

# The 1000 Genomes Project Tutorial

12<sup>th</sup> April 2012  
Laura Clarke



# Updates to slides

[ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120410\\_tutorial\\_docs/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120410_tutorial_docs/)

Documents also available in the shared disk on your desktops



# 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
- **Date Formats**: 1000 genomes dates are always represented as YYYYMMDD



# Outline

- Introduction to the Project
- Data Availability, the FTP Site and File Formats
- Exercise, Finding data and handling data
- The Browser
- The Tools,
- Exercise Tool use





# Introduction to the Project

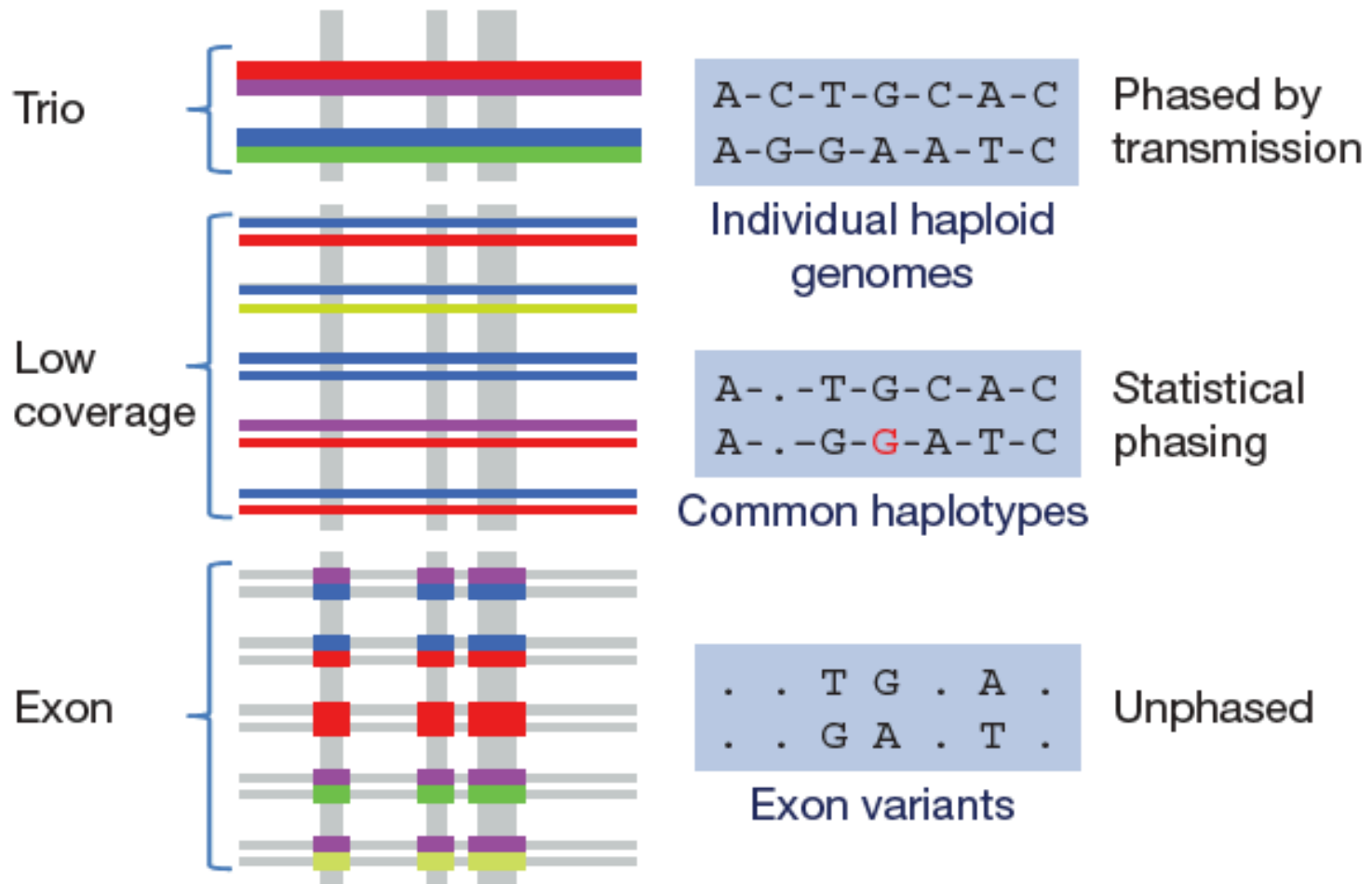


# The 1000 Genomes Project: Overview

- International project to construct a foundational data set for human genetics
  - Discover virtually all common human variations by investigating many genomes at the base pair level
  - Consortium with multiple centers, platforms, funders
- Aims
  - Discover population level human genetic variations of all types (95% of variation  $>$  1% frequency)
  - Define haplotype structure in the human genome
  - Develop sequence analysis methods, tools, and other reagents that can be transferred to other sequencing projects



# 3 pilot coverage strategies



# Main Project Design

- Based on the result of the pilot project, we decided to collect data on more than 2,500 samples from 5 continental groupings
  - Whole-genome low coverage data (>4x)
  - Full exome data at deep coverage (>20x)
  - 500 deep coverage genomes to be sequenced
  - High density genotyping at subsets of sites using both Illumina Omni and Affymetrix Axiom
- Phase 1 Release Integrated Variant Release has been made.



# Phase I (1,150)

# Phase II (1,721)

# Phase III (2,500)

CDX  
17S



CLM (70T); DNA from  
LCL



CHS (100T); DNA from  
LCL



PUR (70T); DNA from  
Blood



FIN (100S); DNA from  
LCL



GBR (96/100S); DNA from



IBS (84/100T); DNA from  
LCL



GWD



GWD



GWD



GWD (target - 100T); DNA from LCL



CDX (100S); DNA: 17 DNA from Bld, 83 from LCL

KHV (82/100) - 15 trios; DNA Bld



45      99 (29T)      23 (7T)

ACB (28/79T) - 14 trios; DNA Bld



PEL (70T); DNA from Blood



3



16 (8T)



PJL (target - 100T); DNA from Blood



15

6

6

195

GIH vs. Sindhi (target - 100T)



Tamil (target -



Sri Lankan (target - 100T)



Bengalee (target - 100T)



Nigeria (target - 100T); DNA from  
Sierra Leone (target - 100T); DNA from LCL



MAB (target - 100T); DNA from  
LCL



AJM (target - 80T); DNA from Bld



270



# Hapmap, The Pilot Project and The Main Project

- **Hapmap**
  - Starting in 2002
  - Last release contained ~3m snps
  - 1400 individuals
  - 11 populations
  - High Throughput genotyping chips
- **1000 Genomes Pilot project**
  - Started in 2008
  - Paper release contained ~14 million snps
  - 179 individuals
  - 4 populations
  - Low coverage next generation sequencing
- **1000 Genomes Phase 1**
  - Started in 2009
  - Phase 1 release has 36.6million snps, 1.5million indels and 14K deletions
  - 1092 individuals
  - 14 populations
  - Low coverage and exome next generation sequencing
- **1000 Genomes Phase 2**
  - Started in 2011
  - 1721 individuals
  - 19 Populations
  - Low coverage and exome next generation sequencing





# Timeline

- **September 2007:** 1000 Genomes project formally proposed Cambridge, UK
- **April 2008:** First Submission of Data to the Short Read Archive.
- **May 2008:** First public data release.
- **October 2008:** SAM/BAM Format Defined.
- **December 2008:** First High Coverage Variants Released.
- **December 2008:** First 1000 genomes browser released
- **May 2009:** First Indel Calls released.
- **July 2009:** VCF Format defined
- **August 2009:** First Large Scale Deletions released.
- **December 2009:** First Main Project Sequence Data Released.
- **March 2010:** Low Coverage Pilot Variant Release made
- **July 2010:** Phased genotypes for 159 Individuals released.
- **October 2010:** A Map of Human Variation from population scale sequencing is published in Nature.
- **January 2011:** Final Phase 1 Low coverage alignments are released
- **May 2011:** @1000genomes appears on Twitter
- **May 2011:** First Variant Release made on more than 1000 individuals
- **October 2011:** Phase 1 integrated variant release made



# Sequencing Data Evolution

- The Project contains data from 3 different providers and multiple platforms

Platform	Min Read Length (bp)	Max Read Length (bp)
454 Roche GS FLX Titanium	70	400
Illumina GA	30	81
Illumina GA II	26	160
Illumina HiSeq	50	102
ABI Solid System 2.0	25	35
ABI Solid System 2.5	50	50
ABI Solid System 3.0	50	50

Fraction of variant sites present in an individual that are NOT already represented in dbSNP

Date	Fraction <u>not</u> in dbSNP
February, 2000	98%
February, 2001	80%
April, 2008	10%
February, 2011	2%
Now	<1%

Ryan Poplin, David Altshuler



# 1000 Genomes Project: Present & Future

- First Phase 2 sequence release 14<sup>th</sup> November 2011
- First Phase 2 alignment release 12<sup>th</sup> March 2012
- First Phase 2 variant site release Summer 2012
  
- Sample collected expected end to June 2012
- Final Phase 3 Sequence release expected December 2012
- 2013 will represent finalization of 1000 genomes analysis results and final data releases



# Pipelines for data processing and variant calling

- Tens of analysis groups have contributed
- Individual pipelines and component tools vary
- Typical main steps:
  - Read mapping
  - Duplicate filtering
  - Base quality score recalibration
  - INDEL realignment
  - Variant Site Discovery
  - Individual Genotype Assignment (sometimes part of site discovery)
  - Variant filtering / call set refinement
  - Variant reporting



# Alignment Data

- The project has made more than 10 releases of Alignment Data
- Pilot Project
  - Aligned to NCBI36
  - Maq and Corona
  - Base Quality Recalibration done
- Phase 1
  - Aligned to GRCh37
  - BWA and Bfast
  - Indel Realignment
- Phase 2
  - Aligned to extended GRCh37
  - Improvements to Base Quality Recalibration





# Methods for Phase 1 Alignments

Platform	Strategy	Aligner	Centre
Solid	Low Coverage	Bfast	TGEN
	Exome	Bfast	Baylor
Illumina	Low Coverage	BWA	Sanger
	Exome	Mosaik	Boston College
454	Low Coverage	SSAHA	Sanger

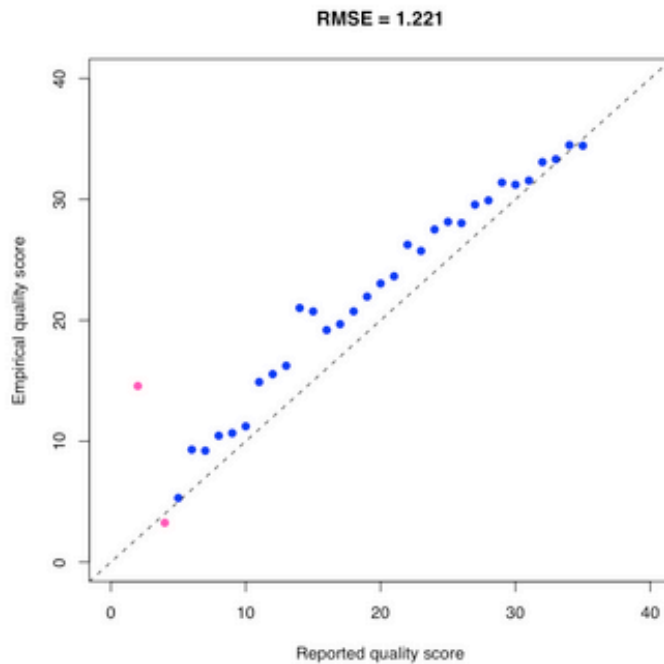
# Base Quality Score Recalibration

- 1000 Genomes Sequence Data is sourced from many different machines across many different institutes
- Each machine may assign Base Quality Values differently
- Base Quality Score Recalibration tests empirical error rates
  - Run alignment
  - Compare mismatches to know variation
- Base Qualities adjusted on basis of empirical measurements

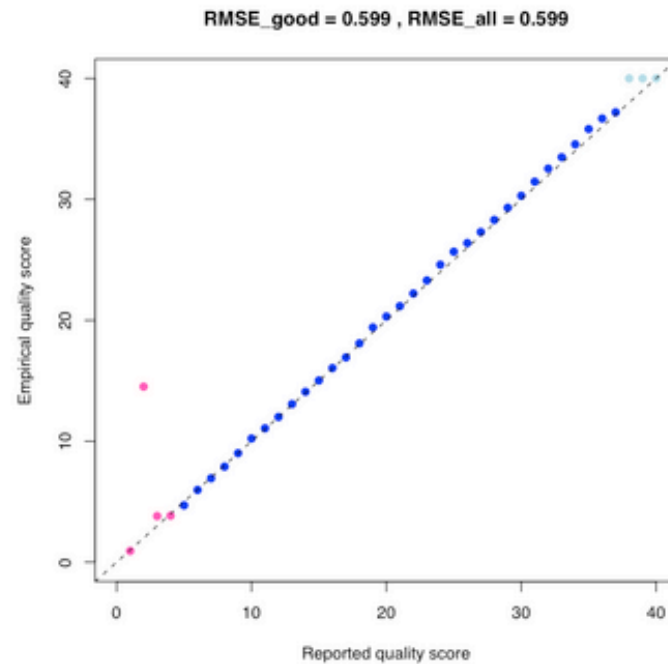


# Base Quality Score Recalibration

## Reported Quality vs. Empirical Quality



Original Data



After GATK Recalibration

# Variant Calling

- Early call sets used a single variant caller
- Intersect approach developed during pilot
- Variant Quality Score Recalibration (VQSR) developed for Phase 1
- Genotype Likelihoods assigned to help with genotype calling
- Integrated genotype calling based on individual variant call sets
- Phase 2 looks to improve site discovery and improve integration



# Variant Quality Score Recalibration

- Multiple Different Variant Callers are used as part of the 1000 Genomes
- Variant Quality Score Recalibration used to define high quality variants from large input set
- Variants as points in a point cloud can be modeled using a Gaussian mixture model
- Model compared to various statistical models to define best set of variants



