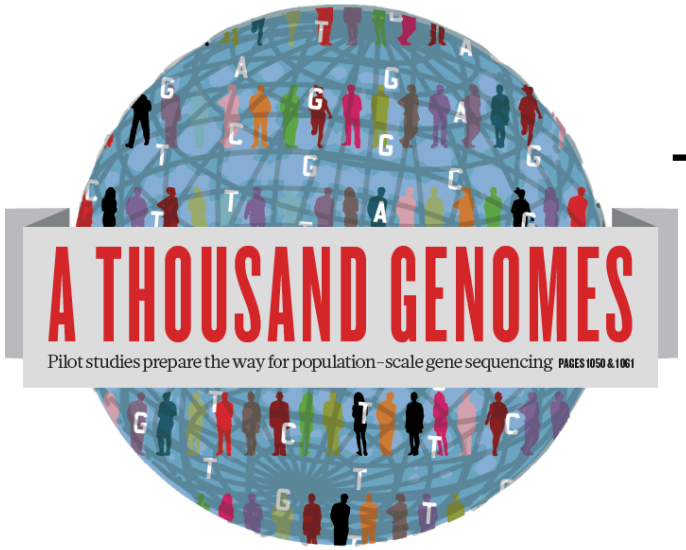



# The 1000 Genomes Project Tutorial

11th April 2012

Laura Clarke

EMBL-EBI

# Updates to slides

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120410_tutorial_docs/

# 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**: In 1000 genomes file/directory names dates are mostly represented as YYYYMMDD

# Outline

## Morning

- Introduction to the Project
- Data Availabilty and the FTP Site
- Exercise, Finding data
- The Browser
- Exercise, Browsing
- The Tools,
- Exercise Tool use

## Afternoon

- File Formats
- Exercise, Finding Data
- The Command Line Tools
- Exercise, Tool use

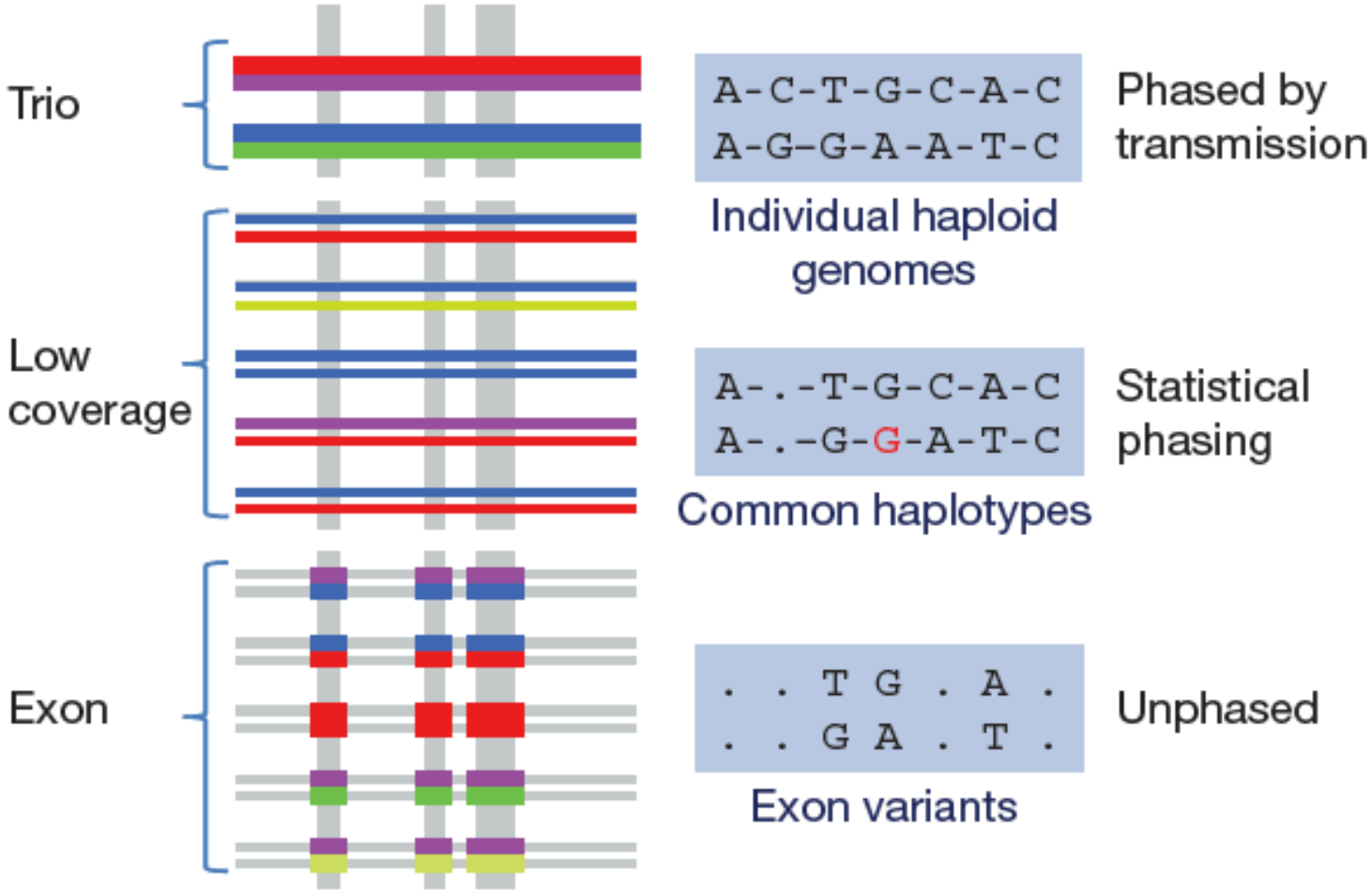
# Introduction to the Project


EMBL-EBI

# 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

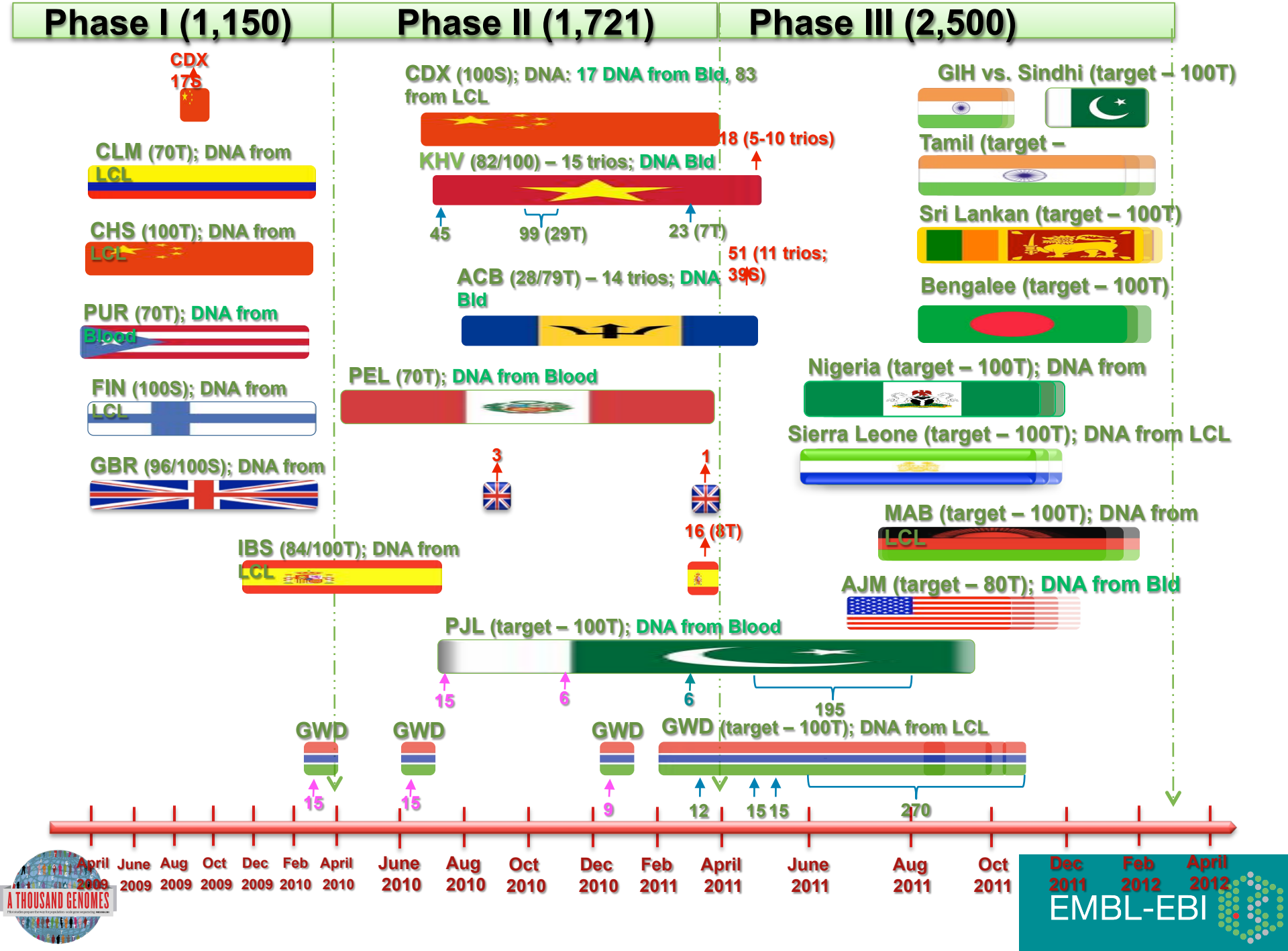
# 3 pilot coverage strategies




EMBL-EBI

# 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 175
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
CDX (100S); DNA: 17 DNA from Bld, 83 from LCL
18 (5-10 trios)
KHV (82/100) – 15 trios; DNA Bld
45
99 (29T)
23 (7T)
51 (11 trios; 39S)
ACB (28/79T) – 14 trios; DNA Bld
PEL (70T); DNA from Blood
3
1
16 (8T)
PJL (target – 100T); DNA from Blood
15
6
6
195
GWD
GWD
GWD
GWD (target – 100T); DNA from LCL
15
15
9
12
15 15
270
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
April 2009
June 2009
Aug 2009
Oct 2009
Dec 2009
Feb 2010
April 2010
June 2010
Aug 2010
Oct 2010
Dec 2010
Feb 2011
April 2011
June 2011
Aug 2011
Oct 2011
Dec 2011
Feb 2012
April 2012
A THOUSAND GENOMES
EMBL-EBI

# 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 36.6millon snps, 1.5millon 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

A THOUSAND GENOMES

EMBL-EBI

# 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

A THOUSAND GENOMES

EMBL-EBI

# 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 not in dbSNP |
|---|---|
| February, 2000 | 98% |
| February, 2001 | 80% |
| April, 2008 | 10% |
| February, 2011 | 2% |
| Now | <1% |

Ryan Poplin, David Altshuler

A THOUSAND GENOMES

EMBL-EBI

# 1000 Genomes Project: Present & Future

- First Phase 2 sequence release 14th November 2011
- First Phase 2 alignment release 12th 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

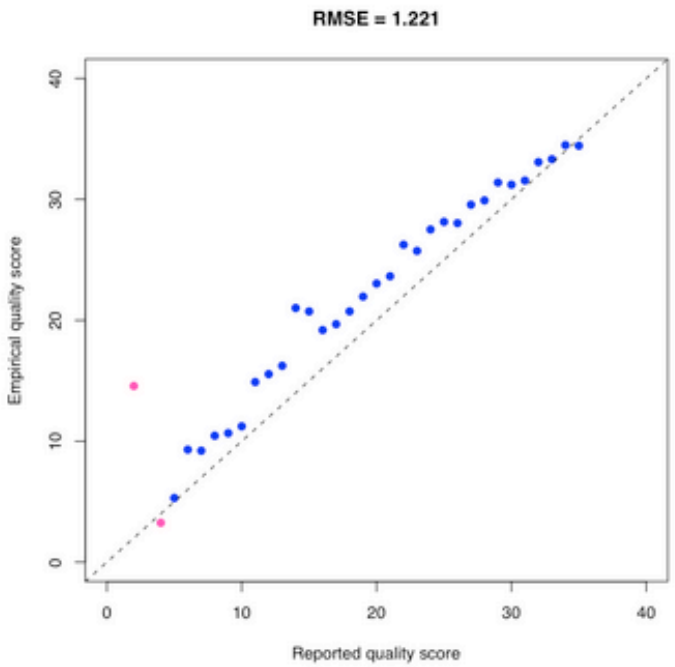
| Platform | Strategy | Aligner | Centre |
|---|---|---|---|
| Solid | Low Coverage | Bfast | TGEN |
| | Exome | Bfast | Baylor |
| Illumina | Low Coverage | BWA | Sanger |
| | Exome | Mosaik | Boston College |
| 454 | Low Coverage | SSAHA | Sanger |

# Base Quality Score Recalibration

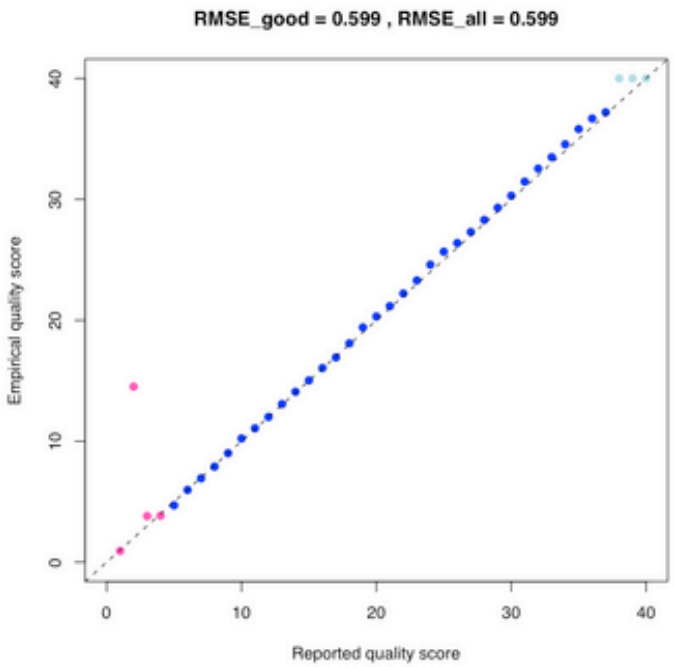
- 1000 Genomes Sequence Data is sourced from many different machines across many different institutes
- Each machine may assign Base Quality Values differently
- Base Quality Score Recalibration tests empirical error rates
  - Run alignment
  - Compare mismatches to know variation
- Base Qualities adjusted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

# 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 | | |

# 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%<br>2.09M / 2.12M | 16.31%<br>9,739 / 59,721 |
| 2 of 6 | 39.11M | 22.24% | 30.41M | 2.153 | 98.55%<br>2.09M / 2.12M | 11.23%<br>6,707 / 59,721 |
| 3 of 6 | 35.69M | 23.62% | 27.26M | 2.219 | 98.09%<br>2.08M / 2.12M | 3.66%<br>2,184 / 59,721 |
| 4 of 6 | 32.55M | 24.82% | 24.48M | 2.263 | 97.39%<br>2.06M / 2.12M | 1.82%<br>1,085 / 59,721 |
| 5 of 6 | 28.45M | 26.72% | 20.85M | 2.286 | 95.93%<br>2.03M / 2.12M | 1.06%<br>634 / 59,721 |
| Intersection | 24.02M | 27.57% | 17.40M | 2.317 | 89.23%<br>1.89M / 2.12M | 0.76%<br>457 / 59,721 |
| | | | | | | |
| **VQSR Project Consensus** | **38.88M** | **21.92%** | **30.36M** | **2.154** | **98.41%**<br>**2.08M / 2.12M** | **2.11%**<br>**1,261 / 59,721** |

