2013

Structured Genome-Scale Variant and Clinical Data Reporting for Meta-Analysis in an Era of Genomic Medicine

Darrell L. Dinwiddie

Carol J. Saunders
Children's Mercy Hospital

Emily G. Farrow
Children's Mercy Hospital

Sarah E. Soden
Children's Mercy Hospital

Neil A. Miller
Children's Mercy Hospital

See next page for additional authors

Recommended Citation

Dinwiddie, D. L., Saunders, C. J., Farrow, E. G., Soden, S. E., Miller, N. A., Kingsmore, S. F. Structured Genome-Scale Variant and Clinical Data Reporting for Meta-Analysis in an Era of Genomic Medicine *Journal of Genomes and Exomes* 2, 31-42 (2013).

Journal of Genomes and Exomes

OPINION

Structured Genome-Scale Variant and Clinical Data Reporting for Meta-Analysis in an Era of Genomic Medicine

Darrell L. Dinwiddie, Carol J. Saunders, Emily G. Farrow, Sarah E. Soden, Neil A. Miller and Stephen F. Kingsmore

Center for Pediatric Genomic Medicine, Children's Mercy Hospital, Kansas City, MO, USA. Department of Pediatrics, University of Missouri-Kansas City School of Medicine, Kansas City, MO, USA.
Corresponding author email: sfkingsmore@cmh.edu

Abstract

Summary: The Journal of Genomes and Exomes is a new, peer-reviewed, open-access, online publication whose scope comprises reporting of high quality genome, exome, and gene panel sequences with attendant, detailed phenotypes. The intent of this journal is to facilitate comparisons between genome, exome and gene panel sequencing studies in order to assist significance testing of the genotype-phenotype associations, particularly those in uncommon genetic diseases. While there is yet to be a consensus regarding these classifications, the definition of an empiric set is helpful in understanding error models. Herein we have suggested structured templates for submissions and the rationale for the data fields in these templates, as well as examples. The editorial board of the Journal of Genomes and Exomes is keen to receive feedback regarding these structured templates and welcomes submissions.

Keywords: genome, exome, DNA, diagnosis, disease, treatment, genomic medicine, nucleotide, variant, genetic disease

Journal of Genomes and Exomes 2013:2 31–42

doi: 10.4137/JGE.S10180

This article is available from http://www.la-press.com.

Introduction

The staggering diversity in human genomes is exemplified by the numerous unique in addition to the many common genetic features and phenotypes present in each individual. Genotype-phenotype associations promise to reveal the basis of many human attributes—both beneficial and deleterious.[1] This is truthful, despite the hubris of genetic essentialism: the belief that genes are deterministic of all phenotypes.[2,3]

The popular concept that biological knowledge is the product of independent research by an investigator working in isolation is no longer unrivaled. In other fields of research, most notably particle physics, the concept is endangered, and almost extinct. There is a growing consensus that the sum of the efforts of a community of investigators working together is much greater than that of the parts in isolation.[4,5] This was noted several millennia ago by King Solomon the Wise.[7] Within biomedical science, human genome analysis has been the forerunner of data sharing and community analysis by virtue of the digital nature of genetic data, which facilitates standardization, compilation, searching, and computation.[1] This has been accelerated by massively parallel next generation sequencing and analysis (NGSA) and systematized funding by the National Institutes of Health.[36,37]

Concomitant compilations or searchable, standardized phenotype descriptions, unfortunately, have lagged far behind genome compilation. The vast majority of human genome and exome sequences are associated either with no phenotypic information or a single bivariable. Efforts are underway to standardize phenotype collections,[7,8] but as yet have not been married with NGSA.[33]

The Journal of Genomes and Exomes is a new forum for structured reporting of rich phenotypic data together with corresponding comprehensive sets of variants culled from high quality NGSA of genomes, exomes, and gene panels.[9] Here we describe the rationale for a working model of the initial standardized data formats and minimal descriptors of human genome sequences and phenotypes for the Journal as well as provide examples.

Results

The primary goal of standardized reporting of genome-scale variation and attendant phenotypes is to allow comparisons to be made seamlessly between studies. In this way, the Journal will facilitate testing of the significance of genotype-phenotype associations, particularly those in rare genetic diseases. The requirements for cancer genomics are somewhat different and are in development. To achieve the goal of cross study comparisons, data formats should be simple, searchable and in common use in order to allow compilation. Flat files of delimited (eg, comma or tab separated) values are preferred. Another prerequisite for data formats is future interoperability with additional layers of genomic complexity (such as haplotypes) and phenotypic complexity (such as quantitative phenotypic descriptions). All datasets must, of course, be de-identified in compliance with the Health Information Privacy Act (HIPAA).[10] The determination of an institutional review board (IRB) regarding whether such datasets constitute research involving human subjects or ought to be waived should be noted. If the former, a statement indicating that the study was approved by an IRB, that informed consent was obtained from all subjects, and that all research was done in accordance with the Declaration of Helsinki, must be included.