# VQSR consensus out performs previous merging strategy

Called In	Total # variants	dbSNP % (129)	# novels	Novel ti/tv	Omni poly sensitivity	Omni mono false discovery
Union	46.26M	19.39%	37.29M	1.998	98.94% 2.09M / 2.12M	16.31% 9,739 / 59,721
2 of 6	39.11M	22.24%	30.41M	2.153	98.55% 2.09M / 2.12M	11.23% 6,707 / 59,721
3 of 6	35.69M	23.62%	27.26M	2.219	98.09% 2.08M / 2.12M	3.66% 2,184 / 59,721
4 of 6	32.55M	24.82%	24.48M	2.263	97.39% 2.06M / 2.12M	1.82% 1,085 / 59,721
5 of 6	28.45M	26.72%	20.85M	2.286	95.93% 2.03M / 2.12M	1.06% 634 / 59,721
Intersection	24.02M	27.57%	17.40M	2.317	89.23% 1.89M / 2.12M	0.76% 457 / 59,721
<b>VQSR Project Consensus</b>	<b>38.88M</b>	<b>21.92%</b>	<b>30.36M</b>	<b>2.154</b>	<b>98.41%</b> 2.08M / 2.12M	<b>2.11%</b> 1,261 / 59,721





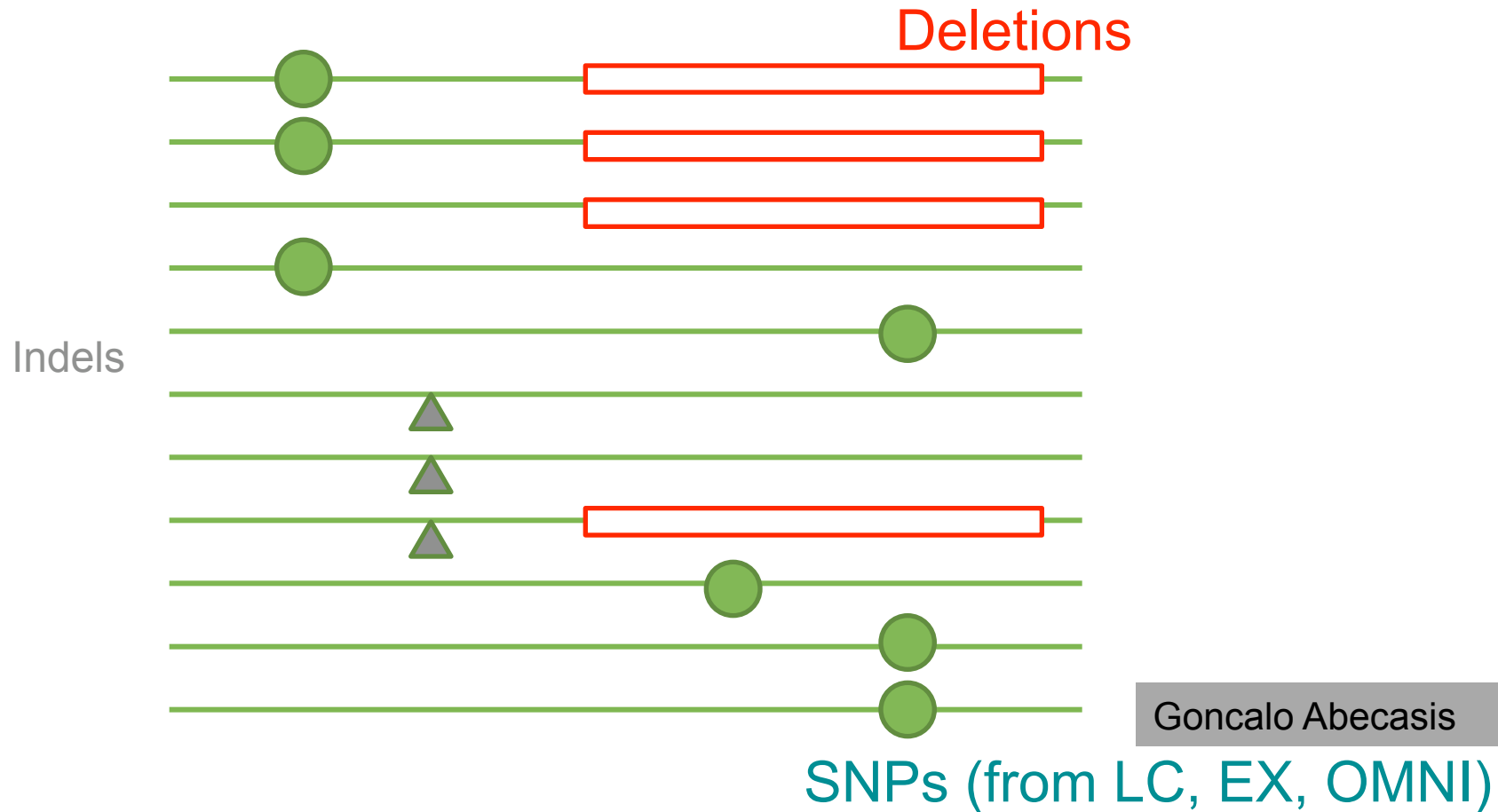
# Methods for integrated genotypes

Components		SNPs	INDELs	SVs
Low-Pass Genomes	Call Sets	BC, BCM, BI NCBI, SI, UM	BC, BI, DI OX, SI	BI, EBI, EMBL UW, Yale
	Consensus	VQSR	VQSR	GenomeSTRiP
Deep Exomes	Call Sets	BC, BCM, BI UM, WCMC	N/A	N/A
	Consensus	SVM	N/A	N/A
Likelihood		BBMM	GATK	GenomeSTRiP
Site Models		Variants are linearly ordered as point mutations		
Haplotyper		MaCH/Thunder with BEAGLE's initial haplotypes		



# Phase 1 analysis goal: an integrated view of human variations

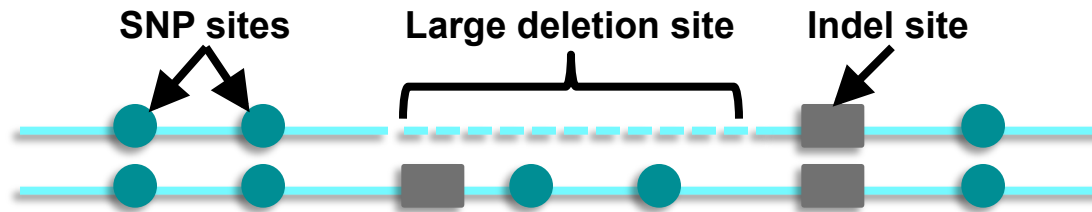
- Reconstruct haplotypes including all variant types, using all datasets



Goncalo Abecasis

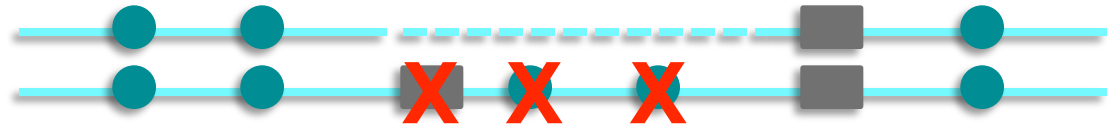


# Strategies for integrating deletions with other types of variation



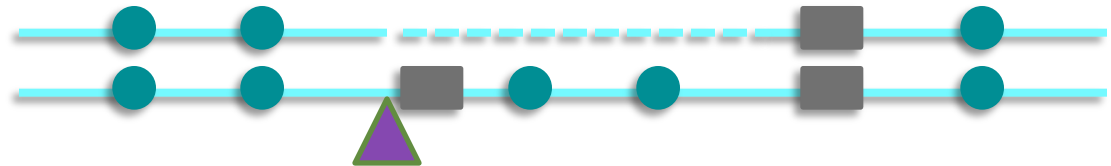
## Previous Approach

Remove SNPs under SVs for imputation  
(1000G pilot, Handsaker et al., 2010)



## Current Approach

Treat SVs as point events  
(1000 Genomes phase 1)



# From PILOT to PHASE1

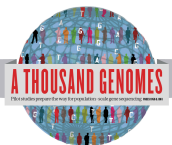
## PILOT

- 14.8M SNPs
- Ts/Tv 2.01
- Includes  
97.8% HapMap3

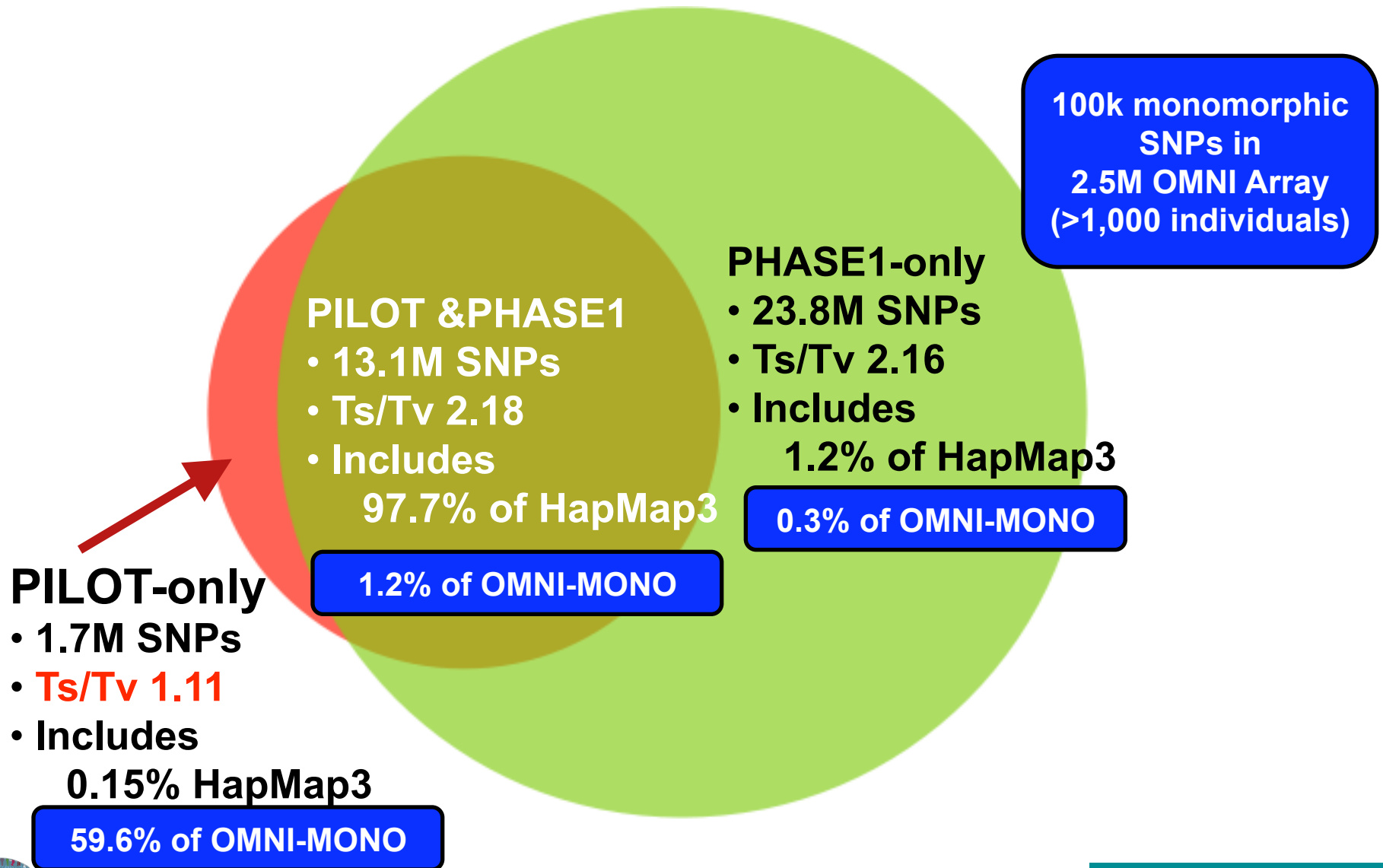
## PHASE1

- 36.8M SNPs
- Ts/Tv 2.17
- Includes  
98.9% HapMap3

*Autosomal chromosomes only*

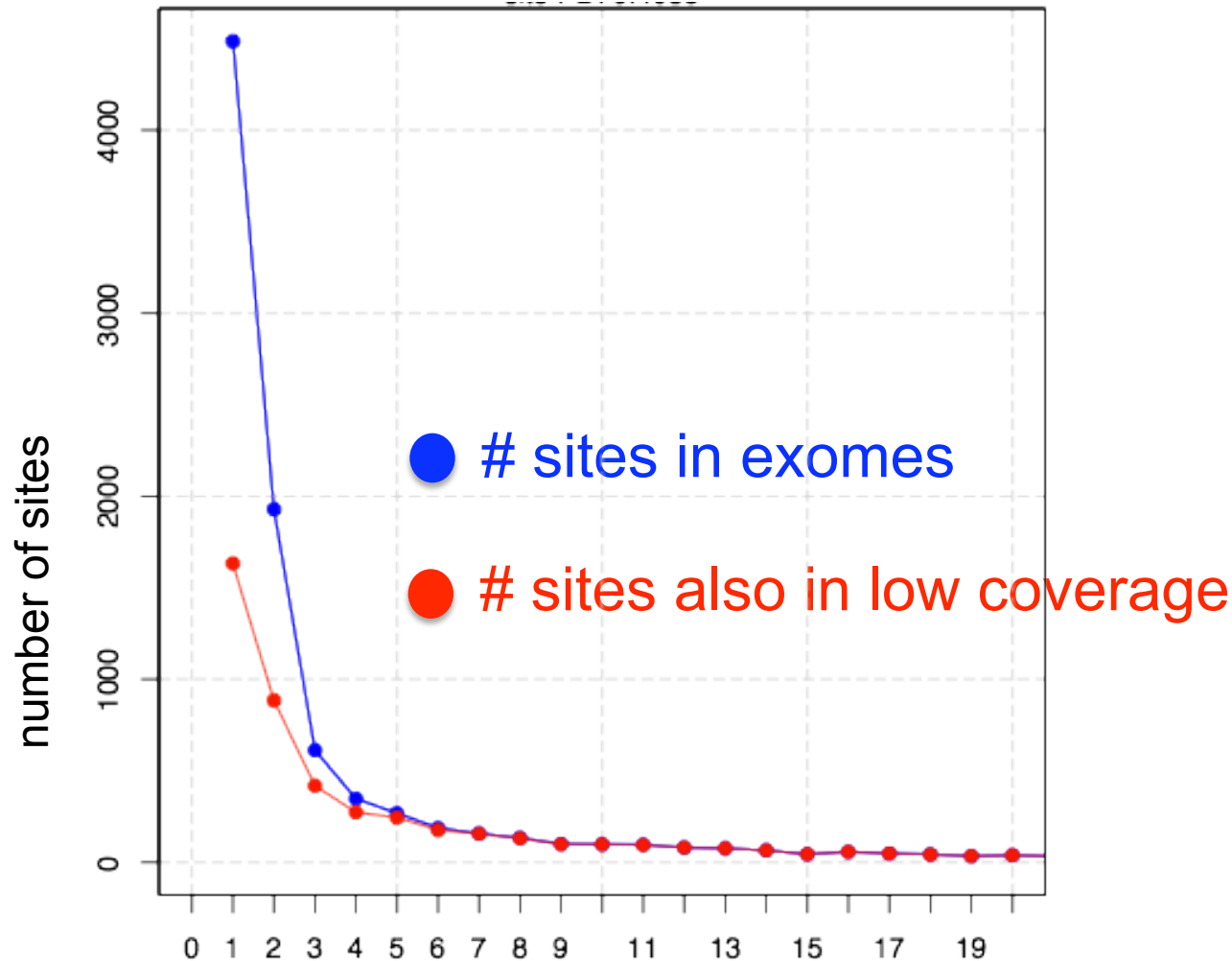


# From PILOT to PHASE1 : Improved SNP calls



OMNI-MONO information was not used in making phase1 variant calls

# Deep coverage exome data is more sensitive to low-frequency variants



Allele count in 766 exomes (chr. 20, exons only)

