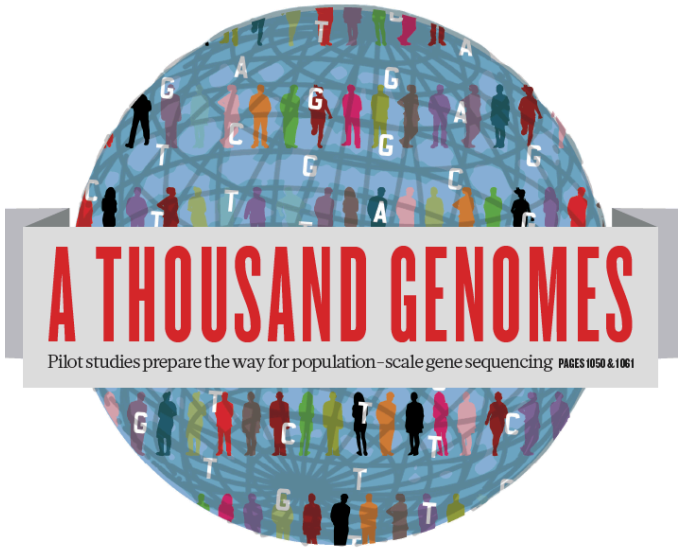


The 1000 Genomes Tutorial
Raw Data and the FTP Site

Laura Clarke

17th February 2012

EMBL-EBI

This Presentation should give an overview of the 1000 Genomes FTP site, the raw data we provide and the formats the data is in.

Glossary

- **Pilot** : The 1000 Genomes project ran a pilot study between 2008 and 2010
- **Phase 1**: The initial round of exome and low coverage sequencing of 1000 individuals
- **Phase 2**: Expanded sequencing of 1700 individuals and method improvement
- **SAM/BAM**: Sequence Alignment/Map Format, an alignment format
- **VCF**: Variant Call Format, a variant format

Summary

- Command Line Tools
- Sequence Data
- Alignment Data
- Variant Call Data
- FTP Site
- Data Slicing
- Data Availability
- Announcements

Command Line Tools

- Samtools http://samtools.sourceforge.net/
- VCFTools http://vcftools.sourceforge.net/
- Tabix http://sourceforge.net/projects/samtools/files/tabix/
  - (Please note it is best to use the trunk svn code for this as the 0.2.5 release has a bug)
  - svn co https://samtools.svn.sourceforge.net/svnroot/samtools/trunk/tabix

Sequence Data

- Fastq files
  - @ERR050087.1 HS18_6628:8:1108:8213:186084#2/1
  - GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
  - +
  - DCDHKHKKIJGNNHIJIIKLLMCLKMAILIJH3K>HL1I=>MK.D
  - http://www.1000genomes.org/faq/what-format-are-your-sequence-files

EMBL-EBI

Alignment Data

- BAM files
- ERR052835 163 11 60239 0 100M = 60609 469
- http://samtools.sourceforge.net/

NAME	DESCRIPTION
QNAME	Query NAME of the read or read pair
FLAG	Bitwise FLAG (pairing, strand, mate strand etc
RNAME	Reference Sequence NAME
POS	1-Based leftmost POSition of clipped alignment
MAPQ	MAPping Quality (Phred-scaled)
CIGAR	Extended CIGAR string (operations: MIDNSHP)
MRNM	Mate Reference NaMe ('=' if same as RNAME)
MPOS	1-Based leftmost Mate POSition
ISIZE	Inferred Insert SIZE
SEQ	Query SEQuence on the same strand as the reference
QUAL	Query QUALity (ASCII-33=Phred base quality)

Alignment data: Extended Cigar Strings

Cigar has been traditionally used as a compact way to represent a sequence alignment. BAM files contain an extended version of this cigar string

Operations include

M - match or mismatch

I - insertion

D - deletion

SAM extends these to include

S - soft clip

H - hard clip

N - skipped bases

P - padding

E.g. Read: ACGCA-TGCAGTtagacgt

Ref: ACTCAGTG----GT

Cigar: 5M1D2M2I2M7S

EMBL-EBI

More Information About BAM Files

- http://samtools.sourceforge.net/
- samtools-help@lists.sourceforge.net

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 25 no. 16 2009, pages 2078–2079*
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