A THOUSAND GENOMES

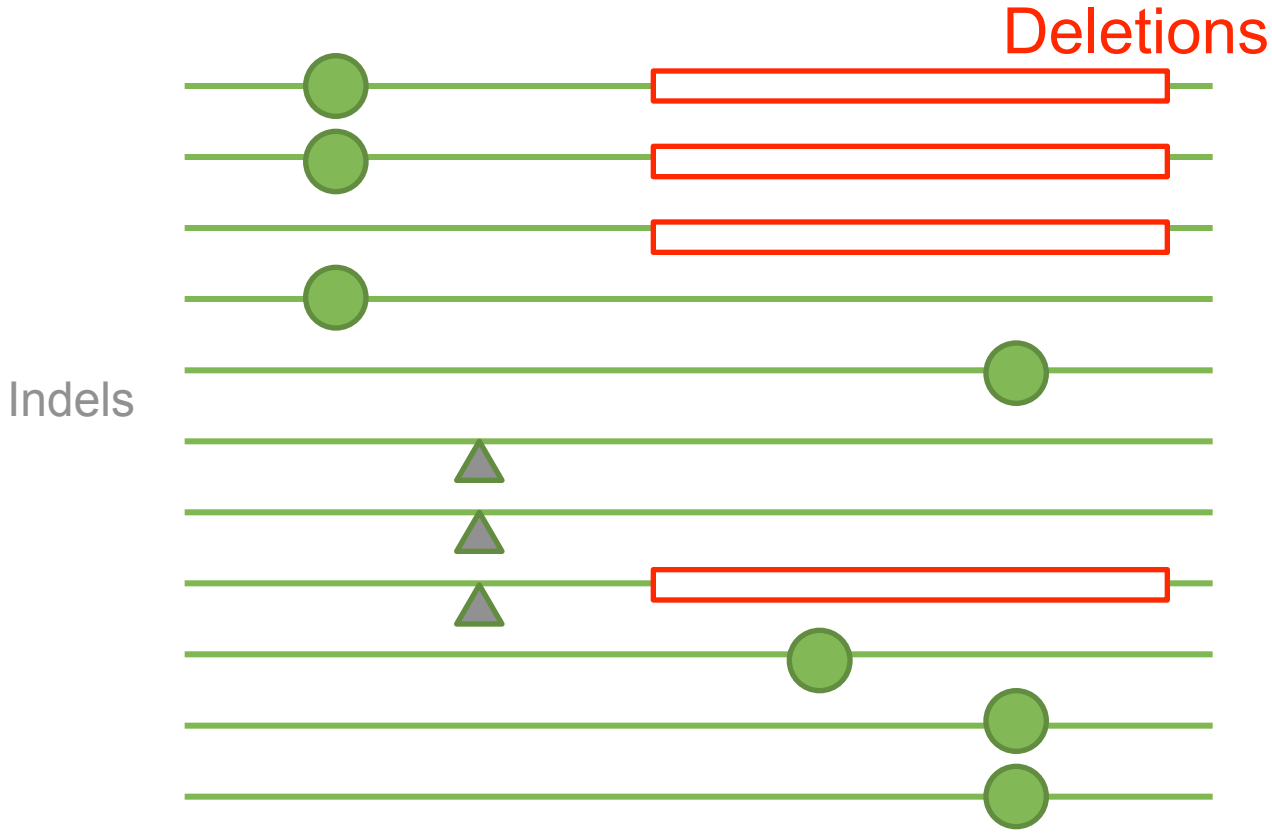
EMBL-EBI

# 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



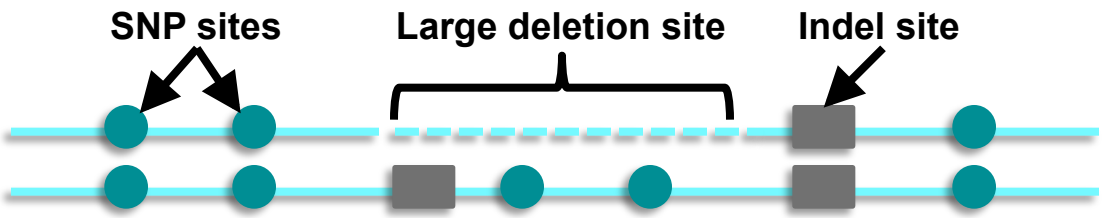

Goncalo Abecasis

SNPs (from LC, EX, OMNI)

A THOUSAND GENOMES

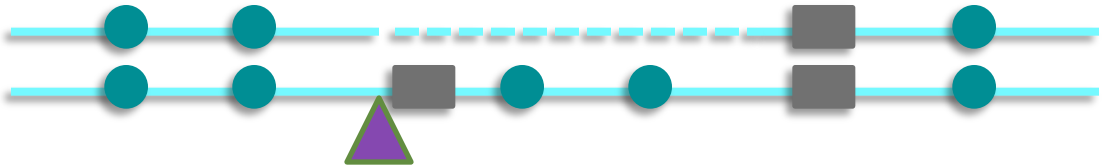
EMBL-EBI

# Strategies for integrating deletions with other types of variation

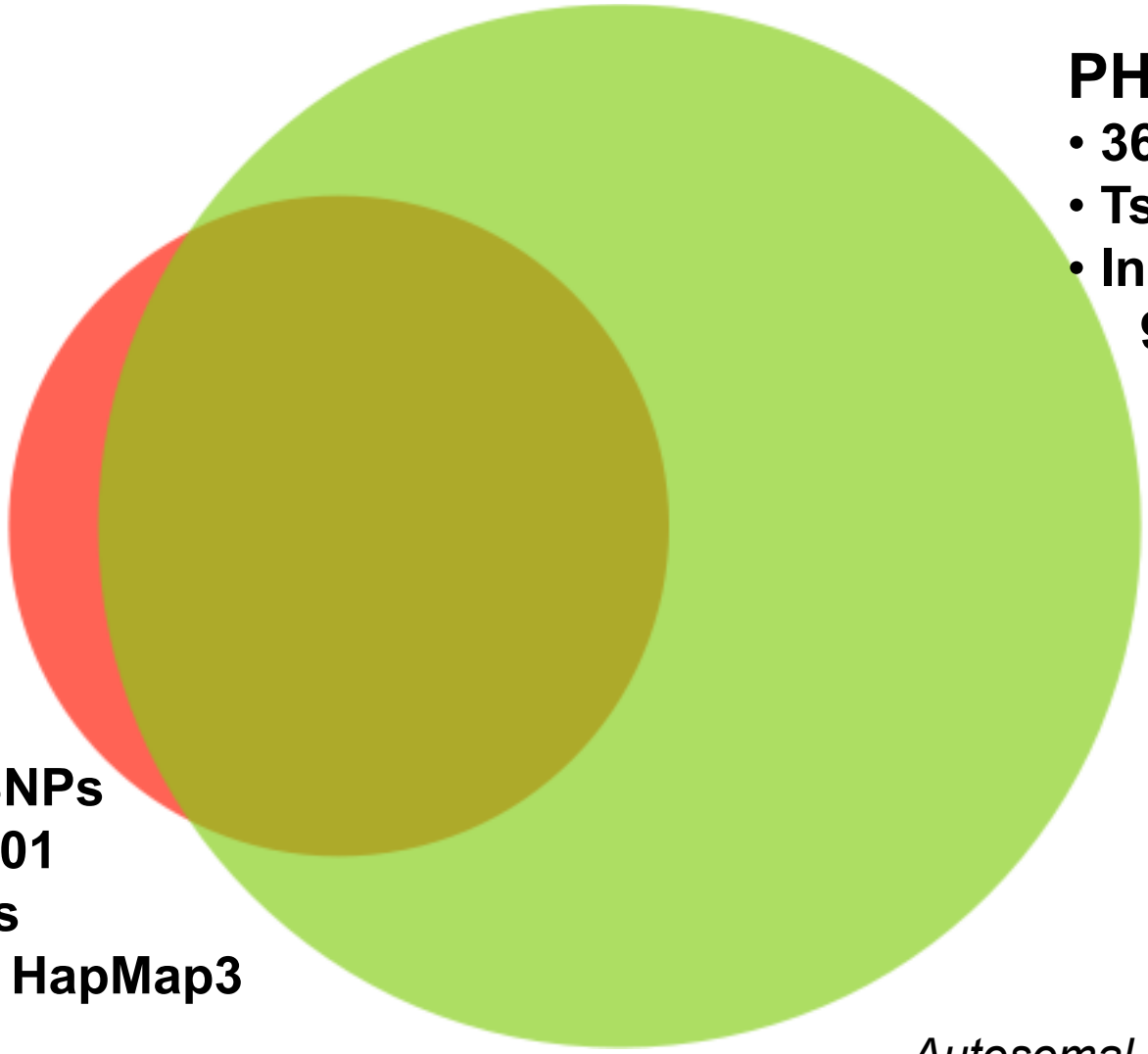



**<u>Previous Approach</u>**
**Remove SNPs under SVs for imputation**
**(1000G pilot, Handsaker et al., 2010)**

**<u>Current Approach</u>**
**Treat SVs as point events**
**(1000 Genomes phase 1)**



# From PILOT to PHASE1



**PHASE1**
- **36.8M SNPs**
- **Ts/Tv 2.17**
- **Includes**
  **98.9% HapMap3**

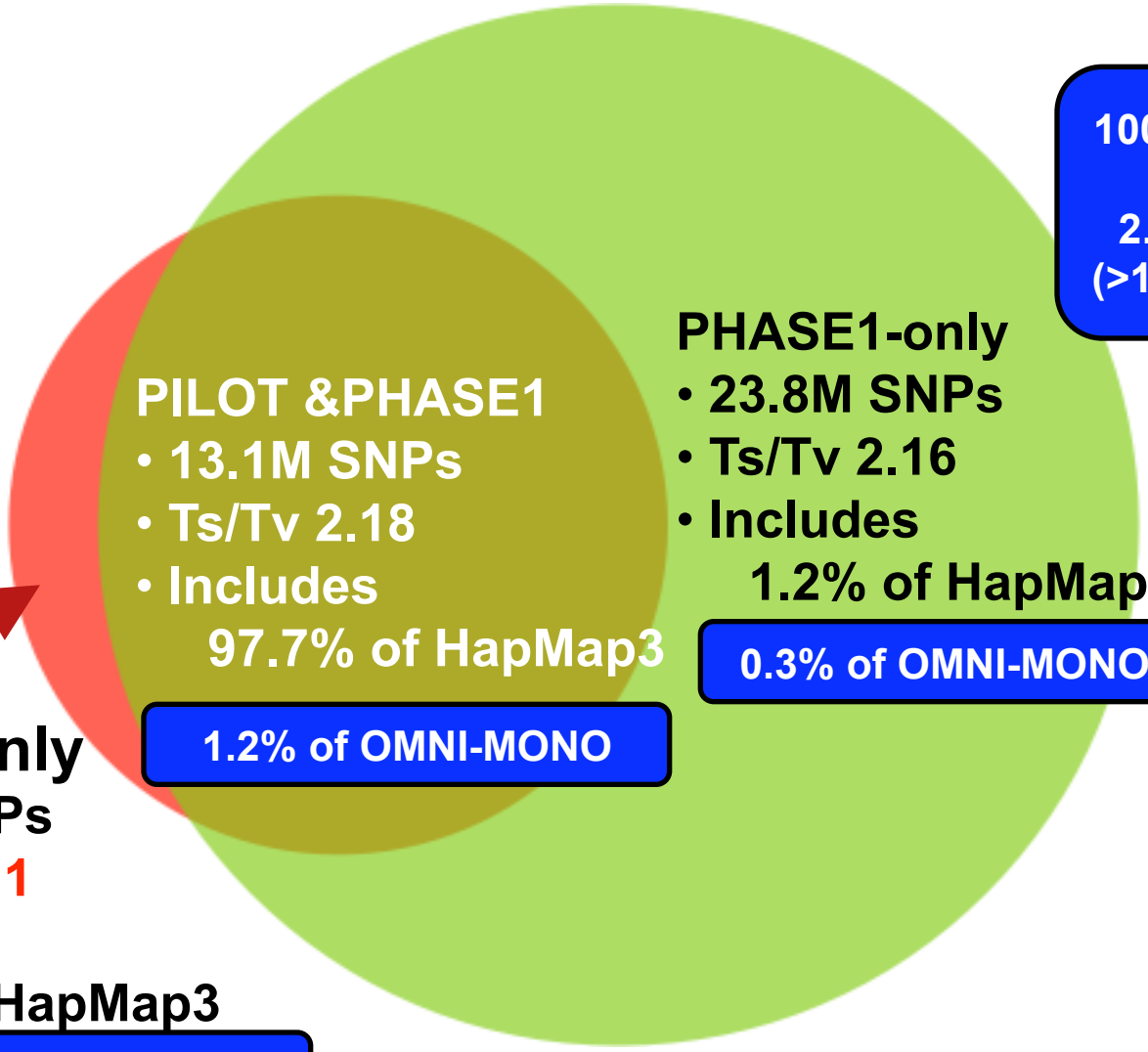
**PILOT**
- **14.8M SNPs**
- **Ts/Tv 2.01**
- **Includes**
  **97.8% HapMap3**

*Autosomal chromosomes only*

A THOUSAND GENOMES

EMBL-EBI

# From PILOT to PHASE1 : Improved SNP calls



**100k monomorphic SNPs in 2.5M OMNI Array (>1,000 individuals)**

**PILOT &PHASE1**
- 13.1M SNPs
- Ts/Tv 2.18
- Includes 97.7% of HapMap3

**1.2% of OMNI-MONO**

**PHASE1-only**
- 23.8M SNPs
- Ts/Tv 2.16
- Includes 1.2% of HapMap3

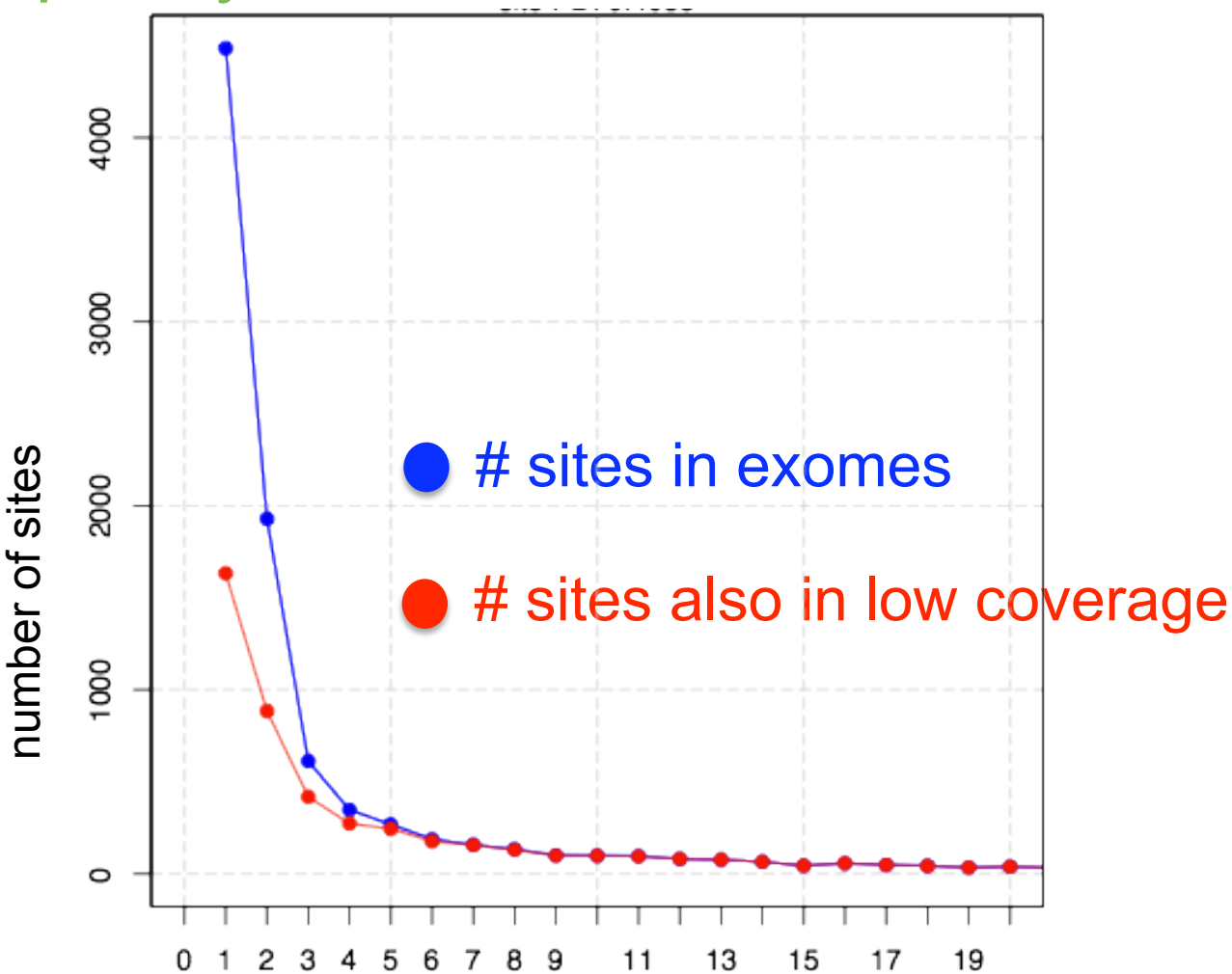
**0.3% of OMNI-MONO**

**PILOT-only**
- 1.7M SNPs
- Ts/Tv 1.11
- Includes 0.15% HapMap3

**59.6% of OMNI-MONO**

*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)