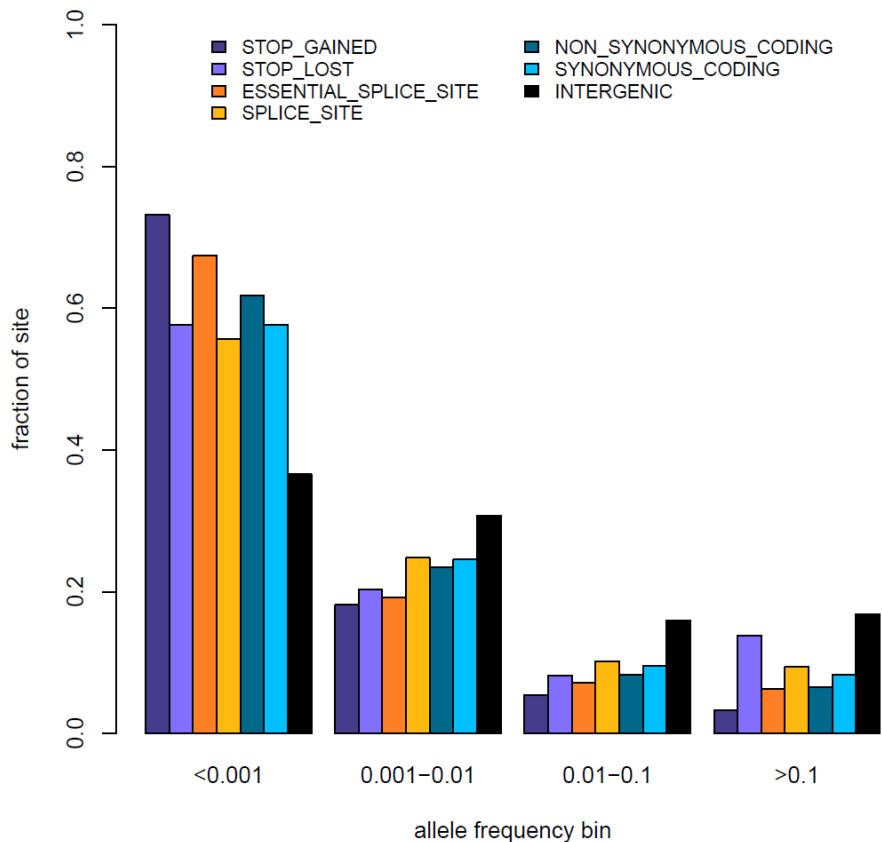
Erik Garrison



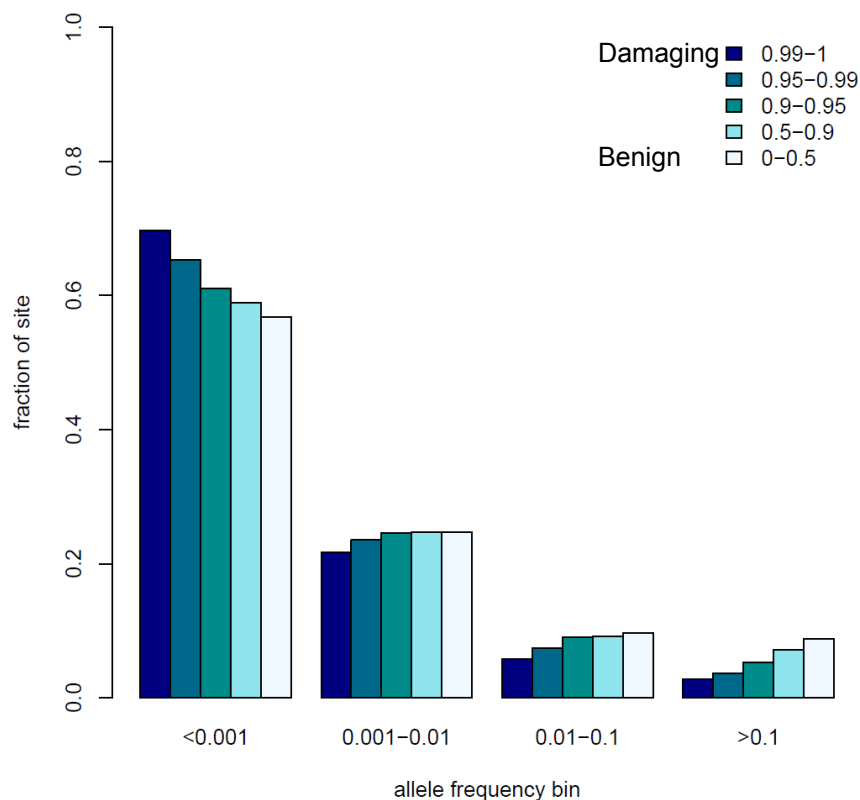


# Newly discovered SNPs are mostly at low frequency and enriched for functional variants

## Functional category

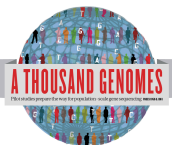


## Non-synonymous: Condel score

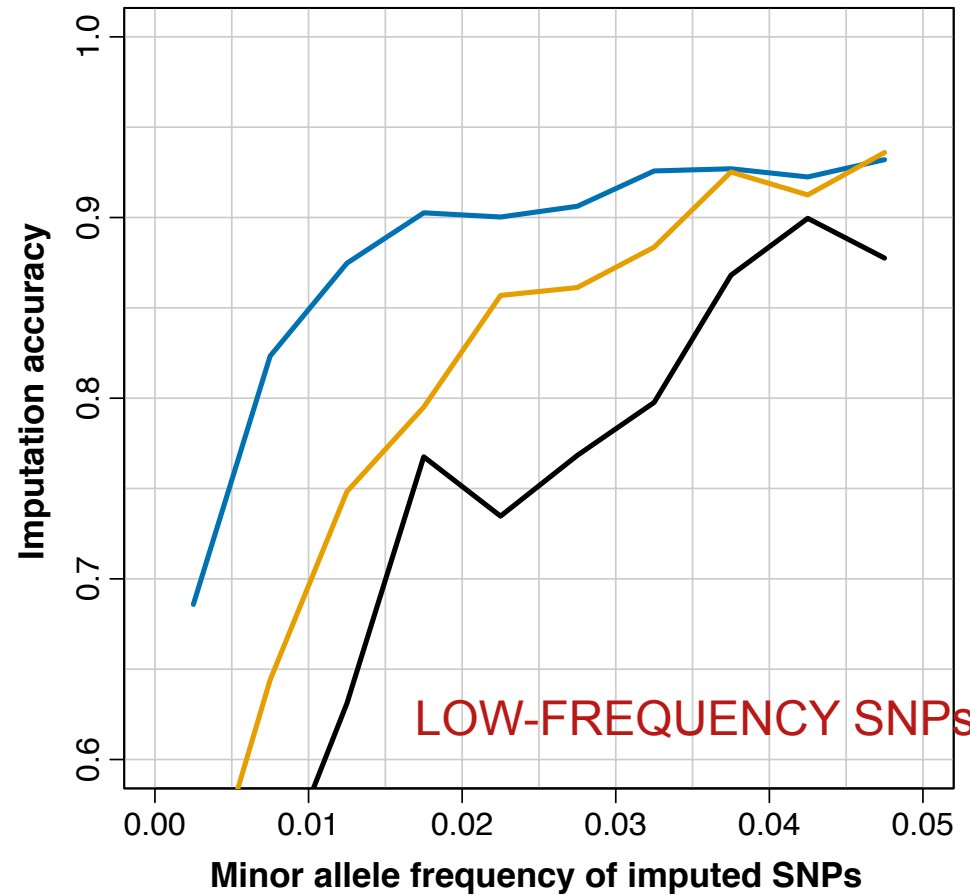
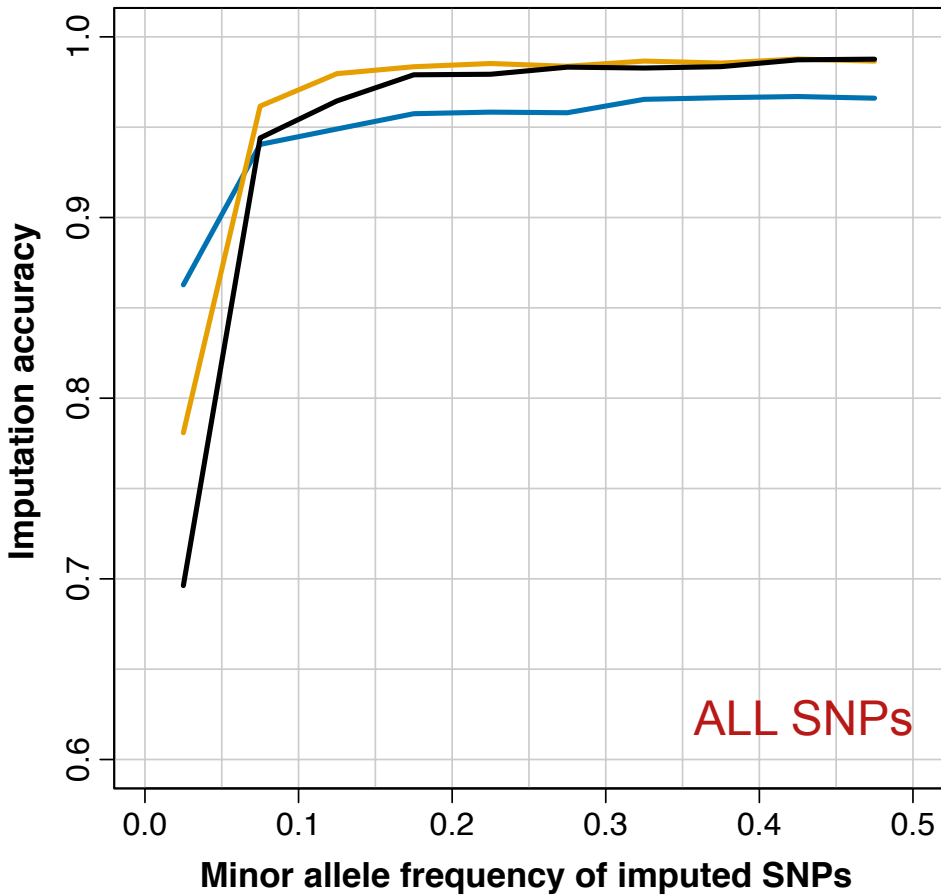


Presentation on using the data for GWAS by Brian Howie

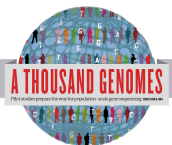
Enza Colonna, Yuan Chen, Yali Xue



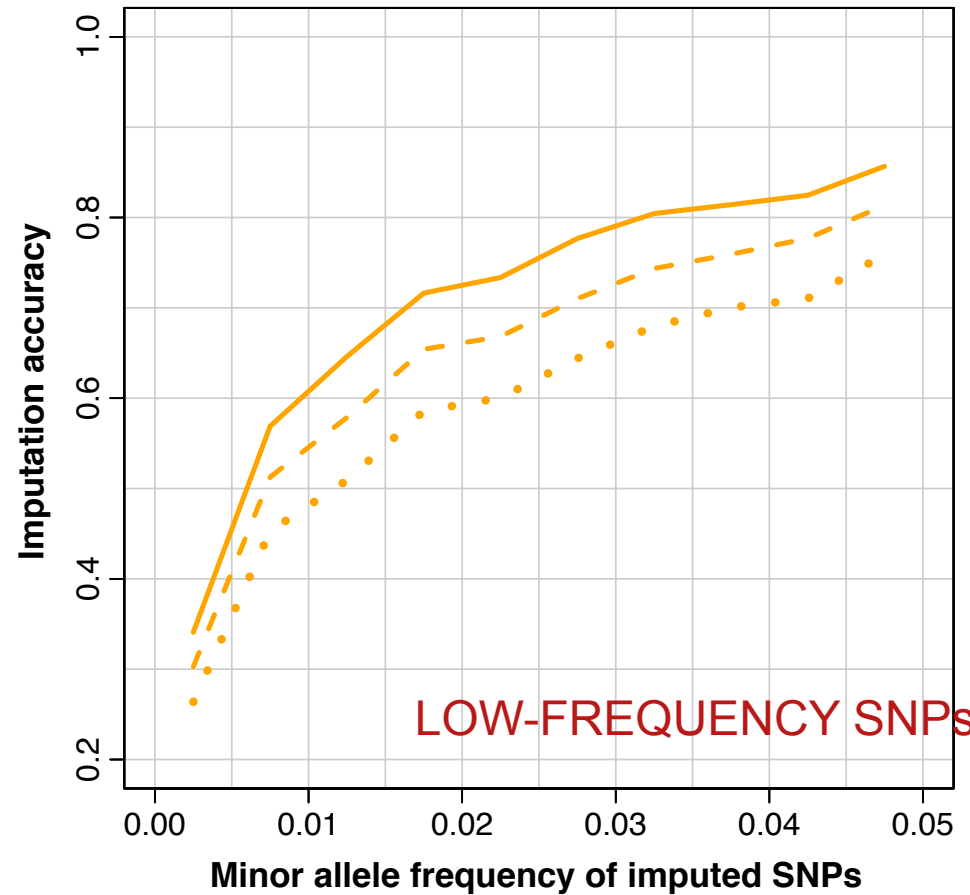
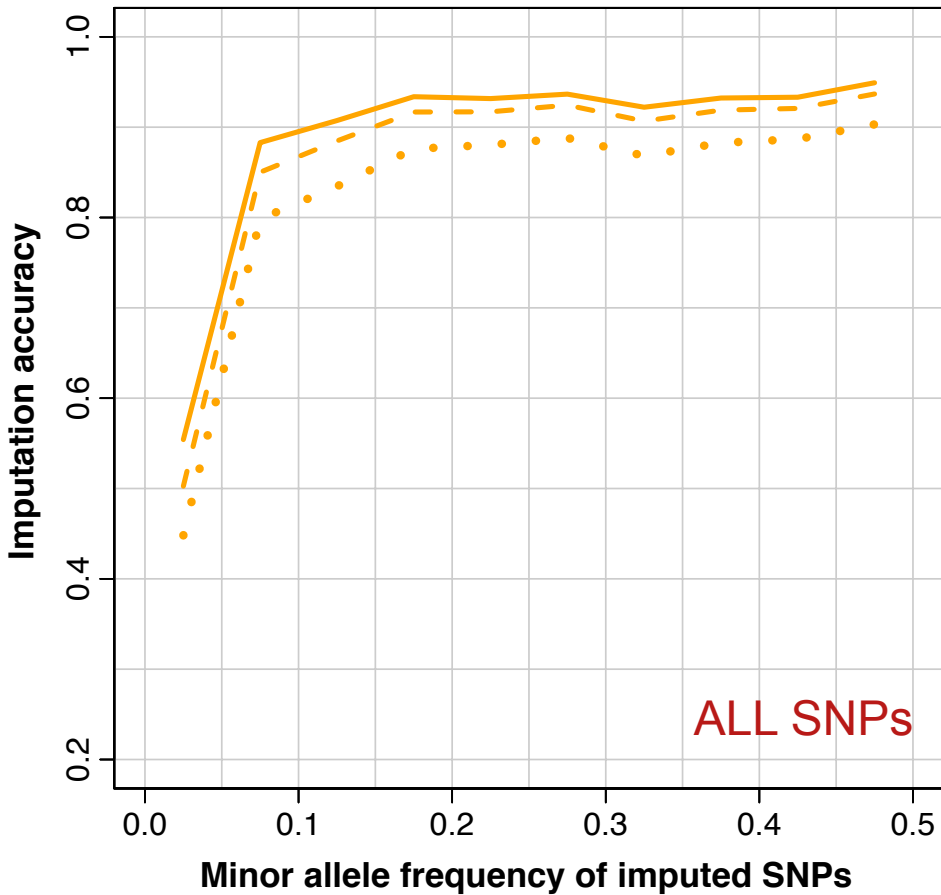
# 1,000 Genomes haplotypes are highly accurate