A THOUSAND GENOMES

EMBL-EBI

Variant Call Data

- VCF Files
- TAB Delimited Text Format

NAME	DESCRIPTION
CHROM	Chromosome name
POS	Position in chromosome
ID	Unique Identifer of variant
REF	Reference Allele
ALT	Alternative Allele
QUAL	Phred scaled quality value
FILTER	Site filter information
INFO	User extensible annotation
FORMAT	Describes the format of the subsequent fields, must always contain Genotype
Individual Genotype Fields	These columns contain the individual genotype data for each individual in the file

Variant Call Data

- Headers

##fileformat=VCFv4.1

##INFO=<ID=RSQ,Number=1,Type=Float,Description="Genotype imputation quality from MaCH/Thunder">

##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate Allele Count">

##INFO=<ID=AN,Number=1,Type=Integer,Description="Total Allele Count">

##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/ancestral_alignments/README">

##INFO=<ID=AF,Number=1,Type=Float,Description="Global Allele Frequency based on AC/AN”>

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

##FORMAT=<ID=DS,Number=1,Type=Float,Description="Genotype dosage from MaCH/Thunder">

##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihoods">

A THOUSAND GENOMES

EMBL-EBI

Variant Call Data

- Example 1000 Genomes Data
- CHROM 4
- POS 42208061
- ID rs186575857
- REF T
- ALT C
- QUAL 100
- FILTER PASS
- INFO AA=T;AN=2184;AC=1;RSQ=0.8138;AF=0.0005;
- FORMAT GT:DS:GL
- GENOTYPE 0|0:0.000:-0.03,-1.19,-5.00

More Information About VCF Files

http://vcftools.sourceforge.net/

vcftools-help@lists.sourceforge.net

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis Advance Access publication June 7, 2011

The variant call format and VCFtools

Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

VCF variant files

All indexed for fast retrieval

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 5 2011, pages 718–719
doi:10.1093/bioinformatics/btq671

Sequence analysis Advance Access publication January 5, 2011

Tabix: fast retrieval of sequence features from generic TAB-delimited files

Heng Li

Program in Medical Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

Associate Editor: Dmitrij Frishman

A THOUSAND GENOMES

EMBL-EBI

FTP Site

- Two mirrored ftp sites
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp
  - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp
- NCBI site is direct mirror of EBI site
- Can be up to 24 hours out of date
- Both also accessible using aspera
- http://asperasoft.com/
- EBI site has http mirror
  - http://ftp.1000genomes.ebi.ac.uk/vol1/ftp

ftp://ftp.1000genomes.ebi.ac.uk

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp

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

Up to higher level directory

Name	Size	Last Modified
CHANGELOG	118 KB	05/01/2012 5/01/201212 :40:00
README.alignment_data	12 KB	26/01/2011 26/01/201112 :00:00
README.ftp_structure	9 KB	04/04/2011 4/04/201112 :00:00
README.pilot_data	3 KB	14/ 912 :00:00
README.populations	2 KB	18/02/2010 18/02/201012 :00:00
README.sequence_data	7 KB	23/ 19 :03:00
alignment_indices		14/07/2011 14/07/201110 :53:00
changelog_details		05/ 12 :40:00
current.tree	29933 KB	05/ 12 :37:00
data		04/ 8 :50:00
phase1		14/ 14 :03:00
pilot_data		27/ 012 :00:00
release		12/ 13 :18:00
sequence.index	27185 KB	20/12/2011 20/12/201112 :26:00
sequence_indices		14/11/2011 14/11/201110 :10:00
technical		13/12/2011 13/12/20111 :05:00

Documentation

Raw Data

Phase 1 Data

Pilot Data

Release Data

Technical Data

A THOUSAND GENOMES

EMBL-EBI

The FTP Site: Data

Index of ftp://ftp.1000genomes...