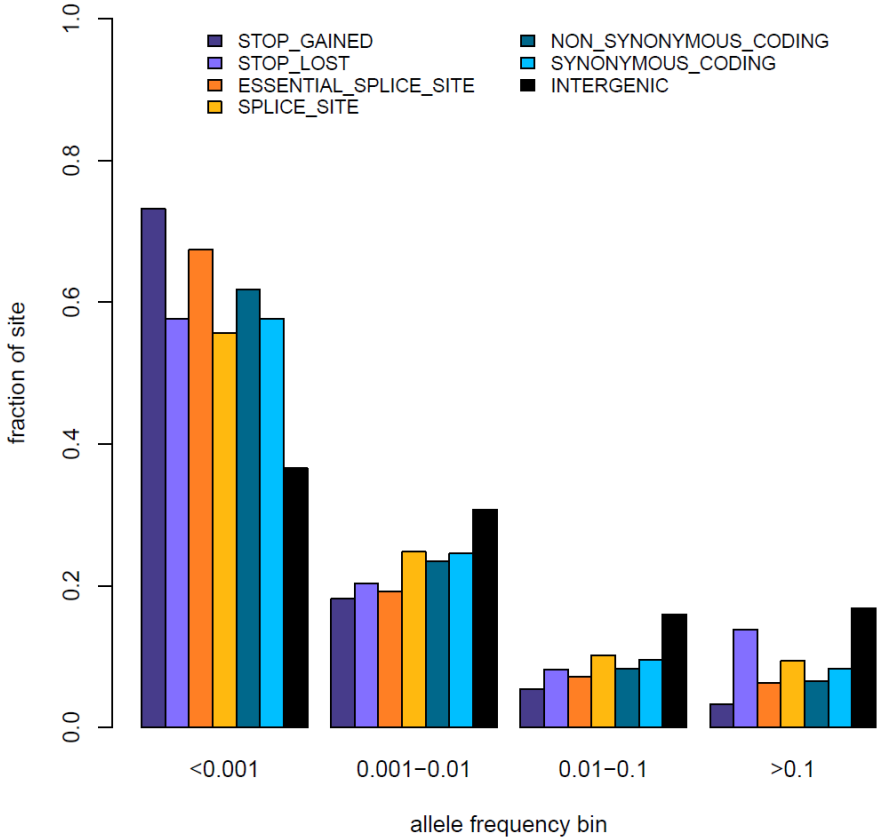
Erik Garrison

A THOUSAND GENOMES

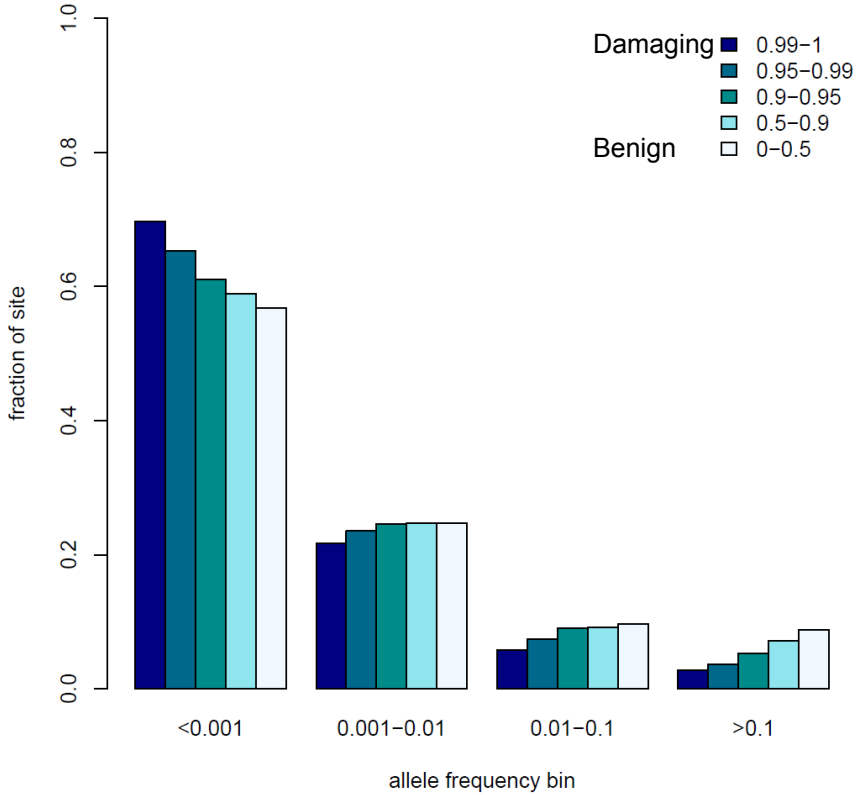
EMBL-EBI

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

## Functional category




## Non-synonymous: Condel score



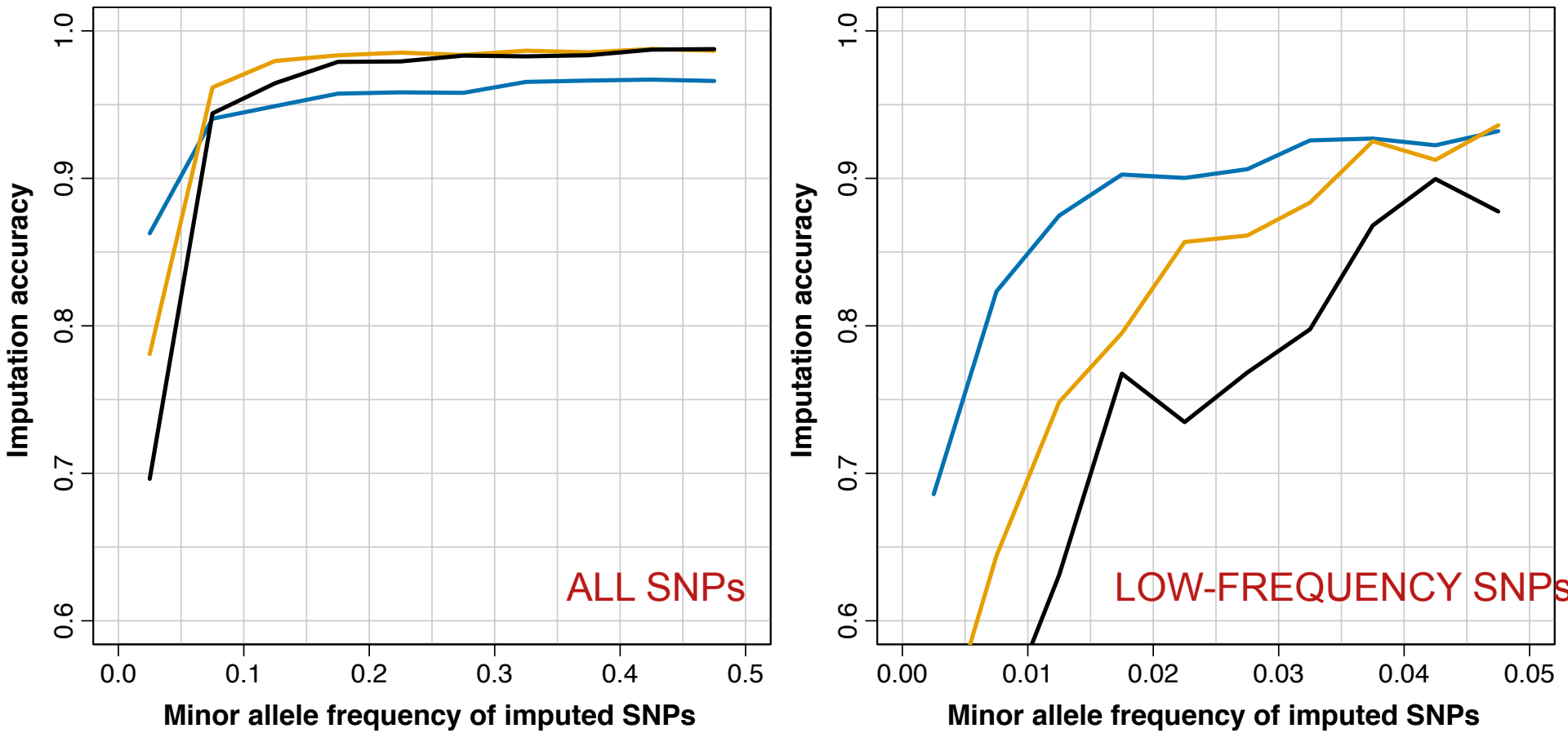

Presentation on using the data for GWAS by Brian Howie

Enza Colonna, Yuan Chen, Yali Xue

A THOUSAND GENOMES

EMBL-EBI

# 1,000 Genomes haplotypes are highly accurate




European ancestry

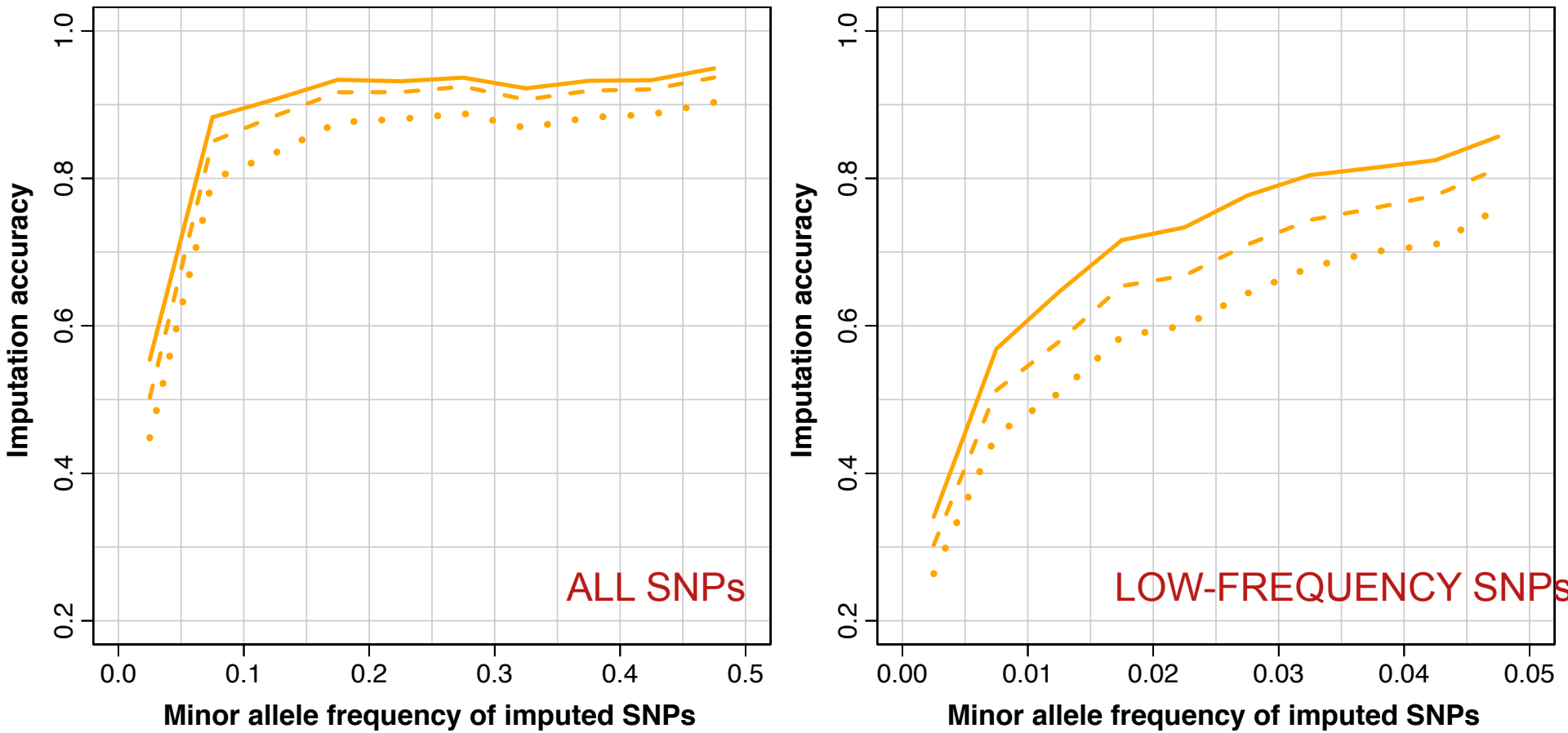
African ancestry

Admixed (Americas)