NGSA Metrics

Deep NGSA is an accurate and sensitive tool to identify and genotype most nucleotide variants at genome scale. NGSA on an Illumina HiSeq 2000 sequencer with an average of 36X and 60X aligned coverage of 100 base pairs (bp) accurately reads genotypes ~95% and ~97%, respectively, of the 3,101,788,170 nucleotide reference genome (the "callable" genome).[13] This was recently recapitulated with 2 × 100 nucleotide HiSeq 2500 NGSA.[11] 100 gigabases (GB) of aligned sequence (average 32X) is becoming a standard for new, reportable genotypes in short read genomes with most NGSA technologies.[13]

Standardized metrics for sequence depth for exomes are less well established. With singleton 100 nuceotide HiSeq 2000 sequencing of Illumina hybrid selection-enriched exomes (approximately 62 Mb of targets), approximately 2% of target nucleotides have no coverage (C0, Fig. 1A). This proportion does not change in the range of 5–20 GB of aligned sequences (Fig. 1A). Fortunately, C0 nucleotides in exome NGSA are highly reproducible,[12] defining a "callable" exome.[13] The proportion of exome nucleotides with 16X coverage (C16), a conservative depth for highly

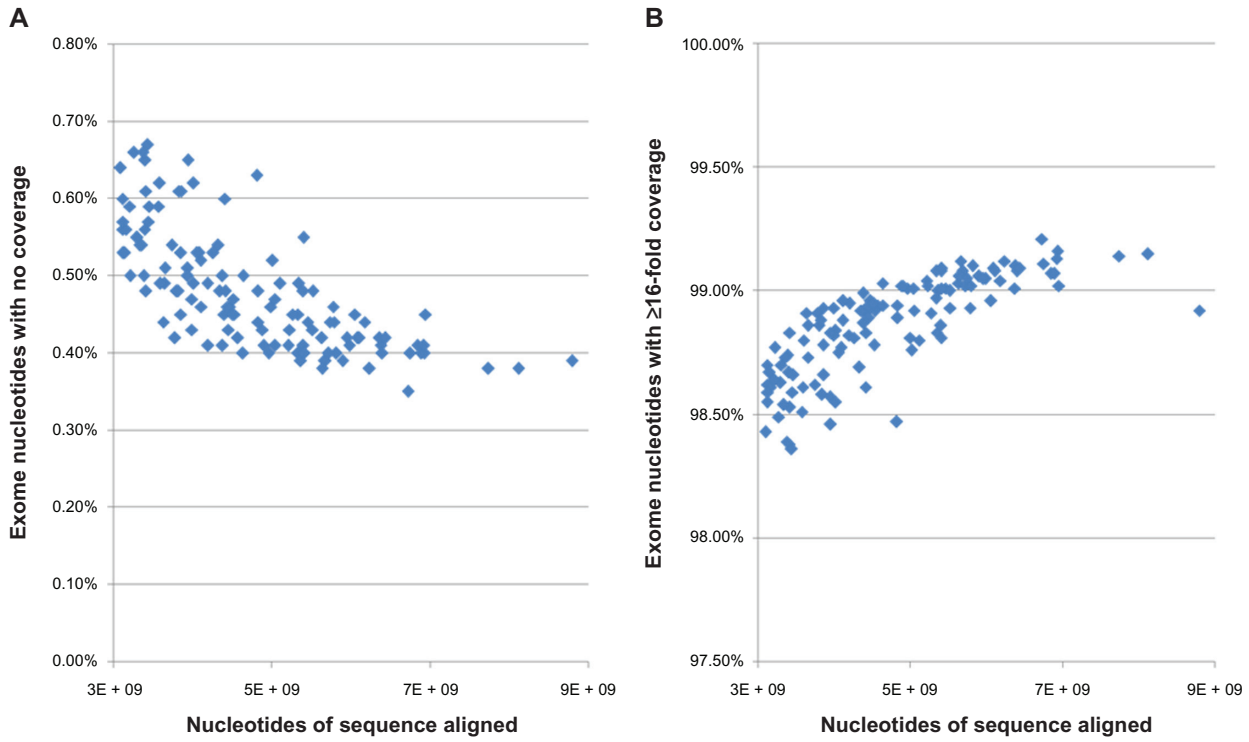

Figure 1. Change in the depth of coverage of the human exome as a function of the amount of aligned sequences.
Notes: Panels show results of singleton 100 nucleotide HiSeq 2000 sequencing of Illumina hybrid selection-enriched exomes (approximately 62 Mb of targets). (**A**) In the range of 5–20 GB of aligned sequences, approximately 2% of exome nucleotides have no coverage (C0). (**B**) There is an approximately linear increase in exome nucleotides with at least 16X coverage (C16), a conservative depth for highly accurate genotyping, as the amount of aligned sequence varies between 5–20 GB.

