Human Genomes and Big Data Challenges

QUANTITY, QUALITY AND QUANDRY

©2013. Gerry Higgins, M.D., Ph.D.
AssureRx Health, Inc.
### Table of Contents

**EXECUTIVE SUMMARY** ................................................................................................................................. 3

I. The Abundance and Diversity of ‘Omics’ Data ........................................................................................................ 4  
   A. The Explosive Growth of Human Genome Data ........................................................................................... 4  
   B. The ENCODE Project and its impact on Medical Genetics ........................................................................... 5  
   C. Omics ................................................................................................................................................................. 7  
   D. Challenges ....................................................................................................................................................... 7  

II. Preliminary Response – Big Data Quality Case Study in Genomics ................................................................. 8  
   A. Overview .......................................................................................................................................................... 8  
      1. Brief Description of the Project ................................................................................................................ 8  
      2. Primary Domain and Data Types ............................................................................................................. 8  
      3. The Four ‘Vs’ ............................................................................................................................................. 9  
         a. Volume .................................................................................................................................................. 9  
         b. Variety ................................................................................................................................................ 9  
         c. Velocity .............................................................................................................................................. 10  
         b. Variability ........................................................................................................................................ 10  
   B. Big Data Quality Applications ....................................................................................................................... 18  
      1. Testing ..................................................................................................................................................... 18  
      2. Data Preparation .................................................................................................................................... 18  
      3. Is Metadata being captured for the Big Data sets? ............................................................................... 19  
      4. Store – Can you store all the data, whether it is persistent or transient? ........................................... 20  
      5. Application Processing – Can the application perform within a reasonable timeframe? .................. 20  
      6. Data Operations - .................................................................................................................................... 21  
      7. Query – Can you search the data? ............................................................................................................ 21  
   C. Big Data Quality Questions ............................................................................................................................ 22  
      1. What type of data validation is currently being performed? .................................................................. 22  
         a. Phred Score ......................................................................................................................................... 22  
         b. Concordance ..................................................................................................................................... 22  
      2. Do you feel that the level of validation is adequate? .............................................................................. 30  
      A. An Interim Solution for Clinical Genomics ............................................................................................. 30  
      B. Flowchart ................................................................................................................................................ 31  
      C. Examples of Quality Control Metric Results .......................................................................................... 34  
      3. Are you experiencing Big Data Quality Problems? ............................................................................... 36  
      4. How is the extent of any of the Data Quality problems determined? .................................................. 38
5. How are Data Quality Problems managed? ................................................................. 39
6. Are samples of Data Quality Problems available? .................................................... 39
D. Big Data Quality Tool Questions .............................................................................. 40
   1. Do you use specialized tools? .................................................................................. 40
   2. Do you have specialized tools? ............................................................................... 40
   3. Is the decision-making context for Big Data a point solution? ............................... 40
   4. Is the decision-making context for use within the organization’s workflow? .......... 40
E. Data Quality Maturity ................................................................................................. 41
   1. Is the framework’s documentation available? ....................................................... 41
   2. Does the framework promote business practices? .................................................. 41
   3. Are testing practices available for use? ................................................................... 41
   4. Are there visualization tools? .................................................................................. 41
   5. Is the framework tailor able? .................................................................................. 41
   6. Does the framework offer enterprise solutions? .................................................... 42
   7. Does the framework offer Big Data expertise? ....................................................... 42
REFERENCES AND ADDITIONAL NOTES ........................................................................ 43
APPENDIX A: OVERVIEW OF SEQUENCING METHODS ............................................... 47
APPENDIX B: TYPES AND NUMBERS OF VARIANTS FOUND ..................................... 65
APPENDIX C: OVERVIEW OF SELECTED BIOINFOMATICS SOFTWARE ......................... 74
APPENDIX D: CHALLENGES: MUTATION NOMENCLATURE AND ETHICAL ISSUES IN THE CLINIC ................................................................. 87
GLOSSARY ...................................................................................................................... 91
EXECUTIVE SUMMARY

This draft report addresses data quality issues associated with a massive dataset of whole human genome sequences. This dataset is the world’s largest collection of whole human sequences, and includes raw ‘short reads’, aligned, and assembled sequences generated by second generation sequencing, as well as annotated variant types. Annotation is an ongoing process, as the pace of discovery in clinical genomics is extremely rapid.