A THOUSAND GENOMES

EMBL-EBI

# Imputation accuracy depends on your GWAS chip




Omni 2.5M

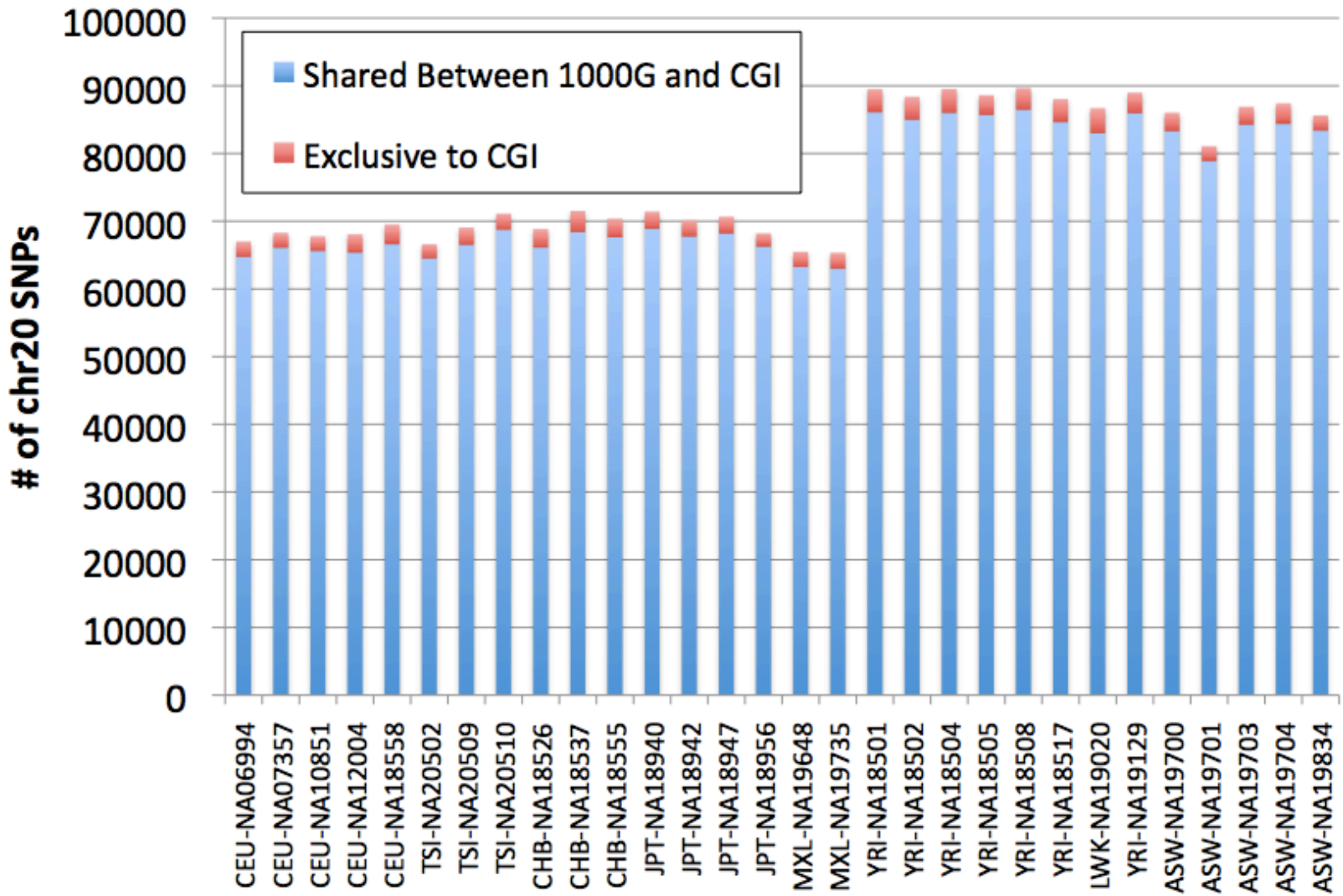
Illumina 550k

Affymetrix 500k

A THOUSAND GENOMES

EMBL-EBI

# >96% SNPs are detected compared to deep genomes




EMBL-EBI

# Data Availability and the FTP site


EMBL-EBI

# File Formats

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

# More Information About BAM Files

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

## 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

**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

EMBL-EBI

# 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

**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

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 |

# 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

# 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

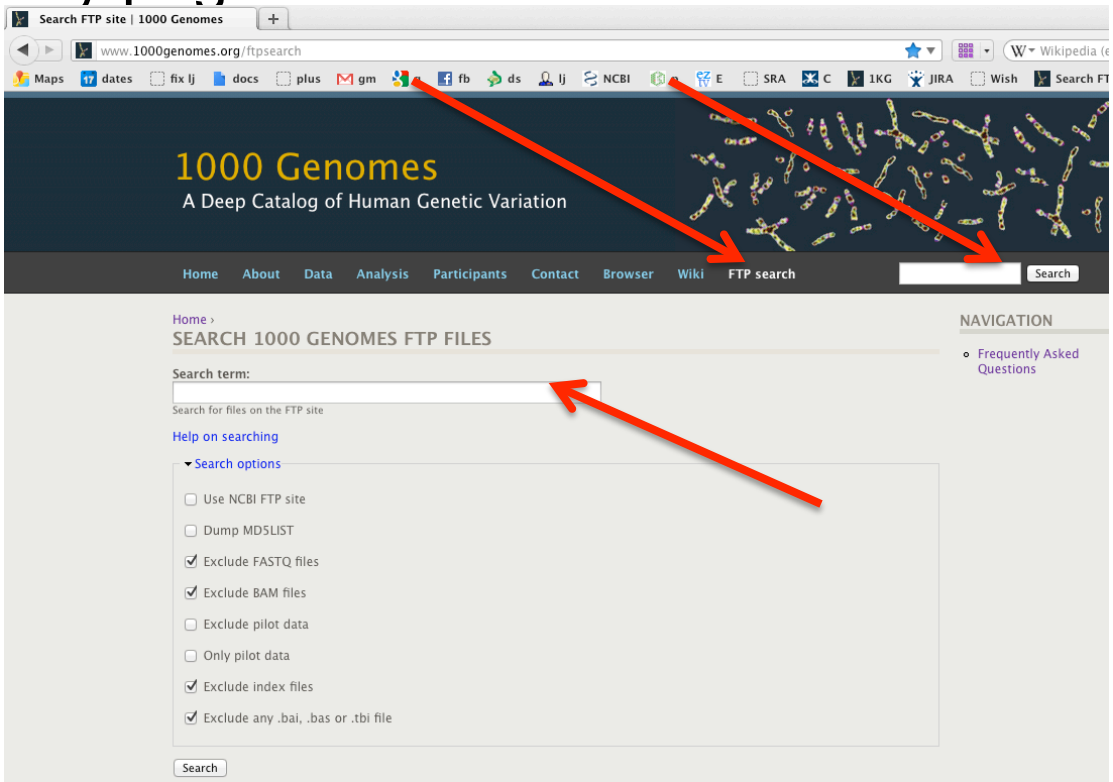
```
ftp     directory       403     Tue Dec 20 16:11:25 2011
ftp/README.ftp_structure        file    8408    Mon Apr  4 14:52:52 2011        2a59a3feb2540c113e10877f3ef1efe5
ftp/README.populations  file    1506    Wed Jan 11 15:12:44 2012        f7c588af82396013c1737e66e58f0f05
ftp/CHANGELOG   file    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

- 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 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

1a. Find what Omni VCF files we have on our ftp site using the website ftp search. (Omni is a high throughput genotyping platform from Illumina on which all 1000 genomes samples are being genotyped)

1b. Find the most recent Omni VCF file on GRCh37 from the 31st January 2012

2. 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.

# Exercise Answers

1a. Put omni*vcf into the ftp site search box

Home ›

SEARCH 1000 GENOMES FTP FILES

Search term:

omni*vcf

Search for files on the FTP site

Help on searching

Search options

Search

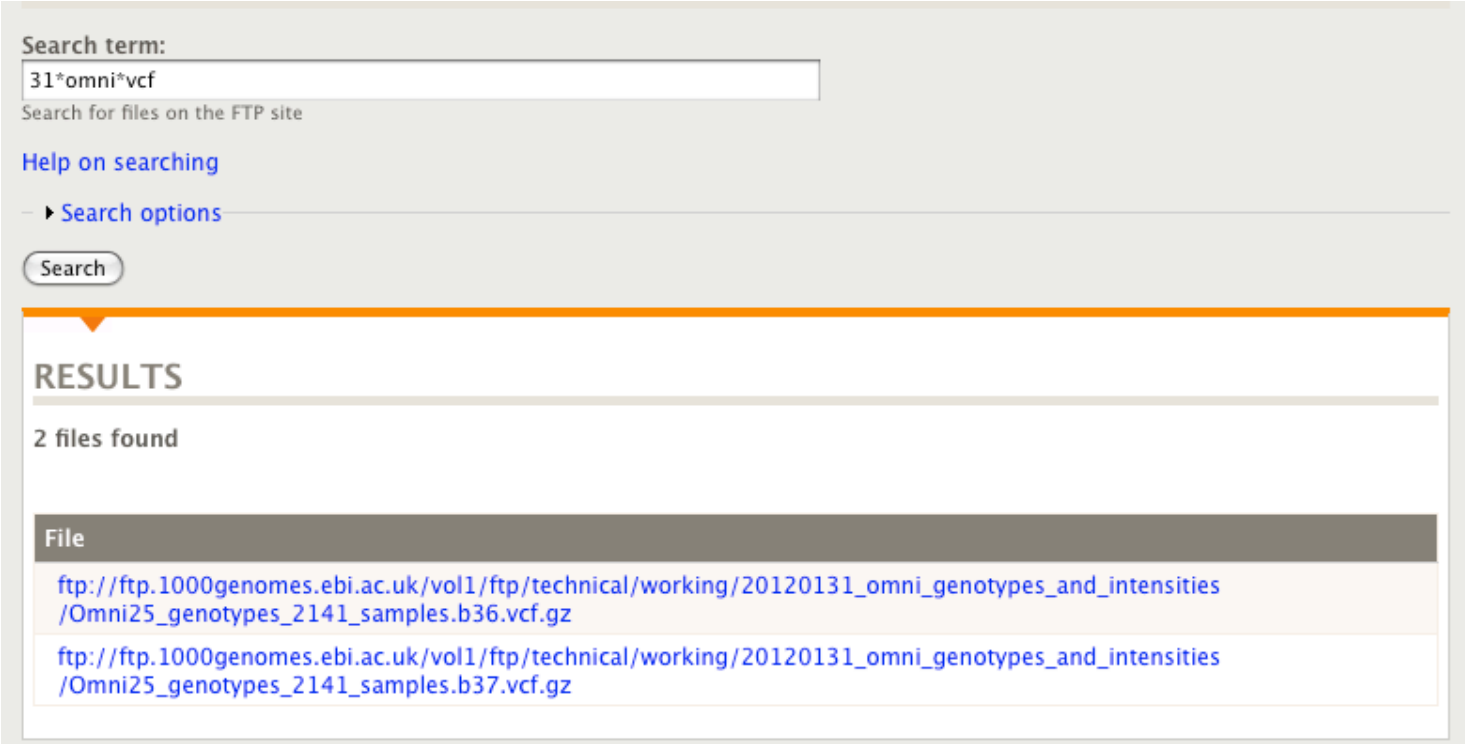
RESULTS

52 files found

| File |
| --- |
| ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b36.vcf.gz |
| ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz |

# Exercise Answers, Finding Data

1b. Use 31*omni*vcf to get results. This should return 2 files. One is labeled b36 and it in NCBI36 coordinates. The other is labeled b37 and is on GRCh37



# Exercise Answers, Finding Data

2. Using the box that is in the top right hand corner of every page of 1000genomes.org with the term sub-section and vcf should return the appropriate FAQ page

Home › Search ›

Content | Users

Enter your keywords:

vcf sub-section [Search]

Advanced search

**Search results**

How do I get a sub-section of a vcf file?

... (Data Access, tabix, tools, variants, **vcf**) ...

FAQ Question - ripley - 2011-10-28 13:43 - 0 comments - 0 attachments

Update to 20110521 Release

... SNPs, short indels and large deletions. Files are in **VCF** format, The sites file represents all the autosomes and chrX but the ... as haploid. The .tbi file associated with each gzipped **vcf** file can be used to extract data for arbitrary chromosome subintervals. ... FAQ http://www.1000genomes.org/faq/how-do-i-get-**sub-section-vcf**-file The **VCF** File is in format 4.1 ...

A THOUSAND GENOMES

EMBL-EBI

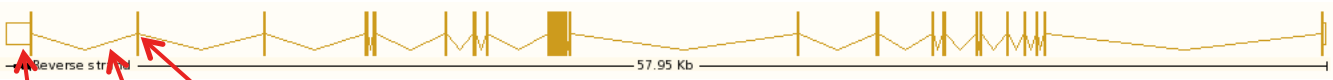
# The 1000 Genomes Browser
# http://browser.1000genomes.org

# Caveats

- 1000 Genomes and Ensembl always define variants on the forward strand

- Allele strings are always reported ref/alt

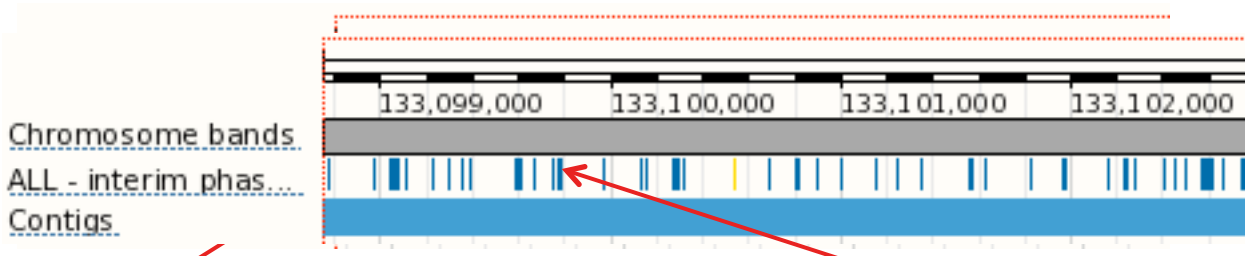
# Genes and SNPs




UTR

Intron

Coding



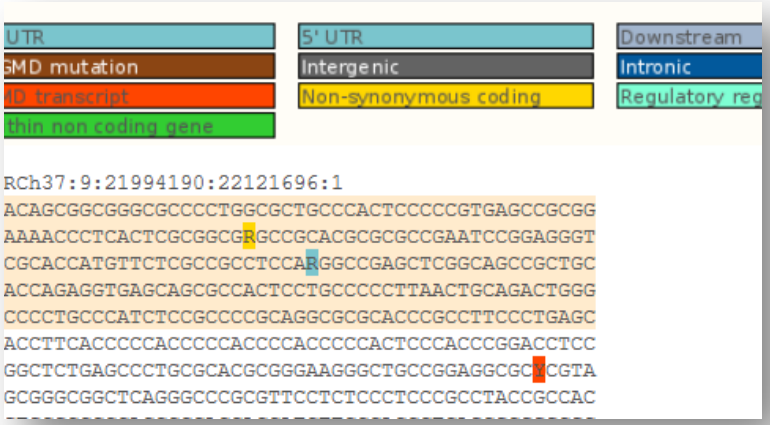

Line indicates number of SNPS

Each Line is One SNP

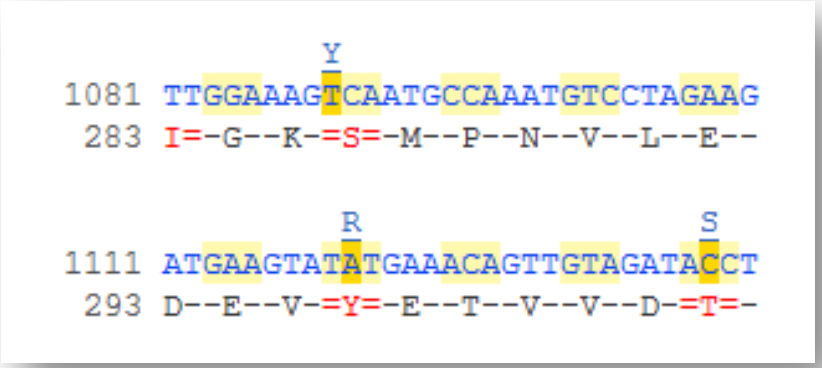
A THOUSAND GENOMES

EMBL-EBI

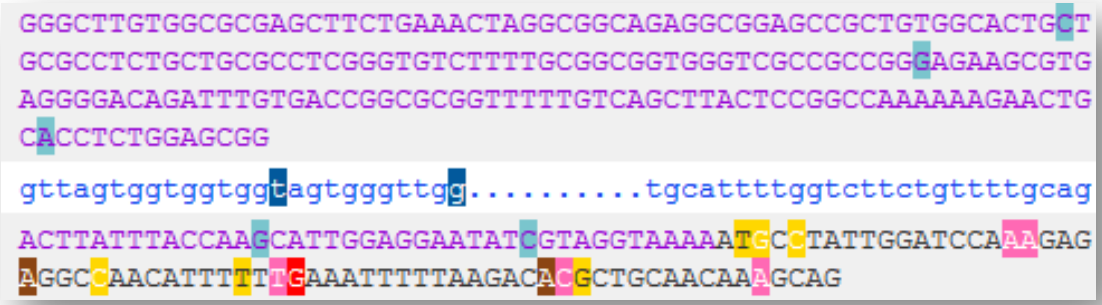
# Sequence and variation displays




Gene: Sequence



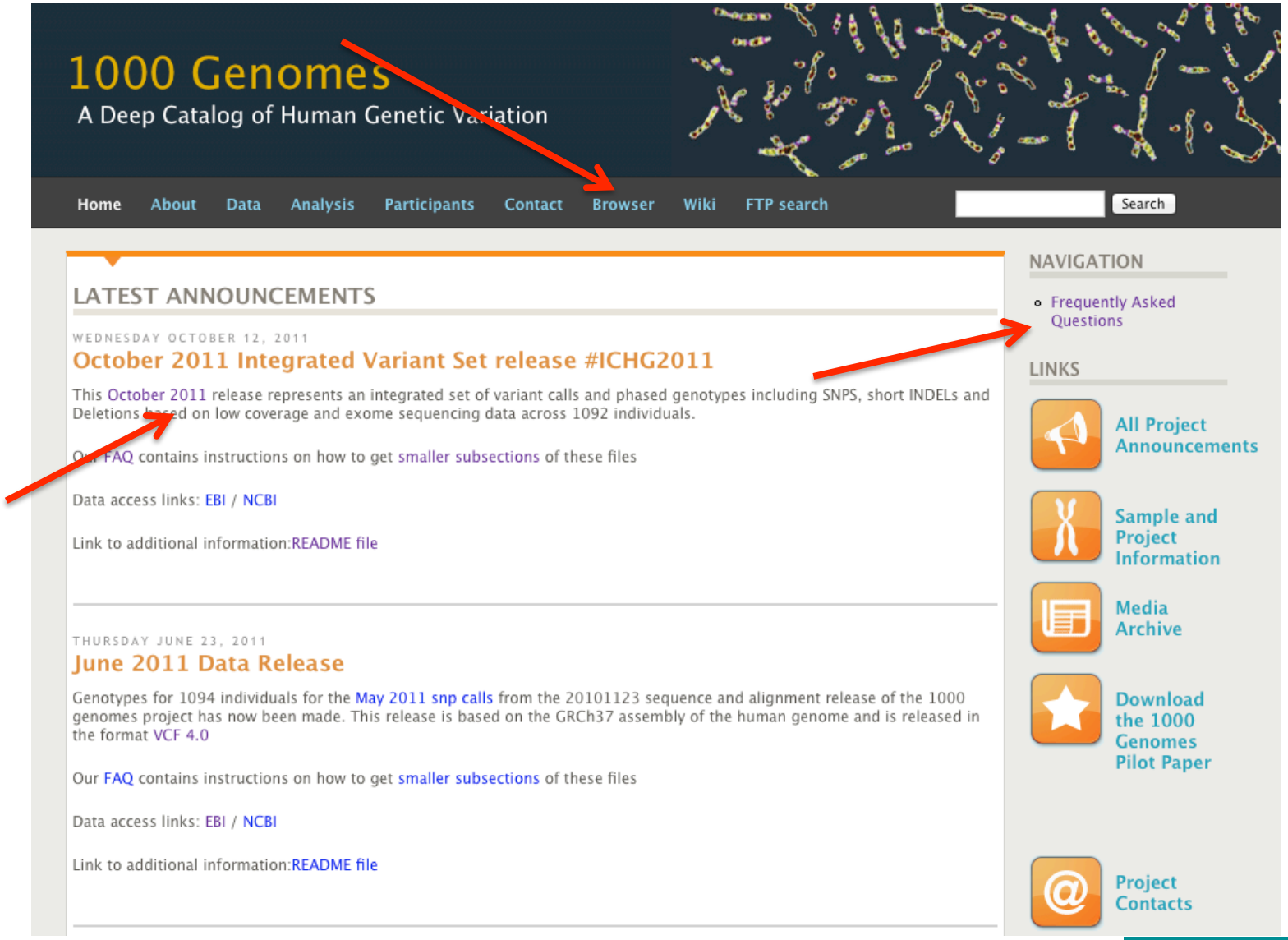

Transcript:cDNA




Transcript: Exons

EMBL-EBI

# http://www.1000genomes.org




A THOUSAND GENOMES

EMBL-EBI

# 1000 Genomes

A Deep Catalog of Human Genetic Variation

Tools | Help

## Search 1000 Genomes

Go

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 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 be 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.

The 1000 Genomes Browser is based on Ensembl web code.

Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute

1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

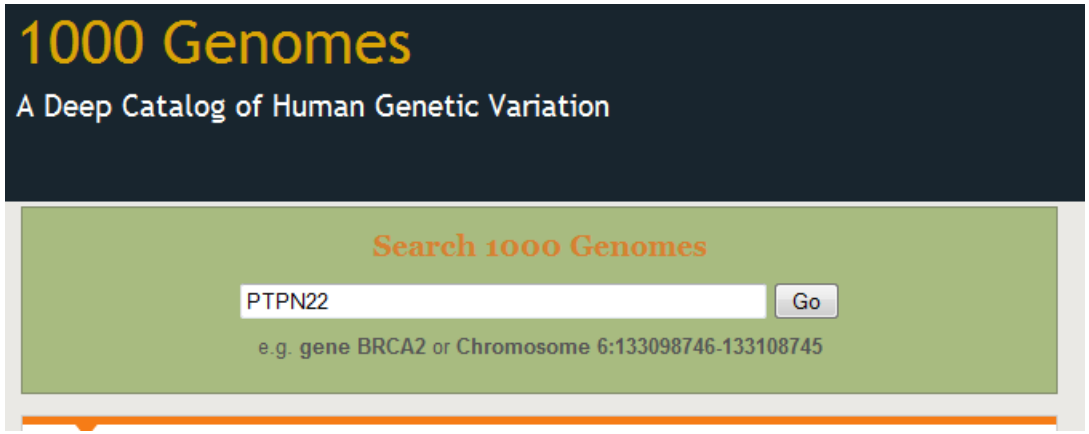
http://browser.1000genomes.org

A THOUSAND GENOMES

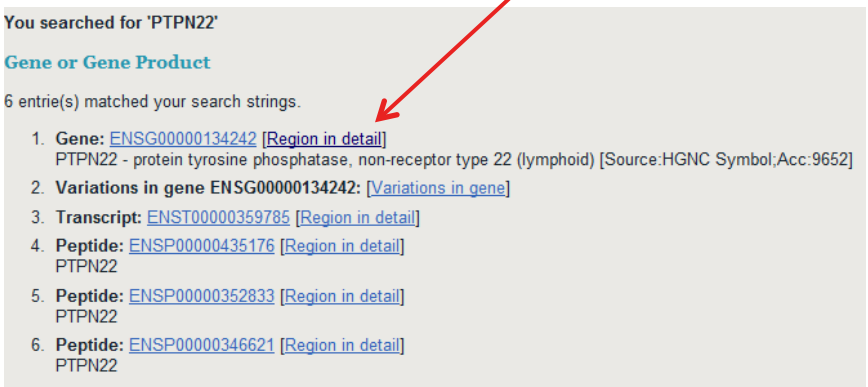
EMBL-EBI

# Searching the Browser

- *http://browser.1000genomes.org*




- Search for PTPN22
- Click ‘Region in Detail’



# Region in Detail

Ensembl genome browser 9: Homo sapiens - Region in detail - Chromosome 1: 114,356,433-114,414,381

browser.1000genomes.org/Homo_sapiens/Location/View?db=core;g=ENSG00000134242;r=1:114356433-114414381

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37) | Location: 1:114,356,433-114,414,381 | Gene: PTPN22 | Tools | Help

Location-based displays
- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Genetic Variation
  - Resequencing (20)
  - Linkage Data
- Markers

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

View in Ensembl

Chromosome 1: 114,356,433-114,414,381

Region in detail help

Location: 1:114356433-114414381 Go

Gene: Go

# Turning on Tracks

Configure this page

Configure Region Image | Configure Overview Image | Custom Data

Configure view

- Image options
  - Active tracks
  - Favourite tracks
  - Track order
  - Search results
  - 1000 Genomes (2/5)
  - 1000 Genomes VCF (0/1)
  - Sequence (2/4)
  - Markers (1/1)
  - Genes (5/5)
  - Prediction transcripts (0/1)
  - Protein alignments (0/5)
  - Protein features (4/5)
  - cDNA/mRNA alignments (0/2)
  - RNA alignments (0/2)

1000 Genomes

- Enable/disable all tracks
- ALL - interim phase 1 - 1000 Genomes variations
- AFR - interim phase 1 - 1000 Genomes variations
- AMR - interim phase 1 - 1000 Genomes variations
- ASN - interim phase 1 - 1000 Genomes variations
- EUR - interim phase 1 - 1000 Genomes variations

Configure Region Image | Configure Overview Image | Custom Data

Configure view

- Image options
  - Active tracks
  - Favourite tracks
  - Track order
  - Search results
  - 1000 Genomes (2/5)
  - 1000 Genomes VCF (0/1)
  - Sequence (1/4)
  - Markers (0/1)
  - Genes (5/5)
  - Prediction transcripts (0/1)

Germline variation

- Structural variants (all sources)
- DGVa structural variations

Key

