Advancing the Clinical Usefulness of Next-Generation Sequencing: Addressing the Need for Genomic Data Standards

Ira M. Lubin, PhD, FACMG and Edward R. Lockhart, PhD on behalf of the Clinical Grade Variant File Workgroup*

Centers for Disease Control and Prevention, Atlanta, GA

Members of the Clinical-Grade Variant File Workgroup: Nazneen Aziz[a], Larry Babb[b], Dennis Ballinger[c], Himani Bisht[d], Deanna M. Church[e], Shaun Cordes[f], Alden A. Dima[g], Karen Eilbeck[h], Timothy Fennell[i], J. Bradley Holmes[g], Fiona Hyland[g], Lisa Kalman[j], Melissa Landrum[g], Edward R. Lockhart[g], Ira M. Lubin[g], Donna Maglott[g], Elliott Margulies[k], Gabor Marth[l], Aki Nakao[g], John Pfeifer[g], Mala Ramaiah[g], Heidi Rehm[g,s], Somak Roy[g], Marc L. Salit[d], Chris Saunders[g], Stephen Sherry[g], Zivana Tezak[g], Rebecca Truty[g], Mollie Ullman-Cullere[g], Karl Voelkerding[g], Elizabeth Worthey[g], Alexander Wait Zaranek[d], and Justin Zook[d]. [a]Phoenix Children's Hospital, Phoenix, Arizona; [b]Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Boston, MA; [c]Complete Genomics, Mountain View, California; [d]Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD; [e]Personalis, Menlo Park, California; [f]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD; [g]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT; [h]Broad Institute, Boston, MA; [i]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD; [j]Life Technologies, Carlsbad, CA; [k]Division of Laboratory Programs, Standards, and Services, Centers for Disease Control and Prevention, Atlanta, Ga; [l]Illumina Cambridge, Saffron Walden, UK; [m]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO; [n]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Cambridge, MA; [o]Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA; [p]Illumina, San Diego, CA; [q]Clinical and Translational Informatics, Dana-Farber Cancer Institute, Boston, MA; [r]Department of Pathology, University of Utah and the Institute for Clinical and Experimental Pathology, Associated Regional and University Pathologists (ARUP) Laboratories, Salt Lake City, Utah; [s]Department of Pediatrics, Medical College of Wisconsin; [t]Personal Genome Project, Harvard Medical School, Boston, MA

Introduction

Systems interoperability that supports genomic applications does not exist because of the absence of a framework for data exchange. To address this challenge, the Centers for Disease Control and Prevention, working with federal partners (National Center for Biotechnology Information, National Institute of Standards and Technology, and the Food and Drug Administration) the HL7 Clinical Genomics Workgroup, and others, established the Clinical Grade Variant File Workgroup to address this shortcoming. This workgroup is proposing a "clinical-grade" variant template that is designed to describe the variant set generated during NGS testing prior to clinical assessment and the clinically relevant findings to support clinical applications. The workgroup identified principles and made recommendations for a constrained dataset that promotes consistency and minimizes ambiguity in the representation of genomic data.

The Clinical Grade Variant Template

Section 1 – General Information

Ref: https://wiki.nci.nih.gov/display/TCGA/TCGA+Variant+Call+Format+%28VCF%29+Specification

1. Patient information[a]
2. Indication for Test[b]
3. Test Name[c]
4. Specimen[d]
5. Method[e]
6. Referring physician and facility
7. Laboratory performing test
8. Date of Initial Test
9. Date of Initial Report
10. Summary of re-analysis[f]

Section 2 -- Clinically Relevant Findings

1. Gene/ Variant/ Haplotype/Amino acid change(s)[g]
2. Classification with supporting references[h]
3. Quality metrics[i]
4. Limitations[j]
5. Re-analyzed data[k]

Section 3 – Sequence Dataset (prior to clinical assessment)

(The Variant Call Format (VCF) is the recommended specification for storing sequence data)

```
##fileformat=VCFv4.1
##fileDate=20090805
##fcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFGeneAnno=<annol.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NORMAL                  TUMOR
20     14370   rs6054257 G   A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51  1|0:48:8:51,51
20     17330   .         T   A      3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50  0|1:3:5:65,3
20     1110696 rs6040355 A   G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;DB       GT:GQ:DP:HQ 1|2:21:6:23,27  2|1:2:0:18,2
```

Fixed fields | INFO meta-information | FILTER meta-information | Optional: FORMAT field specifying data type + Per-sample genotype data

Figures 1. This template is proposed to support systems interoperability by defining and constraining a set of data fields described in three sections:

Section 1 – General Information

- [a]Patient information (name, date of birth, sex)
- [b]Indication for testing (with associated ICD code)
- [c]Test name (Standard, common, LOINC)
- [d]Specimen (Type /site of origin, germline / somatic)
- [e]Method (Platform, Software, etc.), references)
- [f]Summary of Re-analysis:, when applicable