As of early 2013, there have been regulations and guidelines issued recently for how to use next generation sequencing (NGS) in a laboratory that is certified by the Clinical Laboratory Improvements Amendments (CLIA) by the Centers for Medicare and Medicaid Services (CMS). The CLIA-certificated laboratory is the only setting in which such sequences may be generated for clinical applications. Although the recent regulations are informative, the rate of technology acceleration and accompanying discovery in biology and medicine is so fast that no regulatory or professional organizations can hope to keep abreast of such changes using current, highly bureaucratized processes.

There are many quality concerns associated with this massive dataset. These include differences in quality between the NGS platforms that contributed source data, inherent problems associated with short read, 2nd generation sequencing, and issues related to clinical interpretation. It should be understood that there is continuing controversy in genomics concerning metadata standards, mutation nomenclature and quality control metrics, which partly reflects the divide between researchers in the life sciences and those that perform clinical genomic sequencing.

Following a short introductory section, the answers to the ‘Big Data Quality’ questionnaire begin on page 8 of this draft report. It is recommended that the reader pay special attention to the ‘Flowchart’ and related text that begins on page 30, as this is the process that was used for analysis of this specific dataset. Sample preparation using enzymatic amplification always introduces error, as does the use of a human reference sequence, from the National Center for Biotechnology Information (NCBI), National Institutes of Health, for alignment that is based largely on whole genome sequences from Caucasians. However, there are many additional data quality problems, and yet the NGS technology is so powerful that it has led to the greatest number of biomedical advances in the history of medicine.

This report is only the beginning of a process that will hopefully resolve any continuing questions that remain unresolved. Genomics is a domain that is constantly changing, in terms of data quality, variant interpretation and methodology. The next few years will usher in a new era of more accurate sequencing that will focus less on pre-processing informatics, and more emphasis will be placed on engineering issues associated with signal analysis. This shift will transform bioinformatics, and will be equal in impact in biomedicine as were the Coulter counter and the polymerase chain reaction.

Several appendices are provided so that the reader may better understand the complexity of NGS methods and bioinformatics tools for data analysis, as well as specification of the numbers and types of variants discovered in this dataset to date. Also discussed in Appendix D is a case study which focuses on some of the ethical issues associated with clinical genomics. Finally, a Glossary is provided.
I. The Abundance and Diversity of Omics Data

A. The Explosive Growth of Human Genome Data

It has now become a 'cliché' for genomic researchers to present an obligatory slide in their talks, obtained from the web site of the National Human Genome Research Institute, showing the plummeting cost of DNA sequencing [1]. Instead, I would like to compare the amount of data growth from 2nd generation sequencing platforms\(^1\) which produce ‘short read’ sequences, with the growth of other data sources in the realm of ‘big data’. Figure 1 shows a comparison of:

- World human population growth in numbers [2].
- The ‘short read’ archive (SRA) generated from 2nd generation NGS [3].
- Moore’s Law – number of transistors per chip [4].
- Number of whole human genomes sequenced (WGS) [5].
- The number of entries in Genbank [6].
- Global internet traffic [7].
- Sensor data, U.S. unmanned aerial vehicle (UAV)’s, including Global Hawk and Predator [8].

![Figure 1. Approximate growth of different data populations.](image)

Note that the ‘short read archive’ (SRA) has grown faster than what would be predicted by Moore’s law. NGS began with a ‘bang’ in mid-2005 when Jonathan Rothberg’s company, 454, assembled a group of cutting-edge technologies and commercialized the first next-generation sequencer. Curves as indicated by colored text above the graph and as annotated, forecasts in growth indicated by the shaded region of the graph.

\(^1\) For detailed analysis of next generation sequencing methods and technologies, please refer to Appendix A.
B. The ENCODE Project and its impact on Medical Genetics

The domain of human genomics and related molecular data\(^2\) is changing very rapidly, as a consequence of 2 variables:

1. More detailed examination of 1% the human genome, made possible through the coordinated efforts of the investigators working on the ENCODE (Encyclopedia of DNA Elements) project; and
2. The accelerated pace of technology development used for next generation sequencing (NGS), so that these methods can be used for sequencing of other molecular species, including modified genomic variants, RNA, and chromatin as we move towards more accurate, nanopore-based, single molecule sequencing.

Over the past several years, the domain of human genomics has been subject to seismic change, coinciding with the latest release from ENCODE [9]. Among the results arising from a small portion of the human genome, as observed in human cell lines, the following data have turned the specialty of medical genetics on its head. These include, but are not limited to, what is shown in Table 1.