- Track style

ALL - interim phas... 4 0

AFR - interim phas...

Contigs

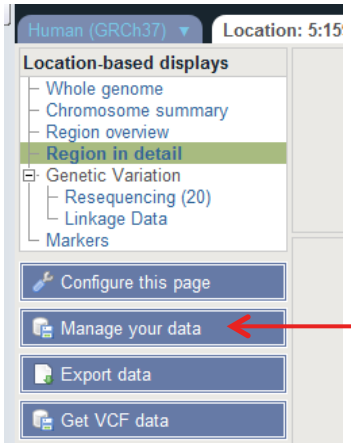
All Structural varia...

AL137856.24 >

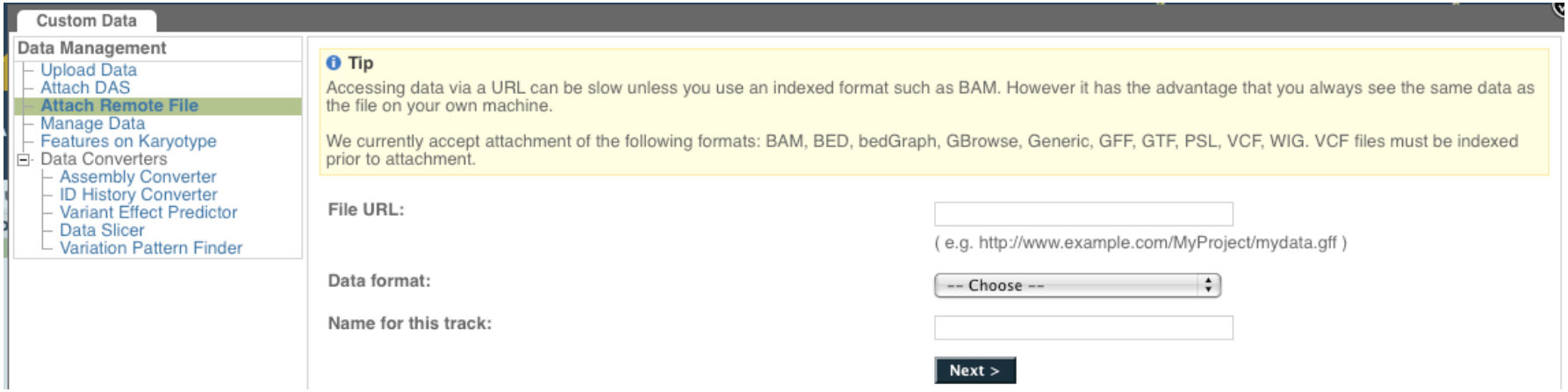
Ensembl/Havana g...

< PTPN22-002

# File upload to view with 1000 Genomes data




Manage your data




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

* VCF must be indexed

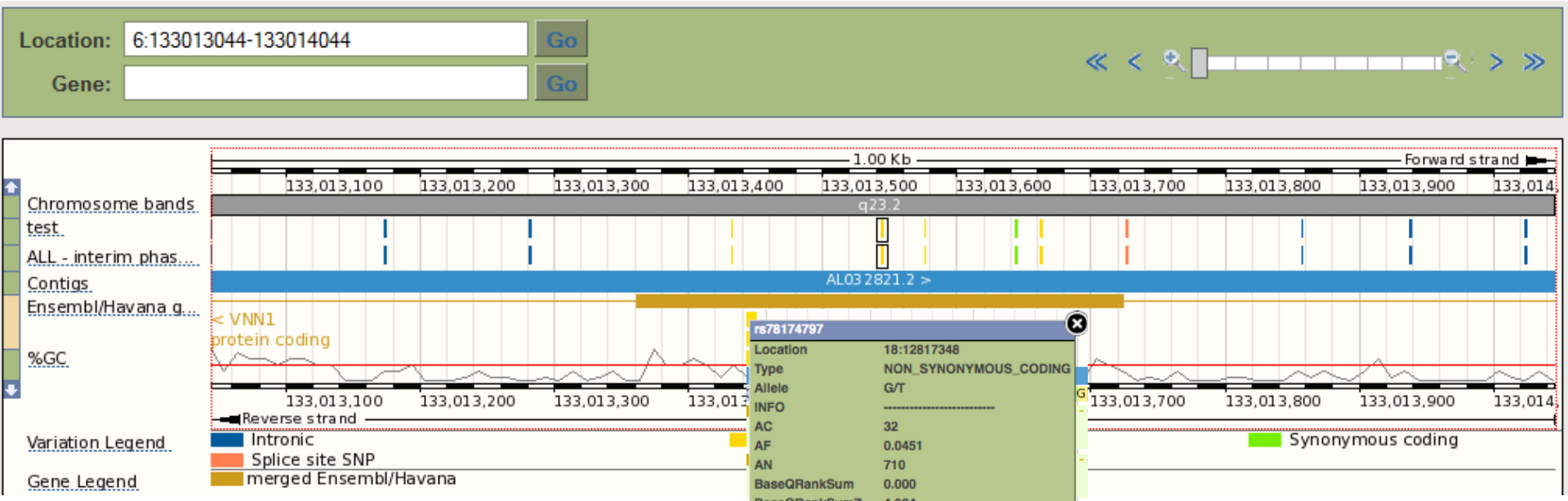
A THOUSAND GENOMES

EMBL-EBI

# 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




| rs78174797 | |
|---|---|
| Location | 18:12817348 |
| Type | NON_SYNONYMOUS_CODING |
| Allele | G/T |
| INFO | ---------- |
| AC | 32 |
| AF | 0.0451 |
| AN | 710 |
| BaseQRankSum | 0.000 |
| BaseQRankSumZ | 4.084 |
| DB | |
| DP | 2210 |
| Dels | 0.00 |
| HRun | 2 |
| HaplotypeScore | 0.2300 |
| MQ | 71.93 |
| MQ0 | 21 |
| MQRankSum | 3.986 |
| MQRankSumZ | -0.255 |
| QD | 9.94 |
| ReadPosRankSum | 11.375 |
| ReadPosRankSumZ | -1.455 |
| SB | -1176.88 |
| VQSLOD | 6.4867 |
| set | ALL5 |

A THOUSAND GENOMES

EMBL-EBI

# Uploaded BAM

Example:

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam


A THOUSAND GENOMES

EMBL-EBI

# Gene View

Click the Gene tab, then ‘Variation Table’ or ‘Variation Image’

Gene Tab

Human (GRCh37) | Location: 1:114,362,205-114,362,276 | Gene: PTPN22 | Tools | Help

**Gene-based displays**

- Gene summary
- Splice variants (12)
- Supporting evidence
- Sequence
- External references
- Regulation
- Genetic Variation
  - Variation Table
  - Structural Variation
  - Variation Image
- External Data
- ID History
  - Gene history

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

**Gene: PTPN22 (ENSG00000134242)**

| | |
|---|---|
| **Description** | protein tyrosine phosphatase, non-receptor type 22 (lymphoid) [Source:HGNC Symbol;Acc:9652] |
| **Location** | Chromosome 1: 114,356,433-114,414,381 reverse strand. |
| **Transcripts** | There are 12 transcripts in this gene |
| | **Click the plus to show the transcript table** |

Download as csv

**Variation Table** help

**Summary of variations in ENSG00000134242 by consequence type**

Show All entries | Filter

| Number of variants | | Type | Description |
|---|---|---|---|
| 19 | Show | Essential splice site | In the first 2 or the last 2 basepairs of an intron |
| 9 | Show | Stop gained | In coding sequence, resulting in the gain of a stop codon |
| 0 | - | Stop lost | In coding sequence, resulting in the loss of a stop codon |
| 0 | - | Complex in/del | Insertion or deletion that spans an exon/intron or coding sequence/UTR border |
| 0 | - | Frameshift coding | In coding sequence, resulting in a frameshift |
| 160 | Show | Non-synonymous coding | In coding sequence and results in an amino acid change in the encoded peptide sequence |
| 65 | Show | Splice site | 1-3 bps into an exon or 3-8 bps into an intron |
| 0 | - | Partial codon | Located within the final, incomplete codon of a transcript whose end coordinate is unknown |
| 83 | Show | Synonymous coding | In coding sequence, not resulting in an amino acid change (silent mutation) |

Get in vcf format

# Structural variation (in the Gene tab)

Human (GRCh37) | Location: 1:114,356,433-114,414,381 | Gene: PTPN22 | Tools | Help

**Gene-based displays**
- Gene summary
- Splice variants (12)
- Supporting evidence
- Sequence
- External references
- Regulation
- Genetic Variation
  - Variation Table
  - **Structural Variation**
  - Variation Image
- External Data
- ID History
  - Gene history

Configure this page
Manage your data
Export data
Get VCF data
Bookmark this page
View in Ensembl

**Gene: PTPN22 (ENSG00000134242)**

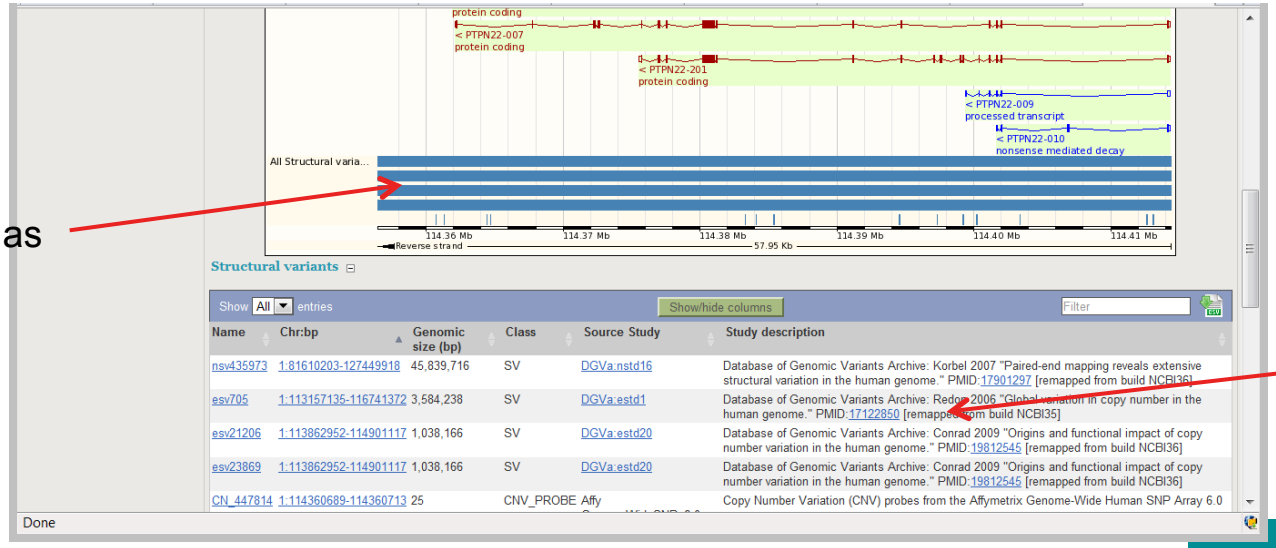
**Description** protein tyrosine phosphatase, non-receptor type 22 (lymphoid) [Source:HGNC Symbol;Acc:9652]
**Location** Chromosome 1: 114,356,433-114,414,381 reverse strand.
**Transcripts** There are 12 transcripts in this gene

Structural Variation

Show All entries | Show/hide columns | Filter

| Name | Transcript ID | Length (bp) | Protein ID | Length (aa) | Biotype | CCDS |
|---|---|---|---|---|---|---|
| PTPN22-001 | ENST00000359785 | 3654 | ENSP00000352833 | 807 | Protein coding | CCDS863 |
| PTPN22-002 | ENST00000460620 | 1794 | ENSP00000433141 | 179 | Protein coding | - |
| PTPN22-004 | ENST00000528414 | 3424 | ENSP00000435176 | 752 | Protein coding | - |
| PTPN22-006 | ENST00000420377 | 2726 | ENSP00000388229 | 795 | Protein coding | - |
| PTPN22-007 | ENST00000525799 | 2118 | ENSP00000432674 | 668 | Protein coding | - |
| PTPN22-201 | ENST00000354605 | 2347 | ENSP00000346621 | 691 | Protein coding | CCDS864 |
| PTPN22-202 | ENST00000538253 | 2414 | ENSP00000439372 | 563 | Protein coding | - |
| PTPN22-008 | ENST00000532224 | 2421 | ENSP00000431249 | 135 | Nonsense mediated decay | - |
| PTPN22-010 | ENST00000529045 | 527 | ENSP00000434932 | 92 | Nonsense mediated decay | - |
| PTPN22-009 | ENST00000534519 | 565 | No protein product | - | Processed transcript | - |
| PTPN22-003 | ENST00000484147 | 2258 | No protein product | - | Retained intron | - |
| PTPN22-005 | ENST00000469077 | 562 | No protein product | - | Retained intron | - |

Structural Variants as boxes



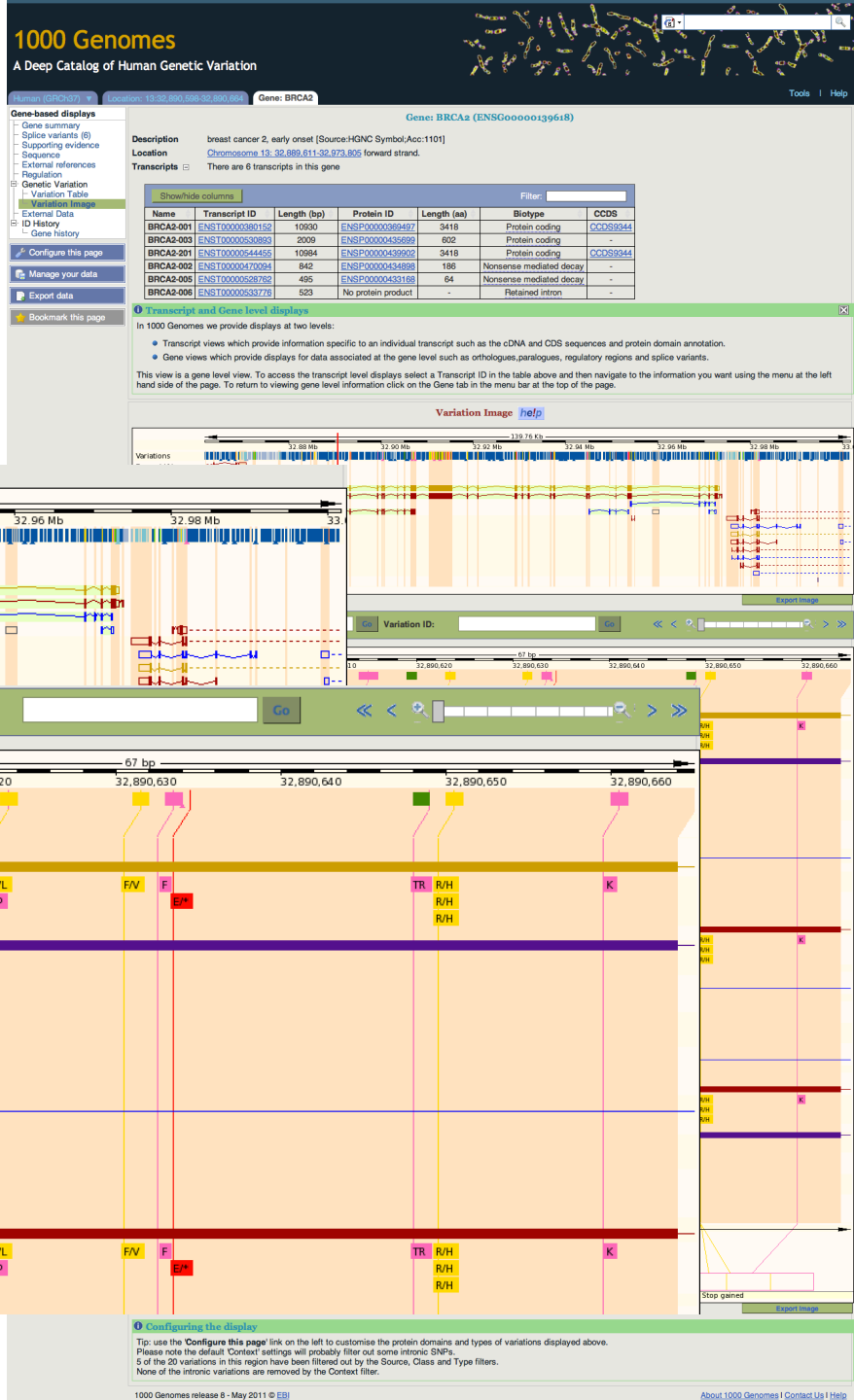

Table

A THOUSAND GENOMES

EMBL-EBI

# Variation Image

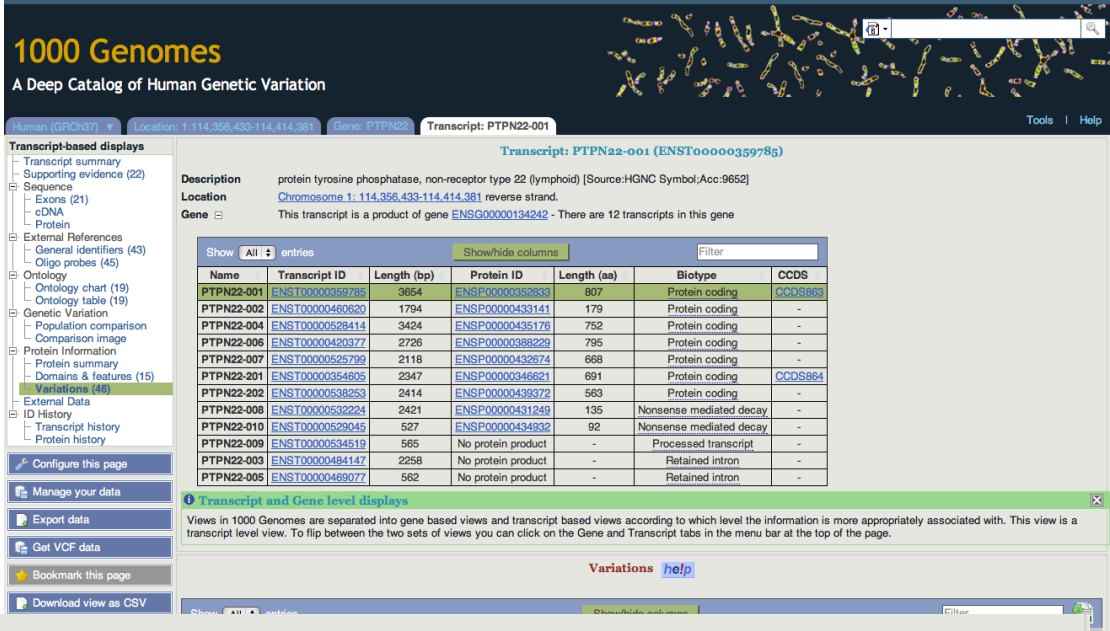
- Gene variation zoom



# Transcript Tab: Variations

Effect on Protein:

- SIFT
- PolyPhen

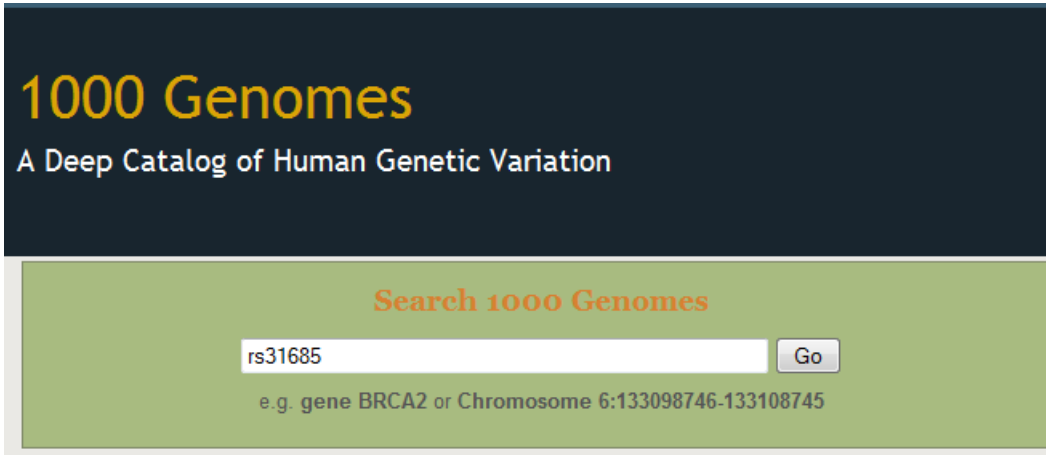



Download view as CSV

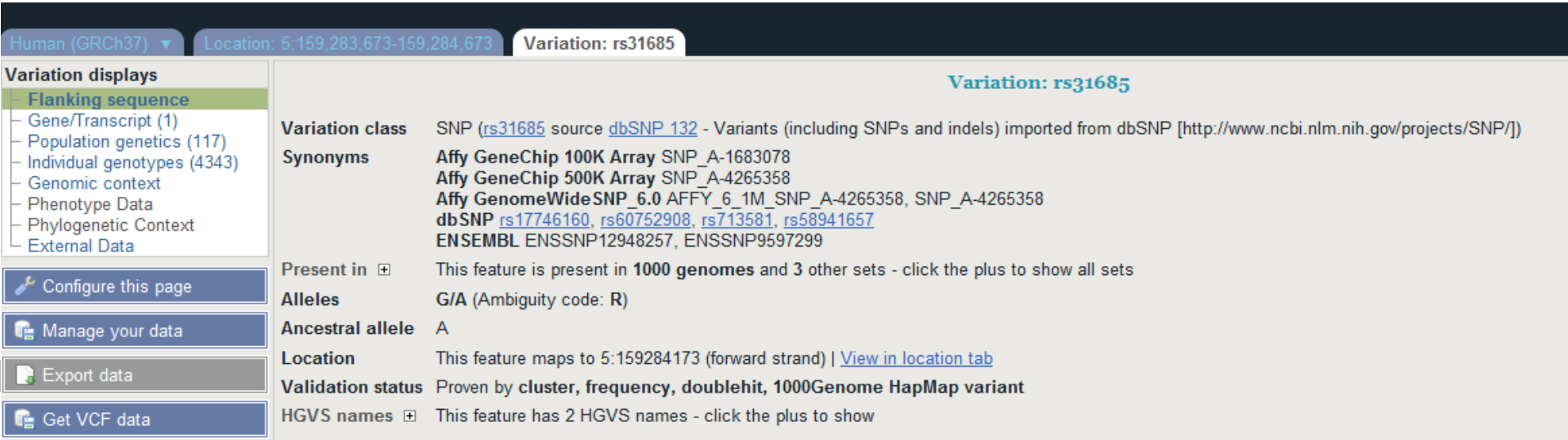
Show All entries | Show/hide columns | Filter

| Residue | Variation ID | Variation type | Alleles | Ambiguity code | Residues | Codons | SIFT | PolyPhen |
|---|---|---|---|---|---|---|---|---|
| 16 | rs74163639 | Synonymous coding | G/A | R | S | AGC, AGT | - | - |
| 49 | rs61745743 | Synonymous coding | A/G | R | A | GCT, GCC | - | - |
| 71 | rs74163642 | Non-synonymous coding | A/G | R | V, A | GTA, GCA | deleterious | probably damaging |
| 141 | rs115552198 | Non-synonymous coding | G/A | R | R, C | CGC, TGC | deleterious | probably damaging |
| 177 | 1KG_1_114399013 | Synonymous coding | C/T | Y | K | AAG, AAA | - | - |
| 183 | rs34590413 | Stop gained | G/A | R | R, * | CGA, TGA | - | - |
| 201 | rs74163647 | Non-synonymous coding | G/A | R | S, F | TCT, TTT | deleterious | probably damaging |
| 206 | rs61738614 | Non-synonymous coding | A/C | M | L, R | CTT, CGT | deleterious | probably damaging |
| 232 | rs78195073 | Synonymous coding | T/C | Y | G | GGA, GGG | - | - |
