Using GVF for Clinical Annotation of Personal Genomes

Barry Moore [1,2], Shawn Rynearson [2], Fiona Cunningham [3], Graham Ritchie [3], Karen Eilbeck [2].

1. Department of Human Genetics, University of Utah, Salt Lake City, Utah USA.
2. Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah USA.
3. Ensembl Variation Group, EMBL-EBI, Genome Campus, Hinxton, UK.

Abstract

Accurately describing the contents of Next Generation Sequencing (NGS) results is vital to both research and clinical analysis of genomic data. Genomics and medicine use different, often incompatible terminologies and standards to describe sequence variants and their functional effects. This creates an information bottleneck that prevents efficient translation of genome scale next-generation sequence (NGS) information into the clinic. While the Variant Call Format (VCF) has met some of these challenges, with regards to describing the results of variant calling pipelines, it lacks the structure needed for detailed annotation of the consequences of sequence alterations. To incorporate genomic results into electronic health records (EHR), the results must also be defined in ways that are compatible with existing medical informatics systems. The Genome Variation Format (GVF) is an extension of the existing genome annotation format GFF3, which uses ontologies to capture the semantic nature of the information on sequence features. GVF uses the Sequence Ontology (SO) to define the type of sequence alteration, the genomic features that are changed and the effect of the change. We have extended and remodeled the Sequence Ontology to include and define more terms that describe the consequence of a variant upon genomic features in support of the Ensemble variation databases. GVF represents genome annotations for clinical applications using existing EHR standards as defined by the international standards consortium: Health Level 7. This means that GVF can describe the information that defines genetic tests, allowing seamless incorporation of genomic data into pre-existing EHR systems. Here we demonstrate the power of GVF to describe, to exchange, and to empower clinical interpretation of personal genome data through an extension of the GVF specification is called GVFClin. The Sequence Ontology Project maintains and updates the specification and provides the underlying structure that describes sequence features, sequence alterations and variant effects and their relationships to each other. The specification is available on the web at http://www.sequenceontology.org/resources/gvfclin.html.

Background

Next generation sequencing (NGS) technologies have provided an enormous expansion in our understanding of the landscape of genetic variation [1, 2] as well as the impact of that variation on human health [3–5]. These datasets create a significant burden in computational analysis and data storage, but established work-flows for analysis are emerging [6] and well established data formats exist for each stage of the process. The original base calls from the sequencer are converted to FASTQ files [7] that contain the sequence data; the SAM format [8] captures the alignment of the sequence to a reference genome and the Variant Call Format [9] has become widely adopted by variant calling tools to report variants and the information needed to call them. However, knowing the type and genomic location of a sequence change is just the first step in understanding it's clinical or biological consequences. Variant annotation then begins the process of adding additional knowledge about the structural and functional consequences of those variants through their impacts on other sequences features and ultimately on phenotype.

In the context of medicine, variant data must flow smoothly and reliably from the sequencer to the physician and formidable barriers currently exist to this flow of information. Significant efforts have been undertaken to standardize the description of genetic variants [10, 11] and the HGVS nomenclature has done much to unify the notation of clinical variants in the literature. NGS

sequencing is providing a wealth of new information about the types of genetic variants that exist [12] and the types of features that those variants impact [13] and thus *ad hoc* descriptions of variants and their effects persist. There is a need for a file format that provides the structure necessary not only to describe sequence variants, but also to bridge the gap between genomics and medicine by providing the structure necessary to capture clinically relevant variant data in a format compatible with EHR standards.