Researchers in the ENCODE project applied a variety of existing and innovative technologies to study the human genome, including:

- **RNA-Seq**: High-throughput sequencing of different fractions of RNA.
- **CAGE**: Used for mapping methylated caps at the 5’ end of RNAs (initiation of transcription).
- **RNA-PET**: Used for mapping full-length RNA [5’ cap + poly (A)].
- **Mass spectrometry**: Used to measure the mass of molecules.
- **ChIP-Seq**: For determination of which of the regions in the genome most often bound by the chromatin-bound protein to which the antibody was directed.
- **DNase-Seq**: Determination of sites ‘hypersensitive’ to DNase I, corresponding to open chromatin.
- **FAIRE-Seq** (Formaldehyde Assisted Isolation of Regulatory Elements): For isolation of nucleosome-depleted genomic regions by exploiting the difference in crosslinking efficiency between nucleosomes (high) and sequence-specific regulatory factors (low).
- **Histone ChIP-Seq**: The application of ChIP-seq to histones.
- **MNase-Seq**: Used to distinguish nucleosome positioning based on the ability of nucleosomes to protect associated DNA from digestion by micrococcal nuclease.
- **RRBS** assay (Reduced Representation Bisulfide Sequencing): Used to determine the methylation status of individual cytosines in a quantitative manner.

Two of the most exciting conclusions from ENCODE would be obvious to an earlier generation of molecular biologists:

1. The genome operates in three-dimensional (3D) space, not the linear representations we have been accustomed to using in bioinformatic analysis.

2. The human genome contains vast numbers of transposons, also called ‘jumping genes.’

\(^2\) Abbreviations and definitions can be found in the Glossary; detailed methodological information can be found in Appendix A.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Classical Medical Genetics</th>
<th>What we know now from ENCODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the unit of heredity?</td>
<td>The gene</td>
<td>The genomic functional element, including distal, cis-acting regulatory sequences.</td>
</tr>
<tr>
<td>How much of human genome is transcribed into RNA?</td>
<td>~2 - 3%, when you add up protein-coding mRNAs with other RNAs.</td>
<td>~60 - 70% is transcribed into RNA over the life of a human – this is called ‘pervasive transcription.’ Many of these non-coding RNAs serve as functional regulatory elements.</td>
</tr>
<tr>
<td>What is the primary regulatory element of a gene that initiates transcription?</td>
<td>A promoter located at the 5’ UTR region of gene, which may be activated by 1 or 2 transcription factors.</td>
<td>Cis-regulatory regions include diverse functional elements (e.g., promoters, enhancers, silencers, and insulators) that collectively modulate the magnitude, timing, and cell-specificity of gene expression. In addition, human cis-regulatory regions characteristically exhibit nuclease (DNase I) hypersensitivity, showing that these sites are ‘open’ areas of chromatin that allow regulation by multiple transcriptional factor complexes, as well as histone proteins.</td>
</tr>
<tr>
<td>Where are promoters located near a genomic functional element?</td>
<td>In the 5’ UTR domain.</td>
<td>Regulatory sequences that surround transcription start sites are symmetrically distributed, with no bias towards upstream regions. So, they can be located in 5’ and 3’ UTR regions, as well as within introns.</td>
</tr>
<tr>
<td>What is an allele?</td>
<td>An allele is an alternative form of a gene (one member of a pair) that is located at a specific position on a specific chromosome. Alleles determine distinct traits that can be passed on from parents to offspring. Allele frequencies can be calculated using the Hardy-Weinberg equilibrium (HWE) equation.</td>
<td>Many genomic functional elements exhibit what is called ‘allelic imbalance,’ that is, allele-specific binding (ASB) and allele-specific expression (ASE). Using traditional methods such as HWE would not be appropriate for these genomic functional elements. HWE also assumes random mating and no genetic drift across generations, which we know to be spurious assumptions for human subpopulations.</td>
</tr>
<tr>
<td>Where are the most pathogenic mutations located in the human genome?</td>
<td>SNPs that occur in the exons of a gene, such as a nonsynonymous or missense change, that would then mean that the gene would encode a different protein product. This is also used bioinformatics software applications such as PolyPhen, to predict which mutations are most deleterious.</td>
<td>There is no a priori reason to believe that mutations within protein coding regions (exons) would prove to be deleterious without confirmation of the negative impact of the altered protein product. Instead, SNPs in regulatory regions that comprise ~80% of the human genome have been shown to be more problematic than SNPs located in protein coding domains.</td>
</tr>
<tr>
<td>What type of mutation differentiates one individual from another?</td>
<td>Single nucleotide polymorphisms (SNPs), which account for the majority of the genetic determinants of disease risk and drug response.</td>
<td>SNPs make a significant contribution to inter-individual genetic variability, but copy number variants (CNVs) and other variants also impact disease risk and drug response.</td>
</tr>
</tbody>
</table>