Section 2 – Clinically Relevant Findings

- [g]Variant/gene and/or haplotype, amino acid changes: (HGNC/HGVS descriptions + common names)
- [H]Pathogenicity classification and supporting data
- [i]Quality metrics (Primarily of clinical relevance)
- [j]Limitations (e.g. evidence of data)
- [k]when new knowledge or different indication of testing presents

Section 3 – Sequence Dataset (prior to clinical assessment)

(Haplotype representation: calls, no-calls, and local phasing)

- [l]Chromosome number or reference to an alternate assembly derived from a GRC reference assembly, when applicable
- [m]Position of nucleotide within a chromosome mapped against a GRC versioned reference assembly (or other standard)
- [n]Identifiers of the represented variant (e.g., dbSNP)
- [o]Reference base of the genome assembly
- [p]Alternate base; non-reference allele
- [q]Quality score for assertion made in the ALT field
- [r]Filter_status; lists whether the site/position has passed filters
- [s]Additional genomic information (i.e. genotype, base call quality, read depth, etc.)

Data Standards are Applied During Testing

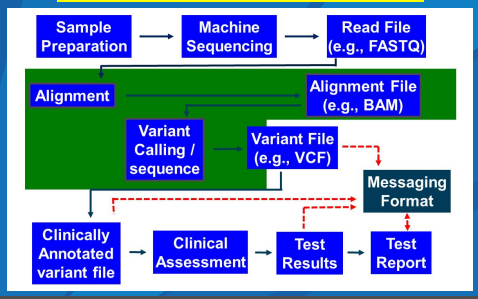

Figure 2. Sequence alignment and variant calling are steps during the informatics analysis that require attention to assure consistent position assignments and descriptions of sequences generated during NGS analysis. The same applies to clinical annotation and assessment but these topics were not addressed by this workgroup.

Clinical Applications Supported by Genomic Data

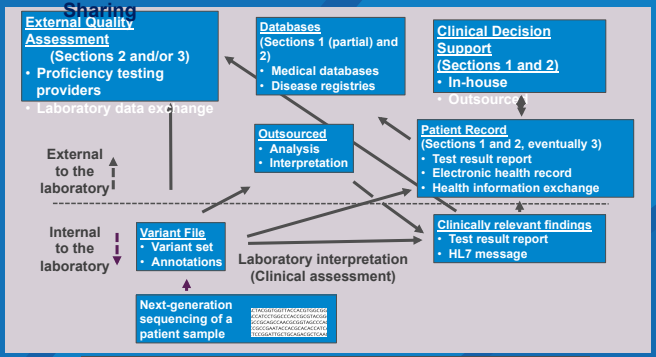

Figure 3. Clinical applications requiring the sharing of genomic data. Sections 1,2, and or 3 of the clinical-grade variant template can be combined and shared as applicable to specific clinical applications. This selected combination is proposed to inform the content of an HL7 message.

Use Case Example: Where Information Sharing is

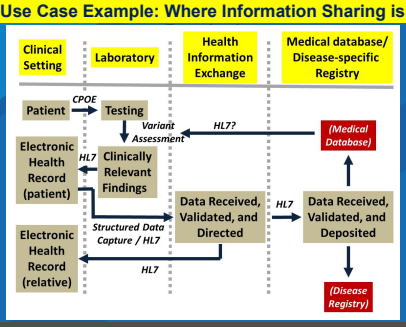

Figure 4. Use case example. The clinical-grade variant template is proposed to inform the development of the HL7 message. Structured data capture protocols are useful for normalizing data among varied practice settings. This example provides a model for communicating genomic data to a relative's medical setting at a distant location from the index patient, deposition of genomic data into a medical database or disease registry, and extraction of genomic data from a medical database to inform a clinical assessment.

Summary

1. Standards for genomic data representation are required to minimize ambiguity in the description of sequencing findings generated from clinical next-generation sequencing.

2. The clinical-grade variant template identifies a constrained genomic dataset that describes the sequence generated from NGS before and after analysis for clinical relevance.

3. The application of standard conventions requires that certain laboratory methods be adopted, such as alignment against a versioned reference assembly.

4. The clinical-grade variant template was developed to inform approaches for genomic data integration into health IT systems using standard messaging conventions (e.g., HL7). These outcomes will support systems interoperability that is essential for accurate and reliable use of genomic data in diverse healthcare settings.

Center for Surveillance, Epidemiology, and Laboratory Services

Division of Laboratory Programs, Standards, and Services

For additional information, please contact:
Ira Lubin, PhD, FACMG at ilubin@cdc.gov