How to Access the Data

Holly Zheng-Bradley
October 8th 2015
Data Availability

• FTP site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
  - Data Files
  - Index files
  - READMEs
  - accessible using aspera and Globus GridFTP
  - http mirror http://ftp.1000genomes.ebi.ac.uk/vol1/ftp

• Browser: http://browser.1000genomes.org
  - Browse 1000 Genomes variants in Genomic Context
  - Tools such as Variant Effect Predictor and Data Slicer

• Ensembl Browser http://grch37.ensembl.org
FTP Site Reorganization

Updated to accommodate multiple projects. Some of the changes are:

- **Added data_collections at top level**
  - The 1000 Genomes Project
  - Human Genome Structural Variation (HGSV)
  - Illumina Platinum Genomes (NA12878 pedigree)

- **Sequence data:**
  - Historic fastq files can be found at pilot, phase1 and phase3 directory
  - The FTP site no longer hosts fastq files for the future project
  - Sequence index file contains path to fastq files in ENA

- **Alignment data:**
  - Data on GRCh37 in phase1 and phase3 directories
  - Data on GRCh38 are now organized by project under data_collections

- **Files previously at top level now in historical_data**
The FTP Site: Top Level

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>changelog_details</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data_collections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>historical_data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phase1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phase3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pilot_data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>release</td>
<td></td>
<td></td>
</tr>
<tr>
<td>technical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHAMELOG</td>
<td>255 KB</td>
<td>04/10/2015 00:43:00</td>
</tr>
<tr>
<td>current.tree</td>
<td>78513 KB</td>
<td>05/10/2015 00:37:00</td>
</tr>
<tr>
<td>README_ebiaspera_info.md</td>
<td>1 KB</td>
<td>17/09/2015 16:45:00</td>
</tr>
<tr>
<td>README_file_formats_and_descriptions.md</td>
<td>7 KB</td>
<td>17/09/2015 16:45:00</td>
</tr>
<tr>
<td>README_ftp_site_structure.md</td>
<td>5 KB</td>
<td>17/09/2015 16:45:00</td>
</tr>
<tr>
<td>README_missing_files.md</td>
<td>1 KB</td>
<td>17/09/2015 16:45:00</td>
</tr>
<tr>
<td>README_populations.md</td>
<td>3 KB</td>
<td>17/09/2015 16:45:00</td>
</tr>
<tr>
<td>README_using_1000genomes_cram.md</td>
<td>5 KB</td>
<td>17/09/2015 16:45:00</td>
</tr>
</tbody>
</table>
FTP Site: Phase3

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/

Name

- integrated_sv_map
- 20130502.phase3.ana
- 20130502.phase3.exo
- 20130502.phase3.low
- 20130502.phase3.seq
- 20130725.phase3.cg_
- 20130820.phase3.cg_
- 20131219.populations
- 20131219.supercolumn
- README_20150504_pl

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/

Name

- HG00096
- HG00097
- HG00099
- HG00100
- HG00101
- HG00102
- HG00103
- HG00104
- HG00105
- HG00106
- HG00107
- HG00108


Name

- alignment
- exome_alignment
- high_coverage_alignment
- sequence_read

Size | Last Modified  | Date       |
-----|---------------|------------|
04/05/2015 22:14:00 | 05/05/2015 01:17:00 |
05/05/2015 03:08:00 | 01/05/2015 08:49:00 |
30/04/2015 14:10:00 | 30/04/2015 14:14:00 |
30/04/2015 12:07:00 | 30/04/2015 12:08:00 |
FTP Site: Latest Variant calls
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502

