Genome-Phenome Analysis for Diagnosis and Gene Discovery
Presented at ACMG 2013

Michael M. Segal MD PhD (SimulConsult, Brookline MA), Marc S. Williams MD (Geisinger Health System, Danville PA), Gerard Tromp (Geisinger Health System, Danville PA),
Joseph G. Gleeson MD (UCSD, La Jolla CA)

The cost of sequencing human genomes has fallen low enough that the main concern about the clinical use is the cost and time of clinical interpretation. The system demonstrated here addresses this concern, both for clinical diagnosis and gene discovery. It integrates the analysis of the patient's clinical findings and the genomic variant table, and compares with known phenotypes of diseases using a diagnostic decision support tool that allows use of both pertinent positives and negatives, and is already in clinical use. Since the assemblage of these phenotypes is known as the "phenome", we refer to the integrated process as "genome-phenome analysis" (www.simulconsult.com/genome/).

Methods

Variant tables consist of thousands of variants found in a proband or trio. The variant table is annotated with gene and sequence information, and with scores relating to type of mutation, population frequency, homozygous shares and heterozygous shares in the lab's database, conservation, functional impact, splice impact prediction, and quality and depth of read. Based on these annotations, severity scores are assigned to variants on a scale from 0 to 4. The severity scores for variants are then used to derive severity scores for each possible zygosity of a gene. To do so, variants for each gene are inter-compared among family members using a novelty and compound heterozygosity analysis. This reduces the number of genes with non-zero severity score typically to <200, depending on the number of individuals included and the ability to keep the number of de novo variant calls low by re-calling their zygosity in trio-based batches. Various cutoffs for frequency and scores are configurable, and the individuals deemed affected can be changes as desired.

The analysis of genes with non-zero severity scores includes:

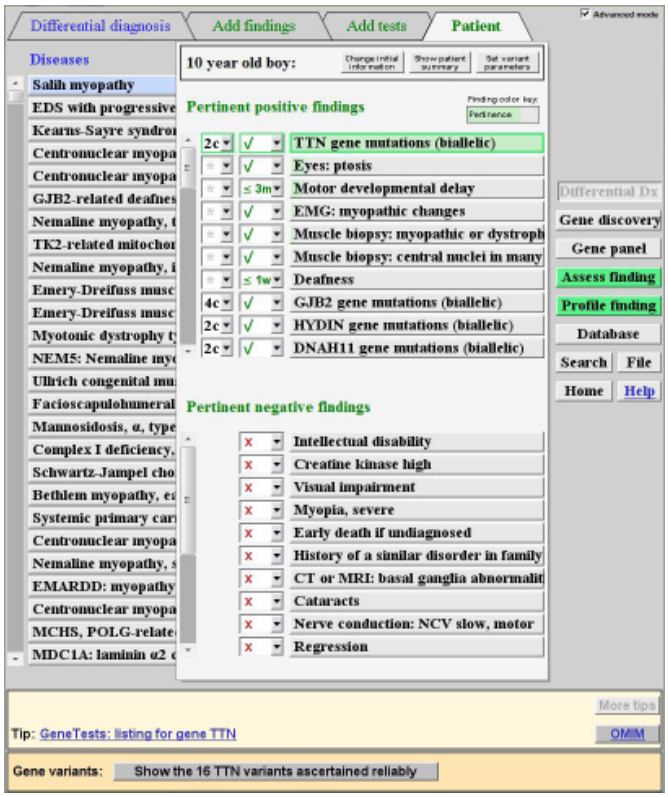
- **Clinical diagnosis mode.** For genes with recognized phenotypes (~15%), two prioritized diagnostic metrics are displayed: ranked by phenotype match and gene pertinence. The gene zygosities are associated in the curated database with the clinical diseases, and the severity scores for gene zygosities are used to weight the probability that the variant is of clinical significance, and thus its influence on the differential diagnosis and gene pertinence.
- **Gene-discovery mode.** For genes with no recognized human phenotype (~85%) or ones in which zygosity doesn't match the known human phenotype, genes are displayed for relevant inheritance models, ranked by severity scores, with hyperlinks to information for gene discovery. The number of genes in this mode is kept low using the same novelty and compound heterozygosity analysis used in the clinical diagnosis mode.

We report here on preliminary results using the genomes from Harvard's CLARITY genome interpretation competition and other genomes that had been analyzed previously using more manual protocols in the Gleeson group at UCSD.

To specify pertinent positive and negative clinical and non-genetic lab findings requires 2-5 minutes for an experienced user of the software, a function ideally done by the referring clinician who knows the patient best and is more clinically-oriented than most lab personnel. Variant table processing with clinical correlation takes ~2 seconds for a trio with ~30,000 variants, and recalculations after changing annotation cutoffs or affected family members takes <2 seconds as well.

