramatically effective, evidence-based nutritional therapy was recognized through the collective work of Lionel Penrose, George Jervis, and Horst Bickel [Berry, 2010]. Although the number of severe metabolic disorders with effective dietary and/or drug therapy continues to increase, identification of more subtle subclinical and preclinical disease states utilizing whole genome or exome data has not yet been explored.

Research findings will eventually move into clinical practice as insight from next-generation sequencing technology is applied to metabolic lessons from the past, and greater correlation between genomic individuality and biochemical individuality is delineated in an expanded number of individuals. Subsequently, identification of subclinical and preclinical phenotypes should lead to effective dietary and drug therapy in individuals exhibiting milder or nonclassic phenotypes of known metabolic diseases. As this will have the effect of broadening both genetic and biochemical screening, a resulting cycle of medical discovery, screening, and treatment recommendations in this area can be expected to accelerate in the coming years.

The Personal Genome Project (PGP) is a Harvard Medical School study with institutional review board approval for the enrollment of 100,000 individuals for complete genomic and phenotypic study (http://www.personalgenomes.org/). Study participants must be at least 21 years of age. Enrollment is entirely online and requires passing an exam testing comprehension of human subject research, PGP protocols, and basic genetics. Study guides and consent forms are available online at: http://www.personalgenomes.org/consent/ and http://www.pgpstudy.org/ [Church, 2005; Lunshof et al., 2010].

Integrated datasets of linked genomic and phenotypic data on each individual are made available publicly as a free resource for the research community and to the study participants themselves. To allow for sequence confirmation and functional studies, participant cell lines are also made available and distributed through the Coriell Institute (http://ccr.coriell.org/). These include fibroblast and Epstein-Barr virus-transformed lymphoblastoid cell lines. Private quarterly questionnaires are used to track safety and prospective clinical outcomes.

More than 1,000 participants have provided phenotype data via personal health records and standardized questionnaires. The project is also actively pursuing the development and administration of new phenotyping tools with help from both the research community and commercial organizations. Immediate phenotyping plans include providing microbiome measurements from several body sites, telomere lengths, and methylation profiles. Participants may then elect to participate in these additional activities as they become available. More than 97% of participants have expressed interest in doing so. More than 85% of participants have also expressed interest in providing discarded surgical samples for analyses and more than 90% of participants have volunteered to provide samples postmortem.

To date, over 1,500 individuals have fully completed enrollment with twice as many at some stage of the enrollment process. From these, 200+ are being selected to have whole-genome sequence at more than 40× coverage from blood- and saliva-derived DNA. Clinical prioritization of participants is aided by a questionnaire designed to enhance for strong genetic etiology. (Table 1)

In this communication, we describe initial plans for metabolic phenotyping in our first 200+ individuals with phenotypically integrated whole-genome sequence datasets. Initial analysis is focused on 88 genes involved in 68 well-established biochemical genetic disorders with known dietary and/or pharmacologic treatment. The vast majority of primary and secondary newborn screening targets recommended by the American College of Medical Genetics (ACMG) are included (Supp. Table S1).