The Genome Variation Format (GVF) [14] is a variant file format for the detailed annotation of genetic variation. GVF is a community supported format that uses established ontologies such as the Sequence Ontology [15] to describe the variant data. GVF does not replace existing variant nomenclature systems such as HGVS [16] and ISCN [17] that provide effective ways to unambiguously describe individual variants in the literature. GVF provides the infrastructure to support inclusion of these nomenclatures along with detailed variant annotation in a format capable of supporting genome scale variant data. GVF is used in the community for exchange of variant annotations (Ensembl: ftp://ftp.ensembl.org/pub/release-67/variation/gvf/ and dbVar: ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_assembly/NCBI36/gvf/) and is compatible with existing GFF3 software [18, 19] as well as emerging domain specific tools [20, 21].
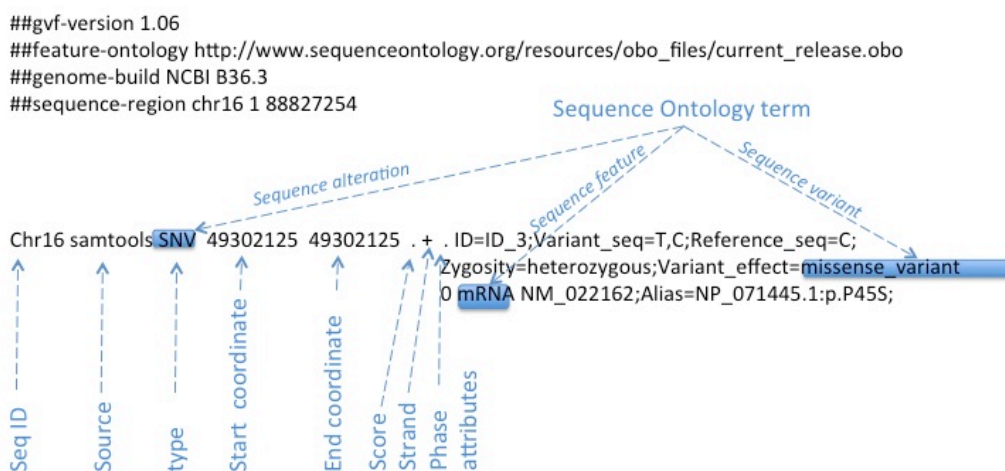

Figure 1. An example GVF file showing a single variant annotation

Implementation
Structurally, GVF is a text based, tab-delimited format modeled on the existing and widely adopted Generic Feature Format - GFF3. GVF describes both file-wide meta data through the use of pragmas, as well as detailed information about individual variants. A very simple GVF file with a single variant is shown in Figure 1. The first 4 lines contain file wide directives while the fifth line describes an SNV on chromosome 16. Here, the reference sequence is C, with a heterozygous variant genotype of T,C. The variant causes a *missense_variant* in the intersecting mRNA. The SO is used three times to type this annotation; the sequence alteration (SNV), the effect (nonsynonymous_codon) and the intersected feature (mRNA).

Describing sequence alterations and their consequences
Sequence alterations are the changes observed in biological sequence when compared to a reference genome. In the SO there are 30 kinds of *sequence _alteration* ranging from very general types such as *substitution* to very specific types such as *purine_to_pyrimidine_transversion*. The relationships defined by the ontology allow users and software to infer more general terms from more specific ones - for example a *purine_to_pyrimidine_transversion* is a *SNV*. Figure 2 shows a summary of the most general *sequence_alteration* terms. The consequence of such alterations fall into two classes: functional variants and structural variants. Annotating functional variants requires a deep

understanding of the underlying biology, while annotations of structural variants can typically be inferred from the reference genome and other genomic features that intersect the variant. The Ensembl Variation group has worked with the SO to produce a classification of the variants found in their database that will allow their users to effectively search variants and their effects.
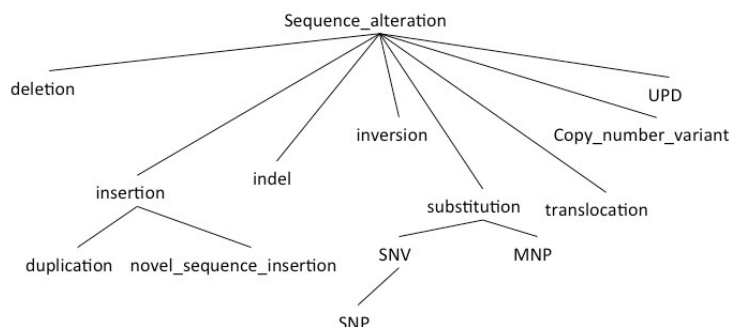

Figure 2. The main subtypes of the term sequence_alteration from the Sequence Ontology.

