1000 genomes tutorial

Data access

Primary project data formats

FASTQ

sequences with base qualities

```
@IL11_193:4:1:878:501
TATTTTGACTTTGAGCGTATCGAGGCTCTTTAACCTGAACGTCAG
+
IIIIIIIIIIIIII1IDII<IIIIIIIIIIIII(I&/97.,8&
```

SAM/BAM

multiple sequence alignments http://samtools.sourceforge.net/swlist.shtml

```
@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 \
AGCTTATCTTGGTCTTGGCCG <<<<<<<<<:<9/,&,22;;<<< RG:Z:L1
read_28881_323b 147 chr20 28834 30 35M = 28701 -168 \
ACCTATATCTGCGGCCTTGCA <<<<<<<<7;:<<<6;<<<<7<< RG:Z:L2
```

Primary project data formats

VCF

variants with genomic location & genotypes

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

```
##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
```

WWW.1000GENOMES.ORG

NCBI

EMBL-EBI

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home | About | Data | Analysis | Participants | Contact | Browser | Wiki

LATEST ANNOUNCEMENTS

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

HOW TO ACCESS 1000 GENOMES DATA

Download data

The sequence and alignment data generated by the 1000genomes project is made available as quickly as possible via our mirrored ftp sites.

EBI FTP: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
NCBI FTP: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/

Users in the Americas should use the NCBI ftp site and users in Europe and the rest of the world should use the EBI ftp site

The data is also available via an aspera server from both sites. To be able to use this service you need to download the Aspera connect software. This provides both a firefox plug in for downloading data and a bulk download client called ascp

The plugin should automatically start when you visit either the EBI Aspera site or the NCBI Aspera site.

An example commandline for the ascp command looks like

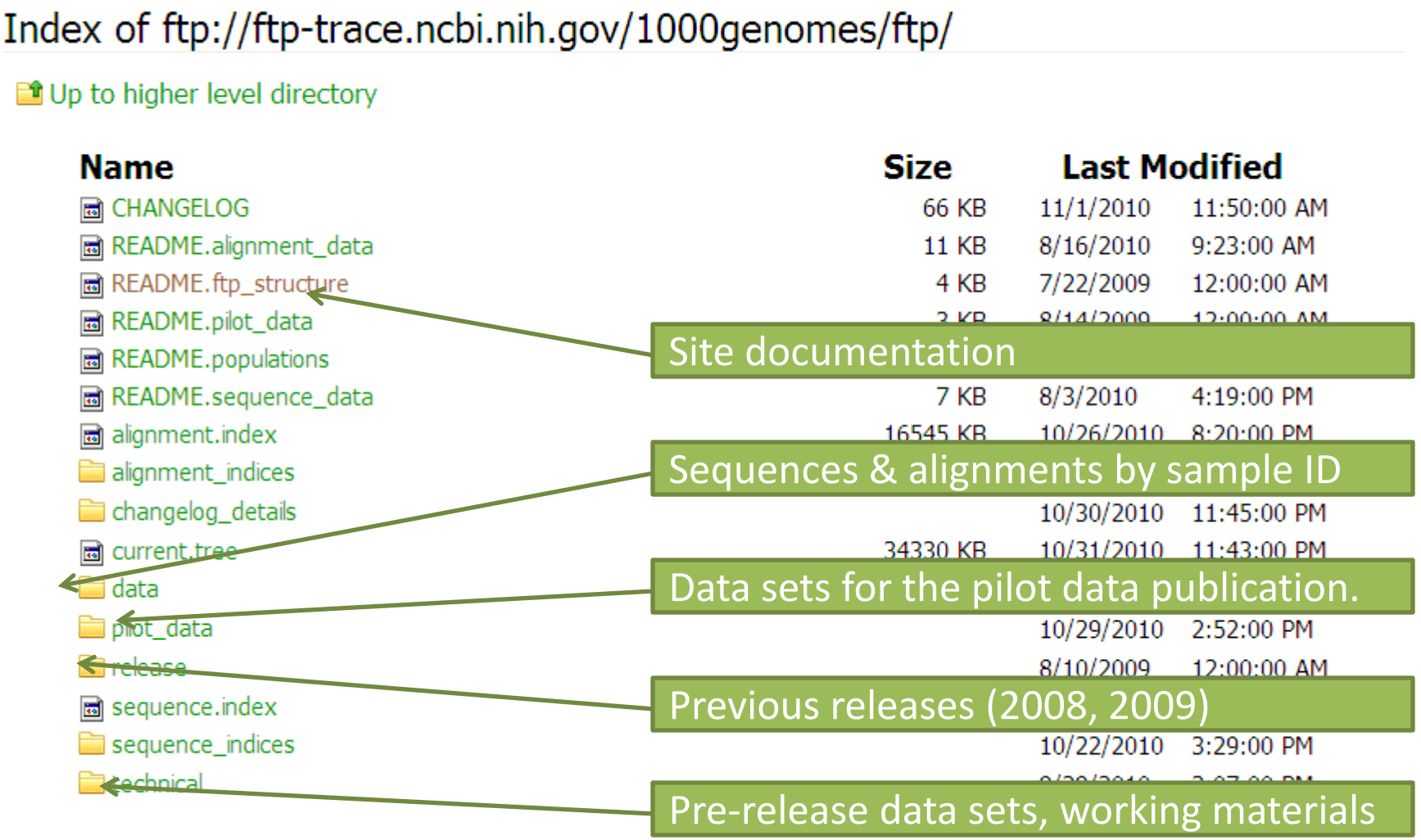
ascp -i bin/aspera/etc/asperaweb_id_dsa.putty -Tr -Q -l 100M -L- fasp-g1k@fas[illegible]00genomes.ebi.ac.uk:vol1/ftp/data/NA12878/alignment /NA12878.chrom10.SLX.SRP000032.2009_04.bam ./