Table 1. Comparison of the results of new research from the ENCODE project, compared to what was previously believed in the specialty of medical genetics. Only selected examples are shown, for details see [9].
C. Omics

The advent of NGS not only ushered in an exciting era of biologic discovery in human genomics, it also spawned new technologies that have led to a better understanding of different molecular and physiologic networks at all levels of spatial and temporal resolution. Figure 2 provides an overview of some of these domains and examples of related technologies.

![Diagram of various 'omics' domains](image)

Figure 2. Examples of some technologies used in various ‘omics’ domains. For a discussion of data standards used in ‘omics’, see Appendix B. Abbreviations- ChIP-Seq: chromatin immunoprecipitation sequencing; RNA-PET: RNA paired-end tags; CAGE: Cap Analysis Gene Expression; MALDI-imaging: Matrix-Assisted Laser Desorption/Ionization imaging mass spectrometry; CyTOF: Mass spectrometry time of flight cytometry; NMR: Nuclear Magnetic Resonance. For details, see the Glossary.

D. Challenges

The advent of NGS has alleviated a major bottleneck for genomics research, namely the speed and cost (and allied accessibility) of the physical process of generating genome sequence data. A byproduct of 2nd generation NGS has been the production of billions of ‘short reads’ from high coverage of just a single human genome. In addition, the success of NGS has created new pressures in other parts of the pipeline, such that many software tools used for bioinformatic derivation and interpretation are immature (bespoke – ‘home-made’), and for clinical application of NGS there is a lack of standards in place to ensure that maximum value is derived from the sequence data generated. Finally, there are no standards for gene mutation nomenclature that are generally accepted, no acceptable metadata in clinical genomics, and very nascent attempts to provide governance for cloud-based storage, archiving and retrieval of data.
II. Preliminary Response – Big Data Quality Case Study in Genomics

A. Overview

1. Brief Description of the Project

The project is focused on the detailed examination of the world’s largest group of whole human genome sequences generated by ‘short read’ next generation sequencing, obtained under federal contract from Complete Genomics, Illumina and Life Technologies. The dataset was obtained under IRB approval. The collection of 17,131 whole genome sequences are identified as to race, age and sex, and are undergoing a rigorous analysis whose first objective is to provide a comprehensive annotation of over 100 different variant types, including single nucleotide variants (SNPs), copy number variants (CNVs), insertions, deletions, rearrangements, enhancers, promoters, coding regions, transposons, splice sites, repeats, transcription factor binding sites, as well as known and unknown genomic functional elements. The second objective of the project, now completed, was performed in collaboration with DDR&E and Johns Hopkins Medicine. An accurate and comprehensive pharmacogenomic classification system was developed using a learning machine trained on ‘synthetic’ data created through artificial forward evolution of this dataset for de-identification purposes, as well as a large de-identified electronic health record database containing over 1 M patient records.

2. Primary domain and data types

The domain is clinical genomics involving whole genome sequencing (WGS) and whole genome analysis (WGA) involving alignment to a human reference sequence. WGS data come in several forms but largely consist of raw output from the sequencers and downstream analysis of the sequence. The raw data output from the sequencers usually comes in the form of images which can be hundreds of gigabytes in size. The raw data output from the sequencer is then processed into sequence. This sequence is aligned to the reference genome which results in information for each sequence about its alignment position and score. After alignment, a variant caller then produces files detailing the variants found in the sequence relative to the reference. In total, the resulting files include the raw sequencer output, sequence called by the software, alignments, and a consensus sequence for all calls at all positions that passed the various bioinformatic filters. Most clinical tests are concerned with being able to detect either heterozygous or homozygous variants, which in massively parallel sequencing are detected by

---

3 See Appendix B.
the software based on a combination of the quality of the raw data, the alignment quality, and the number of times which the position was sequenced.