- European ancestry
- African ancestry
- Admixed (Americas)



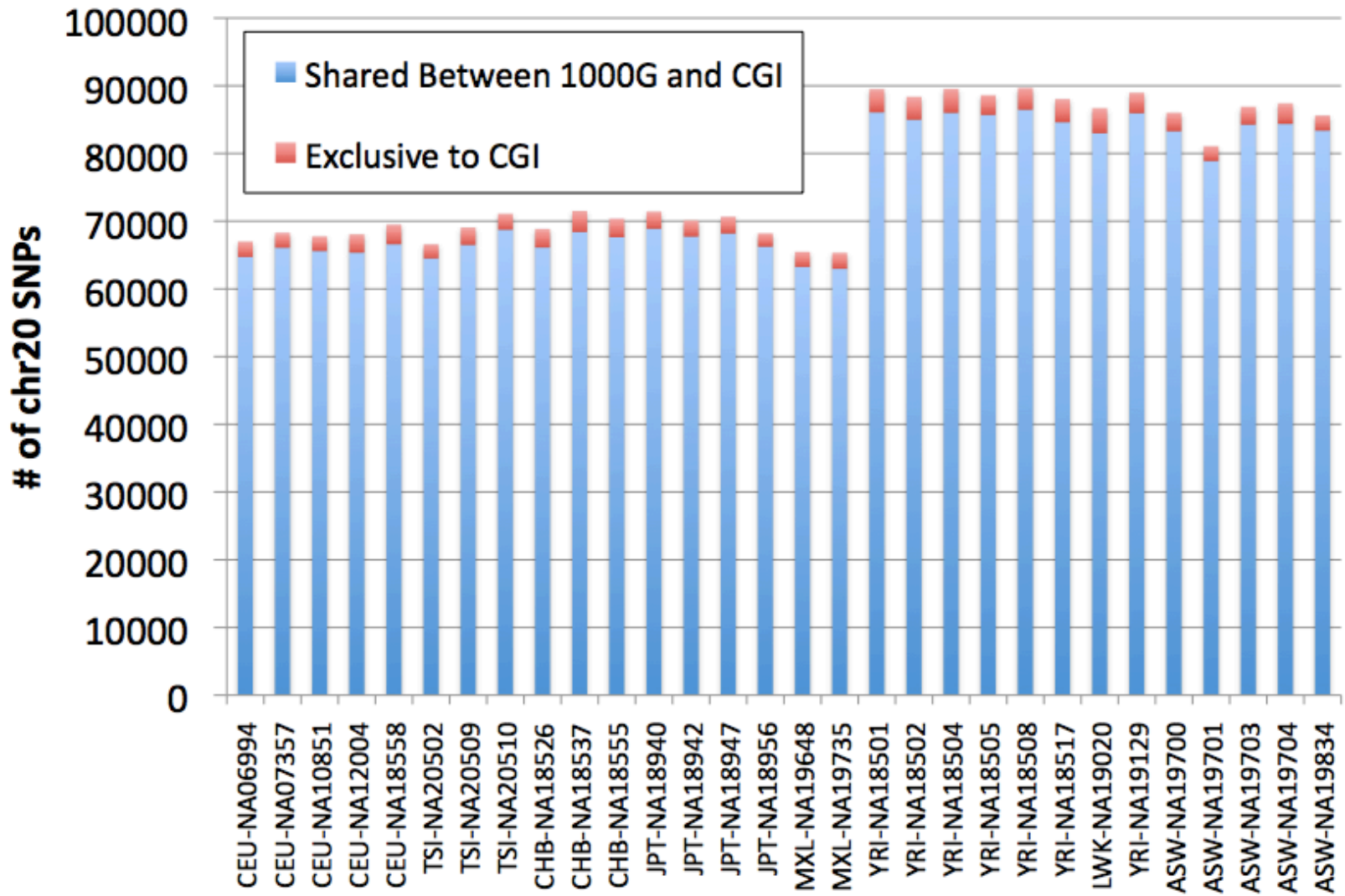
# Imputation accuracy depends on your GWAS chip



- Omni 2.5M
- - - Illumina 550k
- · · Affymetrix 500k



>96% SNPs are detected compared to deep genomes



# Data Availability, FTP site and File Formats



# Command Line Tools

- Samtools <http://samtools.sourceforge.net/>
- VCFTTools <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>



# File Formats

- Sequence in Fastq
- Alignments in SAM/BAM
- Variant Calls in VCF
- Other data
  - ped
  - gff/gtf
  - bed



# 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>





# 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



# More Information About BAM Files

- <http://samtools.sourceforge.net/>
- [samtools-help@lists.sourceforge.net](mailto:samtools-help@lists.sourceforge.net)

## The Sequence Alignment/Map Format and SAMtools

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

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

### ABSTRACT

**Summary:** The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences, supporting short and long reads (up to 128 Mbp) produced by different sequencing platforms. It is flexible in style, compact in size, efficient in random access and is the format in which alignments from the 1000 Genomes Project are released. SAMtools implements various utilities for post-processing alignments in the SAM format, such as indexing, variant caller and alignment viewer,

## 2 METHODS

### 2.1 The SAM format

*2.1.1 Overview of the SAM format* The SAM format consists of one header section and one alignment section. The lines in the header section start with character '@', and lines in the alignment section do not. All lines are TAB delimited. An example is shown in Figure 1b.

In SAM, each alignment line has 11 mandatory fields and a variable number of optional fields. The mandatory fields are briefly described in Table 1. They must be present but their value can be a '\*' or a zero (depending