accurate genotyping,[12] increases somewhat linearly over the same range of aligned sequence (Fig. 1B). Also using these methods, 8 GB of aligned exome sequence corresponds to approximately 70X average coverage and C16 for approximately 99% of target nucleotides (Fig. 1B). Exome capture enrichment is available from multiple vendors and in multiple versions, all covering slightly different targets. Numerous studies have compared depth of coverage and percent of targeted nucleotides covered across exome enrichment from different companies and highlight that each lab may produce may produce different results even with the same enrichment technology.[22–24] Consequently, rather than recommend a specific amount of sequence required for each exome enrichment version, we suggest a minimum average coverage of 70X for targeted regions and C16 for each variant called. The percent of targeted nucleotides covered at C16 and C0 should be reported. Some laboratories apply different coverage minimums for homozygous and heterozygous variants, albeit tools such as GATK do not apply simple coverage filters for calling genotypes, and parameterization is not yet being standardized between centers.

Standardized metrics for sequence depth for gene panels are relatively primitive. The depth of NGSA coverage for accurate and sensitive genotyping of a panel comprising 437 recessive disease genes and 1,978,041 nucleotides enriched by hybrid selection has been extensively evaluated.[12] Agilent hybrid enrichment of these targets, followed by singleton 50 nucleotide Illumina GAIIx or HiSeq 2000 NGSA to aligned sequence depth of 0.75–2.00 GB, gave a highly reproducible subset representing approximately 1% of target nucleotide with C0.[12] The proportion of target nucleotides with 20X coverage (C20) increased linearly over the same range of aligned sequence.[12] 1 GB of sequence corresponded to C20 for approximately 90% of target nucleotide and ~250X average coverage. More recently, we have evaluated the same metrics for a panel comprising 526 recessive disease genes and 2,158,661 nucleotides (Dinwiddie et al, unpublished). Illumina hybrid enrichment of these targets, followed by singleton 100 nucleotide Illumina HiSeq 2000 NGSA to an aligned sequence depth of approximately 3 GB, gave 0.48% of highly reproducible target

nucleotide with C0. 3 GB of sequence corresponded to an average coverage of 850X and to C16 for approximately 98.5% of target nucleotide. 1 GB (or ~350X coverage) is suggested as the interim minimum standard for reportable hybrid selection-enriched panels. Enrichment of targeted panels for NGSA using multiplexed polymerase reaction should theoretically yield only cognate amplicons,[12] but the same interim minimum standard for reporting is desired. Coverage recommendations are much more difficult to standardize in targeted oncology panels, since tumor cell populations can be oligoclonal or polyclonal, differing in somatic mutations.[25]

NGSA technologies are evolving very rapidly. Current technologies and protocols result in different read lengths, raw sequence accuracies, and phasing errors. It is therefore important to record the methods with sufficient detail to allow a future understanding of whether discrepancies between studies were the result of methodological differences. A minimum set of NGSA methodological data fields are the sample preparation (library generation) vendor and version, enrichment technology vendor and type (hybrid enrichment or amplicons), sequencing technology vendor and type (panel, exome, genome), and sequence type (singleton or paired, read length). Average sequence quality scores, alignment algorithm, and parameterization are becoming less material as NGSA technologies mature, but are desired.

Scope of Variant Reporting

100 GB raw genome sequences and 3.1 GB consensus human genome sequences (or 8 GB raw exome sequences and 62 Mb consensus exome sequences) are unwieldy. Provided that the version of the human reference used for alignment is noted, there is little rationale at present for retention of reference nucleotides in most compilations of human genome sequences. Currently, NGSA cannot reliably assemble haplotypes over meaningful genomic intervals at genome scale. When possible, however, retention of phase information will become very important. At present, NGSA is limited in its ability to detect copy number variations (CNV) or structural variations. Thus, the initial minimal descriptors of human genome sequences for the Journal will be nucleotide and polynucleotide substitutions, insertions, and deletions. The cutoff for the size of callable polynucleotide variants will vary for substitutions, insertions, and deletions as well as among NGSA technologies. Typically, in our experience, contiguous substitutions within a read are limited to a maximum size of about five nucleotides, insertions to about fifty nucleotides and deletions to about two kilobases (Dinwiddie et al, unpublished). However, this is highly dependent on the alignment and variant detection methods used. In the future, additional variant categories will be added, as methods are validated for their identification by NGSA, genotyping and imputation of pathogenicity (such as CNV, chromosomal events, regulatory variants, synonymous variants of phenotypic relevance).