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am Later

HG00104	14/12/2011	14/12/201112 :06:00
HG00105	13/12/2011	13/12/20112 :45:00
HG00106	13/12/2011	13/12/20112 :45:00
HG00107	13/12/2011	13/12/20112 :40:00
HG00108	13/12/2011	13/12/20112 :43:00
HG00109		
HG00110	13/12/2011	13/12/20112 :43:00
HG00111	13/12/2011	13/12/20112 :36:00
HG00112		
HG00113	13/12/2011	13/12/20112 :41:00
HG00114	13/12/2011	13/12/20112 :41:00
HG00115		
HG00116		
HG00117	13/12/2011	13/12/20112 :43:00
HG00118	13/12/2011	13/12/20112 :44:00
HG00119	13/12/2011	13/12/20112 :38:00
HG00120	13/12/2011	13/12/20112 :43:00
HG00121	13/12/2011	13/12/20112 :37:00
HG00122	13/12/2011	13/12/20112 :45:00
HG00123	13/12/2011	13/12/20112 :43:00
HG00124	13/12/2011	13/12/20112 :44:00
HG00125	13/12/2011	13/12/20112 :36:00
HG00126	13/12/2011	13/12/20112 :39:00
HG00127	13/12/2011	13/12/20112 :39:00
HG00128	14/12/2011	14/12/201112 :06:00
HG00129	14/12/2011	14/12/201112 :06:00
HG00130	13/12/2011	13/12/20112 :46:00
HG00131	13/12/2011	13/12/20112 :44:00

Sample Level Files

sequence_read

alignment

A THOUSAND GENOMES

EMBL-EBI

FTP Site: Technical

Index of ftp://ftp.1000genomes...

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am Late

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

Up to higher level directory

Name	Size	Last Modified	
README.reference	1 KB	12/10/2009	12/10/200912 :00:00
browser		19/12/2011	19/12/20113 :50:00
method_development		06/06/2011	6/06/201112 :00:00
ncbi_varpipe_data			
other_exome_alignments.alignment_indices		20/07/2011	20/07/201112 :00:00
pilot2_high_cov_GRCh37_bams		11/01/2012	11/01/20125 :56:00
pilot3_exon_targetted_GRCh37_bams			
qc			
reference			
retired_reference			
simulations		04/05/2010	4/05/201012 :00:00
supporting		21/12/2009	21/12/200912 :00:00
working		17/01/2012	17/01/20124 :07:00

Alternative Alignments

Reference Data Sets

Experimental Data

A THOUSAND GENOMES

EMBL-EBI

FTP Site: Release

Index of ftp://ftp.1000genomes...

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am

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

Up to higher level directory

Name	Size	Last Modified
2008_12		2 :00:00
2009_02		21/02/2009 21/02/200912 :00:00
2009_04		07/05/2009 7/05/200912 :00:00
2009_05		08/06/2009 8/06/200912 :00:00
2009_08		10/08/2009 10/08/200912 :00:00
20100804		0:00
20101123		0:00
2010_11		16/02/2011 16/02/201112 :00:00
20110521		16/12/2011 16/12/201110 :09:00

Older Release Dirs

Date Format YYYYMMDD

Sequence Index Dates

20110521.sequence.index	23693 KB	19/07/2011	19/07/201112 :00:00
20110521.sequence.index.exome.stats	48 KB	19/07/2011	19/07/201112 :00:00
20110521.sequence.index.low_coverage.stats	53 KB	21/05/2011	21/05/201112 :00:00
20110521_20110719.exome.stats.csv	2 KB	19/07/2011	19/07/201112 :00:00
20110521_20110719.low_coverage.stats.csv	2 KB	19/07/2011	19/07/201112 :00:00
20110719.sequence.index	23961 KB	19/07/2011	19/07/201112 :00:00
20110719.sequence.index.exome.stats	52 KB	10/10/2011	10/10/201110 :10:00
20110719.sequence.index.low_coverage.stats	54 KB	10/10/2011	10/10/201110 :13:00
20110719_20110920.exome.stats.csv	1 KB	10/10/2011	10/10/20119 :45:00
20110719_20110920.low_coverage.stats.csv	2 KB	10/10/2011	10/10/20119 :45:00

A THOUSAND GENOMES

EMBL-EBI

FTP Site: Pilot Data

Index of ftp://ftp.1000genomes...