# Variant Call Data

- VCF Files
- TAB Delimited Text Format

NAME	DESCRIPTION
CHROM	Chromosome name
POS	Position in chromosome
ID	Unique Identifier 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">
```

# 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](mailto: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<sup>1,†</sup>, Adam Auton<sup>2,†</sup>, Goncalo Abecasis<sup>3</sup>, Cornelis A. Albers<sup>1</sup>, Eric Banks<sup>4</sup>, Mark A. DePristo<sup>4</sup>, Robert E. Handsaker<sup>4</sup>, Gerton Lunter<sup>2</sup>, Gabor T. Marth<sup>5</sup>, Stephen T. Sherry<sup>6</sup>, Gilean McVean<sup>2,7</sup>, Richard Durbin<sup>1,\*</sup> and 1000 Genomes Project Analysis Group<sup>‡</sup>

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

Associate Editor: John Quackenbush

## VCF variant files

### 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

#### ABSTRACT

**Summary:** Tabix is the first generic tool that indexes position sorted files in TAB-delimited formats such as GFF, BED, PSL, SAM and SQL export, and quickly retrieves features overlapping specified regions. Tabix features include few seek function calls per query, data compression with gzip compatibility and direct FTP/HTTP access. Tabix is implemented as a free command-line tool as well as a library in C, Java, Perl and Python. It is particularly useful for manually examining local genomic features on the command line and enables

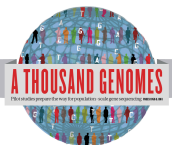
#### 2 METHODS

Tabix indexing is a generalization of BAM indexing for generic TAB-delimited files. It inherits all the advantages of BAM indexing, including data compression and efficient random access in terms of few seek function calls per query.

##### 2.1 Sorting and BGZF compression

Before being indexed, the data file needs to be sorted first by sequence name and then by leftmost coordinate, which can be done with the standard Unix

# All indexed for fast retrieval



















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

Documentation

Raw Data

Phase 1 Data

Pilot Data

Release Data

Technical Data





# Meta Data Formats

- Sequence Index
  - Sequence meta data from ENA
- Alignment Index
  - Location and md5sum for Alignment Files
- BAS
  - Read group level alignment statistics
- HsMetrics
  - Exome alignment statistics based on Picard CalculateHsMetrics



# Sequence Index

- Meta Data File to present information about each fastq file
- Allows easy location of specific subsets of data
- Use to denote specific sequence freezes
- Sequence\_indices directory contains complete history
- Named
  - YYYYMMDD.sequence.index
  - 20120130.sequence.index is most current



Sequence Index	Description	Column	Description
1. Fastq File	Relative path to file	14. Instrument Model	Sequencing Machine Model
2. MD5 checksum	Checksum for file	15. Library Name	
3. Run ID	SRA run id	16. Run Name	
4. Study ID	SRA study id	17. Run Block Name	No Longer used
5. Study Name	SRA study descriptor	18. Insert Size	Estimated Insert Size
6. Center Name	Submission Center	19. Library Layout	Paired or Single ended
7. Submission ID	SRA submission id	20. Paired Fastq	Paired Fastq File
8. Submission Date	Date of Submission	21. Withdrawn	Withdrawn Status
9. Sample ID	SRA Sample ID	22. Withdrawn Date	
10. Sample Name	Coriell Sample name	23. Withdrawn Reason	
11. Population	Population Code	24. Read Count	
12. Experiment ID	SRA Experiment ID	25. Base Count	
13. Instrument Platform	Sequencing Machine Platform	26. Analysis Group	Sequencing Strategy

# Alignment Index

- 6 column file pointing to location of BAM files
- Bam filenames contains majority of information
  - Sample\_name.location.instrument\_platform.alignment\_algorithm.population.analysis\_group.Index\_data.bam
- Alignment index lines contains location and md5 for
  - BAM file
  - BAI file
  - BAS file



# Bas files

- Alignment statistics
- Read group level stats for each alignment
- 21 column file including
  - Read group name
  - Sample name
  - Total Base Count
  - Mapped Base Count
  - Duplicate Base Count



# HsMetrics Files

- Picard Command line tool, CalculateHsMetric
- Used to define completed Exome
- Distributed in gzipped format
- Contains 38 columns like
  - File\_name
  - ON\_BAIT\_BASES
  - MEAN\_BAIT\_COVERAGE
  - PCT\_TARGET\_BASES\_20X

# 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 5/01/2012 12:40:00
 README.alignment_data	12 KB	26/01/2011 26/01/2011 12:00:00
 README.ftp_structure	9 KB	04/04/2011 4/04/2011 12:00:00
 README.pilot_data	3 KB	14/07/2011 14/07/2011 12:00:00
 README.populations	2 KB	18/02/2010 18/02/2010 12:00:00
 README.sequence_data	7 KB	23/07/2011 23/07/2011 19:03:00
 alignment_indices		14/07/2011 14/07/2011 10:53:00
 changelog_details		05/01/2012 05/01/2012 12:40:00
 current.tree	29933 KB	05/01/2012 05/01/2012 12:37:00
 data		04/07/2011 04/07/2011 8:50:00
 phase1		14/07/2011 14/07/2011 14:03:00
 pilot_data		27/07/2011 27/07/2011 12:00:00
 release		12/10/2011 12/10/2011 13:18:00
 sequence.index	27185 KB	20/12/2011 20/12/2011 12:26:00
 sequence_indices		14/11/2011 14/11/2011 10:10:00
 technical		13/12/2011 13/12/2011 10:05:00

Documentation

Raw Data

Phase 1 Data

Pilot Data

Release Data

Technical Data



# The FTP Site: Data

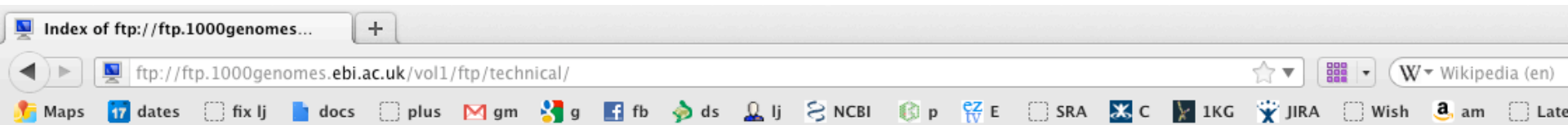
Index of ftp://ftp.1000genomes...  
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/

Sample ID	Date 1	Date 2	Time
HG00104	14/12/2011	14/12/2011	12:06:00
HG00105	13/12/2011	13/12/2011	12:45:00
HG00106	13/12/2011	13/12/2011	12:45:00
HG00107	13/12/2011	13/12/2011	12:40:00
HG00108	13/12/2011	13/12/2011	12:43:00
HG00109	13/12/2011	13/12/2011	12:43:00
HG00110	13/12/2011	13/12/2011	12:43:00
HG00111	13/12/2011	13/12/2011	12:36:00
HG00112	13/12/2011	13/12/2011	12:41:00
HG00113	13/12/2011	13/12/2011	12:41:00
HG00114	13/12/2011	13/12/2011	12:41:00
HG00115	13/12/2011	13/12/2011	12:43:00
HG00116	13/12/2011	13/12/2011	12:44:00
HG00117	13/12/2011	13/12/2011	12:38:00
HG00118	13/12/2011	13/12/2011	12:43:00
HG00119	13/12/2011	13/12/2011	12:37:00
HG00120	13/12/2011	13/12/2011	12:45:00
HG00121	13/12/2011	13/12/2011	12:43:00
HG00122	13/12/2011	13/12/2011	12:44:00
HG00123	13/12/2011	13/12/2011	12:36:00
HG00124	13/12/2011	13/12/2011	12:39:00
HG00125	13/12/2011	13/12/2011	12:39:00
HG00126	14/12/2011	14/12/2011	12:06:00
HG00127	14/12/2011	14/12/2011	12:06:00
HG00128	13/12/2011	13/12/2011	12:46:00
HG00129	13/12/2011	13/12/2011	12:44:00
HG00130	13/12/2011	13/12/2011	12:44:00
HG00131	13/12/2011	13/12/2011	12:44:00

Sample Level Files  
sequence\_read  
alignment



# FTP Site: Technical



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

[Up to higher level directory](#)

Name	Size	Last Modified
<a href="#">README.reference</a>	1 KB	12/10/2009 12/10/2009 12 :00:00
<a href="#">browser</a>		19/12/2011 19/12/2011 3 :50:00
<a href="#">method_development</a>		06/06/2011 6/06/2011 12 :00:00
<a href="#">ncbi_varpipe_data</a>		
<a href="#">other_exome_alignments.alignment_indices</a>		20/07/2011 20/07/2011 12 :00:00
<a href="#">pilot2_high_cov_GRCh37_bams</a>		11/01/2012 11/01/2012 5 :56:00
<a href="#">pilot3_exon_targetted_GRCh37_bams</a>		
<a href="#">qc</a>		
<a href="#">reference</a>		
<a href="#">retired_reference</a>		
<a href="#">simulations</a>		04/05/2010 4/05/2010 12 :00:00
<a href="#">supporting</a>		21/12/2009 21/12/2009 12 :00:00
<a href="#">working</a>		17/01/2012 17/01/2012 4 :07:00

Alternative Alignments

Reference Data Sets

Experimental Data



# FTP Site: Release



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

[Up to higher level directory](#)

Name

Size

Last Modified

2008\_12

Older Release Dirs

2009\_02

21/02/2009 21/02/2009 12:00:00

2009\_04

07/05/2009 7/05/2009 12:00:00

2009\_05

08/06/2009 8/06/2009 12:00:00

2009\_08

10/08/2009 10/08/2009 12:00:00

20100804

20101123

Sequence Index Dates

2010\_11

16/02/2011 16/02/2011 12:00:00

20110521

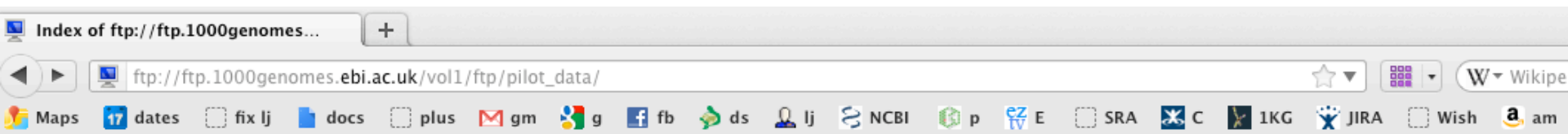
16/12/2011 16/12/2011 10:09:00

Date Format YYYYMMDD

20110521.sequence.index	23693 KB	19/07/2011	19/07/2011 12:00:00
20110521.sequence.index.exome.stats	48 KB	19/07/2011	19/07/2011 12:00:00
20110521.sequence.index.low_coverage.stats	53 KB	21/05/2011	21/05/2011 12:00:00
20110521_20110719.exome.stats.csv	2 KB	19/07/2011	19/07/2011 12:00:00
20110521_20110719.low_coverage.stats.csv	2 KB	19/07/2011	19/07/2011 12:00:00
20110719.sequence.index	23961 KB	19/07/2011	19/07/2011 12:00:00
20110719.sequence.index.exome.stats	52 KB	10/10/2011	10/10/2011 10:10:00
20110719.sequence.index.low_coverage.stats	54 KB	10/10/2011	10/10/2011 10:13:00
20110719_20110920.exome.stats.csv	1 KB	10/10/2011	10/10/2011 9:45:00
20110719_20110920.low_coverage.stats.csv	2 KB	10/10/2011	10/10/2011 9:45:00



# FTP Site: Pilot Data



## Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot\_data/

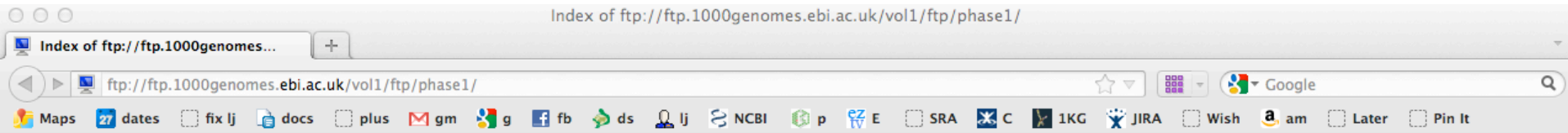
[Up to higher level directory](#)

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

Pilot Paper Data



# FTP Site: Phase 1



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

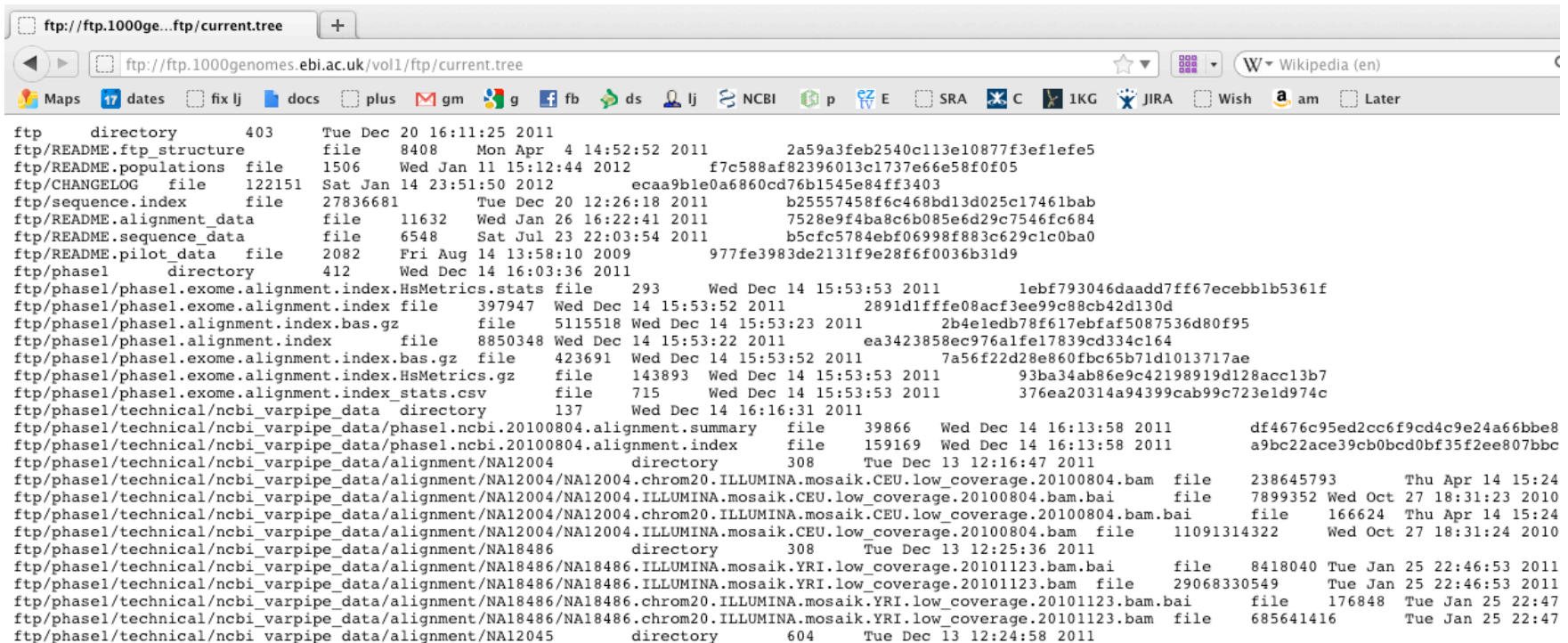
[Up to higher level directory](#)

Name	Size		
<a href="#">README.phase1_alignment_data</a>	11 KB	08	
<a href="#">data</a>		13/12/2011	13/12/2011 12:54:00
<a href="#">phase1.alignment.index</a>	8643 KB	14/12/2011	14/12/2011 13:53:00
<a href="#">phase1.alignment.index.bas.gz</a>	4996 KB	14/12/2011	14/12/2011 13:53:00
<a href="#">phase1.exome.alignment.index</a>	389 KB	14/12/2011	14/12/2011 13:53:00
<a href="#">phase1.exome.alignment.index.HsMetrics.gz</a>	141 KB	14/12/2011	14/12/2011 13:53:00
<a href="#">phase1.exome.alignment.index.HsMetrics.stats</a>	1 KB	14/12/2011	14/12/2011 13:53:00
<a href="#">phase1.exome.alignment.index.bas.gz</a>	414 KB	14/12/2011	14/12/2011 13:53:00
<a href="#">phase1.exome.alignment.index_stats.csv</a>	1 KB	14/12/2011	14/12/2011 13:53:00
<a href="#">technical</a>		14/12/2011	14/12/2011 14:11:00

Frozen Phase1  
Alignments

# 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 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 b2557458f6c468bd13d025c17461bab
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 b5cfc5784ebf06998f883c629c10ba0
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 1ebf793046daadd7ff67ececbb1b5361f
ftp/phase1/phase1.exome.alignment.index file 397947 Wed Dec 14 15:53:52 2011 2891d1ffffe08acf3ee99c88cb42d130d
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.chrom20.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
```



# 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

The screenshot shows a web browser window with the URL [www.1000genomes.org/ftpsearch](http://www.1000genomes.org/ftpsearch). The page features a dark blue header with the text "1000 Genomes" and "A Deep Catalog of Human Genetic Variation". Below the header is a navigation menu with links for Home, About, Data, Analysis, Participants, Contact, Browser, Wiki, and FTP search. A search box is located in the top right corner of the navigation menu. The main content area is titled "SEARCH 1000 GENOMES FTP FILES" and contains a search form with a "Search term:" input field and a "Search" button. Below the search form are search options, including checkboxes for "Use NCBI FTP site", "Dump MD5LIST", "Exclude FASTQ files", "Exclude BAM files", "Exclude pilot data", "Only pilot data", "Exclude index files", and "Exclude any .bai, .bas or .tbi file". A "Search" button is also present at the bottom of the search options. Red arrows point from the search box in the navigation menu to the search term input field, and from the search options section to the search term input field.





# 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](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](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 | vcf-subset -c HG00737**
- You can also use the web Data Slicer interface to do this
- [http://browser.1000genomes.org/Homo\\_sapiens/UserData/SelectSlice](http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice)



# Data Slicing

- VCFtools provides some useful additional functionality on the command line including:
- vcf-compare, comparison 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-stats, return statistics about a vcf file
- vcf-validator, will validate a particular



# 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



# Exercises

1. How many Omni VCF files can you find on the ftp site (Omni is a high throughput genotyping platform from Illumina on which all 1000 genomes samples are being genotyped)
2. Find the most recent Omni VCF file on GRCh37 from the 31st January 2012
3. Use the Website search box found in the top right hand corner of all pages to find the FAQ question about getting subsections of VCF files.
4. Which exome sample from 20110521 has the highest percentage of targets covered at 20x or greater using the 20110521.exome.alignment.index.HsMetrics.gz file and PCT\_TARGET\_BASES\_20X column
5. Find the exome bam file for this sample
6. Get a slice of this exome bam file between 7:114173990-114175942 (exon of FOXP2)



# Exercise Answers, Finding Data

```
> wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree
> grep omni current.tree | cut -f1 | grep vcf | grep -v tbi | wc
-|
> 52
> grep omni current.tree | cut -f1 | grep vcf | grep -v tbi |
grep 20120131 | grep b37 | awk '{print "ftp://ftp.
1000genomes.ebi.ac.uk/vol1/"$1}'
> ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/
20120131_omni_genotypes_and_intensities/
Omni25_genotypes_2141_samples.b37.vcf.gz
> zcat 20110521.exome.alignment.index.HsMetrics.gz | cut
-f1,31 | sort -k2 -n | tail -n1
> HG00737.mapped.illumina.mosaik.PUR.exome.
20110411.bam 0.932651
```



# Exercise Answers, Finding Data

```
> grep HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam  
current.tree | grep -v "bam\  
1000genomes.ebi.ac.uk/vol1/"$1}'
```

```
> ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/HG00737/  
exome_alignment/HG00737.mapped.illumina.mosaik.PUR.exome.  
20110411.bam
```

```
> samtools view ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/  
HG00737/exome_alignment/  
HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam  
7:114173990-114175942 | tail -n1
```

```
> SRR099984.44615561 83 7 114174990 65 76M =  
114174660 -405  
GAACCATATTTGGTGTACATAGGCATAAAGAATTTTGCATAAAACCC  
CCTTGTGGGATTTTATTCATACATAGGTT  
SD@GIB>BFDDHDCDBBJCAFHHJBBDDEHDBFFDCHJB<CCC4IIHHI  
ECGCGGGAEEE@AEBH??@H@?CFDBS RG:Z:SRR099984 NM:i:0  
OQ:Z:DE@DEE?
```

```
EEBEGEDEGFHHFGHHHHGHHFHGHHDHHHHHHGHHDHHGGGGHH  
HHHHHHHHHHHHHHHGFHHHHGHHHHH
```



# The 1000 Genomes Browser

<http://browser.1000genomes.org>



**1000 Genomes**  
A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants Contact **Browse** Wiki FTP search  Search

### LATEST ANNOUNCEMENTS

WEDNESDAY OCTOBER 12, 2011

#### October 2011 Integrated Variant Set release #ICHG2011

This **October 2011** release represents an integrated set of variant calls and phased genotypes including SNPS, short INDELS and Deletions based on low coverage and exome sequencing data across 1092 individuals.