In the SO the sequence alteration and the effects of the alteration are separated in the ontology, and in the annotation. For example, historically a small deletion may be referred to as a microindel, where as a much larger deletion might be described as a copy number variant (CNV). In the SO however a single term, *deletion*, is used to describe all instances where a region of sequence is removed. The kinds of *sequence_alteration* are shown in Figure 2. The effect of the *deletion* on the structure of the genome is either a kind of *feature_variant* where by the internals of the feature such as an exon are changed, or a *feature_ablation* where a region comprising one or more features is removed. Thus the effect of a small deletion is annotated using the appropriate child of *feature_variant* such as *frameshift_truncation* and the effect of large deletions are annotated with the appropriate child term of *feature_ablation* such as *transcript_ablation*. *Sequence_variant* and child terms that categorize the effect of a sequence alteration are depicted in Figure 3.
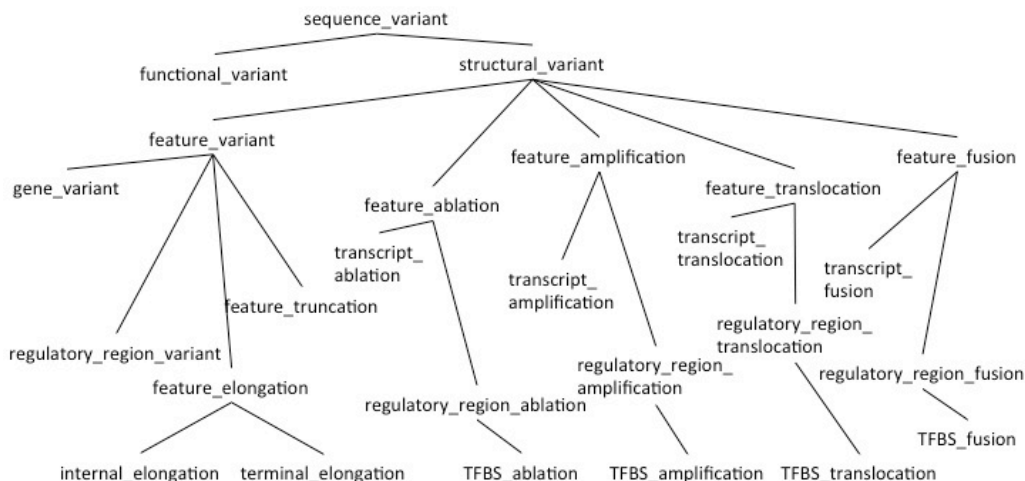

Figure 3. The subtypes of the term sequence_variant in the Sequence Ontology highlighting the kinds of structural_variants. Structural variants are sequence alterations that change the structure of one or more features. Not all subtypes are shown

The majority of the sequence alterations annotated by the EBI group cause *feature_variants*. These feature variants are shown in Figure 4, where the terms used in EBI annotations are highlighted in blue. There are four main subtypes: *upstream_gene_variant*, *downstream_gene_variant*, *gene_variant* and *regulatory_region_variant*. Of these terms, *gene_variant* has 77 direct and indirect subtypes and includes most of the terms that describe structural sequence variants caused by substitutions and small inserts and deletions. This portion of the Sequence Ontology contains terms with multiple parents, to allow for effective querying of the annotations. For example, the term