Variant Annotation Standardization

Standards for the annotation of nucleotide variants and their likely functional consequence(s) are relatively well established:

- The Variant Call Format (VCF) for nucleotide variant description;
- The Human Genome Variation Society (HGVS) format for recording the coordinates and identities of nucleotide variants at the levels of chromosome, transcript(s), and predicted protein(s) sequences;[14]
- For variants at gene loci, the HUGO Gene Nomenclature Committee (HGNC) nomenclature for gene names;[15]
- For variants in monogenic phenotypes, an American College of Medical Genetics (ACMG) pathogenicity category[16] (Table 1);
- Human Gene Mutation Database (HGMD),[17] NCBI dbSNP, NCBI ClinVar,[33] Leiden Open Variation Database (LOVD),[34] and/or MutaDATABASE[35] accession numbers, if present;
- For monogenic phenotypes, the Online Mendelian Inheritance in Man (OMIM) accession number, if available;
- For phenotypes other than monogenic disorders, a controlled vocabulary, such as SNOMED CT (Systematized Nomenclature of Medicine) or Human Phenotype Ontology (HPO) terms;[7,8,27,28]
- For phenotypes other than monogenic disorders, in silico prediction of variant consequences, for example using the ENSEMBL Variant Effect Predictor, or ANNOVAR with ENSEMBL or RefSeq/UCSC gene annotations;[18–20]

Table 1. The example of the American College of Medical Genetics (ACMG) categories for description of the pathogenicity of nucleotide variants in monogenic diseases.

Category	Description	Criteria
1	Known to be causative of disease	HGMD "disease mutant" OR dbSNP "pathogenic" clinical significance AND allele frequency <1%
2	Novel but expected to be causative of disease	Loss of initiation codon OR premature stop codon OR loss of stop codon OR whole transcript deleted OR frameshifting indel OR affects splice donor/acceptor site OR disrupts splicing by deletion causing coding domain/intron fusion AND allele frequency <1%
3	Previously unreported; may or may not be causative of disease	Non-synonymous substition OR in-frame indel OR disruption of polypyrimidine tract OR overlap with 5′ exonic, %′ flank or 3′ exonic splice contexts AND allele frequency <1%
4	Probably not causative of disease	Synonymous variants unlikely to affect splicing, deep intronic variants, etc
5	Previously reported; recognized neutral variant	Review of literature and central mutation databases to assess degree of certainty that variant is not disease causing. For severe recessive diseases, homozygosity in unaffected individuals is strong negative evidence; For severe dominant diseases, presence in unaffected individuals is strong negative evidence; For rare genetic disorders, allele frequency >1% is strong negative evidence

Note: The exclusion of variants as pathogenic on the basis of allele frequencies greater than 1% is well accepted but not definitive.

- Where available, the variant allele frequency. There are several public resources of such information,[31,32] however, allele frequencies from other populations are welcomed. This is particularly important since many variants annotated as causative of uncommon monogenic diseases have allele frequencies that are too high to be causative. Allele frequency >1% and homozygosity in healthy individuals useful for distinguishing variants annotated as causative of uncommon monogenic diseases from misannotated common polymorphisms.[21] Known exceptions exist including Factor V Leiden (frequency 3%–8% in general US and European populations), Hemoglobin S and C (7.4% and 1.8% in African Americans, respectively) and hemochromatosis *HFE* p.C282Y (11% in European populations);
- For non-synonymous variants, scores predictive of deleteriousness (such as SIFT, PolyPhen)[26] or tests of evolutionary conservation. The use of multiple prediction tools can yield conflicting evidence. However, many newer tools are available (PANTHER, FATHMM, Hansa, nsSNPAnalyzer, SNPs&GO and MutPred). A recent comparison suggested that SNPs&GO and MutPred may be the best of these, and superior to PolyPhen or SIFT.[29,30]

A standardized data format that combines these elements is shown in Table 2, where individual variations are rows and descriptors as columns. The magnitude of variant reporting of this type is shown in Table 3. Genome, exome and targeted gene panel NGSA at depths of 120 GB, 8 GB and 3 GB, respectively, yield, on average, 4,079,138, 87,542 and 8,510 variants, respectively. Since files with 4 million rows are not trivial to search, we suggest reporting only of gene-associated variants, or variants that may have a functional consequence (ACMG Categories 1–3, thus omitting most synonymous and intronic variants). For causative variants other than nucleotide substitutions in reports of genetic diseases, confirmatory studies in trios are requested using established, traditional methods.

