1000 genomes tutorial

Data access
Primary project data formats

**FASTQ**
sequences with base qualities

```
@IL11_193:4:1:878:501
TATTTTGACTTTGAGCGTATCGAGGCTCTTTAACCTGAACGTCAG
+
I I I I I I I I I I I I I I I I I I (I&/97.,8&
```

**SAM/BAM**
multiple sequence alignments

```
@HD VN:1.0
@SQ SN:chr20 LN:62435964
@RG ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_29006_6945 99 chr20 28833 20 3M1D25M = 28993 195 \ 
AGCTTTATCTTGGTGCTTGGCCG <<<<<<<<<:<9/,.22_;<<<< RG:Z:L1
read_28881_323b 147 chr20 28834 30 35M = 28701 -168 \ 
ACCTATATATCTGCAGCCTTGCA <<<<<<<<<7;<<<<6;<<<<7<< RG:Z:L2
```

http://samtools.sourceforge.net/swlist.shtml
## VCF
variants with genomic location & genotypes

http://vcftools.sourceforge.net/index.html

```plaintext
#fileformat=VCFv4.0
#fileDate=20100721
#source=VCFtools
#reference=NCBI36  (preferred use is assembly accession.version)
##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=GQ,     Number=1, Type=Integer, Description="Genotype Quality">  
##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  40   PASS   .                  GT:DP    1/1:13  2/2:29
1   2   .   C    T,CT  .    PASS   H2;AA=T            GT       0|1     2/2
1   5   rs12 A    G     67   PASS   .                  GT:DP    1|0:16  2/2:20
X   100  .   T    <DEL> .    PASS   SVTYPE=DEL;END=300 GT:GQ:DP 1:12:15 0/0:20:13
```

Primary project data formats
1000 Genomes Pilot Paper Published

27 October 2010

The 1000 Genomes Project Consortium has published the results of the pilot project analysis in the journal Nature in an article appearing on line today. The paper A map of human genome variation from population-scale sequencing is available from the Nature web site and is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence to ensure wide distribution. The paper is also available directly from this link. Please share our paper.

July 2010 Data Release
1000 GENOMES DATA AND SAMPLE INFORMATION

The 1000 Genomes Project is a community resource project that aims to release data rapidly for the benefit of the scientific community.

Description of data released by the project
How to Access 1000 Genomes Data
Data Release Policy
Sample Availability
Use of the Project data, presentations and publications, and authorship

DATA RELEASED BY THE 1000 GENOMES PROJECT

Sample lists and sequencing progress

A summary of sequencing done for each of the three pilot projects is available here. The list of samples and allocations is provided in a spreadsheet.

Variant Calls

The pilot variant calls are available in vcf format from EBI|NCBI

Alignments

The main project alignments are available in BAM format. A list of the files currently available can be found in the alignment index EBI|NCBI. Alignment statistics can be found in the alignment_indices directory EBI|NCBI. There is also a README which explains the alignment process and file layout

Raw sequence files

The main project raw sequence data is available in fastq format. A list of files currently available can be found in the sequence.index EBI|NCBI. Sequence statistics can be found in the sequence_indices directory EBI|NCBI. There is also a README which explains the sequence processing and the file layout.
Aspera is ~10x faster than FTP.
Sequences & alignments by sample ID

Data sets for the pilot data publication.

Previous releases (2008, 2009)

Pre-release data sets, working materials

Site documentation
## dbSNP build 132

<table>
<thead>
<tr>
<th>SnpClass</th>
<th>SnpClassCode</th>
<th>rsCount – uniquely placed</th>
<th>rsCount – Other weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>single base</td>
<td>23,665,960</td>
<td>1,142,880</td>
</tr>
<tr>
<td>2</td>
<td>dips</td>
<td>5,035,890</td>
<td>83,535</td>
</tr>
<tr>
<td>3</td>
<td>HETEROZYGOUS</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Microsatellite</td>
<td>4,462</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Named snp</td>
<td>38,674</td>
<td>984</td>
</tr>
<tr>
<td>7</td>
<td>mixed</td>
<td>116,257</td>
<td>1,406</td>
</tr>
<tr>
<td>8</td>
<td>multi-base</td>
<td>43,250</td>
<td>10,824</td>
</tr>
</tbody>
</table>

In dbSNP VCF file

<table>
<thead>
<tr>
<th>49 SNP INFO TAGS</th>
<th>To indicate...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core properties</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; appearance, variation type</td>
</tr>
<tr>
<td>Frequency</td>
<td>Common in populations</td>
</tr>
<tr>
<td>Discovery</td>
<td>1000 genomes ascertainment</td>
</tr>
<tr>
<td>Functional / Clinical</td>
<td>cSNP, intron, splice, LSDB, GTR, protein structure</td>
</tr>
<tr>
<td>Validation</td>
<td>Validation / withdrawn status</td>
</tr>
<tr>
<td>Sequence annotation</td>
<td>Orientation, specific assemblies, missing alleles, conflicts</td>
</tr>
<tr>
<td>Genotyping</td>
<td>Genotypes available, feature on a typing platform, conflicts in genotypes, typed by HapMap</td>
</tr>
<tr>
<td>Other</td>
<td>Extra data available, links to PubMed articles, micro-attribution available, third party annotation, Inconsistent submissions</td>
</tr>
</tbody>
</table>

Annotation of build 132

• Available in latest RefSeq release
  – Chromosomes
  – mRNAs & proteins
  – RefSeqGene / LRG records
Results: 6

1. Homo sapiens SRA sample SRS000092
   SRA:SRS000092 Coriell:GM12692 HapMap:NA12692
   ID: 1575

2. Homo sapiens SRA sample SRS000091
   SRA:SRS000091 Coriell:GM12691 HapMap:NA12691
   ID: 1574

3. Homo sapiens SRA sample SRS000090
   SRA:SRS000090 Coriell:GM12878 HapMap:NA12878
   ID: 1573

4. Homo sapiens SRA sample SRS000212
   SRA:SRS000212 Coriell:GM19238 HapMap:NA19238
   ID: 1694

5. Homo sapiens SRA sample SRS000214
   SRA:SRS000214 Coriell:GM19240 HapMap:NA19240
   ID: 1696

6. Homo sapiens SRA sample SRS000213
Homo sapiens SRA sample SRS000092

Identifiers
- SRA:SRS000092
- Coriell:GM12892
- HapMap:NA12892

Organism
- Homo sapiens (human)
  - Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
  - Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Attributes
- population: CEU_1
- family id: 1463-16
- sex: Female
- relationship: Mother

Additional attributes
- Coriell plate: HAPMAPPT01
- Coriell cell culture ID: GM12892
- HapMap sample ID: NA12892

Description
- Human HapMap individual NA12892

Links
- Individual record in dbSNP

ID: 1575
1000 Genomes is in the Amazon cloud

1KG pilot content (BAM) is available at s3://1000genomes.s3.amazonaws.com

You can see the XML at http://1000genomes.s3.amazonaws.com