Gene pertinence is calculated in a manner similar to the calculation of usefulness of findings (Segal 2004), except that instead of applying the calculation prospectively to findings being considered and determining their expected effect on the differential diagnosis, the pertinence calculation is applied retrospectively to findings to determine the effect they had on the differential diagnosis.

The resulting display is shown in the Figure. Pertinent positive and pertinent negative clinical and lab findings are shown for one of the CLARITY cases, with pertinence shown as green shading. A differential diagnosis is also shown on the left, with disease probabilities shown as blue shading. Note that clinical and lab information consist of individual observable pertinent positive and pertinent negative findings, with age included where appropriate. No judgments as to mechanism of inheritance or disease classification are needed, in contrast to other approaches in which such information is used as filters to restrict the analysis to standardized subsets of diseases.



Results

Nine families have been analyzed. In 6 of those, a known diagnosis allowed us to measure performance; those results are shown in Table 1. Five of the cases had autosomal recessive inheritance, and one had autosomal dominant inheritance, but no such hypotheses about inheritance needed to be specified as input. All families have data from the trio, and the AD dominant family also has data from a maternal uncle.

The 3 of 9 cases not included in the tables included one of the 3 CLARITY families and one of the 6 Gleeson cases not included in the tables because the causative gene was not clear to the CLARITY organizers or the Gleeson lab, even after their and our analysis. Another one of the Gleeson cases is not included in the tables, since although the diagnostic software arrived at the same gene as chosen by the Gleeson lab, this was a Gene Discovery situation, and thus not comparable using the metrics in Table 2.

The 6 cases identified 8 diseases corresponding to 8 genes with autosomal recessive or autosomal dominant inheritance. In one of the CLARITY cases there was consensus among the organizers that two genes, *TTN* and *GJB2*, combined to produce the phenotype. In another, there was consensus that *TRPM4* produced the cardiac electrophysiological phenotype, and our assessment, shared also by the CLARITY organizers, that in addition *GJA1* was a variant of uncertain significance that may have produced the

cardiac structural phenotype, which is not accounted for by known data about *TRPM4*.

Table 1: Cases

	CLARITY 1	CLARITY 2	Gleeson 1642	Gleeson 1572	Gleeson 1492	Gleeson 1536
Gene decision	*TTN* and *GJB2*	*TRPM4* (and *GJA1*)	*WDR62*	*RPGR1P1L*	*CC2D2A*	*CEP290*
Inheritance	AR and AR	AD (and AD)	AR	AR	AR	AR
Diagnosis decision	TTN myopathy and GJB2 deafness	Cardiac conduction defect AD (and atrioventricular septal defect)	MCPH2: microcephaly, primary AR 2	MKS5: Meckel syndrome 5	COACH/ Joubert	COACH/ Joubert
Findings used (positive)	6	2	4	5	3	5
Findings used (negative)	17	5	0	0	0	0
Variants	32387	36215	5736	3900	14210	4189
Variants with severity scores > 0	3674	3357	1504	1220	4905	1208

Trends in the data (Table 2):

Phenome-only ranking is good but not excellent (P1 in Table 2): Using just the clinical and non-gene laboratory information, the #1 ranked disease phenotype was the correct one for 4 of the 8 diseases in the 6 probands (lumping diseases that were indistinguishable except by genetic testing (e.g., Meckel syndrome types attributable to different genes). To reach all diagnoses by going through the differential diagnoses lists, including the secondary diagnoses in the 2 CLARITY cases, an extra 167 diseases would need to be considered.

Genome-only ranking is good but not excellent (G2): Using just the genome information from the family genomes (trio, and in the AD case, also the maternal uncle), the #1 ranked gene in pertinence was the correct one for 5 of 8 genes for the 6 probands (also counting rank #2 for cases with 2 genes). To reach all diagnoses by going through the gene pertinence lists, including the secondary diagnoses in the 2 CLARITY cases, an extra 38 genes would need to be considered. (To facilitate comparison to other conditions in the software, the genome-only analysis shown here made use of the default of the software of taking into account age of the patient and frequency of diseases, but an analysis based only on ranking of severity scores produced similar results.)

Phenotype rank from integrated genome-phenome analysis improves as more genome information is added: Using clinical information alone, the phenotype rank of #1 was correct for 4 of the 8 diseases (row P1), which increased to 5 with proband genome information (row P2) and 6 using family genome information (row P3). Similarly, the number of extra diseases to consider fell from 167 to 25 to 6.