Our [FAQ](#) contains instructions on how to get [smaller subsections](#) of these files

Data access links: [EBI](#) / [NCBI](#)

Link to additional information: [README file](#)

---

THURSDAY JUNE 23, 2011

#### June 2011 Data Release

Genotypes for 1094 individuals for the [May 2011 snp calls](#) from the 20101123 sequence and alignment release of the 1000 genomes project has now been made. This release is based on the GRCh37 assembly of the human genome and is released in the format [VCF 4.0](#)

Our [FAQ](#) contains instructions on how to get [smaller subsections](#) of these files

Data access links: [EBI](#) / [NCBI](#)

Link to additional information: [README file](#)

### NAVIGATION

- [Frequently Asked Questions](#)

### LINKS

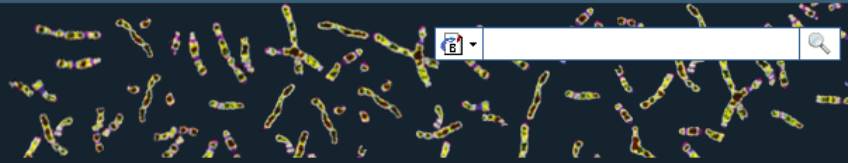
- [All Project Announcements](#)
- [Sample and Project Information](#)
- [Media Archive](#)
- [Download the 1000 Genomes Pilot Paper](#)
- [Project Contacts](#)





# 1000 Genomes

A Deep Catalog of Human Genetic Variation



Tools | Help

## Search 1000 Genomes

e.g. gene BRCA2 or Chromosome 6:133098746-133108745

## Start Browsing 1000 Genomes data



[Browse Human](#) →  
GRCh37

[Protein variations](#) →  
View the consequences of sequence variation at the level of each protein in the genome.

[Individual genotypes](#) →  
Show different individual's genotype, for a variant.

## Browser update September 2011

based on interim Main project data from 20101123 for 1094 individuals and ensembl release 63. The data can be found on [the ftp site](#).

Please see [www.1000genomes.org](http://www.1000genomes.org) for more information about the data presented here and instructions for downloading the complete data set.

- [View sample data](#)

## The 1000 Genomes Browser

### Ensembl-based browser provides early access to 1000genomes data

In order to facilitate immediate analysis of the 1000 Genomes Project data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls from an [interim release 20101123](#). This data has been submitted to dbSNP, and once rsid's have been allocated, will be absorbed into the UCSC and Ensembl browsers according to their respective release cycles. Until that point any non rs SNP id's on this site are temporary and will NOT be maintained.

### Links



[1000 Genomes](#) →  
More information about the 1000 Genomes Project on the 1000 genomes main site.



[Pilot browser](#) →  
This browser is based on Ensembl release 60 and represents the variant set analysed as part of [A map of human genome variation from population-scale sequencing](#), Nature 467, 1061.1073.



[Tutorial](#) →  
The 1000 Genomes Browser Tutorial.

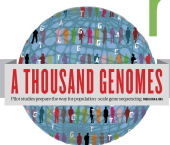
The 1000 Genomes Project is an international collaborative project described at [www.1000genomes.org](http://www.1000genomes.org).

The 1000 Genomes Browser is based on Ensembl web code.

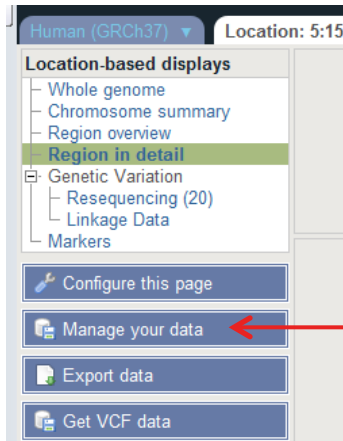
Ensembl is a joint project of EMBL-EBI  and the [Wellcome Trust Sanger Institute](#)



<http://browser.1000genomes.org>



# File upload to view with 1000 Genomes data



Manage your data

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

**Tip**  
Accessing data via a URL can be slow unless you use an indexed format such as BAM. However it has the advantage that you always see the same data as the file on your own machine.

We currently accept attachment of the following formats: BAM, BED, bedGraph, GBrowse, Generic, GFF, GTF, PSL, VCF, WIG. VCF files must be indexed prior to attachment.

File URL:   
( e.g. http://www.example.com/MyProject/mydata.gff )

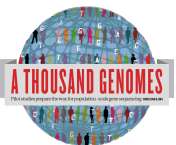
Data format:

Name for this track:

Next >

- Supports popular file types:
  - BAM, BED, bedGraph, BigWig, GBrowse, Generic, GFF, GTF, PSL, VCF\*, WIG

\* VCF must be indexed



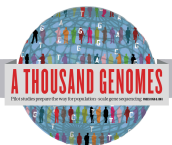
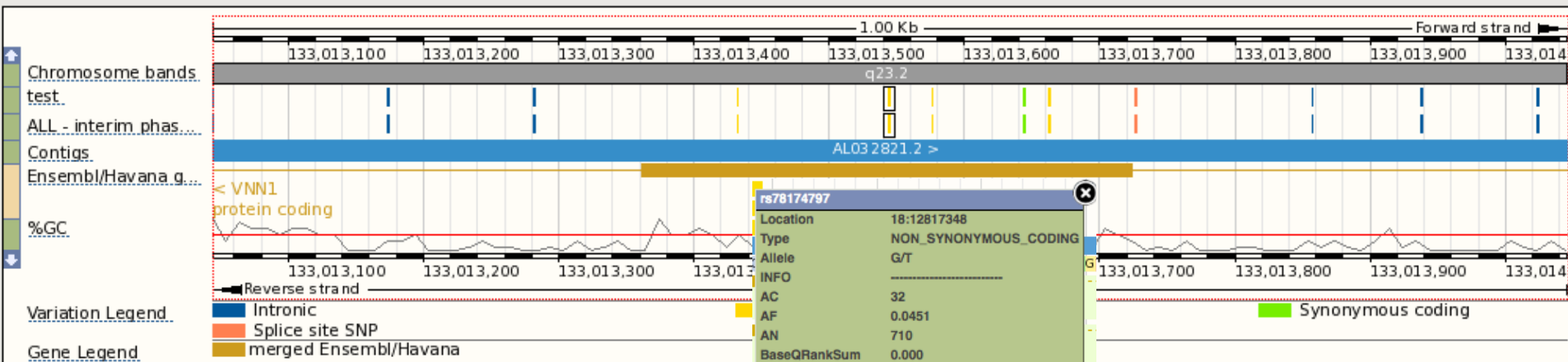

# Uploaded VCF

Example:

[ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.wgs.phase1\\_release\\_v2.20101123.snps\\_indels\\_sv.sites.vcf.gz](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.wgs.phase1_release_v2.20101123.snps_indels_sv.sites.vcf.gz)

Location:

Gene:



# Uploaded BAM

Example:

[http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low\\_coverage.20111114.bam](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam)



# Start again- search for a variation (rs31685)

**1000 Genomes**  
A Deep Catalog of Human Genetic Variation

**Search 1000 Genomes**

rs31685

e.g. gene BRCA2 or Chromosome 6:133098746-133108745

- The Variation tab- left hand links take you to more information

Human (GRCh37) Location: 5:159,283,673-159,284,673 Variation: rs31685

**Variation displays**

- Flanking sequence
- Gene/Transcript (1)
- Population genetics (117)
- Individual genotypes (4343)
- Genomic context
- Phenotype Data
- Phylogenetic Context
- External Data

**Variation: rs31685**

**Variation class** SNP ([rs31685](#) source [dbSNP\\_132](#) - Variants (including SNPs and indels) imported from dbSNP [<http://www.ncbi.nlm.nih.gov/projects/SNP/>])

**Synonyms** Affy GeneChip 100K Array SNP\_A-1683078  
Affy GeneChip 500K Array SNP\_A-4265358  
Affy GenomeWideSNP\_6.0 AFFY\_6\_1M\_SNP\_A-4265358, SNP\_A-4265358  
dbSNP [rs17746160](#), [rs60752908](#), [rs713581](#), [rs58941657](#)  
ENSEMBL ENSSNP12948257, ENSSNP9597299

**Present in**  This feature is present in **1000 genomes** and 3 other sets - click the plus to show all sets

**Alleles** G/A (Ambiguity code: R)

**Ancestral allele** A

**Location** This feature maps to 5:159284173 (forward strand) | [View in location tab](#)

**Validation status** Proven by **cluster, frequency, doublehit, 1000Genome HapMap variant**

**HGVS names**  This feature has 2 HGVS names - click the plus to show

- Population

# 1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37) Location: 6:74,125,388-74,126,388 Variation: rs311685

**Variation displays**

- Flanking sequence
- Gene/Transcript (3)
- Population genetics (46)**
- Individual genotypes (2769)
- Genomic context
- Phenotype Data
- Phylogenetic Context
- External Data

**Variation class** SNP (rs311685 source dbSNP\_132 - Variants (including SNPs and indels) imported from dbSNP [http://www.ncbi.nlm.nih.gov/projects/SNP/])

**Synonyms** Affy GeneChip 100K Array SNP\_A-1679873  
Affy GenomeWideSNP\_6.0 AFFY\_6\_1M\_SNP\_A-8668494, SNP\_A-8668494  
dbSNP rs58378291, rs17756820, rs52794514, rs524803, rs3173186, rs11567000, rs17421786  
ENSEMBL ENSNP9062281  
Illumina\_Human1M-duoV3 rs311685  
Uniprot VAR\_057235

**Present in** 1000 genomes - High coverage - Trios (1000 genomes - High coverage - Trios - CEU, 1000 genomes - High coverage - Trios - YRI), 1000 genomes - Low coverage (1000 genomes - Low coverage - CEU, 1000 genomes - Low coverage - CHB+JPT, 1000 genomes - Low coverage - YRI), ALL - interim phase 1 - 1000 Genomes (AFR - interim phase 1 - 1000 Genomes, AMR - interim phase 1 - 1000 Genomes, ASN - interim phase 1 - 1000 Genomes, EUR - interim phase 1 - 1000 Genomes), ENSEMBL:Venter,HapMap

**Alleles** A/G (Ambiguity code: R)

**Ancestral allele** A

**Location** This feature maps to 6:74125888 (forward strand) | [View in location tab](#)

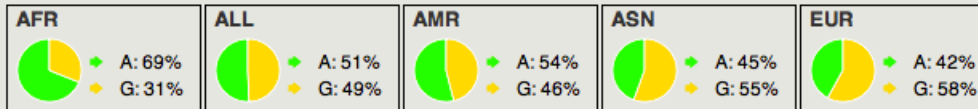
**Validation status** Proven by cluster, frequency, doublehit, 1000Genome HapMap variant

**HGVS names** This feature has 4 HGVS names - click the plus to show

[Population genetics help](#)



## 1000 genomes alleles frequencies



## 1000 genomes

Show/hide columns Filter

Population	Alleles A	Alleles G	Genotypes A/A	Genotypes A/G	Genotypes G/G	Count
1000GENOMES:AFR	0.689	0.311	0.463	0.451	0.085	114
1000GENOMES:ALL	0.507	0.493	0.269	0.477	0.254	294
1000GENOMES:AMR	0.539	0.461	0.293	0.492	0.215	53
1000GENOMES:ASN	0.446	0.554	0.199	0.493	0.308	57
1000GENOMES:EUR	0.421	0.579	0.184	0.475	0.341	70

## 1000 genomes pilot

Show/hide columns Filter

Population	ssID	Submitter	Alleles A	Alleles G	Count
<a href="#">1000GENOMES:pilot 1 CEU low coverage panel</a>	<a href="#">ss233534774</a>	<a href="#">1000GENOMES</a>	0.458	0.542	
<a href="#">1000GENOMES:pilot 1 CHB+JPT low coverage panel</a>	<a href="#">ss240577229</a>	<a href="#">1000GENOMES</a>	0.400	0.600	
<a href="#">1000GENOMES:pilot 1 YRI low coverage panel</a>	<a href="#">ss222470667</a>	<a href="#">1000GENOMES</a>	0.729	0.271	

# Coming Soon Ensembl 65

**e!Ensembl** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors Login · Register

Human (GRCh37) | Location: 9:22,125,003-22,126,003 | Variation: rs1333049

### Variation displays

- Explore this variation
- Genomic context
  - Gene/Transcript (2)
- Population genetics (28)
- Individual genotypes (1737)
- Linkage disequilibrium
- Phenotype Data (8)
- Phylogenetic Context (4)
- Flanking sequence
- External Data

**rs1333049** SNP

**Source** [dbSNP 134](#) - Variants (including SNPs and indels) imported from [dbSNP](#)

**Alleles** Reference/Alternative: **G/C** | Ancestral: **C** | Ambiguity code: **S** | MAF: **0.40** (C)

**Location** Chromosome **9:22125503** (forward strand) | [View in location tab](#)

**Validation status** This variation is validated by **1000 Genomes**, **HapMap** and also cluster, doublehit, frequency, precious, submitter

**Synonyms** This feature has **7** synonyms - click the plus to show

**HGVS name** [g.22125503G>C](#)

[Configure this page](#) | [Manage your data](#) | [Export data](#) | [Bookmark this page](#)

### Explore this variation [help](#)

- Genomic context**
- Gene / Transcript**
- Population genetics**
- Individual genotypes**
- Linkage disequilibrium**
- Phenotype data**
- Phylogenetic context**
- Flanking sequence**

**Help with variations**

**YouTube videos**

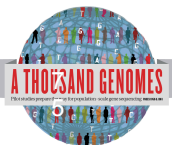
- [SNPs and other Variations - 1 of 2](#)
- [SNPs and other Variations - 2 of 2](#)
- [Clip: Genome Variation](#)
- [BioMart: Variation IDs to HGNC Symbols](#)

**Reference materials**

- [Ensembl variation data: background and terminology](#)
- [Variation Quick Reference card](#)

**Additional resources**

- [Accessing variation data with the Variation API](#)
- [Genomes and SNPs in Malaria](#)



Should arrive in May



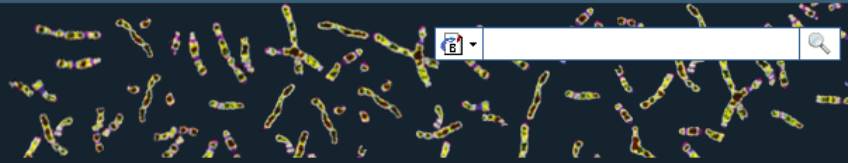
# 1000 Genomes Tools





# 1000 Genomes

A Deep Catalog of Human Genetic Variation



Tools | Help

## Search 1000 Genomes

e.g. gene BRCA2 or Chromosome 6:133098746-133108745

## Start Browsing 1000 Genomes data



[Browse Human](#) →  
GRCh37

[Protein variations](#) →  
View the consequences of sequence variation at the level of each protein in the genome.