FTP Hierachy

The FTP site follows a specific data hierachy to enable data discovery

- CHANGLOG, This file gives summaries and dates for changes
- changelog_details, This directory contains file lists specifying pa[illegible]me format is explained *Need link to filename convention text*
- current.tree, This file describes the current file hierachy and the size of all the files in it.
- data, This directory contains another directory per individual named for the sample name like NA12878. The contents of each individual directory is described below in the data directory section

Aspera is ~10x faster than FTP.


Index of ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/
Up to higher level directory
Name
Size
Last Modified
CHANGELOG
66 KB
11/1/2010
11:50:00 AM
README.alignment_data
11 KB
8/16/2010
9:23:00 AM
README.ftp_structure
4 KB
7/22/2009
12:00:00 AM
README.pilot_data
README.populations
README.sequence_data
7 KB
8/3/2010
4:19:00 PM
alignment.index
16545 KB
10/26/2010
8:20:00 PM
alignment_indices
changelog_details
10/30/2010
11:45:00 PM
current.tree
34330 KB
10/31/2010
11:43:00 PM
data
pilot_data
10/29/2010
2:52:00 PM
release
8/10/2009
12:00:00 AM
sequence.index
sequence_indices
10/22/2010
3:29:00 PM
technical
Site documentation
Sequences & alignments by sample ID
Data sets for the pilot data publication.
Previous releases (2008, 2009)
Pre-release data sets, working materials

dbSNP build 132

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

In dbSNP VCF file

ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/...
…VCF/v4.0/ByChromosomeNoGeno/00-All.vcf.gz

Tag summary online: ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/...
...VCF/v4.0/ByChromosomeNoGeno/snp_info_tag.xslx

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

Annotation of build 132

- Available in latest RefSeq release
  - Chromosomes
  - mRNAs & proteins
  - RefSeqGene / LRG records

BioSample

National Center for Biotechnology Information

Search: BioSample Save search Limits Advanced search Help

1000Genomes_pilot2[filter] Search Clear

Display Settings: Summary, 20 per page Send to:

Filters: Manage Filters

Results: 6

1. Homo sapiens SRA sample SRS000092
   Homo sapiens SRA sample
   SRA:SRS000092 Coriell:GM12892 HapMap:NA12892
   ID: 1575

2. Homo sapiens SRA sample SRS000091
   Homo sapiens SRA sample
   SRA:SRS000091 Coriell:GM12891 HapMap:NA12891
   ID: 1574

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

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

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

Homo sapiens SRA sample SRS000213

Find related data

Database: Select

Find items

Search details

1000Genomes_pilot2[filter]

Search See more...

Recent activity

Turn Off Clear

1000Genomes_pilot2[filter] (6)
BioSample

1000Genomes_pilot1[filter] (179)
BioSample

1000Genomes_pilot3[filter] (757)
BioSample

Display Settings: Full Send to:

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 DNA source	

ID: 1575

All links from this record
SRA
Taxonomy

Recent activity
Turn Off Clear
1000Genomes_pilot2[filter] (6) BioSample
1000Genomes_pilot1[filter] (179) BioSample
1000Genomes_pilot3[filter] (757) BioSample
1000Genomes_pilot1 (0) BioSample
1000Genomes (0) BioSample
See more...

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