| 247 | rs35910094 | Synonymous coding | T/G | K | L | CTA, CTC | - | - |
| 263 | rs33996649 | Non-synonymous coding | C/T | Y | R, Q | CGG, CAG | tolerated | benign |
| 266 | rs72650670 | Non-synonymous coding | G/A | R | R, W | CGG, TGG | deleterious | probably damaging |
| 277 | rs72483511 | Stop gained, Splice site | C/A | M | E, * | GAA, TAA | - | - |
| 324 | rs113984534 | Synonymous coding | A/G | R | Y | TAT, TAC | - | - |
| 366 | rs74163654 | Synonymous coding | C/T | Y | E | GAG, GAA | - | - |
| 370 | rs72650671 | Non-synonymous coding | G/T | K | H, N | CAC, AAC | deleterious | possibly damaging |
| 388 | rs77913785 | Non-synonymous coding | G/T | K | D, E | GAC, GAA | deleterious | benign |
| 413 | 1KG_1_114380784 | Non-synonymous coding | T/G | K | Q, P | CAA, CCA | deleterious | benign |
| 414 | 1KG_1_114380780 | Synonymous coding | A/G | R | S | AGT, AGC | - | - |
| 427 | rs112873647 | Non-synonymous coding | -/ATT | - | -, N | -, AAT | - | - |
| 444 | rs74163655 | Non-synonymous coding | T/A | W | I, L | ATA, TTA | tolerated | benign |
| 447 | rs112191110 | Non-synonymous coding | G/A | R | T, I | ACC, ATC | deleterious | probably damaging |
| 452 | rs56174946 | Synonymous coding | A/G | R | F | TTT, TTC | - | - |
| 456 | rs72650672 | Non-synonymous coding | G/C | S | Q, E | CAG, GAG | deleterious | possibly damaging |
| 477 | rs74163656 | Synonymous coding | A/G | R | H | CAT, CAC | - | - |
| 778 | rs41313296 | Non-synonymous coding | T/A | W | N, I | AAT, ATT | deleterious | probably damaging |

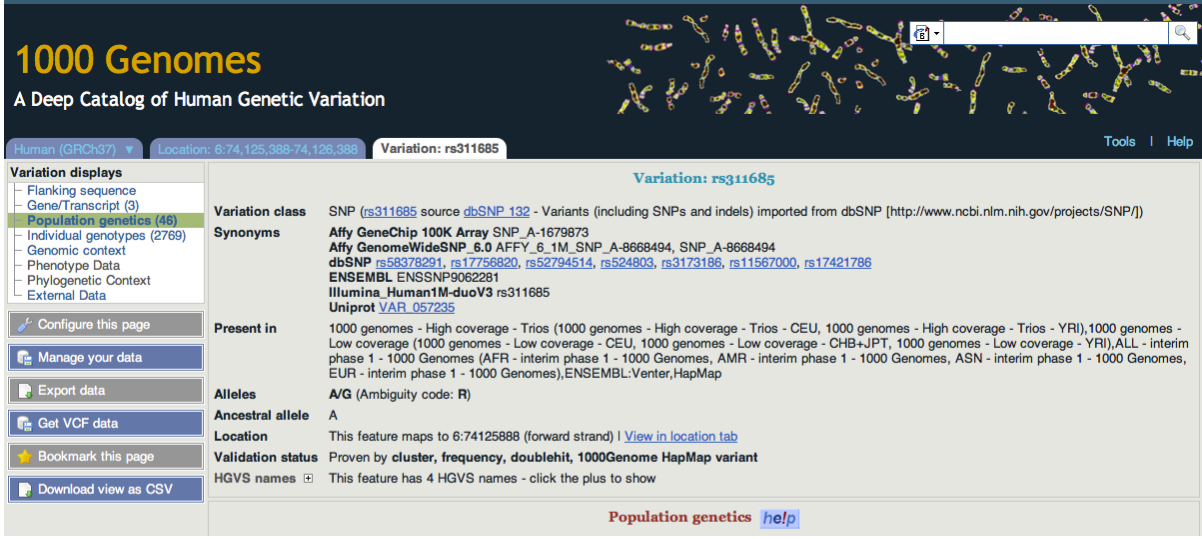
# Start again- search for a variation (rs31685)




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



- Population




## 1000 genomes alleles frequencies

| AFR | ALL | AMR | ASN | EUR |
|---|---|---|---|---|
| A: 69% | A: 51% | A: 54% | A: 45% | A: 42% |
| G: 31% | G: 49% | G: 46% | G: 55% | G: 58% |

## 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 |
|---|---|---|---|---|---|
| 1000GENOMES:pilot_1_CEU_low_coverage_panel | ss233534774 | 1000GENOMES | 0.458 | 0.542 | |
| 1000GENOMES:pilot_1_CHB+JPT_low_coverage_panel | ss240577229 | 1000GENOMES | 0.400 | 0.600 | |
| 1000GENOMES:pilot_1_YRI_low_coverage_panel | ss222470667 | 1000GENOMES | 0.729 | 0.271 | |

# Phenotype for one variant

2,027 Variation: rs420259

**Variation: rs420259**

| | |
|---|---|
| Variation class | (source dbSNP) |
| Synonyms | Affy GeneChip 500K Array SNP_A-2248415<br>Affy GenomeWideSNP_6.0 SNP_A-2248415 |
| Alleles | A/G (Type: Unknown)<br>*Ancestral allele*: G |
| Location | This feature maps to 1 genomic location(s). show locations |

« Context | **Phenotype Data** | Evolutionary or Phylogenetic

| Disease/Trait | Source | Study | Associated Gene(s) | Strongest risk allele | Associated variant | P value |
|---|---|---|---|---|---|---|
| Bipolar Disorder (BD) | [EGA] | | | | rs420259 | |
| | [NHGRI_GWAS_catalog] | pubmed/17554300 | PALB2,NDUFAB1,DCTN5 | rs420259-A | rs420259 | 6.00E-08 |

| | |
|---|---|
| EGA | http://www.ebi.ac.uk/ega |
| NHGRI | http://www.genome.gov/gwastudies/ |
| Open GWAS DB | http://www.biomedcentral.com/1471-2350/10/6 |
| COSMIC | http://www.sanger.ac.uk/genetics/CGP/cosmic/ |
| OMIM | http://www.ncbi.nlm.nih.gov/omim |
| HGMD-Public | http://www.hgmd.cf.ac.uk/ac/index.php |
| UniProt | http://www.uniprot.org/ |

# Coming Soon Ensembl 65


Should arrive in May

EMBL-EBI

# Exercise, Browser

3. Find the variant rs45562238 using http://browser.1000genomes.org.

4. In what 1000 Genomes Super Population is this variant detected?

5. What are its global allele frequencies in the 1000 Genomes Data set?

6. In which gene is the variant found?

# Exercise Answers, Browser

3


## SNP

1 entrie(s) matched your search strings.

1. dbSNP SNP: rs45562238

## Interpro Domain

0 entrie(s) matched your search strings.

# Exercise Answers

## 4. In what 1000 Genomes Super Population is this variant detected?

American and European

## 5. What are its global allele frequencies in the 1000 Genomes Data set?

0.02 is the global allele frequency, this is also the American Allele Frequency but it rises to 0.04 in the Europeans. The absence of Asians or Africans in this chart means that the variant was not found in any of our Asian of African individuals.

## 6. In which gene is the variant found?

ENSG00000112299, Vanin 1

# 1000 Genomes Tools

# 1000 Genomes

A Deep Catalog of Human Genetic Variation

Tools | Help

## Search 1000 Genomes

Go

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 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 be 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.

The 1000 Genomes Browser is based on Ensembl web code.

Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute

1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

http://browser.1000genomes.org

EMBL-EBI

# 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: GFF, GTF, BED, PSL N.B. Export is currently in GFF only | Online version | API script |
| ID History converter | Convert a set of Ensembl IDs from a previous release into their current equivalents. | Online version (max 30 ids) | API script |
| Variant Effect Predictor | (Formerly SNP Effect Predictor). Upload a set of SNPs in our standard format and export a file containing consequence types. Uploaded tracks can also be viewed on Location pages. | Online version (max 750 SNPs) | API script |
| Data Slicer | Get a subset of data from a BAM or VCF file. | Online version (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 here for more extensive documentation. | Online version | API script |
| 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 ld visualization tools like Haploview. Click here for more extensive documentation. | Online version | API script |

A THOUSAND GENOMES

EMBL-EBI

# Data Slicer

- Remote Bam or VCF files
- Genomic Location
- Returns subsection of given file
- VCF files can be subset by
  - Population
  - Individual
  - Must provide a panel file to map individual to population

# Data Slicing

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:**

6:46620015-46620998

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 Slicer Example screens

## VCF filter by population(s)

**Select one or more populations from the scrollable list:**

ASW
CEU
CHB
CHS
CLM
FIN
GBR
IBS
JPT
LWK

< Back  Next >

Thank you - your VCF file [filtered_6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz] [Size: 7529] has been generated.
Right click on the file name and choose "Save link as .." from the menu

## Preview