20140625_related_individuals.txt  25-Jun-2014 11:47  1.1K
ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz  27-May-2015 16:15  1.1G
ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi  27-May-2015 16:50  219K
...
ALL.chrY.phase3_integrated_v1b.20130502.genotypes.vcf.gz  18-Aug-2015 14:38  2.3M
ALL.chrY.phase3_integrated_v1b.20130502.genotypes.vcf.gz.tbi  18-Aug-2015 14:38  2.3M
ALL.wgs.phase3_shapeit2_mvncall_integrated_v5b.20130502.sites.vcf.gz  18-Aug-2015 14:33  1.8G
ALL.wgs.phase3_shapeit2_mvncall_integrated_v5b.20130502.sites.vcf.gz.tbi  18-Aug-2015 14:33  1.8G
README_known_issues_20150813  01-Sep-2015 12:42  3.1K
README_phase3_callset_20150220  06-May-2015 18:37  6.4K
README_phase3_chrY_calls_20141104  15-Nov-2014 02:44  5.2K
README_vcf_info_annotation.20141104  04-Nov-2014 12:20  2.8K
integrated_call_male_samples_v3.20130502.ALL.panel  03-Nov-2014 12:12  25K
integrated_call_samples.20130502.ALL.ped  20-Jun-2014 18:11  176K
integrated_call_samples_v3.20130502.ALL.panel  09-Sep-2014 12:00  54K
supporting/
FTP site: the data_collections directory

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000_genomes_project</td>
<td>02/10/2015 15:02:00</td>
<td></td>
</tr>
<tr>
<td>hgsv_sv_discovery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>illumina_platinum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>README_data_collec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>README.100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Size</td>
<td>Last Modified</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>ACB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG01879</td>
<td>14/09/2015 12:19:00</td>
<td></td>
</tr>
<tr>
<td>HG01880</td>
<td>16/09/2015 11:16:00</td>
<td></td>
</tr>
<tr>
<td>HG01882</td>
<td>13/09/2015 14:59:00</td>
<td></td>
</tr>
<tr>
<td>HG01883</td>
<td>14/09/2015 09:40:00</td>
<td></td>
</tr>
<tr>
<td>HG01885</td>
<td>13/09/2015 14:53:00</td>
<td></td>
</tr>
<tr>
<td>HG01886</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG01889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG01890</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>exome_alignment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Finding Data

• Current.tree file
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree
  - Current Tree is updated nightly so can be upto 24 hours out of date

• Web based search tool
  - http://www.1000genomes.org/ftpsearch
Tools
http://browser.1000genomes.org/tools.html

- Data Slicer
- Variation Pattern Finder
- VCF to PED Converter
- Variant Effect Predictor
- Allele Frequency Calculator
- Forge
Beyond the 1000 Genomes Project

IGSR

• The International Genome and Sample Resource
• Based in EMBL-EBI
• Launched in Jan. 2015, funded by Wellcome Trust
• Goals:
  ▪ Maintain the existing 1000 genomes data and move to GRCh38
  ▪ Collect other data sets generated on the Coriell Cell Lines including Geuvadis
  ▪ Add new populations to expand the global diversity of the variant catalog
Align G1K Reads to GRCh38

• Reference genome available at
  ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
  GRCh38_reference_genome/

• 1000 Genomes Project low coverage data has been aligned

• High coverage and exome data to follow

• Alignments being distributed as lossy CRAM files
  ▪ New format - appreciate feedback

<table>
<thead>
<tr>
<th></th>
<th>LC WGS</th>
<th>HC WXS</th>
<th>HC WGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>#_of_samples</td>
<td>2691</td>
<td>2693</td>
<td>27</td>
</tr>
<tr>
<td>#_of_reads (billion)</td>
<td>329</td>
<td>151</td>
<td>12</td>
</tr>
</tbody>
</table>
Announcements and Contact Info

http://1000genomes.org

http://www.1000genomes.org/1000-genomes-announcement-mailing-list

Twitter @1000genomes

Please send questions to info@1000genomes.org
Acknowledgements

The 1000 Genomes Consortium

IGSR DCC
Paul Flicek
Laura Clark
Susan Fairley

Julia Khobova
Emily Pritchard
Ian Streeter
Avik Datta
David Richardson

Ensembl Variation Team
Fiona Cunningham
Will McLaren
Sarah Hunt
Anja Thormann

The rest of Ensembl