[Individual genotypes](#) →  
Show different individual's genotype, for a variant.

## Browser update September 2011

based on interim Main project data from 20101123 for 1094 individuals and ensembl release 63. The data can be found on [the ftp site](#).

Please see [www.1000genomes.org](http://www.1000genomes.org) for more information about the data presented here and instructions for downloading the complete data set.

- [View sample data](#)

## The 1000 Genomes Browser

### Ensembl-based browser provides early access to 1000genomes data

In order to facilitate immediate analysis of the 1000 Genomes Project data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls from an [interim release 20101123](#). This data has been submitted to dbSNP, and once rsid's have been allocated, will be absorbed into the UCSC and Ensembl browsers according to their respective release cycles. Until that point any non rs SNP id's on this site are temporary and will NOT be maintained.

### Links



[1000 Genomes](#) →  
More information about the 1000 Genomes Project on the 1000 genomes main site.



[Pilot browser](#) →  
This browser is based on Ensembl release 60 and represents the variant set analysed as part of [A map of human genome variation from population-scale sequencing](#), Nature 467, 1061.1073.

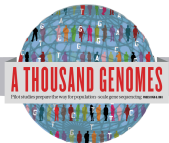
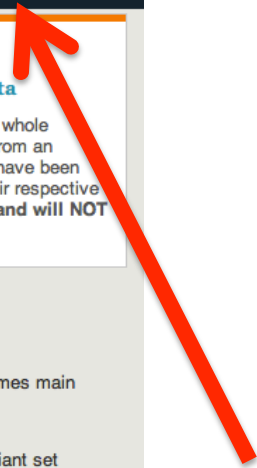


[Tutorial](#) →  
The 1000 Genomes Browser Tutorial.

The 1000 Genomes Project is an international collaborative project described at [www.1000genomes.org](http://www.1000genomes.org).

The 1000 Genomes Browser is based on Ensembl web code.

Ensembl is a joint project of EMBL-EBI  and the [Wellcome Trust Sanger Institute](#)



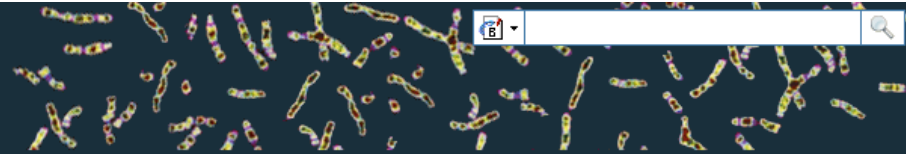
<http://browser.1000genomes.org>



# Tools page

## 1000 Genomes

A Deep Catalog of Human Genetic Variation



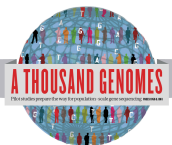
Tools | Help

We provide a number of ready-made tools for processing your data. At the moment, small datasets can be uploaded to our servers and processed online; for larger datasets, we provide an API script that can be downloaded (you will also need to [install our Perl API](#) to use these).

In the near future we aim to offer an intermediate service, whereby medium-to-large data sets can be submitted to a queue, similar to BLAST.

Currently available:

Tool	Description		
Assembly converter	Map your data to the current assembly. Accepted file formats: <a href="#">GFF</a> , <a href="#">GTF</a> , <a href="#">BED</a> , <a href="#">PSL</a> . N.B. Export is currently in GFF only	<a href="#">Online version</a>	<a href="#">API script</a>
ID History converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	<a href="#">Online version</a> (max 30 ids)	<a href="#">API script</a>
Variant Effect Predictor	(Formerly SNP Effect Predictor). Upload a set of SNPs in our <a href="#">standard format</a> and export a file containing consequence types. Uploaded tracks can also be viewed on Location pages.	<a href="#">Online version</a> (max 750 SNPs)	<a href="#">API script</a>
Data Slicer	Get a subset of data from a BAM or VCF file.	<a href="#">Online version</a> (max 10K region)	
Variation Pattern Finder	Identify variation patterns in a chromosomal region of interest for different individuals. Only variations with functional significance such non-synonymous coding, splice site will be reported by the tool. Click <a href="#">here</a> for more extensive documentation.	<a href="#">Online version</a>	<a href="#">API script</a>
VCF to PED converter	The VCF to PED converter allows users to parse a vcf file to create a linkage pedigree file (.ped) and a marker information file, which together may be loaded into Id visualization tools like Haploview. Click <a href="#">here</a> for more extensive documentation.	<a href="#">Online version</a>	<a href="#">API script</a>

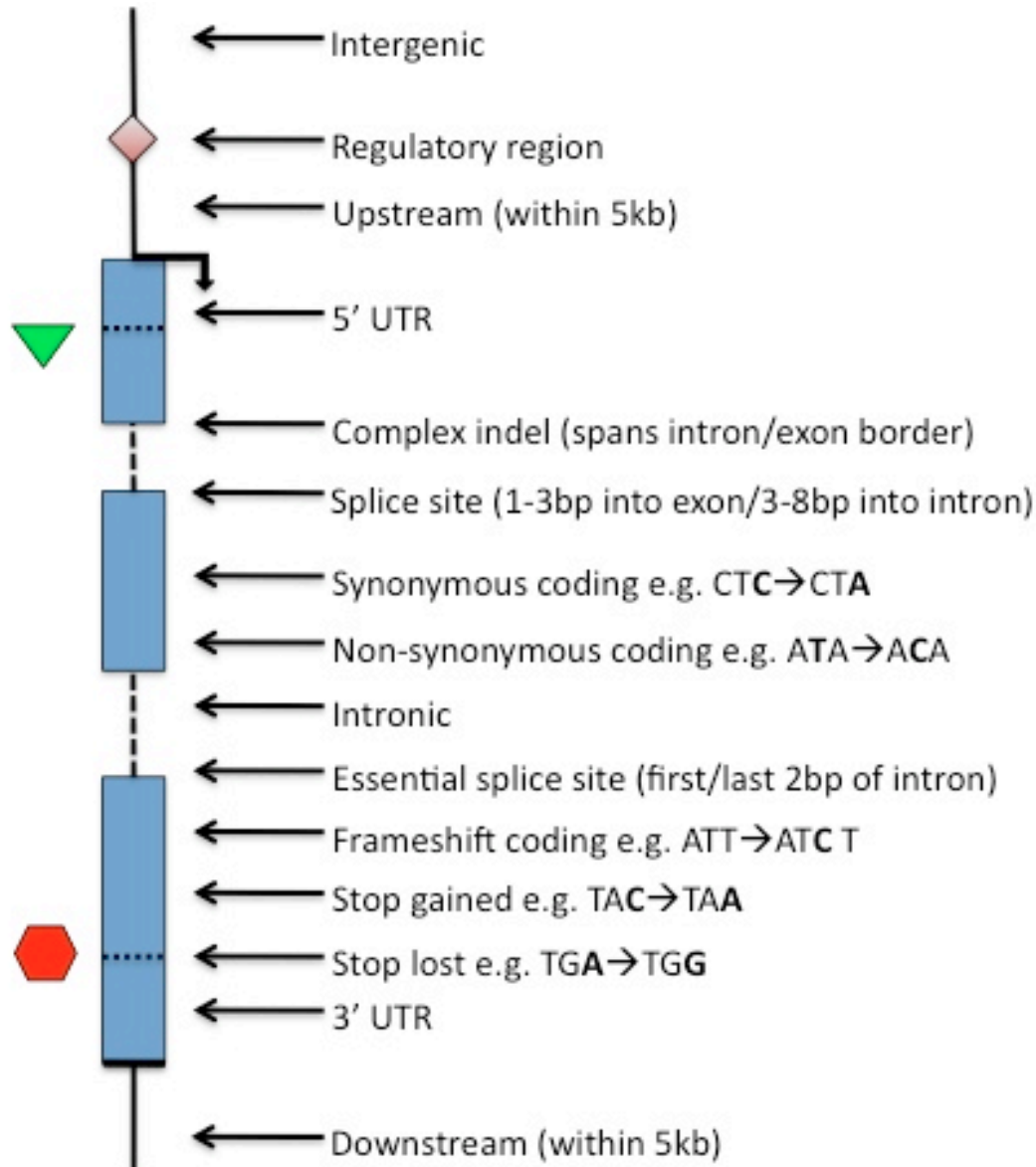


# Variant Effect Predictor

- Predicts Functional Consequences of Variants
- Both Web Front end and API script
- Can provide
  - sift/polyphen/condel consequences
  - Refseq gene names
  - HGVS output
- Can run from a cache as well as Database
- Convert from one input format to another
- Script available for download from:
- [ftp://ftp.ensembl.org/pub/misc-scripts/  
Variant\\_effect\\_predictor/](ftp://ftp.ensembl.org/pub/misc-scripts/Variant_effect_predictor/)
- [http://browser.1000genomes.org/Homo\\_sapiens/  
UserData/UploadVariations](http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations)



# Variant Effect Predictor



Others: Within non-coding gene, Within mature miRNA, NMD transcript

# Variant Effect Predictor

- **perl variant\_effect\_predictor.pl** -input  
6\_381831625\_3184704.vcf -sift p -polyphen p –  
check\_existing
- less variant\_effect\_output.txt

```
#Uploaded_variation Location Allele Gene Feature Feature_type Consequence
cDNA_position CDS_position Protein_position Amino_acids Codons Exi
sting_variation Extra
rs138094825 6:31831667 A ENSG00000204385 ENST00000414427 Transcript
DOWNSTREAM - - - - - rs138094825 -
rs138094825 6:31831667 A ENSG00000204385 ENST00000229729 Transcript
INTRONIC - - - - - rs138094825 -
6_31832657_C/T 6:31832657 T ENSG00000204385 ENST00000229729
Transcript NON_SYNONYMOUS_CODING 1883 1862 621 R/H cGc/cAc -
PolyPhen=possibly_damaging;SIFT=deleterious
```

# Variation Effect Predictor Output

6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204386</a>	<a href="#">ENST00000480384</a>	Transcript	UPSTREAM	-	-	-	-	-	-	
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204386</a>	<a href="#">ENST00000491768</a>	Transcript	UPSTREAM	-	-	-	-	-	-	
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204386</a>	<a href="#">ENST00000375631</a>	Transcript	UPSTREAM	-	-	-	-	-	-	
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204386</a>	<a href="#">ENST00000479533</a>	Transcript	UPSTREAM	-	-	-	-	-	-	
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204385</a>	<a href="#">ENST00000229729</a>	Transcript	NON_SYNONYMOUS_CODING	1625	1604	535	R/H	cGc/cAc	<a href="#">1KG 6 31833357</a>	SIFT=deleterious; PolyPhen=probably_damaging
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204385</a>	<a href="#">ENST00000375562</a>	Transcript	NON_SYNONYMOUS_CODING	1544	1478	493	R/H	cGc/cAc	<a href="#">1KG 6 31833357</a>	SIFT=deleterious; PolyPhen=possibly_damaging
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204385</a>	<a href="#">ENST00000544672</a>	Transcript	NON_SYNONYMOUS_CODING	1673	1376	459	R/H	cGc/cAc	<a href="#">1KG 6 31833357</a>	SIFT=deleterious; PolyPhen=probably_damaging
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204385</a>	<a href="#">ENST00000487680</a>	Transcript	UPSTREAM	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204385</a>	<a href="#">ENST00000414427</a>	Transcript	DOWNSTREAM	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204385</a>	<a href="#">ENST00000479777</a>	Transcript	DOWNSTREAM	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
6_31833357_C/T	<a href="#">6:31833357</a>	T	<a href="#">ENSG00000204385</a>	<a href="#">ENST00000475563</a>	Transcript	DOWNSTREAM	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
<a href="#">0204386</a>	<a href="#">ENST00000491768</a>	Transcript	UPSTREAM	-	-	-	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
<a href="#">0204386</a>	<a href="#">ENST00000375631</a>	Transcript	UPSTREAM	-	-	-	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
<a href="#">0204386</a>	<a href="#">ENST00000479533</a>	Transcript	UPSTREAM	-	-	-	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
<a href="#">0204385</a>	<a href="#">ENST00000229729</a>	Transcript	NON_SYNONYMOUS_CODING	1625	1604	535	R/H	cGc/cAc	<a href="#">1KG 6 31833357</a>	SIFT=deleterious; PolyPhen=probably_damaging			
<a href="#">0204385</a>	<a href="#">ENST00000375562</a>	Transcript	NON_SYNONYMOUS_CODING	1544	1478	493	R/H	cGc/cAc	<a href="#">1KG 6 31833357</a>	SIFT=deleterious; PolyPhen=possibly_damaging			
<a href="#">0204385</a>	<a href="#">ENST00000544672</a>	Transcript	NON_SYNONYMOUS_CODING	1673	1376	459	R/H	cGc/cAc	<a href="#">1KG 6 31833357</a>	SIFT=deleterious; PolyPhen=probably_damaging			
<a href="#">0204385</a>	<a href="#">ENST00000487680</a>	Transcript	UPSTREAM	-	-	-	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
<a href="#">0204385</a>	<a href="#">ENST00000414427</a>	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
<a href="#">0204385</a>	<a href="#">ENST00000479777</a>	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-
<a href="#">0204385</a>	<a href="#">ENST00000475563</a>	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-	<a href="#">1KG 6 31833357</a>	-



# Variation Pattern Finder

- Remote or local tabix indexed VCF input
  - Web version must use remote files
- Discovers patterns of Shared Inheritance
- Variants with functional consequences considered by default
- Web output with CSV and Excel downloads
- [http://browser.1000genomes.org/Homo\\_sapiens/UserData/VariationsMapVCF](http://browser.1000genomes.org/Homo_sapiens/UserData/VariationsMapVCF)



# Variation Pattern Finder

- **perl variant\_pattern\_finder.pl** -vcf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1\_integrated\_calls.20101123.snps\_indels\_svsvs.genotypes.vcf.gz -sample\_panel\_file ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1\_integrated\_calls.20101123.ALL.panel -region 6:31830969-31846823 -expand





# Variation Pattern Finder Output

freq	6:31833647_[T]	6:31833660_rs6915800[G]	samples	
freq	ENST00000414427- SPlice_SITE[],ENST0000054 4672- SPlice_SITE[],ENST0000022 9729- SPlice_SITE[],ENST0000037 5562-SPlice_SITE[]	ENST00000414427- NON_SYNONYMOUS_CODING[R/ C],ENST00000229729- NON_SYNONYMOUS_CODING[R/ C],ENST00000544672- NON_SYNONYMOUS_CODING[R/ C],ENST00000375562- NON_SYNONYMOUS_CODING[R/C]	samples	
0.73	REF REF	G A	YRI(3)	NA18933, NA19149, NA19098 and 0 others.
0.27	REF REF	A G	YRI(2)	NA19146, NA19198
0.18	REF REF	A A	LWK(1)	NA19372
0.09	C T	REF REF	CHB(1)	NA18592