```
##fileformat=VCFv4.0
##source=BCM:SNPTools:hapfuse
##reference=1000Genomes-NCBI37
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AP,Number=2,Type=Float,Description="Allelic Probability, P(Allele=1
##source_20120302.1=/nfs/public/rw/ensembl/vcftools/bin/vcf-subset -c HG01112,HG(
#CHROM  POS      ID       REF      ALT      QUAL     FILTER  INFO     FORMAT   HG01112 I


6       31831159         rs3869144        C        T        100      PASS     .        (
```

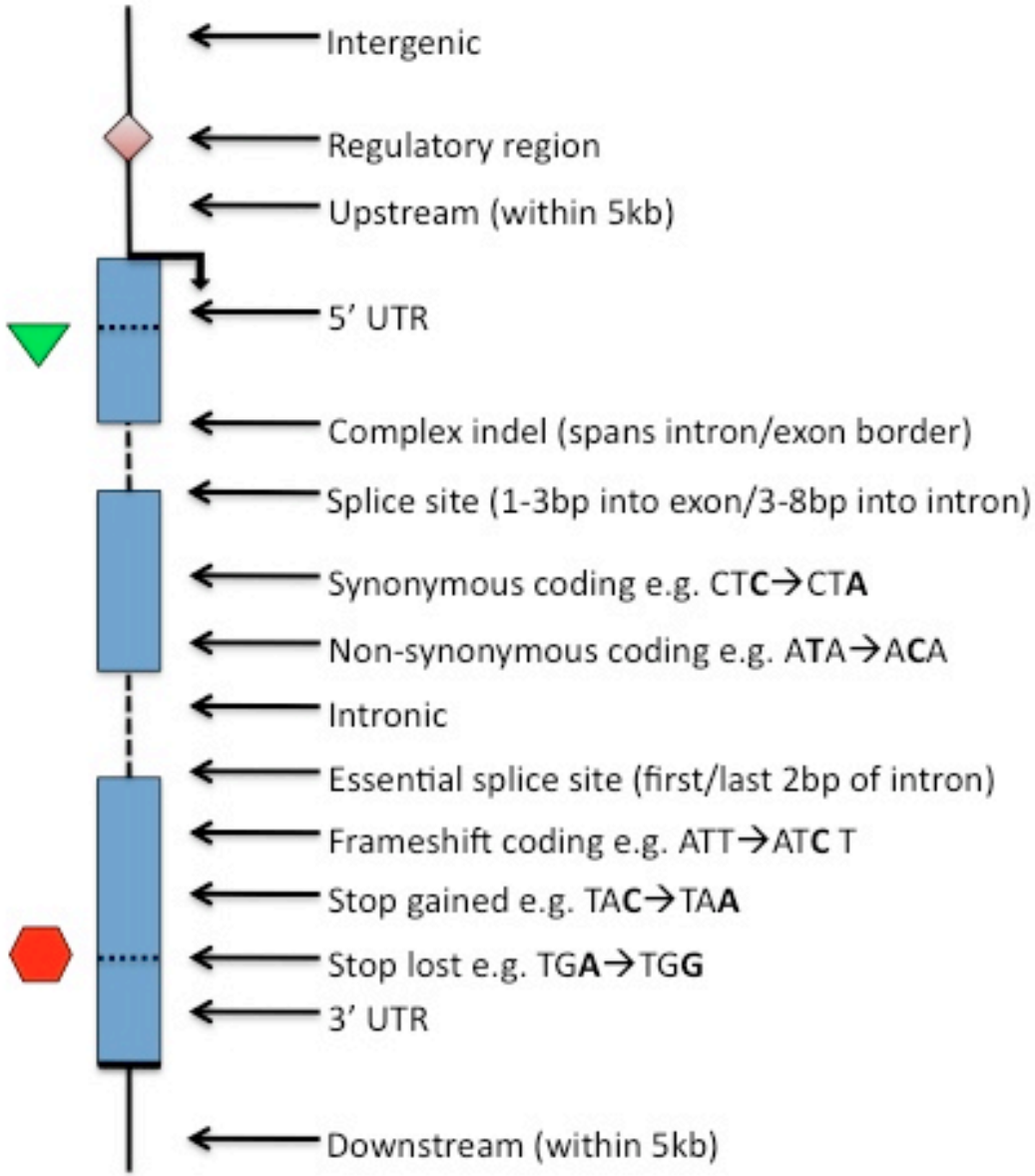
A THOUSAND GENOMES

EMBL-EBI

# 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/
- http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations

# Variant Effect Predictor




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

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

**Variant Effect Predictor:**
This tool takes a list of variant positions and alleles, and predicts the effects of each of these on overlapping transcripts and regulatory regions annotated in Ensembl. The tool accepts substitutions, insertions and deletions as input, uploaded as a list of tab separated values, VCF or Pileup format input.

Upload is limited to 750 variants; lines after the limit will be ignored. Users with more than 750 variations can split files into smaller chunks, use the standalone perl script or the variation API. See also full documentation

## Input file

Species: Human (Homo sapiens): GRCh37

Name for this upload (optional):

Paste file:

Upload file: Choose File no file selected

or provide file URL:

Input file format: Ensembl default

## Options

Get regulatory region consequences: ☑

Type of consequences to display: Ensembl terms

Check for existing co-located variants: Yes

Return results for variants in coding regions only: ☐

Show HGNC identifier for genes where available: ☐

Show Ensembl protein identifiers where available: ☐

Show HGVS identifiers for variants where available: No

## Non-synonymous SNP predictions (human only)

SIFT predictions: No

PolyPhen predictions: No

Condel consensus (SIFT/PolyPhen) predictions: No

## Frequency filtering of existing variants (human only)

Filter variants by frequency: ☐

NB: Enabling frequency filtering may be very slow for large datasets

Filter: Exclude | variants with MAF greater than | 0.1 | in any 1KG low coverage population

Next >

# Variation Effect Predictor Output

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000480384 | Transcript | UPSTREAM | - | - | - | | | | | | | | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000491768 | Transcript | UPSTREAM | - | - | - | 0204386 | ENST00000491768 | Transcript | UPSTREAM | - | - | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000375631 | Transcript | UPSTREAM | - | - | - | 0204386 | ENST00000375631 | Transcript | UPSTREAM | - | - | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000479533 | Transcript | UPSTREAM | - | - | - | 0204386 | ENST00000479533 | Transcript | UPSTREAM | - | - | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000229729 | Transcript | NON_SYNONYMOUS_CODING | 1625 | 1604 | 535 | 0204385 | ENST00000229729 | Transcript | NON_SYNONYMOUS_CODING | 1625 | 1604 | 535 | R/H | cGc/cAc | 1KG_6_31833357 | SIFT=deleterious; PolyPhen=probably_damaging |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000375562 | Transcript | NON_SYNONYMOUS_CODING | 1544 | 1478 | 493 | 0204385 | ENST00000375562 | Transcript | NON_SYNONYMOUS_CODING | 1544 | 1478 | 493 | R/H | cGc/cAc | 1KG_6_31833357 | SIFT=deleterious; PolyPhen=possibly_damaging |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000544672 | Transcript | NON_SYNONYMOUS_CODING | 1673 | 1376 | 459 | 0204385 | ENST00000544672 | Transcript | NON_SYNONYMOUS_CODING | 1673 | 1376 | 459 | R/H | cGc/cAc | 1KG_6_31833357 | SIFT=deleterious; PolyPhen=probably_damaging |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000487680 | Transcript | UPSTREAM | - | - | - | 0204385 | ENST00000487680 | Transcript | UPSTREAM | - | - | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000414427 | Transcript | DOWNSTREAM | - | - | - | 0204385 | ENST00000414427 | Transcript | DOWNSTREAM | - | - | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000479777 | Transcript | DOWNSTREAM | - | - | - | 0204385 | ENST00000479777 | Transcript | DOWNSTREAM | - | - | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000475563 | Transcript | DOWNSTREAM | - | - | - | 0204385 | ENST00000475563 | Transcript | DOWNSTREAM | - | - | - | - | - | 1KG_6_31833357 | - | |

# Variation Pattern Finder

- Remote or local tabix indexed VCF input
- 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

# Variation Pattern Finder

Tools | Help

**Variation Pattern Finder:**

The Variation Pattern Finder allows one to look for patterns of shared variation between individuals in the same vcf file. The finder looks for distinct variation combinations within the region, as well as individuals associated with each variation combination pattern. Only variants which have potentially functional consequences are considered, both intergenic and intronic snps are excluded. Click here for more extensive documentation.

The search will be performed on any VCF file you provided. It should be a URL for the file location. Please refer to http://vcftools.sourceforge.net/specs.html for VCF format specification. A URL for the latest VCF file for variation calls and genotypes released by the 1000 Genomes Project is displayed as an example below the input box. A mapping file between individual sample and population is required as well. The latest mapping file between individual sample and population released by the 1000 Genomes Project is displayed as well below the input box.

## Upload files

**VCF File URL:**

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123
/interim_phase1_release
/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz
```