stop_retained_variant is both a *synonymous_variant* and a *terminator_codon_variant*. Users are thus able to query the Ensembl databases for all terminator codon variants or all synonymous variants. Variant genome annotations for 19 organisms, typed using SO and available in GVF are available within the Ensembl databases (http://www.ensembl.org/) and for download (ftp://ftp.ensembl.org/pub/release-67/variation/gvf/).
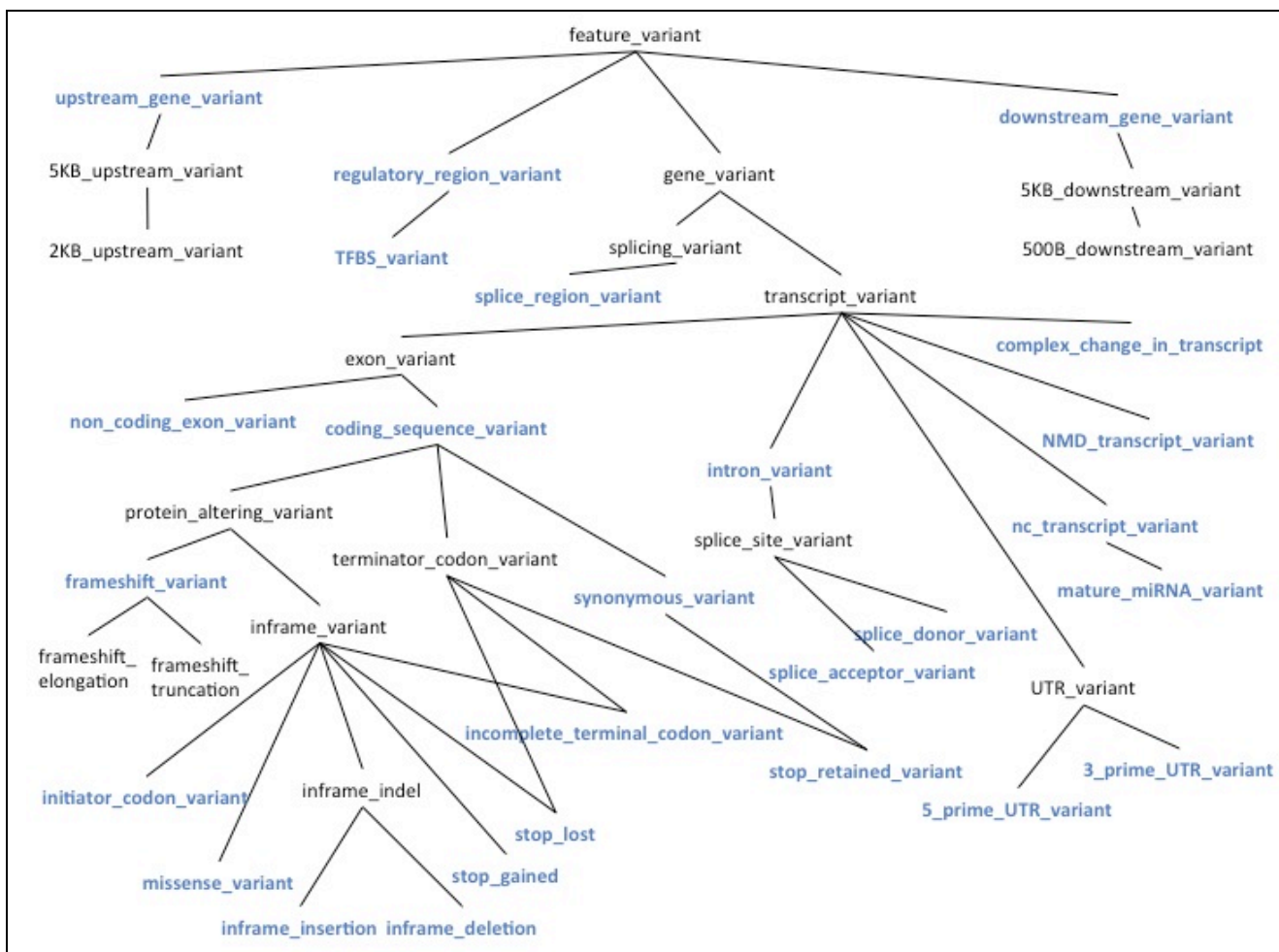

Figure 4. Subtypes of the Sequence Ontology term feature_variant. Feature variants result from sequence alterations that change the of sequence features such as exons and regulatory regions. The terms used in Ensembl annotations are highlighted in blue and defined at http://www.ensembl.org/info/docs/variation/predicted_data.html.