# Variation Pattern Finder Output

## Variation Pattern Finder

Export data: [CSV](#) [Excel](#)

**Go to collapsed view**

CEU	CI Freq	rs12661281:T/A	6:31843711:C/T	6:31845340:C/T	rs2075798:C/A
		6:31842598	6:31843711	6:31845340	6:31846741
		DING:N/S ENST00000229729 NON_SYNONYMOUS_CODING:D/V	ENST00000229729 SPLICE_SITE	ENST00000544672 SPLICE_SITE	ENST00000229729 NON_SYNONYMOU
		DING:N/S ENST00000544672 NON_SYNONYMOUS_CODING:D/V	ENST00000375562 SPLICE_SITE	ENST00000544672 5PRIME_UTR	ENST00000375562 NON_SYNONYMOU
		DING:N/S ENST00000414427 NON_SYNONYMOUS_CODING:D/V	ENST00000544672 SPLICE_SITE		ENST00000414427 NON_SYNONYMOU
			ENST00000414427 SPLICE_SITE		
			ENST00000465707 SPLICE_SITE		
			ENST00000462671 SPLICE_SITE		
NA12872, NA07000 and 1 other(s)	N 0.032	TIA	CIC	CIC	CIC
NA12874, NA12717	N 0.028	TIT	CIC	CIC	AIC
NA07346	N 0.027	TIT	CIC	CIC	CIA
	N 0.027	TIT	CIC	CIC	CIC
NA10851, NA12342 and 5 other(s)	N 0.024	AIT	CIC	CIC	CIC
NA12058, NA12273 and 1 other(s)	N 0.020	AIA	CIC	CIC	CIC
	N 0.018	TIT	CIC	CIC	CIC
	N 0.015	AIT	CIC	CIC	CIA
	N 0.014	TIT	CIC	CIC	AIA
	N 0.013	TIT	CIC	CIC	CIC
NA10847	N 0.011	TIA	CIC	CIC	AIC
NA12286, NA11892 and 2 other(s)	N 0.009	TIT	CIC	CIC	CIC

# VCF to PED

- LD Visualization tools like Haploview require PED files
- VCF to PED converts VCF to PED
- Will a file divide by individual or population
- [http://browser.1000genomes.org/Homo\\_sapiens/UserData/Haploview](http://browser.1000genomes.org/Homo_sapiens/UserData/Haploview)



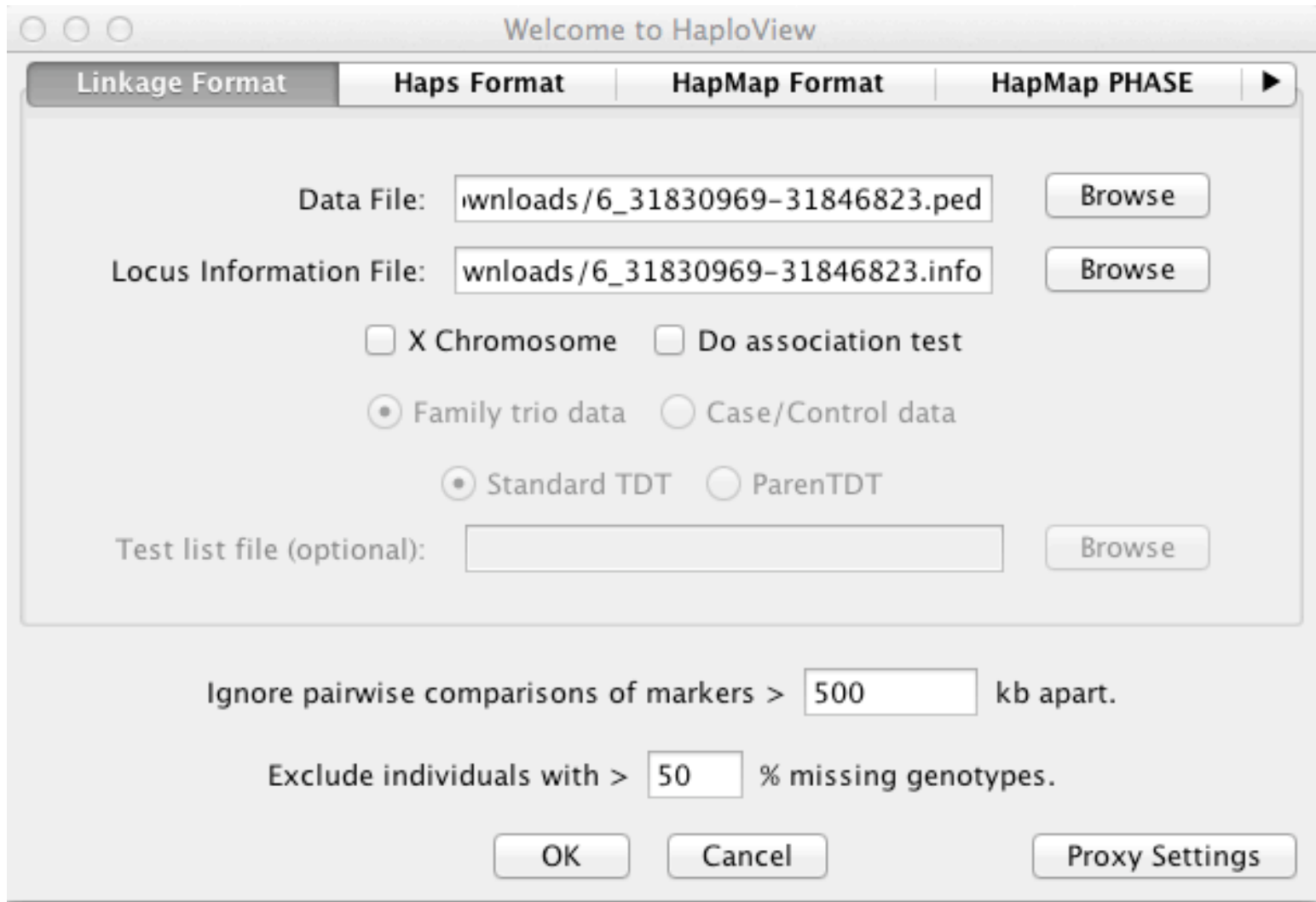
# VCF to PED

- **perl vcf\_to\_ped\_convert.pl** -vcf [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1\\_integrated\\_calls.20101123.snps\\_indels\\_svsvs.genotypes.vcf.gz](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_integrated_calls.20101123.snps_indels_svsvs.genotypes.vcf.gz) -sample\_panel\_file [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1\\_integrated\\_calls.20101123.ALL.panel](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel) -region **6:31830969-31846823** -population **CEU**
- Output should be two files
- 6\_31830969-31846823.info
- 6\_31830969-31846823.ped



# Haplotype example input

```
java -jar Haploview.jar
```



Welcome to HaploView

Linkage Format | **Haps Format** | HapMap Format | HapMap PHASE ▶

Data File:

Locus Information File:

X Chromosome     Do association test

Family trio data     Case/Control data

Standard TDT     ParentTDT

Test list file (optional):

Ignore pairwise comparisons of markers >  kb apart.

Exclude individuals with >  % missing genotypes.





# Access to backend Ensembl databases

- Public MySQL database at
  - `mysql-db.1000genomes.org` port 4272
- Full programmatic access with Ensembl API
  - The 1000 Genomes Pilot uses Ensembl v60 databases and the NCBI36 assembly (this is frozen)
  - The 1000 Genomes main project currently uses Ensembl v63 databases
- <http://jun2011.archive.ensembl.org/info/docs/api/variation/index.html>
- <http://www.ensembl.org/info/docs/api/variation/index.html>
- <http://www.1000genomes.org/node/517>



# Amazon Web Service Cloud

- 1000 Genomes Alignments and Variant files are available in AWS
- AMI image available to run 1000 Genomes Tutorial
- <http://www.1000genomes.org/using-1000-genomes-data-amazon-web-service-cloud>





## Exercises, Command Line Tools

7. Get the 7:114304000-114305000 (FoxP2 exon) section of the 20110521 release chr 7 genotypes file
8. Use vcftools vcf-stats to specify which SNP transition happens most in this section. You should look at the statistics for all variants/samples.
9. Use this piece with tools, the variant effect predictor, the vcf pattern finder
10. Are there any snps with deleterious sift/polyphen consequences?
11. What is the most common pattern of variation in this region?
12. Use the vcf to ped script with 6:31830700-31840700
13. How many different haplotype blocks does the section contain?



# Exercise Answers, Command Line Tools

```
> tabix -h ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/
20110521/
ALL.chr7.phase1_release_v3.20101123.snps_indels_svsvs.g
notypes.vcf.gz 7:114304000-114305000 > 20110521.vcf
> vcf-stats 20110521.vcf
```

```
'all' => {
  'snp_count' => 16,
  'count' => 16,
  'snp' => {
    'A>C' => 1,
    'A>T' => 1,
    'G>T' => 1,
    'A>G' => 2,
    'T>G' => 1,
    'T>C' => 2,
    'C>T' => 3,
    'G>A' => 5
  },
  'nalt_1' => 16
}
```



# Exercise Answers, Command Line Tools

```
>perl variant_effect_predictor.pl -input ~/20110521.vcf -sift p  
-polyphen p --force_overwrite
```

```
> grep SIFT variant_effect_output.txt
```

```
> rs182138317 7:114304331 A
```

```
ENSG00000128573 ENST00000393489 Transcript
```

```
NON_SYNONYMOUS_CODING 1949 1567 523 A/T
```

```
Gcc/Acc -
```

```
PolyPhen=possibly_damaging;SIFT=deleterious
```



# Exercise Answers, Command Line Tool

```
> perl variant_pattern_finder.pl -vcf ~/20110521.vcf -sample  
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/  
phase1_integrated_calls.20101123.ALL.panel -region  
7:114304000-114305000
```

This produces a tsv file which can be view in a spreadsheet program

```
7:114304563_rs1378771[C] 7:114304630_rs1378772[A] 7:114304969_rs2396765[T]  
> 14.38 - - C|T - A|T - - - - - T|C TSI(30)
```



# Exercise Answers: Command Line Tools

```
> perl vcf_to_ped_convert.pl -vcf 20110521.vcf -sample  
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/  
phase1_integrated_calls.20101123.ALL.panel -region  
6:31830700-31840700 -population CEU
```

```
> ls ./
```

```
> 6_31830700-31840700.info
```

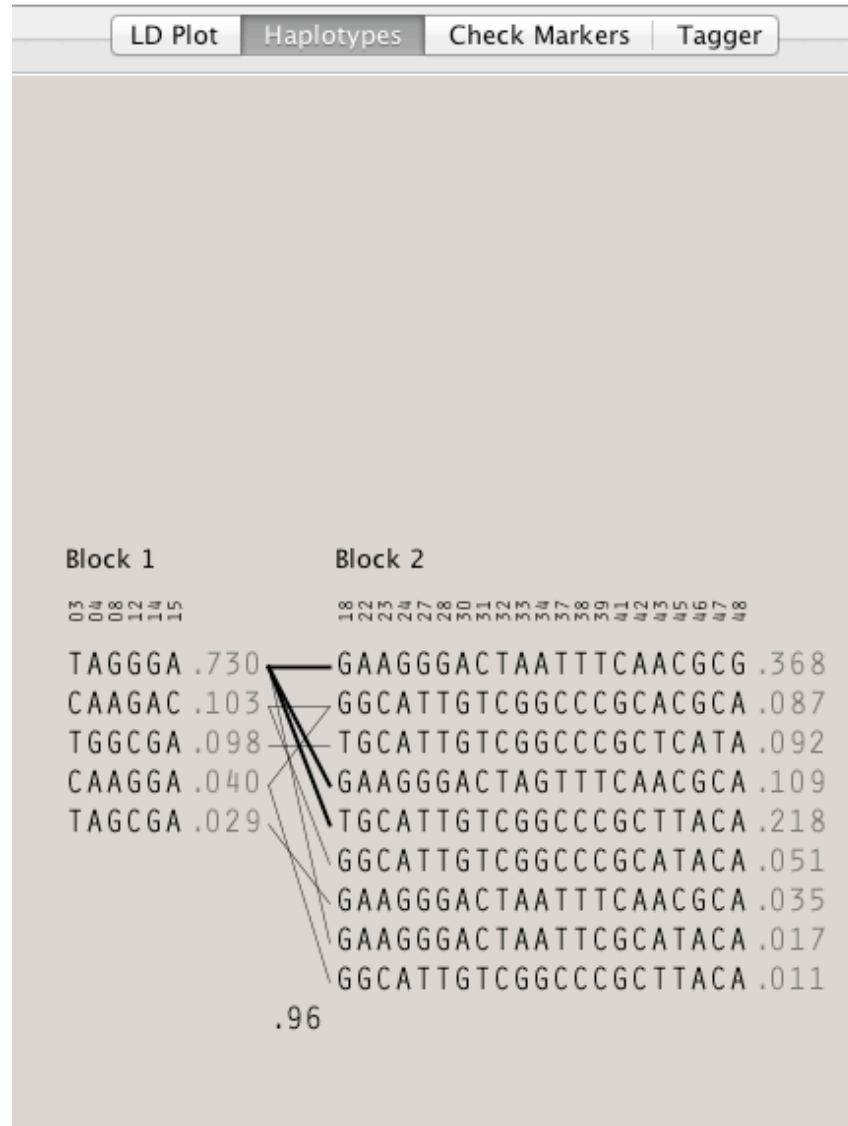
```
> 6_31830700-31840700.ped
```



# Exercise Answers



# Exercise Answers



# 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](mailto:1000announce@1000genomes.org)
- <http://www.1000genomes.org/1000-genomes-announcement-mailing-list>
- <http://www.1000genomes.org/announcements/rss.xml>
- <http://twitter.com/#!/1000genomes>



# Questions

Please send any future questions about this presentation and any other material on our website to [info@1000genomes.org](mailto:info@1000genomes.org)



<http://www.1000genomes.org/using-1000-genomes-data>



# 1000 Genomes Community Meeting

- University of Michigan, Ann Arbor on the 12th and 13th of July 2012
- Showcase Advances made by the Project
- Generate Discussion about the next round of Human Genome Sequencing
- Registration closes May 15th
- <http://1000gconference.sph.umich.edu/>



# Thanks

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



# Questionnaire

<http://goo.gl/VxK2c>