Table 1. PGP Screening Questions Enriching for Genetic Etiologies

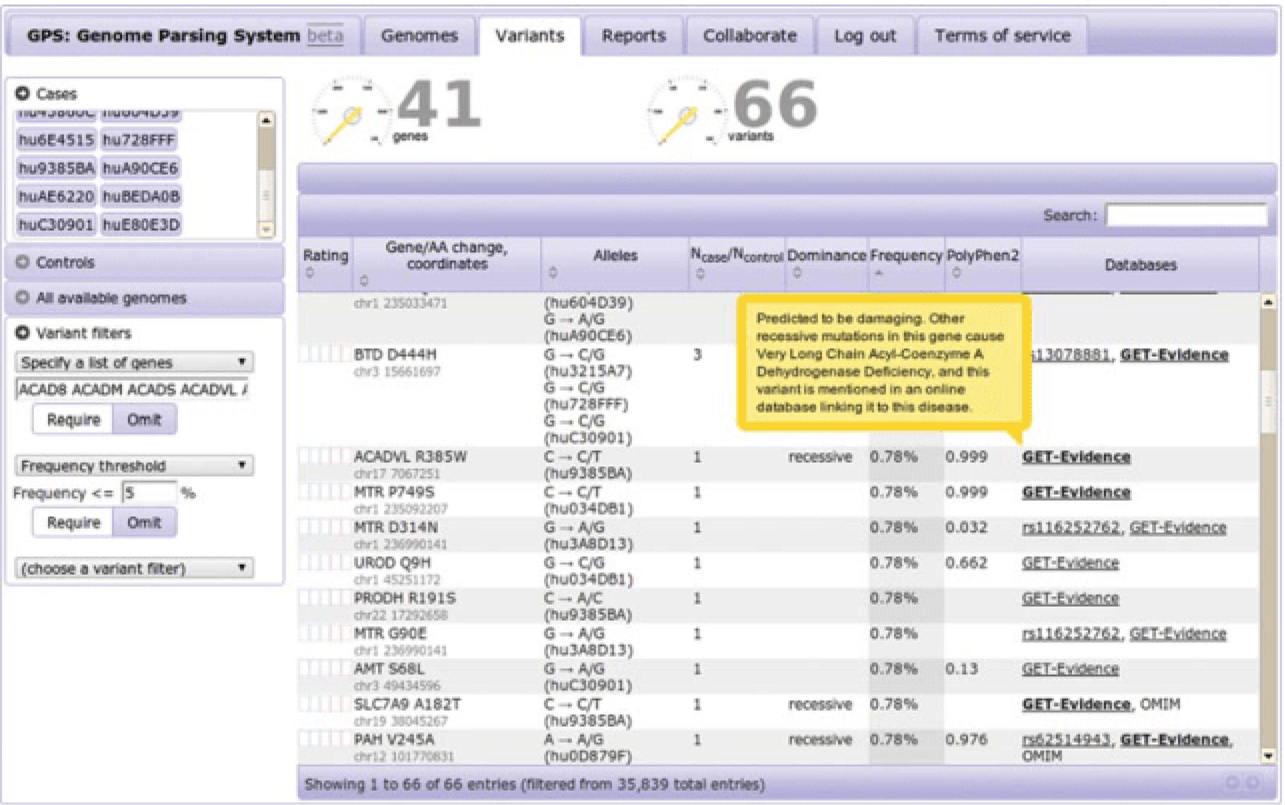
Question type(s)	Purpose
1. Age	Prioritizes for both early-onset disease and advanced age controls with retrospective data.
2. Presence of severe or rare disease phenotype (self-reported).	Prioritize by condition or suspected genetic etiology (free text permitted for detailed responses).
3. If yes to #2, disease onset, rarity, severity, and presence of family history are assessed.	Prioritize further within the disease category of interest.
4. Is objective disease evidence from physician diagnosis and/or medical testing available?	Prioritize diseases with evidence beyond self-reporting and/or with supporting laboratory, imaging, or genetic data.
5. Will data from #4 be uploaded into participant PGP profiles?	Prioritize by accessible medical phenotype data.
6. Demographics: geographic (from local to continent level), as well as ethnic (i.e., "ethnicity" will not always be concordant with "geography") and gender. Geographic and ethnic data (both voluntary to answer) can be provided for all four grandparents.	Provide flexibility in rapid hypothesis-driven prioritization of already enrolled cohorts. Enable ancestry, epigenetic, environmental studies. Apply appropriate population frequency thresholds when interpreting "-omic" variants and other datasets.
7. Co-enrollment with affected or unaffected family members? State disease(s), affected status, and familial relationship.	Prioritize on feasibility of familial-based genomic or other analyses.
8. What type of biological samples will be provided (e.g., blood, saliva, "normal" flora (for microbiomes), skin, or other tissues)?	Prioritize based on available tissue/cell types or feasibility of somatic versus germline comparative studies.

Methods

Purified DNA from saliva or blood on over 200 PGP participants are slated for library preparation and sequencing by Complete Genomics, Inc. Data are annotated using their 2.X pipeline matching against the National Center for Biotechnology Information (NCBI) build 37 reference genome. A preliminary interpretation derived from this data is provided privately to participants and becomes public after they are allotted 30 days for review. Individual datasets are linked to the participant ID and are published in the public domain under the Creative Commons CC0 waiver.

We have developed the GET-Evidence system to produce reports and make datasets available to the study participants and to the public. The purpose of GET-Evidence is to build up a public database of variant annotations that will ultimately be used to assist in clinical analysis. GET-Evidence prioritizes variants for review based on allele frequency, protein-predicted pathogenicity, and presence in clinical gene and variant databases. As more variants are reviewed, the participants' reports are updated to reflect the newer annotations.

For user-specified analyses, Clinical Future (founded by J.V.T. and A.W.Z. with support from G.M.C.) has developed the Genome Parsing System "GPS"—a secure, private Web service for genomic and phenotypic data management and filtration. A sample GPS analysis of the PGP pilot genomes is found in Figure 1. The system has been used to effectively filter variants for high-clinical importance parsing