Electronic health record compliant data with GVFClin

GVF was initially developed for exchange of variant annotations in personal genomes. To empower clinical use of personal genomic data we have specified the format to adhere to existing EHR standards defined by the HL7 (http://www.hl7.org) clinical genomics working group including LOINC® (http://www.loinc.org), and the SNOMED [22], RxNorm [23] and HGVS [24] vocabularies and nomenclature. Use of Locus Reference Genomic (LRG) [25] sequences provide a stable genomic sequence reference set within these standards which stabilizes the description of variants relative to permanent sequence feature coordinates. We have added 14 additional attributes (Table 1) to support annotation of clinical variants and refer to this extension of the standard as GVFClin. The extensions which define a GVFClin document may be found online at http://www.sequenceontology.org/resources/gvfclin.html.

Table 1. Examples of attributes added to the GVFClin specification for the clinical annotation of variants.

Attribute	Definition and example
Clin_gene	**HGNC** [26] gene identifier set by the Human Genome Organization Nomenclature Committee. `Clin_gene=HEXA;`
Clin_genomic_reference	The **NCBI RefSeq** accession.version for the genomic sequence that the sequence_alteration is located. This is an NC_ style sequence ID for a chromosome. `Clin_genomic_reference=NC_000001.9;`
Clin_transcript	An **NCBI RefSeq** or **LRG** transcript accession.version. LRG accessions do not contain versions. `Clin_transcript=NM_003482.3;`
Clin_variant_id	A **dbSNP** rs number for the sequence_alteration. `Clin_variant_id=rs35731153;`
Clin_HGVS_DNA	A valid **HGVS** nomenclature description of the sequence_alteration at the DNA level. `Clin_HGVS_DNA=NC_000016.9:g.23360165C>G;`
Clin_HGVS_protein	A valid **HGVS** nomenclature description of the sequence_alteration at the DNA level. `Clin_HGVS_protein=NP_001128727.1:p.Val209Ile;`
Clin_disease_interpret	An interpretation of the pathogenicity of a the given sequence_alteration with regards to the assessed disease. With values constrained by the answer list associated with the **LOINC** code 53037-8. Positive, Negative, Inconclusive, Failure `Clin_disease_interpret=Positive;`
Clin_drug_metabolism_interpret	An interpretation of the metabolism rate due to a given sequence_alteration with regards to the assessed drug. With values constrained by the answer list associated with this **LOINC** code 53040-2. Ultrarapid metabolizer, Extensive metabolizer, Intermediate metabolizer, Poor metabolizer `Clin_disease_interpret=Ultrarapid metabolizer;`
Clin_drug_efficacy_interpret	An interpretation of the efficacy of a drug, due to the sequence_alteration. With values constrained by the answer list associated with this **LOINC** code 51961-1. Resistant, Responsive, Presumed resistant, Presumed responsive, Unknown Significance, Benign, Presumed Benign, Presumed non-responsive `Clin_drug_efficacy_interpret=non-responsive;`

Conclusions

Next generation sequencing technologies have provided unprecedented opportunities for low cost and large-scale analysis of human genetic variation and it's consequences. The ability of the emerging field of personal genomics to provide genome wide information on genetic variation for an individual promises more accurate and effective health care. The ability to deliver on this promise is currently hampered by the inability of existing formats to annotate genome scale genetic variation data in a way that is compatible with EHRs. The Genome Variation Format builds on an established genome annotation standard, with additional structure for describing the genetic variation in personal genomes.