Phenotypic Description and Standardization

Rich description of the components of phenotypes is necessary for meta-analysis of genotype-phenotype associations. Standardized Human Phenotype Ontology (HPO) or SNOMED CT (Systematized Nomenclature of Medicine) terms are becoming the consensus for this purpose.[7,8,27,28] Most SNOMED CT terms are qualitative clinical findings derived from human diseases. They have limitations for

Table 2. An example of a standardized format for reporting of nucleotide variants.

Chr	Variant start	Variant stop	Variant type	Reference nucleotide	Variant nucleotide	Gene(s)	HGVS cDNA
19	282753	282753	Substitution	G	A	PPAP2C	ENST00000269812.1:c.539C>T; ENST00000434325.1:c.371C>T; ENST00000327790.1:c.602C>T
19	287970	287971	Insertion	–	T	PPAP2C	ENST00000269812.1: c.204+49_204+50insA; ENST00000327790.1: c.267+49_267+50insA; ENST00000434325.1: c.36+49_36+50insA
19	288329	288330	Insertion	–	C	PPAP2C	ENST00000434325.1: c.-116-159dupG; ENST00000327790.1: c.116-159dupG; ENST00000269812.1: c.53-159dupG
19	288374	288374	Substitution	T	C	PPAP2C	ENST00000269812.1: c.53-203A>G; ENST00000434325.1: c.-116-203A>G; ENST00000327790.1: c.116-203A>G
19	307037	307037	Substitution	C	T	MIER2	ENST00000264819.3: c.1616+82G>A
19	308681	308681	Substitution	G	A	MIER2	ENST00000264819.3: c.1110-16C>T
19	311708	311708	Substitution	A	G	MIER2	ENST00000264819.3: c.984+137T>C
19	311787	311787	Substitution	C	G	MIER2	ENST00000264819.3: c.984+58G>C
19	312026	312026	Substitution	C	T	MIER2	ENST00000264819.3: c.890-87G>A
19	312143	312143	Substitution	T	C	MIER2	ENST00000264819.3: c.889+48A>G

description of certain relevant findings, such as dysmorphology terms, specific laboratory, pathology, or imaging findings. In these cases, a parent descriptor should be used. HPO terms are superior in this regard and have recently been mapped to the non-structured, but widely used, terms of the London Dysmorphology Database. HPO terms also have the advantage that they are in the public domain. If necessary, additional detail can be provided in the text. The burden of phenotype description of this magnitude is important to assess. Table 4 shows the results of detailed translation of the medical records of 8 individuals with monogenic disorders into SNOMED CT terms. There was an average of 14 terms per individual (range 7–21). Material negative findings may also be added. A model for a standardized data format that combines these elements is shown in Table 4.

Ideally phenotypic descriptions of genetic diseases would include pedigrees. Illustrations of pedigrees, such as progeny, should be provided, or at least the initial rows of the phenotypic description describe the sample label, accession number, gender, clinical status, accession number(s) of sample(s) from related individual(s), relationships between those samples,

HGVS protein	AA change	BLOSUM	Impact	Geno-type	dbSNP accession	CMH allele frequency	Classification
ENSP00000388565.1: p.Ala124Val; ENSP00000269812.1: p.Ala180Val; ENSP00000329697.1: p.Ala201Val	A>V	0	Non-synonymous	2	rs1138439	126/651	4
				2	rs61624925	119/651	4
				2	rs35895757	32/651	4
				2	rs12981067	34/651	4
				1	rs72982402	33/651	4
				1	rs59415447	2/651	4
				1	rs72984427	33/651	4
				1	rs111820777	35/651	4
				1	rs60667274	33/651	4
				2	rs10416918	152/651	4

Notes: Variant characteristics are listed as columns. Variants are rows.
Abbreviations: Chr, chromosome; Ref, reference; HGVS, Human Genome Variation Society.

summary of phenotype (OMIM), and primary causative locus (HGNC).

In the future it will be desirable to add modifiers to the terms, such as age of onset, frequency, severity, duration, complications, and outcomes. It will also be very important to add treatments and responses to treatments. It is envisaged that these innovations will be added in time.

