NCBI Genotype Archive

Douglas J. Slotta

Stanford, CA
Sept 11, 2012
What is a Genotype?

**Wikipedia**

The genotype is the genetic makeup of a cell, an organism, or an individual (i.e. the specific allele makeup of the individual) usually with reference to a specific character under consideration.

**Genotype Archive**

The experimentally observed sequence for a sample at a given position on a reference sequence.
Data Size

**Submitted BAM**
- read IDs as strings
- original quality & recalibrated quality scores
- additional analysis tags

**cSRA (lossless)**
- read IDs as integers
- 40-level read qualities using recalibrated quality scores

**cSRA (lossy)**
- 8 level qualities for all sites
- uniform binning of recalibrated quality scores

**VCF**
- genotype likelihoods for all variants
- 0.1T

**Component**
- total project size
- lossless cSRA
- lossy cSRA
- analysis genotypes
The current total size of 1000G genotype data is around 0.25 TB.

<table>
<thead>
<tr>
<th>1000G Interim</th>
<th>VCF Files (gzip’d)</th>
<th>Variations (millions)</th>
<th>Samples (people)</th>
<th>Data Points (billions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>v. 1</td>
<td>66GB</td>
<td>28.2</td>
<td>629</td>
<td>44.6</td>
</tr>
<tr>
<td>v. 2</td>
<td>18GB</td>
<td>38.8</td>
<td>1094</td>
<td>42.5</td>
</tr>
<tr>
<td>v. 3</td>
<td>163GB</td>
<td>39.6</td>
<td>1092</td>
<td>130.1</td>
</tr>
</tbody>
</table>
Studies, and versions, and references! Oh, my!

- **Study Accession**: GTYP00001
- **Submitted Version**: v.1 August 2010, v.2 November 2010, v.3 May 2011
- **Sequence Reference Version**: v.1 GRCh36, v.1 GRCh37, v.1 GRCh37, v.2 GRCh38
- **SciDB Array**: GTYP00001_1_1, GTYP00001_2_1, GTYP00001_3_1, GTYP00001_3_2

Douglas J. Slotta
Genotype Archive
VCF File Format

### Example

```plaintext
#fileformat=VCFv4.0
#fileDate=20100707
#source=VCFtools
#reference=NCBI36
#INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">  #INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">  
#FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">  
#FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">  
#FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">  
#FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">  
#ALT=<ID=DEL,Description="Deletion">  
#INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">  
#INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">  
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
# 1   1   ACG A,AT .   PASS .   GT:DP 1/2:13 0/0:29
# 1   2   rs1 C T,CT .   PASS H2;AA=T GT:GQ 0/1:100 2/2:70
# 1   5   A G .   PASS .   GT:GQ 1/0:77 1/1:95
# 1 100   T <DEL> .   PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
```

- **Mandatory header lines**
  - `#fileformat=VCFv4.0`
  - `#fileDate=20100707`
  - `#source=VCFtools`
  - `#reference=NCBI36`
  - `#INFO` and `#FORMAT` entries define the information in the VCF file.
  - `#ALT` defines the presence of deletions.
  - `#INFO` and `#FORMAT` entries define the annotations in the VCF file.
  - `#CHROM`, `POS`, `ID`, `REF`, `ALT`, `QUAL`, `FILTER`, and `INFO` are mandatory.

- **Optional header lines**
  - Meta-data lines define additional information.

- **Reference alleles** (GT=0)
  - Reference alleles are denoted by a genotype value of zero.

- **Alternate alleles** (GT>0)
  - Alternate alleles are denoted by a genotype value greater than zero.

- **Phased data**
  - Phased data indicates that the alleles are phased.

- **Deletion**
  - Deletions are indicated by a special symbol, such as `<DEL>`.

- **SNP**
  - Single nucleotide polymorphisms are indicated.

- **Insertion**
  - Insertions are indicated by a special symbol, such as `T`.

- **Large SV**
  - Large structural variants are indicated.

- **Other event**
  - Other events are indicated.

### Just what the world needed, another tab-delimited text file format.
A genotype dataset is a very large matrix with orthogonal access patterns:

- List all variations for a subject
- List all genotypes for a given variation
Introduction
Genotype Architecture
Results and Conclusions
Data Organization
VCF Format
1000 Genome Browser

1000 Genome Browser

Douglas J. Slotta
Genotype Archive
How Doth the Little SciDB?

Adding new data to SciDB

<table>
<thead>
<tr>
<th>1000G</th>
<th>VCF Size</th>
<th>Load</th>
<th>Sort</th>
<th>Analyze</th>
</tr>
</thead>
<tbody>
<tr>
<td>v. 1</td>
<td>66GB</td>
<td>20.8h</td>
<td>18.5h</td>
<td>6h 41m</td>
</tr>
<tr>
<td>v. 2</td>
<td>18GB</td>
<td>2d 5h</td>
<td>1d 4h</td>
<td>13h 12m</td>
</tr>
<tr>
<td>v. 3</td>
<td>163GB</td>
<td>2d 3h</td>
<td>2d 3h</td>
<td>14h 41m</td>
</tr>
</tbody>
</table>

Fetching Data from SciDB

<table>
<thead>
<tr>
<th>Variations</th>
<th>Units</th>
<th>Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.5</td>
<td>Genes/sec</td>
<td>13.6</td>
</tr>
<tr>
<td>1,678</td>
<td>Variations/sec</td>
<td>1,038</td>
</tr>
<tr>
<td>—</td>
<td>Genotypes/sec</td>
<td>1.13 million</td>
</tr>
<tr>
<td>13.7</td>
<td>Mbit/sec</td>
<td>17.7</td>
</tr>
</tbody>
</table>

Douglas J. Slotta
Genotype Archive
Acknowledgments

Douglas Slotta  Lead Developer
Charlie Liu    Developer
Chris Grim    Systems Administrator
Don Preuss    Systems Facilitator
Victor Ananiev  Web Frontend
Ken Addess  Quality Assurance

Advice and Consent:
Tanya Barrett
Deanna Church
Mike Feolo
Timothy Hefferson
Doug Hoffman
Donna Maglott
Jim Ostell
Justin Paschall
Lon Phan
Steve Sherry
Karl Sirotkin
Ming Ward
Eugene Yaschenko

NCBI