Gene identification from integrated genome-phenome analysis improves as more genome information is added: Using the clinical information together with only proband genetic information, the gene pertinence rank was correct for 4 of the 8 genes (G3). Adding the genetic information from the other family members resulted in 7 of the 8 genes topping the lists (G4). Similarly, the number of extra genes to consider fell from 28 to 1. The one case in which the identification was not perfect was one in which two genes with broad phenotypes were involved, and the gene ranked #2 in pertinence was one that had elements of both cardiac electrophysiological and structural phenotypes.

Adding clinical information to a proband has a similar effect to adding other family genomes: Measures of extra genes to check and gene pertinence rank correctness were similar for proband + family (G2) and proband + clinical (G3).

Gene pertinence provides a better metric than phenotype rank: Gene pertinence ranks the answers as correctly as possible in 7 of 8 instances (row G4) while phenotype rank does so only in 6 of 8 instances (row P3). Similarly, the number of extra items was 1 for the gene pertinence rank, while it was 6 for the phenotype rank.

Integrated use of clinical information to prioritize genome analysis reduces entities needing clinical review: For a proband analysis, adding the clinical information resulted in gene rankings being as correct as possible going from 2 instances (in row G1) to 4 (row G3), and reducing extra genes to consider from 124 to 28. For a trio/family analysis, adding the clinical information resulted in gene rankings being as correct as possible going from 4 instances (row G2) to 7 (row G4), and reducing extra genes to consider from 38 to 1.

Table 2: Ranks in phenotype and gene pertinence lists

					Rank of correct diagnosis						
		Findings used			CLARITY		Gleeson				
	Row	Clinical	Proband variants	Parents' variants	1	2	1642	1572	1492	1536	Extra items to check
Phenotype rank	P1	✓			1 and 88	13 (and 73)	>100 but 11 for MCPH5	3, but 1 for MKS3	1	1	167
	P2	✓	✓		1 and 11	2 (and 16)	3	1	1	1	25
	P3	✓	✓	✓	1 and 6	1 (and 4)	1	1	1	1	6
Gene pertinence rank	G1		✓		6 and 11	15 (and 34)	5	1	78	1	124
	G2		✓	✓	27 and 1	10 (and 2)	1	1	6	1	38
	G3	✓	✓		1 and 6	6 (and 24)	3	1	1	1	28
	G4	✓	✓	✓	1 and 2	1 (and 3)	1	1	1	1	1

Conclusions

Integrated Genome-Phenome analysis addresses the cost and time concerns about interpreting genomic testing accurately: Dr. Bruce Korf summarized the concerns about the clinical usefulness of genome sequencing when he stated, "We are close to having a $1,000 genome sequence, but this may be accompanied by a $1 million interpretation" (Davies 2010). Here, we demonstrate a system with processing times for this clinical correlation for a typical exome of ~2 seconds, which performs adequately on the phenotype rank and excellently on the gene pertinence rank. The effect of adding clinical information or family genomes to proband genomic information were similar, and adding both was highly accurate. In practice, clinicians look below #1 in the gene pertinence when signing off on a genomic analysis, but in no case, including ones with 2 pathogenic genes, did we conclude it was necessary to go below #3 in the gene pertinence list. However, going through the whole list of ~30 genes took ~15 minutes per trio, using the curated information and links of the software. Even when the ~30 minutes needed to formulate a write-up is included, the analysis and reporting times remained < 1 hour, going a long way to address the concerns about the difficulty of clinical interpretation. For cases of gene discovery, where no good match was found for known human phenotypes, following the OMIM links in the gene discovery part of the software took ~2 hours per case.

The metric of "gene pertinence" solves one of the key issues in diagnostic decision support: A long-standing concern about diagnostic decision support software has been the "two diagnosis" problem, the difficulty of choosing a single known phenotype when the patient's clinical picture is composed of two or more causes. This problem has long been considered the major argument against real-world effectiveness of diagnostic decision support. The ability demonstrated here to attach a pertinence metric to a gold standard, a genetic variant, provides a way of teasing apart the diagnoses and solving the "two diagnosis" problem for the subset of situations in which the diseases have a genetic cause or some other pathognomonic finding. It solves the "two diagnosis" problem in a way that doesn't depend on manual parsing of findings into separate bins by the clinician interpreting the test, thus avoiding concerns about tractability of analysis. In addition, the pertinence measure was superior to the differential diagnosis in identifying the correct disease entities in situations in which the clinical presentation was not typical. Thus, use of the gene pertinence metric is an advance in applying decision support to the interpretation of genomes.

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number 1R43HG006974-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References: Davies K (2010) A Grand Vision for Genomic Medicine. Bio-IT World, 28 September, http://www.bio-itworld.com/BioIT_Article.aspx?id=101670
Segal MM (2004) Systems and methods for diagnosing medical conditions. US Patent 6,754,655