GVFClin provides and additional layer of constraints designed to make compliant documents readily interpretable in a clinical context compatible with EHR standards. The terms used by GVF and GVFClin to describe sequence alterations, their effects and the affected sequence features are constrained by the Sequence Ontology through an open community supported process. The Vertebrate Genomics group at the EBI and the dbVar group at the NCBI have adopted GVF for

distribution of genetic variation data. In addition to the existing software tools that support GFF3 format (and thus by extension support the fully compatible GVF format), domain specific software tools have been published which natively support GVF files.

Acknowledgements

References

1. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061–73.

2. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers C a, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo M a, Banks E, Hu M, Handsaker RE, Rosenfeld J a, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner M-M, Hunt T, Barnes IH a, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs R a, McCarroll S a, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C: **A systematic survey of loss-of-function variants in human protein-coding genes.** *Science* 2012, **335**:823–8.

3. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nature Reviews Genetics* 2011, **12**:745–55.

4. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder.** *Nature Genetics* 2010, **42**:30–5.

5. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM, Carey JC, Opitz JM, Stevens C a, Jiang T, Schank C, Fain HD, Robison R, Dalley B, Chin S, South ST, Pysher TJ, Jorde LB, Hakonarson H, Lillehaug JR, Biesecker LG, Yandell M, Arnesen T, Lyon GJ: **Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency.** *American Journal of Human Genetics* 2011, **89**:28–43.

6. Koboldt DC, Ding L, Mardis ER, Wilson RK: **Challenges of sequencing human genomes.** *Briefings in Bioinformatics* 2010, **11**:484–98.

7. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Research* 2010, **38**:1767–1771.

8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–9.

9. Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**:2156–8.

10. Ogino S, Gulley ML, den Dunnen JT, Wilson RB: **Standard mutation nomenclature in molecular diagnostics: practical and educational challenges.** *The Journal of Molecular Diagnostics* 2007, **9**:1–6.

11. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM: **Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker.** *Human Mutation* 2008, **29**:6–13.

12. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L: **Challenges and standards in integrating surveys of structural variation.** *Nature Genetics* 2007, **39**:S7–15.

13. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, Crawford GE: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biology* 2011, **9**:e1001046.

14. Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K: **A standard variation file format for human genome sequences.** *Genome Biology* 2010, **11**:R88.

15. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biology* 2005, **6**:R44.

16. Cotton RGH, Horaitis O: **Human Genome Variation Society.** In *Nature Encyclopedia of the Human Genome*. London: Nature Publishing Group; 2003:361–362.

17. *ISCN 2009: An International System for Human Cytogenetic Nomenclature*. Basel: Karger AG; 2009:138.

18. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Research* 2002, **12**:1611–8.

19. **Generic Model Organism Database**.

20. Song T, Hwang K-B, Hsing M, Lee K, Bohn J, Kong SW: **gSearch: a fast and flexible general search tool for whole-genome sequencing.** *Bioinformatics* 2012.

21. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG: **A probabilistic disease-gene finder for personal genomes.** *Genome Research* 2011, **21**:1529–42.

22. Stearns MQ, Price C, Spackman K a, Wang a Y: **SNOMED clinical terms: overview of the development process and project status.** *Proceedings / AMIA Annual Symposium* 2001:662–6.

23. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R: **Normalized names for clinical drugs: RxNorm at 6 years.** *Journal of the American Medical Informatics Association* 2011, **18**:441–8.

24. Horaitis O, Cotton RGH: **The challenge of documenting mutation across the genome: the human genome variation society approach.** *Human Mutation* 2004, **23**:447–52.

25. Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Béroud C, Dobson G, Lehväslaiho H, Taschner PE, den Dunnen JT, Devereau A, Birney E, Brookes AJ, Maglott DR: **Locus Reference Genomic sequences: an improved basis for describing human DNA variants.** *Genome Medicine* 2010, **2**:24.

26. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA: **genenames.org: the HGNC resources in 2011.** *Nucleic Acids Research* 2011, **39**:D514–9.