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am

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

Up to higher level directory

Name	Size	Last Modified	
README.alignment.index	2 KB	26/08/2009	26/08/200912 :00:00
README.bas	3 KB	27/08/2009	27/08/200912 :00:00
README.sequence.index	2 KB	22/07/2009	22/07/200912 :00:00
SRP000031.sequence.index	7365 KB	12/07/2010	12/07/201012 :00:00
SRP000032.sequence.index	2181 KB	12/07/2010	12/07/201012 :00:00
SRP000033.sequence.index	480 KB	12/07/2010	12/07/201012 :00:00
data			0:00
paper_data_sets		03/02/2011	3/02/201112 :00:00
pilot_data.alignment.index	795 KB	06/05/2010	6/05/201012 :00:00
pilot_data.alignment.index.bas.gz	1740 KB	14/06/2010	14/06/201012 :00:00
pilot_data.sequence.index	10025 KB	12/07/2010	12/07/201012 :00:00
release		20/07/2010	20/07/201012 :00:00
technical		29/07/2010	29/07/201012 :00:00

Pilot Paper Data

A THOUSAND GENOMES

EMBL-EBI

FTP Site: Phase 1

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/

Index of ftp://ftp.1000genomes...

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am Later Pin It

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/

Up to higher level directory

Name	Size		
README.phase1_alignment_data	11 KB	08	
data		13/12/2011	13/12/20112 :34:00
phase1.alignment.index	8643 KB	14/12/2011	14/12/20113 :53:00
phase1.alignment.index.bas.gz	4996 KB	14/12/2011	14/12/20113 :53:00
phase1.exome.alignment.index	389 KB	14/12/2011	14/12/20113 :53:00
phase1.exome.alignment.index.HsMetrics.gz	141 KB	14/12/2011	14/12/20113 :53:00
phase1.exome.alignment.index.HsMetrics.stats	1 KB	14/12/2011	14/12/20113 :53:00
phase1.exome.alignment.index.bas.gz	414 KB	14/12/2011	14/12/20113 :53:00
phase1.exome.alignment.index_stats.csv	1 KB	14/12/2011	14/12/20113 :53:00
technical		14/12/2011	14/12/20114 :11:00

Frozen Phase1 Alignments

A THOUSAND GENOMES

EMBL-EBI

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

ftp://ftp.1000ge...ftp/current.tree

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am Later

```
ftp     directory       403     Tue Dec 20 16:11:25 2011
ftp/README.ftp_structure        file    8408    Mon Apr  4 14:52:52 2011        2a59a3feb2540c113e10877f3ef1efe5
ftp/README.populations  file    1506    Wed Jan 11 15:12:44 2012        f7c588af82396013c1737e66e58f0f05
ftp/CHANGELOG   file    122151  Sat Jan 14 23:51:50 2012        ecaa9b1e0a6860cd76b1545e84ff3403
ftp/sequence.index      file    27836681        Tue Dec 20 12:26:18 2011        b25557458f6c468bd13d025c17461bab
ftp/README.alignment_data       file    11632   Wed Jan 26 16:22:41 2011        7528e9f4ba8c6b085e6d29c7546fc684
ftp/README.sequence_data        file    6548    Sat Jul 23 22:03:54 2011        b5cfc5784ebf06998f883c629c1c0ba0
ftp/README.pilot_data   file    2082    Fri Aug 14 13:58:10 2009        977fe3983de2131f9e28f6f0036b31d9
ftp/phase1      directory       412     Wed Dec 14 16:03:36 2011
ftp/phase1/phase1.exome.alignment.index.HsMetrics.stats file    293     Wed Dec 14 15:53:53 2011        1ebf793046daadd7ff67ecebb1b5361f
ftp/phase1/phase1.exome.alignment.index file    397947  Wed Dec 14 15:53:52 2011        2891d1fffe08acf3ee99c88cb42d130d
ftp/phase1/phase1.alignment.index.bas.gz        file    5115518 Wed Dec 14 15:53:23 2011        2b4e1edb78f617ebfaf5087536d80f95
ftp/phase1/phase1.alignment.index       file    8850348 Wed Dec 14 15:53:22 2011        ea3423858ec976a1fe17839cd334c164
ftp/phase1/phase1.exome.alignment.index.bas.gz  file    423691  Wed Dec 14 15:53:52 2011        7a56f22d28e860fbc65b71d1013717ae
ftp/phase1/phase1.exome.alignment.index.HsMetrics.gz    file    143893  Wed Dec 14 15:53:53 2011        93ba34ab86e9c42198919d128acc13b7
ftp/phase1/phase1.exome.alignment.index_stats.csv       file    715     Wed Dec 14 15:53:53 2011        376ea20314a94399cab99c723e1d974c
ftp/phase1/technical/ncbi_varpipe_data  directory       137     Wed Dec 14 16:16:31 2011
ftp/phase1/technical/ncbi_varpipe_data/phase1.ncbi.20100804.alignment.summary   file    39866   Wed Dec 14 16:13:58 2011        df4676c95ed2cc6f9cd4c9e24a66bbe8
ftp/phase1/technical/ncbi_varpipe_data/phase1.ncbi.20100804.alignment.index     file    159169  Wed Dec 14 16:13:58 2011        a9bc22ace39cb0bcd0bf35f2ee807bbc
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004        directory       308     Tue Dec 13 12:16:47 2011
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.chrom20.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam  file    238645793       Thu Apr 14 15:24
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam.bai      file    7899352 Wed Oct 27 18:31:23 2010
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.chrom20.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam.bai      file    166624  Thu Apr 14 15:24
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam  file    11091314322     Wed Oct 27 18:31:24 2010
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486        directory       308     Tue Dec 13 12:25:36 2011
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam.bai      file    8418040 Tue Jan 25 22:46:53 2011
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam  file    29068330549     Tue Jan 25 22:46:53 2011
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.chrom20.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam.bai      file    176848  Tue Jan 25 22:47
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.chrom20.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam  file    685641416       Tue Jan 25 22:47
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12045        directory       604     Tue Dec 13 12:24:58 2011
```