ACAD8 ACADM ACADS ACADVL

Frequency <= 5 %

Rating	Gene/AA change, coordinates	Alleles	$N_{case}/N_{control}$	Dominance	Frequency	PolyPhen2	Databases
	chr1 235033471	(hu604D39) G → A/G (huA90CE6)					
	BTD D444H chr3 15661697	G → C/G (hu3215A7) G → C/G (hu728FFF) G → C/G (huC30901)	3				s13078881, **GET-Evidence**
	ACADVL R385W chr17 7067251	C → C/T (hu9385BA)	1	recessive	0.78%	0.999	**GET-Evidence**
	MTR P749S chr1 235092207	C → C/T (hu034DB1)	1		0.78%	0.999	**GET-Evidence**
	MTR D314N chr1 236990141	G → A/G (hu3A8D13)	1		0.78%	0.032	rs116252762, GET-Evidence
	UROD Q9H chr1 45251172	G → C/G (hu034DB1)	1		0.78%	0.662	GET-Evidence
	PRODH R191S chr22 17292658	C → A/C (hu9385BA)	1		0.78%		GET-Evidence
	MTR G90E chr1 236990141	G → A/G (hu3A8D13)	1		0.78%		rs116252762, GET-Evidence
	AMT S68L chr3 49434596	G → A/G (huC30901)	1		0.78%	0.13	GET-Evidence
	SLC7A9 A182T chr19 38045267	C → C/T (hu9385BA)	1	recessive	0.78%		**GET-Evidence**, OMIM
	PAH V245A chr12 101770831	A → A/G (hu0D879F)	1	recessive	0.78%	0.976	rs62514943, **GET-Evidence**, OMIM

Predicted to be damaging. Other recessive mutations in this gene cause Very Long Chain Acyl-Coenzyme A Dehydrogenase Deficiency, and this variant is mentioned in an online database linking it to this disease.

Showing 1 to 66 of 66 entries (filtered from 35,839 total entries)

Figure 1. Genome Parsing System (GPS) screenshot: Whole-genome data from 16 Personal Genome Project (PGP) participants parsed against 88 metabolic disease genes show an average of four to nine variants per genome, are less than 5% in frequency, and appear in OMIM and/or are protein predicted to be damaging. N.B.: the predominance of the MAF of 0.0078 in these rarest variants occurs because each variant occurs only once in a limited frequency database of 64 public genomes used for this analysis.

genomic data against clinical gene and variant databases, filtering by low allele frequency and protein-predicted pathogenicity [Adzhubei et al., 2010]. By analyzing aggregate data from 5,400 individual exomes, available from the NHLBI Exome Variant Server, we find four to nine variants with frequency less than 10% specifically from the 88 genes associated with the targeted disorders from Supp. Table S1. In the PGP pilot data, each participant has four to nine variants with frequency less than 5% and zero to one variants in OMIM (www.omim.org) specifically from the 88 genes associated with the targeted disorders from Supp. Table S1. When analysis is extended to the NHLBI Exome Variant dataset, we find slightly fewer variants, three to seven on average per exome, with a frequency less than 5% [Exome Variant Server, 2012].

Consensus from several publications also indicates that an average of 10–30 variants per genome are present heterozygously for autosomal-recessive disorders. One or more of these typically involve established metabolic disorders. Furthermore, we avoided the summation due to the wide population-specific variability for each disorder, but adding up estimated carrier rates for all 88 disorders should also support the hypothesis of finding at least one biochemical disorder of interest, simply on the basis of carrier status for at least one treatable metabolic disorder listed in Supp. Table S1 [Lupski et al., 2010].

All 200+ participants will have the following laboratory studies performed in a CLIA certified clinical laboratory for biochemical phenotyping that are relevant to the treatable disorders listed in Supp. Table S1: plasma amino acids, urine organic acids, plasma acylcarnitines, urine acylglycines, basic chem7, NH_4 level, carnitine profile (free and total), folate level, zinc level, B12 level, urine-reducing substances, lipid profile, hemoglobin electrophoresis, pyridoxine level, biotin level, urine galactitol, galactose-1-phosphate, copper level, ceruloplasmin, magnesium level, carbohydrate-deficient transferrin, urine and plasma porphobilinogen, urine and plasma delta-aminolevulinic acid, RBC plasmalogens, pipecolic acid, and plasma very-long-chain fatty acids. The majority of these biochemical tests will be performed in-house at Children's Hospital Boston and Massachusetts General Hospital with some highly specialized tests being performed by outside clinical collaborators (Table 2).

After identification of both known and potentially pathogenic mutations within the targeted 88 biochemical genes with the GPS platform (Supp. Table S1), we will analyze participant metabolite values and ratios in which mutation status suggests possible deviation from normal values using Mann–Whitney and Kolmogorov–Smirnov tests. Analyses for statistically significant and pathophysiologically consistent differences observed against matched controls will be aided by performing the same biochemical testing on all participants and allowing each participant to also serve as control for the biochemical disorders and pathways in which they are not found to have potentially pathogenic mutations.