Discussion

Genomic medicine is a new, structured approach to disease discovery, diagnosis, and management that prominently features NGSA.[4] Over the next several years, genomic medicine is anticipated to discover the genes that underpin ~3500 Mendelian disorders of unknown cause. It will also identify genotype-phenotype relationships and on an unparalleled scale. In addition, it promises to deliver simultaneous, comprehensive differential diagnostic testing of likely genetic illnesses at time of presentation, accelerating molecular diagnosis, increasing rates of ascertainment, minimizing duration of empiric treatment, and time-to-genetic counseling. In the longer term, genomic medicine will help

Table 3. Nucleotide variants yielded by genome, exome and targeted gene panel NGSA at depths of 120 GB, 8 GB and 3 GB, respectively, and categorization according to likelihood of being causative of genetic disease.

Sample	Sequencing type	Total variants	Gene-assoc. variants	Variants with allele frequency >1% (n = 662)	Cat. 1 variants	Cat. 2 variants	Cat. 3 variants	Cat. 4 variants	Mis-annotated as disease causative variants (Cat. 5)
UDT1	Targeted panel	9,090	8,943	486	4	0	44	438	9
UDT2	Targeted panel	8,922	8,744	516	5	1	42	468	9
UDT3	Targeted panel	8,216	8,031	205	2	0	31	172	9
UDT173	Targeted panel	7,142	6,980	123	0	0	9	114	5
UDT4	Targeted panel	9,181	8,962	440	2	0	32	406	14
CMH001	Exome	91,119	88,990	2,870	14	16	417	2,423	12
CMH002	Exome	93,542	91,368	2,881	11	25	393	2,452	15
CMH006	Exome	100,761	98,548	3,965	6	27	474	3,458	12
CMH007	Exome	92,566	90,471	3,768	5	17	438	3,308	14
CMH064	Exome	109,720	106,982	4,202	14	26	451	3,711	23
CMH064	Genome	3,985,315	1,869,515	1,249,633	24	260	1,446	1,247,903	9
CMH076	Genome	4,497,940	2,098,886	1,479,793	19	281	1,930	1,477,563	7
UDT2	Genome	4,014,036	1,888,650	691,123	22	292	2,647	688,162	14
UDT173	Genome	3,976,271	1,859,095	668,922	15	265	1,339	667,303	12
CMH184	Genome	3,922,130	1,840,738	516,549	9	93	844	515,612	17
Average	Targeted panel	8,510	8,332	354	3	0	32	320	9
Average	Exome	97,542	95,272	3,537	10	22	435	3,070	15
Average	Genome	4,079,138	1,911,377	921,204	18	238	1,641	919,309	12

Table 4. An example of a standardized format for reporting of phenotypes.

Sample feature	SNOMED term	CMH001	CMH002	CMH006	CMH007	09_956	CMH172	CMH184	CMH185	PMLD1
Phenotype status		A	A	A	A	A	A	A	A	A
OMIM ID		208920	208920	612940	612940	614171	614498	-	-	607694
Causative gene		606350	606350	179035	179035	604310	614506	609004	609004	614258
Related other sample(s)		CMH002	CMH001	CMH007	CMH006	-	-	CMH185	CMH184	-
Relationship(s) with other samples		Sibling	Sibling	Sibling	Sibling	-	-	Sibling	Sibling	-
Gender		F	F	M	M	F	F	M	M	M
Sequencing scope		E	E	E	E	E	G	G	G	E
GB sequence aligned		14	14	10	19	27	113	137	117	13
Read Phred score > 30 (%)		nk	nk	nk	nk	nk	91	90	93	nk
Sequence type (HiSeq)		2000	2000	2000	2000	2000	500	2500	2500	2000
Aligner		GSNAP	GSNAP	GSNAP	GSNAP	GSNAP	ELAND	ELAND	ELAND	GSNAP
Read length (nucleotides)		100	100	100	100	2 × 130	2 × 100	2 × 100	2 × 100	100
Average coverage										104
Ataxia	20262006	1	1							1
Hypotonia	398152000	1	1	0	1					
Gait disturbance	22325002	1	1							1
Dysarthria	8011004	1	1							1
Fatigue	84229001	1	0							
Cerebellar atrophy	371313002	1	1							
Chorea	271700006	1	1							
Dysmetria	32566006	1	0							
Decreased deep tendon reflexes	37280007	1	1							
Developmental delay	248290002	0	1	1	1	0				1
Delayed speech development	229721007	0	1	1	0					
Hip dysplasia	52781008			1	1					
Cryptorchidism	204878001			1	1					1
Loose skin	58588007			1	1					
Hyperextensible joints	298203008			1	1					
Scoliosis	298382003			1	0					
Syndactyly	373413006			1	1					
Macrocephaly	19410003			1	0					
Frontal bossing	90145001			1	1					
Low set ears	95515009			1	1					
Very low birth weight	276611006			1	1					
Tremor	26079004			1	1		0			
Epilepsy	84757009			1	0		1			
Gynecomastia	4754008			0	1					
Pes planus	53226007			0	1					
Agenesis corpus callosum	5102002			0	1					
Colpocephaly	253160006			0	1					
Obesity	414916001			1	0					