A THOUSAND GENOMES

EMBL-EBI

Finding Data

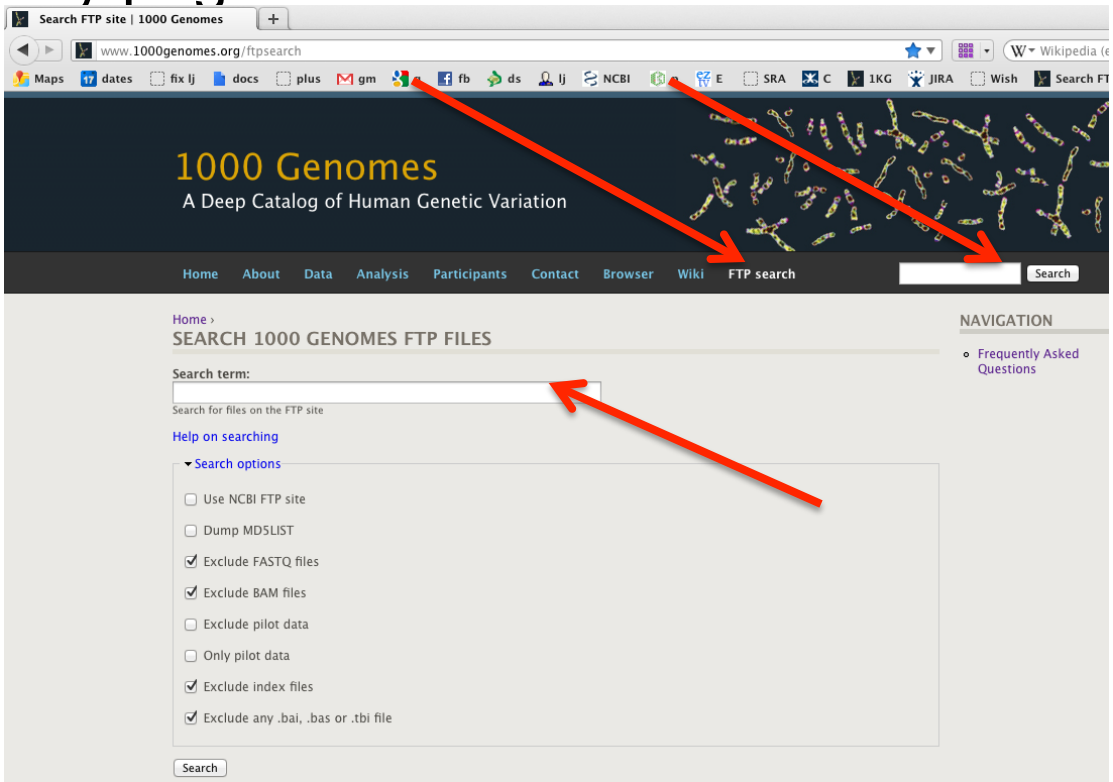
- Current tree file

Description	Example
Relative Path	ftp/data/NA21091/alignment/ NA21091.chrom20.ILLUMINA.bwa.GIH.low_coverage. 20111114.bam
Type (file/directory)	file
Size in bytes	297914382
Last Updated Time Stamp	Thu Jan 26 00:26:52 2012
MD5 checksum	3fd679acc8c92cdc838aa0e5c1849d58

- Relative path does not contain the complete ftp path
- ftp://ftp.1000genomes.ebi.ac.uk/vol1/
- ftp://ftp-trace.ncbi.nih.gov/1000genomes/

Finding Data

- FTP search
- http://www.1000genomes.org/ftpsearch
- Search on the current.tree file
- Provides full ftp paths and md5 checksums
- Every page also has a website search box



EMBL-EBI

Data Slicing

- All alignment and variant files are indexed so subsections can be downloaded remotely
- Use samtools to get subsections of bam files
  - **samtools view** http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam 6:31833200-31834200
- Use tabix to get subsections of vcf files
  - **tabix -h** ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31833200-31834200
- You can also use the web Data Slicer interface to do this

Data Slicing

- VCFtools provides some useful additional functionality on the command line including:
- vcf-compare, comparision and stats about two or more vcf files
- vcf-isec, creates an intersection of two or more vcf files
- vcf-subset, will subset a vcf file only retaining the specified individual columns
- vcf-validator, will validate a particular

Data Slicing

- http://browser.1000genomes.org/tools.html
- http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice

Custom Data

Data Management
- Upload Data
- Attach DAS
- Attach Remote File
- Manage Data
- Features on Karyotype
- Data Converters
  - Assembly Converter
  - ID History Converter
  - Variant Effect Predictor
  - Data Slicer
  - Variation Pattern Finder
  - VCF to PED converter

Data Slicer:
When slicing a VCF or BAM file, both the data file and its index file should be present on the web server and named correctly.
The VCF file should have a ".vcf.gz" extension, and the index file should have a ".vcf.gz.tbi" extension, E.g: MyData.vcf.gz, MyData.vcf.gz.tbi
The BAM file should have a ".bam" extension, and the index file should have a ".bam.bai" extension, E.g: MyData.bam, MyData.bam.bai

Click here for more extensive documentation.

Upload files

VCF File URL:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr1.phase1.projectConsensus.genotypes.vcf.gz
Clear box
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr1.phase1.projectConsensus.genotypes.vcf.gz

Region:
e.g. 1:1-50000

Use VCF filters (this doesn't apply to BAM files):
- None
- By individual(s)
- By population(s) *

(to filter by populations please provide URL to a Sample-Population Mapping File in the box below)

Sample-Population Mapping File URL:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel
Clear box
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel

Next >

A THOUSAND GENOMES

EMBL-EBI

Data Availability

- FTP site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
  - Raw Data Files
- Web site: http://www.1000genomes.org
  - Release Announcements
  - Documentation
- Ensembl Style Browser: http://browser.1000genomes.org
  - Browse 1000 Genomes variants in Genomic Context
  - Variant Effect Predictor
  - Data Slicer
  - Other Tools

Announcements

- http://1000genomes.org
- 1000announce@1000genomes.org
- http://www.1000genomes.org/1000-genomes-annoucement-mailing-list
- http://www.1000genomes.org/announcements/rss.xml
- http://twitter.com/#!/1000genomes

Questions

Please send any questions about this presentation and any other material on our website to info@1000genomes.org

A THOUSAND GENOMES

EMBL-EBI

Thanks

- The 1000 Genomes Project Consortium
- Paul Flicek
- Richard Smith
- Holly Zheng Bradley
- Ian Streeter

A THOUSAND GENOMES

EMBL-EBI