Discussion

The concept of biochemical individuality first introduced by Garrod has had enormous impact on modern medicine and human

Table 2. Planned Biochemical Phenotyping for 200+ PGP Participants with Whole-Genome Data

Plasma amino acids
Urine organic acids
Plasma acylcarnitines
Urine acylglycines
Sodium
Potassium
Chloride
Bicarbonate
Blood urea nitrogen
Creatinine
Glucose
NH_4 level
Carnitine profile (free and total)
Folate level
Zinc level
B12 level
Urine-reducing substances
Lipid profile
Hemoglobin electrophoresis
Pyridoxine level
Biotin level
Urine galactitol
Galactose-1-phosphate
Copper level
Ceruloplasmin
Magnesium level
Carbohydrate-deficient transferrin
Urine and plasma porphobilinogen
Urine and plasma delta-aminolevulinic acid
RBC plasmalogens
Pipecolic acid
Plasma very-long-chain fatty acids

genetics. In contrast, due to direct observation of familial similarities, especially physical similarities in the case of monozygotic twins, "genomic individuality" has not only been assumed since before the term "genome" was coined but also could correctly be considered a redundant term. Yet, only recently, with the deep sequencing of multiple whole genomes, exomes, and targeted sequencing of genes in the tens to thousands becoming more practical in clinical research, are we able to systematically study and correlate three critical axes of medical research: genomic, metabolomic, and phenomic. Additional axes, such as functional data on an individual's cell line, will also aid in supporting hypothesis of causality. Four decades worth of observational data on the natural history of treated patients for some of these disorders that were the first to be biochemically screened for in the 1960s is also extremely informative.

We expect to see correlations between rarer variants and larger deviations from normal (in the expected direction for the specific disorder and biochemical metabolites). The frequency and degree to which analyte deviations are in the expected direction for the particular disorder will also be biostatistically analyzed. Since all 200+ participants will have the full range of biochemical studies relevant to 88 genes involved in 68 treatable biochemical disorders, those without suspected pathogenic variants in a specific gene(s) or disorder will serve as controls for those who are biochemically studied based on sequence data for the same specific disorder.

Achieving statistical significance correlating relevant biochemical analytes with genomic data in individuals found to have one or more potentially pathogenic mutations across these 68 biochemical disorders in over 200 individuals will be challenging because of multiple hypothesis testing. We still expect to see interesting data trends supporting known biochemical pathophysiology even in this cohort size when targeting the rarest protein altering variants. In some instances, statistically significant differences should eventually be observed once a critical mass of individuals with matching genotype, metabolic profile, and phenotype is reached.

Neither the metabolic diseases we have chosen to study in our initial metabolic analysis nor the laboratory tests we will perform on all 200+ individuals are comprehensive of treatable metabolic disorders or available clinical biochemical testing, respectively, but it should generate helpful pilot data and lay the foundation for future trials studying an expanded number of genes, metabolic disorders, and individuals.

Our finding of four to nine rare variants predicted to be pathogenic variants per genome on average within 88 genes causing metabolic disease with established dietary and/or pharmacologic therapy is highly dependent on the filtering algorithm. This low figure is also bounded by the limited number of genes studied and our current understanding of metabolic diseases. Regardless, at 10 or less variants per person with our current algorithm, the prospect of systematic development of individualized dietary and/or medical data informed by genomic and metabolomic data finally comes into practical view.

We anticipate the biochemical interrogation of 200+ whole genomes guided by genomic individuality, and linked to a process of individual phenotype data gathering guided by the known natural history of a subset of clinically well-characterized metabolic disorders will prove valuable.

Identifying the genomic and metabolomic circumstances under which subclinical or preclinical states exist for these same disorders may eventually lead to the first evidence-based efficacy studies for nutrigenomics in these patients who would now otherwise go untreated and undetected by current methods and standard practices.

Acknowledgments

Disclosure Statement: J.V.T. and A.W.Z. declare potential conflict of interest as cofounders of Clinical Future, Inc., Somerville, MA.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev, SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

American College of Medical Genetics. 2006. Health Resources and Services Administration Commissioned Report. Newborn screening: toward a uniform screening panel and system. Genet Med 8:1S–252S.

Berry GT. 2008. Metabolic profiling. Nestle Nutr Workshop Ser Pediatr Program 62:55–75.

Church GC. 2005. Personal genome project. Mol Syst Biol 1–3.

Exome Variant Server. NHLBI Exome Sequencing Project (ESP). Seattle, WA. Available at: http://evs.gs.washington.edu/EVS/. (Accessed February, 2012).

Lunshof JE, Bobe J, Aach J, Angrist M, Thakuria JV, Vorhaus DB, Hoehe MR, Church GM. Personal genomes in progress: from the human genome project to the personal genome project. 2010. Dialogues Clin Neurosci 12:47–60.

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, and others. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362:1181–1191.

Scriver CR, Beaudet AL, Sly WS, Valle D, Childs B, Kinzler KW, Vogelstein B. 2012. Metabolic and molecular bases of inherited disease. New York: McGraw-Hill.