(*Continued*)

Table 4. (*Continued*)

Sample feature	SNOMED term	CMH001	CMH002	CMH006	CMH007	09_956	CMH172	CMH184	CMH185	PMLD1
Albinism	15890002					1				
Nystagmus	563001					1				
Skin infection	108365000					1				
Thrombocytopenia	302215000					1				
Leukopenia	84828003					1				
Lymphocyte disorder	3239007					1				
Recurrent infection	428875002					1				1
Small head	271611007						1			
Birth length ≤ 3rd centile	169887003						1			
Meconium stained amniotic fluid	168092006						1			
Status epilepticus	230456007						1			
Patent foramen ovale	204317008						1			
Tricuspid valve regurgitation	111287006						1			
Micrognathia	32958008						1			
Abnormal shape of nose	249321001						1			
Hypoxemia	389087006						1			
Bradycardia	48867003						1			
Upslanted palpebral fissures	246799009						1			
Congenital abnormality of external ear	282038006						1			
Clinodactyly	17268007						1			
Increased muscle tone	56731001						1			
Hyperreflexia	86854008						1			
Clonus	36649002						1			
Dextro transposition of the great arteries	399216004							1	0	
Total anomalous pulmonary venous return	111323005							1		
Ventricular septal defect	30288003							1	1	
Atrial septal defect	70142008							1		
Pulmonary valve atresia	204342004							1	0	
Patent ductus arteriosus	83330001							1		
Situs inversus	43876007							1	1	
Dextrocardia	27637000							1	1	
Pulmonary valve stenosis	67278007									
Azygous continuation of the inferior vena cava	253315002									
Double outlet right ventricle	7484005								1	
Hernia	414403008									1
External ophthalmoplegia	19373007									1
Hypometric saccades	246768008									1
High myopia	34187009									1
Dysmyelination	125495003									1

Notes: Individuals are shown as columns. Phenotypes are rows.

pharmacogenetically-informed treatment regimens to be implemented.[5–7] Lastly, it will increasingly provide molecular diagnoses and potential drug/dosing selections that could not have been ascertained by conventional approaches by virtue of pleiotropic clinical presentation and genetic heterogeneity.[8–11] This is anticipated to transform the diagnosis and treatment of genetic diseases from phenotype-driven, and genotype-assisted, to genotype-driven and phenotype-assisted.[12]

The imminence of genomic medicine has been substantially hastened by inexpensive sequencing of exomes (all protein coding exons) and targeted gene panels.[5,6,8,10,17,20–22] Exomes are about ten-fold less costly than whole genomes. Targeted gene panels, in turn, are less costly than whole exomes. In addition, their interpretation and, thus, actionability are much simpler. Besides the discovery and clinical testing of genetic disease and pharmacoligically relevant genes, these technologies are also expanding the applicability of sequence analysis. Examples include oligogenotype-phenotype relationships, such as epistasis, and ascertainment of the breadth of clinical and genetic heterogeneity in diseases.

The Journal of Genomes and Exomes seeks to assist in the implementation of genomic medicine by scalable reporting of high quality genome, exome, and gene panel sequences with attendant, detailed phenotypes. Through such reports, the Journal seeks to be an international forum for community-based confirmation or rebuttal of preliminary genotype-phenotype relationships by requiring the submission of supplementary, structured information in a flat file format. Herein we have described the initial structured templates for submission of such information, the rationale for these templates and examples. The Journal of Genomes and Exomes is keen to receive feedback regarding these structured templates and examples. This is intended to be a responsive community resource. The greater the number of high quality exomes and genomes we publish, the more valuable this resource for discoveries and refinements in genomic medicine will be.