The steps included in this project included:

1. Raw genomic DNA reads from 2nd generation platforms, including the Illumina HiSeq 2000 and Life Technologies’ 5500xl SOLiD™ machines. These instruments first captured images of many parallel reactions and ultimately yielded at least two and usually three pieces of information for each DNA sequencing step: the called base (which may be no-call), a quality score, and intensity values for all four possible bases. Collectively, this process is referred to as “primary” analysis or base calling. Once base calling has been performed, the raw image files were retained in this project. These are massive datasets, consisting of several primary data output files including image files such as *.tiff, *.csfasta, [CY3|CY3|CY5|FTC|TXR].fasta (color space reads), *.fasta, *.fastq, –QV.qual and *.stats. Every file format was converted to *.fastq.

2. After sequences were generated, they were aligned (‘mapped’) to a known human reference sequence (NCBI build 36-hg18) or GRCh37 (hg19) using BWA or SOAP, with base alignment and quality files including *.sam and *.bam.

3. Initial variant analysis was performed using SAMtools on all of the sequences from the Illumina and SOLiD platforms. Many different file formats, including *.vcf and *.gff. Variant calls from the large Complete Genomics data were output as *.var files. Every file format was converted to *.vcf.

3. The Four ‘V’s

A. Volume – How much data (number of bytes) is in the ‘Big Data’ dataset?

I have broken it down into various elements, based on type:

<table>
<thead>
<tr>
<th>NGS Stage</th>
<th>Bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data assembly with base calling at the level of individual reads.</td>
</tr>
<tr>
<td>2</td>
<td>Alignment of the assembled sequence to a reference sequence.</td>
</tr>
<tr>
<td>3</td>
<td>Variant calls.</td>
</tr>
<tr>
<td>4</td>
<td>First level of advanced annotation.</td>
</tr>
<tr>
<td>5</td>
<td>Advanced annotation for clinical applications.</td>
</tr>
</tbody>
</table>

**TOTAL SIZE OF PRIMARY DATASET ALONG WITH ANNOTATIONS ON 01/03/13** 516.14 PB

---

$^4$Petabyte: 10$^{15}$, also expressed as PiB (2$^{50}$)
$^5$Terabyte: 10$^{12}$ or TiB (2$^{40}$)
$^6$Gigabyte: 10$^9$ or GiB (2$^{30}$)
B. Variety – What are the sources of the dataset and how many are there?

The dataset consists of ‘whole’ human genome sequences, including a variety of formats, depending on NGS stage (see Table above), from 3 different companies.

Original sources:

<table>
<thead>
<tr>
<th>SEQUENCER or SERVICE</th>
<th>Illumina HiSeq 2000</th>
<th>ABI SOLiD 5500xl Life Technologies</th>
<th>Complete Genomics (sequencing-as-a-service)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER OF SEQUENCES</td>
<td>3,255</td>
<td>1,542</td>
<td>12,334</td>
</tr>
<tr>
<td>PERCENTAGE OF DATASET</td>
<td>19%</td>
<td>9%</td>
<td>72%</td>
</tr>
<tr>
<td>COVERAGE PER GENOME SEQUENCE</td>
<td>average: 45-fold</td>
<td>average: 62-fold</td>
<td>average: 58-fold</td>
</tr>
</tbody>
</table>

C. Velocity - How often is the Big Data being generated, collected/sampled, and received?

The original dataset is relatively static; however it has been updated by obtaining more raw reads and assemblies over time. The variant analysis is not static – as new findings emerge about the clinical relevance of variants, especially with regards to pharmacogenomics, new analyses are being performed every day, increasing the size of the annotation files.

D. Variability – How many different formats are involved in the Big Data collection? How complex are the formats?

There are a dozen different file formats in the dataset. The list below provides examples of the most common and most unusual file formats, only a few examples of the actual file formats are provided.

*.bam - see [10]

A BAM file (Binary Alignment/Map) is the binary version of a SAM file. BAM is compressed in the BGZF format. Many analysis tasks involve detecting small differences (substitutions and small insertions/deletions) relative to a reference human genome. For this, unassembled reads from a new genome sequence data set only need to be mapped (aligned) to the reference. Current state of the art in read mapping can map one billion reads in about 500 cpu-hours. Mapped reads are stored in a binary format called BAM, requiring 1 byte/base and a total of 100 GB for 1 billion reads. All multi-byte numbers in BAM are little-endian, regardless of the machine endianness. The format is formally described in the following table where values in brackets are the default when the corresponding information is not available; an underlined word in uppercase denotes a field in the SAM format.