Author Contributions

Conceived and designed the experiments: DLD, SFK. Developed programs to analyse data: NAM, SES. Generated sequence data: DLD, EGF. Analysed the data: DLD, CJS, EGF, SES, NAM. Wrote the first draft of the manuscript: SFK. Contributed to the writing of the manuscript: DLD. Agree with manuscript results and conclusions: DLD, CJS, EGF, SES, NAM, SFK. Made critical revisions and approved final version: DLD, SFK. All authors reviewed and approved of the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

DLD has received fees for speaking and travel funding from Illumina. Other authors disclose no competing interests.

Acknowledgements

A deo lumen, ab amicis auxilium. We thank the reviewers for their assistance in describing structured templates.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Peltonen L. The molecular dissection of human diseases after the human genome project. *Pharmacogenomics J*. 2001;1(1):5–6.
2. Dar-Nimrod I, Heine SJ. Genetic essentialism: On the deceptive determinism of DNA. *Psychol Bull*. 2011;137(5):800–18.
3. Kluger J. Too many one-night stands? Blame your genes. *Time*. 2010. Available from: http://healthland.time.com/2010/12/02/too-many-one-night-stands-blame-zyour-genes/. Accessed May 8, 2012.
4. Fischer BA, Zigmond MJ. The essential nature of sharing in science. *Sci Eng Ethics*. 2010;16(4):783–99.
5. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*. 2007;2(3):e308.

6. Ecclesiastes 4:12, Holy Bible.
7. Friemer N, Sabatti C. The human phenome project. *Nature Genetics.* 2003;34(1):15–21.
8. Pathak J, Wang J, Kashyap S, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc*. Jul–Aug 2011; 18(4):376–86.
9. Kingsmore SF. A new journal and a new model for structured data dissemination for an era of genomic medicine. *J Genome Exome*. 2012;1:1–5.
10. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf.
11. Miller NA, Soden SE, Saunders CJ, et al. STATseq: Rapid whole genome sequencing for monogenic disease diagnosis in neonatal intensive care units. (Submitted)
12. Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next generation sequencing. *Sci Transl Med*. 2011;3(65):65ra4.
13. Ajay SS, Parker SC, Abaan HO, Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. *Genome Res*. 2011;21(9): 1498–505.
14. den Dunnen JT, Antonarakis SE. Mutation Nomenclature Extensions and Suggestions to Describe Complex Mutations: A Discussion. *Hum Mutat*. 2000;15(1):7–12.
15. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*. 2011;39(Database issue): D514–9.
16. Richards CS, Bale S, Bellissimo DB, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. *Genet Med*. 2008;10(4):294–300.
17. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, Cooper DN. The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics*. 2009;4,(2):69–72.
18. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics.* 2010;26(16):2069–70.
19. Flicek P, Amode MR, Barrell D, et al. Ensembl 2012. *Nucleic Acids Res*. 2012;40(Database issue):D84–90.
20. Dreszer TR, Karolchik D, Zweig AS, et al. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res*. 2012; 40(Database issue):D918–23.
21. Cotton RG, Scriver CR. Proof of "disease causing" mutation. *Hum Mutat*. 1998;12(1):1–3.
22. Clark MJ, Chen R, Lam HY, et al. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol*. Sep 25, 2011;29(10):908–14.
23. Asan, Xu Y, Jiang H, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol.* Sep 28, 2011; 12(9):R95.
24. Sulonen AM, Ellonen P, Almusa H, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol*. Sep 28, 2011;12(9):R94.
25. Zhang G, Beck BB, Luo W, Wu F, Kingsmore SF, Dai D. Development of a phylogenetic tree model to investigate the role of genetic mutations in endometrial tumors. *Oncol Rep*. May 2011;25(5):1447–54.
26. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*. Aug 18, 2011;12(9): 628–40.
27. Köhler S, Doelken SC, Rath A, Aymé S, Robinson PN. Ontological phenotype standards for neurogenetics. *Hum Mutat*. Sep 2012;33(9):1333–9.
28. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*. 2009;85(4):457–64.
29. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*. 2011;32(4):358–68.
30. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum Mutat*. Jan 2013;34(1):57–65.
31. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. Nov 1, 2012;491(7422):56–65.
32. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. Jul 6, 2012;337(6090):64–9.
33. www.ncbi.nlm.nih.gov/clinvar/.
34. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat*. May 2011;32(5):557–63.
35. Bale S, Devisscher M, Van Criekinge W, et al. MutaDATABASE: a centralized and standardized DNA variation database. *Nat Biotechnol.* 2011;29(2):117–18.
36. Green ED, Guyer MS, National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature*. Feb 10, 2011;470(7333):204–13.
37. Collins FS. No longer just looking under the lamppost. *Am J Hum Genet*. Sep 2006;79(3):421–6.