*.csfasta

The CSFASTA is a variant of the *fasta file format that is a proprietary version from Life Technologies for the SOLiD 5500xl sequencer. ABI SOLiD sequencing works in color space not sequence space, leading ABI to introduce Color Space FASTA (CSFASTA) files with matching QUAL files, and also Color Space FASTQ (CSFASTQ) files. These use the digits 0–3 to encode the color calls (base transitions).
FASTA is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The definition line and sequence character format used by NCBI. All database entries from Entrez are available in this format. A sequence in Fasta format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">" ) symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length to facilitate viewing and editing. Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue).

**Example:**

```>Human
ATGGCACATGCAGCGCAAGTAGTCTACAAGACGCTACTTCCCTATATAGAAGAGTTATCACCTTTCATGATCAGGCC
CTCATATAATTTTCTTTATCTGCTCTCTATGTATGCCCTTTTCCTCAAACATCAACACAAAAACACACTAACAT
ACATCTGACGCTACAAAGGAAATGAAACGTCTGAATCATCTGCGCCATCATCTCTATGCTCTCGCTATTCGCCC
AGCATCTTTACATAACAGACGGTCAACGATCCTCCCTTACCATAACCATCAAATGCTGAGATCTAGACTACG
AGTACACGGACTAGGGGACTAATCTCAACTCTCACATATCACTTCCCATTGTAACAGCGACCTGGGACTCC
TTGACGGTGGACATCGAGTCTGACATCCGGATTGAAATCCGCACCAATCTGTCATCGATTGCAAATTGGCC
AGCTGTCCCCACATTAGGCTTTAAAAACAGATGCAATTGGGAATCCTGCTGCTGCTGCTGCTGCTGCTG
AAATCTTTGAAATA------------------GGGCCCTGATTACCCTATAG```

**.fastq**

The FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity.
Mapping Phred scores directly on a 2nd generation sequence:

Example of quality control in 2nd generation sequence data. The graph above shows an example of a DNA sequence with the Phred score (bars) mapped onto each colored peak. For example, thymine (T) bases are represented in red, and this sequence contains four T's in a row, as indicated by the four red peaks in a row. The aqua horizontal bar placed across the Phred bars represents a Phred score of 20. As indicated on the previous page, a Phred score of 20, corresponding to 90% accuracy in base calls. Therefore, bars above this line indicate base calls that have a higher than 99% accuracy of being correct.

Calculation of a Phred score: Phred Quality Scores \( Q \) are defined as property which is logarithmically related to the base-calling error probabilities \( P \).

\[
Q = -10 \log_{10} P \quad \text{or} \quad P = 10^{-\frac{Q}{10}}
\]

For example, if Phred assigns a Quality Score of 30, the chances that this base is called incorrectly are 1 in 1000. It is general practice to only count bases with a Phred score of 20 or above (1 in 100 incorrect calls).

FASTQ was originally developed at the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data, but has recently become the de facto standard for storing the output of high throughput sequencing instruments such as those made from Illumina sequencing machines. One run on an Illumina HiSeq 2000 will produce over 200 GBs of data.
Typical results will look like the following from an Illumina HiSeq 2000 sequencer, and one human genome sequence read in this format will circle the diameter of the earth 1.5 times:

There are four line types in the FASTQ format. First a ‘@’ title line which often holds just a record identifier. This is a free format field with no length limit—allowing arbitrary annotation or comments to be included. The second line is the sequence line(s), which as in the FASTA format, and can be line wrapped. Also like FASTA format, there is no explicit limitation on the characters expected, but restriction to the IUPAC single letter codes for (ambiguous) DNA or RNA is wise, and upper case is conventional. In some contexts, the use of lower or mixed case or the inclusion of a gap character may make sense. White space such as tabs or spaces is not permitted. Third, to signal the end of the sequence lines and the start of the quality string, comes the ‘+’ line. Originally this also included a full repeat of the title line text; however, by common usage, this is optional and the ‘+’ line can contain just this one character, reducing the file size significantly. Finally, come quality line(s) which again can be wrapped. This is a subset of the ASCII printable characters (at most ASCII 33–126 inclusive) with a simple offset mapping. Crucially, after concatenation (removing line breaks), the quality string must be equal in length to the sequence string.
The genome variation format (GVF) is an extension of the generic feature format version 3 (GFF3), and is a simple tab-delimited format for DNA variant files, which uses sequence ontology to describe genome variation data. It allows bioinformatics software to easily parse variant data, and uses Sequence Ontology (SO) - an ontology developed by the Gene Ontology Consortium to describe the parts of genomic annotations, and how these parts relate to each other [12]. Using SO to type both the features and the consequences of a variation gives GVF files the flexibility necessary to capture a variety of variation data, maintaining unified semantics and a simple file format.

<table>
<thead>
<tr>
<th>TAG</th>
<th>VALUE</th>
<th>NECESSITY</th>
<th>DESCRIPTION</th>
</tr>
</thead>
</table>
| ID             | String                 | Mandatory | While the GFF3 specification considers the ID tag to be optional, GVF requires it. As in GFF3 this ID must be unique within the file and is not required to have meaning outside of the file  
ID = chr1:Soap:SNP:12345;  
ID = rs10399749; |
| Variant_seq    | String                 | Optional  | All sequences found in this individual (or group of individuals) at a variant location are given with the Variant_seq tag. If the sequence is longer than 50 nucleotides, the sequence may be abbreviated as ‘~’. In the case where the variant represents a deletion of sequence relative to the reference, the Variant_seq is given as ‘-’  
Variant_seq = A,T; |
| Reference_seq  | String                 | Optional  | The reference sequence corresponding to the start and end coordinates of this feature  
Reference_seq = G; |
| Variant_reads  | Integer                | Optional  | The number of reads supporting each variant at this location  
Variant_reads = 34, 23; |
| Total_reads    | Integer                | Optional  | The total number of reads covering a variant  
Total_reads = 57; |
| Genotype       | String                 | Optional  | The genotype of this variant, either heterozygous, homozygous, or hemizygous: Genotype = heterozygous; |
| Variant_freq   | Real number between 0 and 1 | Optional | A real number describing the frequency of the variant in a population. The details of the source of the frequency should be described in an attribute-method pragma. The order of the values given must be in the same order that the corresponding sequences occur in the Variant_seq tag: Variant_freq = 0.05; |
| Variant_effect | [1]String: SO term sequence_variant  
[2]Integer-index  
[4]String feature ID | Optional | The effect of a variant on sequence features that overlap it. It is a four part, space delimited tag, The sequence_variant describes the effect of the alteration on the sequence features that follow. Both are typed by SO. The 0-based index corresponds to the causative sequence in the Variant_seq tag. The feature ID lists the IDs of affected features. A variant may have more than one variant effect depending on the intersected features  
Variant_effect = sequence_variant 0 mRNA NM_012345, NM_098765; |
| Variant_copy_number | Integer | Optional | For regions on the variant genome that exist in multiple copies, this tag represents the copy number of the region as an integer value  
Variant_copy_number = 7; |
| Reference_copy_number | Integer | Optional | For regions on the reference genome that exist in multiple copies, this tag represents the copy number of the region as an integer in the form: Reference_copy_number = 5; |
| Nomenclature   | String                 | Optional  | A tag to capture the given nomenclature of the variant, as described by an authority such as the Human Genome Variation Society  
Nomenclature = HGVS: p.Trp26Cys; |
SAM (Sequence Alignment/Map) is a flexible generic format for storing nucleotide sequence alignment. SAM tools provide efficient utilities on manipulating alignments in the SAM format. SAM (Sequence Alignment/Map) format is the most commonly used generic format for storing large nucleotide sequence alignments. SAM is a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with `@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional and variable number of optional fields for flexible or aligner-specific information.

**Example:** Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment that is the following:

```
Coor  12345678901234 56789012345678901234567890123456
ref   AGCATGT*TAGATAAGGATACGTAGCAGTCAGCGCCAT
+r001/1 TTAGATAAAGGATACTG
+r002   aaaAGATAA*GGATA
+r003   gcctaAGCTAA
+r004   ATAGCT..............TCAGC
-r003   CAGCGCCAT
-r001/2   ttagctTAGGC
```

The corresponding SAM format is:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACGT*
r002  0   ref 9 30 3S6M1P11M  *  0  0 AAAAGATAA*GGATA*
r003  0   ref 9 30 5H6M  *  0  0 AGCTAA * NM:i:1
r004  0   ref 16 30 6M14N5M  *  0  0 ATAGCTTCAGC*
r003  16   ref 29 30 6H5M  *  0  0 TAGGC * NM:i:0
r001  83   ref 37 30 9M  = 7 -39 CAGCGCCCAT*
```
Each header line begins with character ‘@’ followed by a two-letter record type code. In the header, each line is TAB-delimited and except the @CO lines, each data field follows a format ‘TAG:VALUE’ where TAG is a two-letter string that defines the content and the format of VALUE. Each header line should match: \/^@[A-Za-z][A-Za-z](t[A-Za-z][A-Za-z0-9]:[^-]+)+$/ or /^@CO\t.*/. Tags containing lowercase letters are reserved for end users.

Terminologies and Concepts – SAM:

Template: A DNA/RNA sequence part of which is sequenced on a sequencing machine or assembled from raw sequences.

Segment: A contiguous (sub) sequence on a template which is sequenced or assembled. For sequencing data, segments are indexed by the order in which they are sequenced. For segments of an assembled sequence, they are indexed by the order of the leftmost coordinate on the assembled sequence.

Read: A raw sequence that comes off a sequencing machine. A read may consist of multiple segments, which are sometimes called sub-reads.

1-based coordinate system: A coordinate system where the first base of a sequence is one. In this coordinate system, a region is specified by a closed interval. The SAM, GFF and Wiggle formats use the 1-based coordinate system.

0-based coordinate system: A coordinate system where the first base of a sequence is zero. In this coordinate system, a region is specified by a half-closed-half-open interval. The BAM, BED, and PSL formats use the 0-based coordinate system.

Phred scale: Given a probability 0 \(<\ p \leq\ 1\), the phred scale of p equals \(-10\ \log_{10}\ p\), rounded to the closest integer.

*.tar

TAR is short for Tape ARchiver. The Unix TAR program is an archiver program which stores files in a single archive without compression. Each file archived is represented by a header record which describes the file, followed by zero or more records which give the contents of the file. At the end of the archive file there may be a record filled with binary zeros as an end-of-file marker. A reasonable system should write a record of zeros at the end, but must not assume that such a record exists when reading an archive. All characters in header records are represented by using 8-bit characters in the local variant of ASCII. Each field within the structure is contiguous; that is, there is no padding used within the structure. Each character on the archive medium is stored contiguously.

*.tiff

The ‘tagged image file format’ (TIFF) is one of the most popular image file formats. Since 2\textsuperscript{nd} generation NGS machines produce a large set of tiled fluorescence or luminescence images of a flow-cell surface, these images need to be recorded after each iterative round of sequencing. The TIFF image file format is based on the lossless compression method, allowing for an image to be stored without losing any original data. Moreover, an image in TIFF format can undergo multiple edits with no negative effect on its quality parameters.
The VARIANT file format is proprietary for NGS data from Complete Genomics. The VAR-files archives are approximately 1 GB each and contain the variation calls (including SNPs, indels, copy number variations, structural variations, and mobile element insertions), scores, and annotations. These files are approximately 300 GB per genome, or larger.

69 Genomes Data: A diverse dataset of whole human genome sequences that is freely available for public use to enhance any genomic study or evaluate Complete Genomics data results and file formats. These include 69 DNA samples sequenced using our Standard Sequencing Service, which includes whole genome sequencing, mapping of the resulting reads to a human reference genome, comprehensive detection of variations, scoring, and informative annotation. The 69 genome data set includes:

- A Yoruban trio
- A Puerto Rican trio
- A 17-member CEPH pedigree across three generations
- A diversity panel representing unrelated individuals from nine different populations
- The CEPH samples within the pedigree and diversity sets are from the NIGMS Repository; the remainder of the samples is from the NHGRI Repository, and both are housed at the Coriell Institute for Medical Research. These samples were sequenced with an average genome-wide coverage of 80X (a range of 51X to 89X).

*.vcf

The ‘variant call format’ (VCF) is a flexible and extendable format for variation data such as single nucleotide variants, insertions/deletions, copy number variants and structural variants. When a VCF file is compressed and indexed using tabix, and made web-accessible, the Genome Browser can fetch only the portions of the file necessary to display items in the viewed region. VCF line-oriented text format was developed by the 1000 Genomes Project for releases of single nucleotide variants, indels, copy number variants and structural variants discovered by the project. When a VCF file is compressed and indexed using tabix, and made web-accessible, the Genome Browser can fetch only the portions of the file necessary to display items in the viewed region. This makes it possible to display variants from files that are so large that the connection to genome browsers such the UCSC browser would time out when attempting to upload the whole file to UCSC. Both the VCF file and its tabix index file remain on your web-accessible server (http, https, or ftp), not on the UCSC server. UCSC temporarily caches the accessed portions of the files to speed up interactive display.