Clear box

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz

**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

**Region:**

e.g. 6:46620015-46620998

Next >

A THOUSAND GENOMES

EMBL-EBI

# Variation Pattern Finder Output

## Variation Pattern Finder

Export data: CSV Excel

Go to collapsed view

| CEU | C | 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 |

A THOUSAND GENOMES

EMBL-EBI

# 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

# VCF to PED

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

**VCF to PED converter:**

When providing a VCF 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
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.chr6.phase1.projectConsensus.genotypes.vcf.gz
```

Clear box

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz

**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

**Region:**

e.g. 6:46620015-46620998

Next >

A THOUSAND GENOMES

EMBL-EBI

# VCF to PED example output

## VCF filter by population(s)

**Select one or more populations from the scrollable list:**

CHS
CLM
FIN
GBR
IBS
JPT
LWK
MXL
PUR
TSI

**Next >**

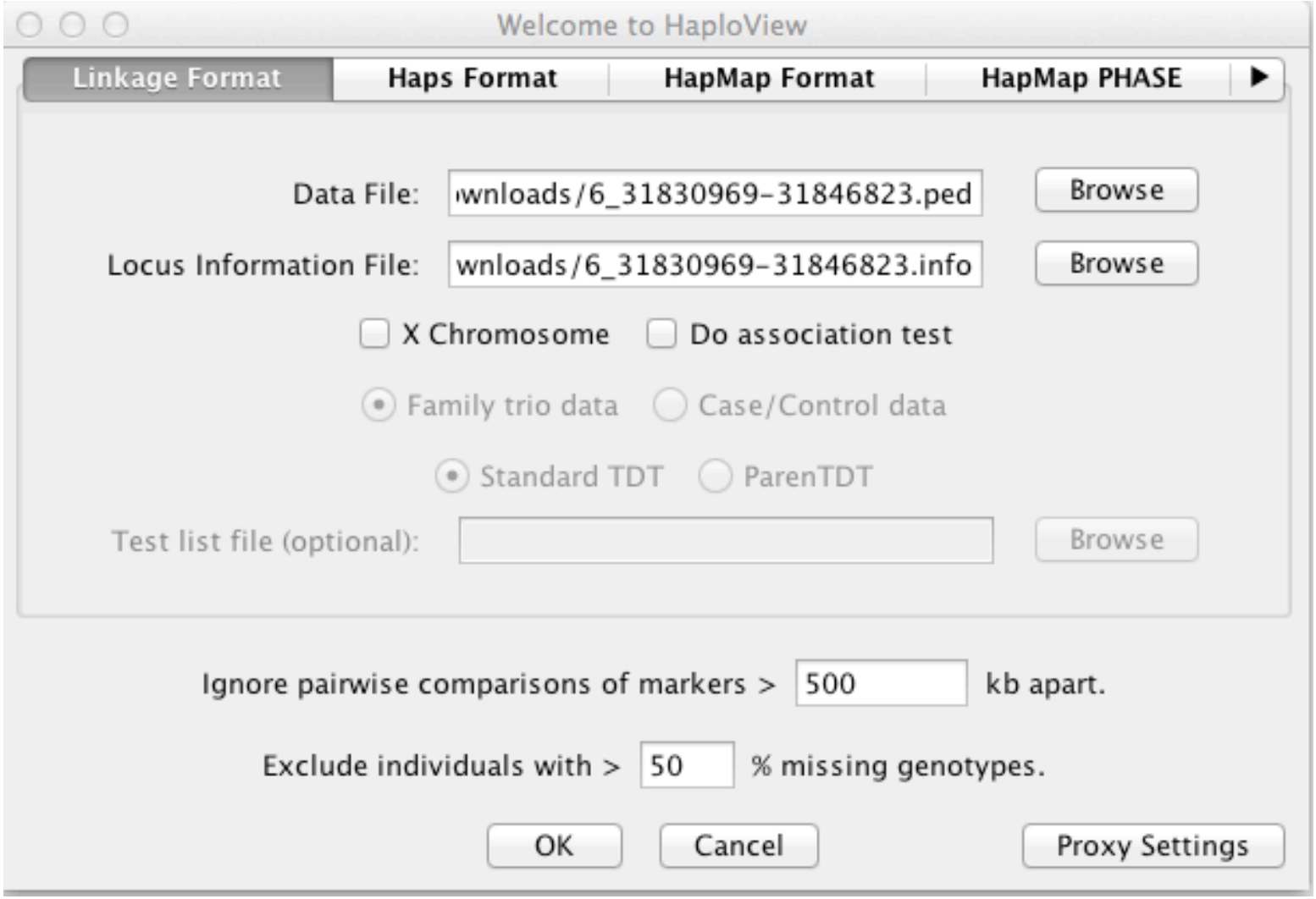
Your linkage pedigree and marker information files have been generated:
Right click on the file name and choose "Save link as .." from the menu:
Marker Information File  Linkage Pedigree File

A THOUSAND GENOMES

EMBL-EBI

# Haplotype example input

java –jar Haploview.jar




EMBL-EBI

# Haploview

- haploview


http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview

# Exercises

Use the browser to find the SLC44A4 gene.

7. Use the get VCF button in the left hand menu on the gene page to get a slice of a vcf file for this Gene.

8. Unzip this VCF file using a tool like winzip or Archive Utility.

9. Upload this VCF file to the Variant Effect Predictor.
http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations

10. Do any of the variants have negative Sift or Polyphen predictions?

11. Using the example URLs on the Variation Pattern Finder tool menu look at the patterns of inheritance for this region: 6:31830700-31840700
http://browser.1000genomes.org/Homo_sapiens/UserData/VariationsMapVCF

12. For the same region use the VCF to PED tool to produce a ped and info file for the CEU population.

13. Look at these files in haploview.

14. How many haplotype blocks does haploview think there are in this section?

# Exercise Answers

1000 Genomes

A Deep Catalog of Human Genetic Variation

SLC44A4

Tools | Help

Search 1000 Genomes

The 1000 Genomes Browser

Ensembl-based browser provides early access to 1000genomes data

1000 Genomes

A Deep Catalog of Human Genetic Variation

Tools | Help

Human (GRCh37)

Search 1000 Genomes

New Search

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

Results Summary

You searched for 'SLC44A4'

Gene or Gene Product

10 entrie(s) matched your search strings.

1. Gene: ENSG00000204385 [Region in detail]
   SLC44A4 - solute carrier family 44, member 4 [Source:HGNC Symbol;Acc:13941]
2. Transcript: ENST00000229729 [Region in detail]
3. Peptide: ENSP00000398764 [Region in detail]
   SLC44A4
4. Peptide: ENSP00000392054 [Region in detail]
   SLC44A4
5. Peptide: ENSP00000404572 [Region in detail]
   SLC44A4
6. Peptide: ENSP00000398901 [Region in detail]
   SLC44A4
7. Peptide: ENSP00000415708 [Region in detail]
   SLC44A4
8. Peptide: ENSP00000400263 [Region in detail]
   SLC44A4
9. Peptide: ENSP00000414296 [Region in detail]
   SLC44A4
10. Peptide: ENSP00000399161 [Region in detail]
    SLC44A4

# Exercise Answers

Human (GRCh37) ▼ | Location: 6:31,830,969-31,846,823 | Gene: SLC44A4 | Tools | Help

**Gene-based displays**

- Gene summary
- Splice variants (9)
- Supporting evidence
- Sequence
- External references
- Regulation
- Genetic Variation
  - Variation Table
  - Structural Variation
  - Variation Image
- External Data
- ID History
  - Gene history

Configure this page
Manage your data
Export data
Get VCF data
Bookmark this page

**Gene: SLC44A4 (ENSG00000204385)**

**Description** solute carrier family 44, member 4 [Source:HGNC Symbol;Acc:13941]
**Location** Chromosome 6: 31,830,969-31,846,823 reverse strand.
**Transcripts** There are 9 transcripts in this gene

Show/hide columns | Filter

| Name | Transcript ID | Length (bp) | Protein ID | Length (aa) | Biotype | CCDS |
|---|---|---|---|---|---|---|
| SLC44A4-001 | ENST00000229729 | 2589 | ENSP00000229729 | 710 | Protein coding | CCDS4724 |
| SLC44A4-004 | ENST00000 | | | | | |
| SLC44A4-201 | ENST00000 | | | | | |
| SLC44A4-202 | ENST00000 | | | | | |
| SLC44A4-002 | ENST00000 | | | | | |
| SLC44A4-003 | ENST00000 | | | | | |
| SLC44A4-007 | ENST00000 | | | | | |
| SLC44A4-005 | ENST00000 | | | | | |
| SLC44A4-006 | ENST00000 | | | | | |

**VCF / BAM File URL:**
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr1.phase1.projectConsensus.genotypes.vcf.gz

**Region:**
6:31830969-31846823
( 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)

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel

< Back | Next >

Configure Page | 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

Thank you - your VCF file [6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz] [Size: 83436] has been generated.
Right click on the file name and choose "Save link as .." from the menu

**Preview**

```
##fileformat=VCFv4.0
##source=BCM:SNPTools:hapfuse
##reference=1000Genomes-NCBI37
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AP,Number=2,Type=Float,Description="Allelic Probability, P(Allele=1
#CHROM  POS      ID       REF      ALT      QUAL     FILTER  INFO     FORMAT  HG00096

6       31831159         rs3869144        C        T        100      PASS     .
6       31831167         .        T        C        100      PASS     .        GT:AP
```

A THOUSAND GENOMES

EMBL-EBI

# Exercise Answers

Custom Data

- Assembly Converter
- ID History Converter
- Variant Effect Predictor
- Data Slicer
- Variation Pattern Finder

## Input file

Species: Human (Homo sapiens): GRCh37

Name for this upload (optional): SLC44A4

Paste file:

Upload file: /Users/laura/Downloads/6. Browse...

or provide file URL:

Input file format: VCF

## Options

Get regulatory region consequences: ☑

Type of consequences to display: Ensembl terms

Check for existing co-located variants: Yes

Return results for variants in coding regions only: ☐

Show HGNC identifier for genes where available: ☐

Show Ensembl protein identifiers where available: ☐

Show HGVS identifiers for variants where available: No

## Non-synonymous SNP predictions (human only)

SIFT predictions: Prediction only

PolyPhen predictions: Prediction only

Condel consensus (SIFT/PolyPhen) predictions: No

# Exercise Answers

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6_31833249_A/G | 6:31833249 | G | ENSG00000204385 | ENST00000487680 | Transcript | UPSTREAM | - | - | - | - |
| 6_31833249_A/G | 6:31833249 | G | ENSG00000204385 | ENST00000414427 | Transcript | DOWNSTREAM | - | - | - | - |
| 6_31833249_A/G | 6:31833249 | G | ENSG00000204385 | ENST00000479777 | Transcript | DOWNSTREAM | - | - | - | - |
| 6_31833249_A/G | 6:31833249 | G | ENSG00000204385 | ENST00000475563 | Transcript | DOWNSTREAM | - | - | - | - |
| 6_31833357_C/T | 6:31833357 | T | - | ENSR00000487922 | RegulatoryFeature | REGULATORY_REGION | - | - | - | - |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000495807 | Transcript | UPSTREAM | - | - | - | - |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000480384 | Transcript | UPSTREAM | - | - | - | - |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000491768 | Transcript | UPSTREAM | - | - | - | - |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000375631 | Transcript | UPSTREAM | - | - | - | - |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204386 | ENST00000479533 | Transcript | UPSTREAM | - | - | - | - |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000229729 | Transcript | NON_SYNONYMOUS_CODING | 1625 | 1604 | 535 | R/H |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000375562 | Transcript | NON_SYNONYMOUS_CODING | 1544 | 1478 | 493 | R/H |
| 6_31833357_C/T | 6:31833357 | T | ENSG00000204385 | ENST00000544672 | Transcript | NON_SYNONYMOUS_CODING | 1673 | 1376 | 459 | R/H |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6_31833357_C/T | 6:31833357 | T | ENSG00 | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00 | | | | | | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00 | - | - | - | 1KG_6_31833357 | - | |
| 6_31833357_C/T | 6:31833357 | T | ENSG00 | 535 | R/H | cGc/cAc | 1KG_6_31833357 | SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious | |
| 6_31833612_C/G | 6:31833612 | G | | | | | | | |
| 6_31833612_C/G | 6:31833612 | G | ENSG00 | | | | | | |
| 6_31833612_C/G | 6:31833612 | G | ENSG00 | 493 | R/H | cGc/cAc | 1KG_6_31833357 | SIFT=deleterious; PolyPhen=possibly_damaging; Condel=deleterious | |
| | | | | 459 | R/H | cGc/cAc | 1KG_6_31833357 | SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious | |
| | | | | - | - | - | 1KG_6_31833357 | - | |
| | | | | - | - | - | 1KG_6_31833357 | - | |
| | | | | - | - | - | 1KG_6_31833357 | - | |

A THOUSAND GENOMES

EMBL-EBI

# Exercise Answers

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

**Variation Pattern Finder:**

The Variation Pattern Finder allows one to look for patterns of shared variation between individuals in the same vcf file. The finder looks for distinct variation combinations within the region, as well as individuals associated with each variation combination pattern. Only variants which have potentially functional consequences are considered, both intergenic and intronic snps are excluded. Click here for more extensive documentation.

The search will be performed on any VCF file you provided. It should be a URL for the file location. Please refer to http://vcftools.sourceforge.net/specs.html for VCF format specification. A URL for the latest VCF file for variation calls and genotypes released by the 1000 Genomes Project is displayed as an example below the input box. A mapping file between individual sample and population is required as well. The latest mapping file between individual sample and population released by the 1000 Genomes Project is displayed as well below the input box.

## Upload files

**VCF File URL:**

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123
/interim_phase1_release
/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz
```

Clear box

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz

**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

**Region:**

6:31830700-31840700

e.g. 6:46620015-46620998

Next >

A THOUSAND GENOMES

EMBL-EBI

# Exercise Answers

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

## Variation Pattern Finder

Export data: CSV Excel

Go to collapsed view

| *Population* ASW | CEU | Freq | | rs116706632:G/A | rs117127493:G/C | rs644827:T/C |
|---|---|---|---|---|---|---|
| | | | | 6:31836976 | 6:31837009 | 6:31838441 |
| | | | _CODING:R/C | ENST00000229729 NON_SYNONYMOUS_CODING:P/S | ENST00000229729 NON_SYNONYMOUS_CODING:Q/E | ENST00000229729 NON_SYNONYMOU |
| | | | _CODING:R/C | ENST00000375562 NON_SYNONYMOUS_CODING:P/S | ENST00000375562 NON_SYNONYMOUS_CODING:Q/E | ENST00000375562 NON_SYNONYMOU |
| | | | _CODING:R/C | ENST00000544672 NON_SYNONYMOUS_CODING:P/S | ENST00000544672 NON_SYNONYMOUS_CODING:Q/E | ENST00000544672 NON_SYNONYMOU |
| | | | _CODING:R/C | ENST00000414427 NON_SYNONYMOUS_CODING:P/S | ENST00000414427 NON_SYNONYMOUS_CODING:Q/E | |
| NA20289, NA20296 and 13 other(s) | NA069 | 0.293 | | GIG | GIG | CIC |
| NA20127, NA19703 and 9 other(s) | NA125 | 0.203 | | GIG | GIG | CIT |
| NA20314, NA20317 and 6 other(s) | NA120 | 0.195 | | GIG | GIG | TIC |
| NA19920, NA19700 and 2 other(s) | | 0.032 | | GIG | GIG | CIC |
| NA19819, NA20281 and 2 other(s) | | 0.026 | | GIG | GIG | CIC |
| NA20291, NA20356 and 3 other(s) | | 0.016 | | GIG | GIG | TIC |
| NA19908 | NA122 | 0.013 | | GIG | GIG | CIT |
| | | 0.008 | | GIG | CIG | CIC |
| | | 0.005 | | GIG | GIC | TIC |
| | NA119 | 0.005 | | GIG | GIC | CIC |
| NA19916 | | 0.004 | | GIG | GIG | CIC |
| NA19711, NA20340 | | 0.003 | | GIG | GIG | CIC |
| | | 0.003 | | GIG | GIG | CIT |
| | NA119 | 0.003 | | GIA | GIG | CIC |
| | | 0.003 | | GIG | CIG | CIT |

A THOUSAND GENOMES

EMBL-EBI

# Exercise Answers

**VCF to PED converter:**

When providing a VCF 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
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.chr6.phase1.projectConsensus.genotypes.vcf.gz
```

Clear box

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz

**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

**Region:**

6:31830700-31840700

e.g. 6:46620015-46620998

Next >

A THOUSAND GENOMES

EMBL-EBI

# Exercise Answers

## VCF filter by population(s)

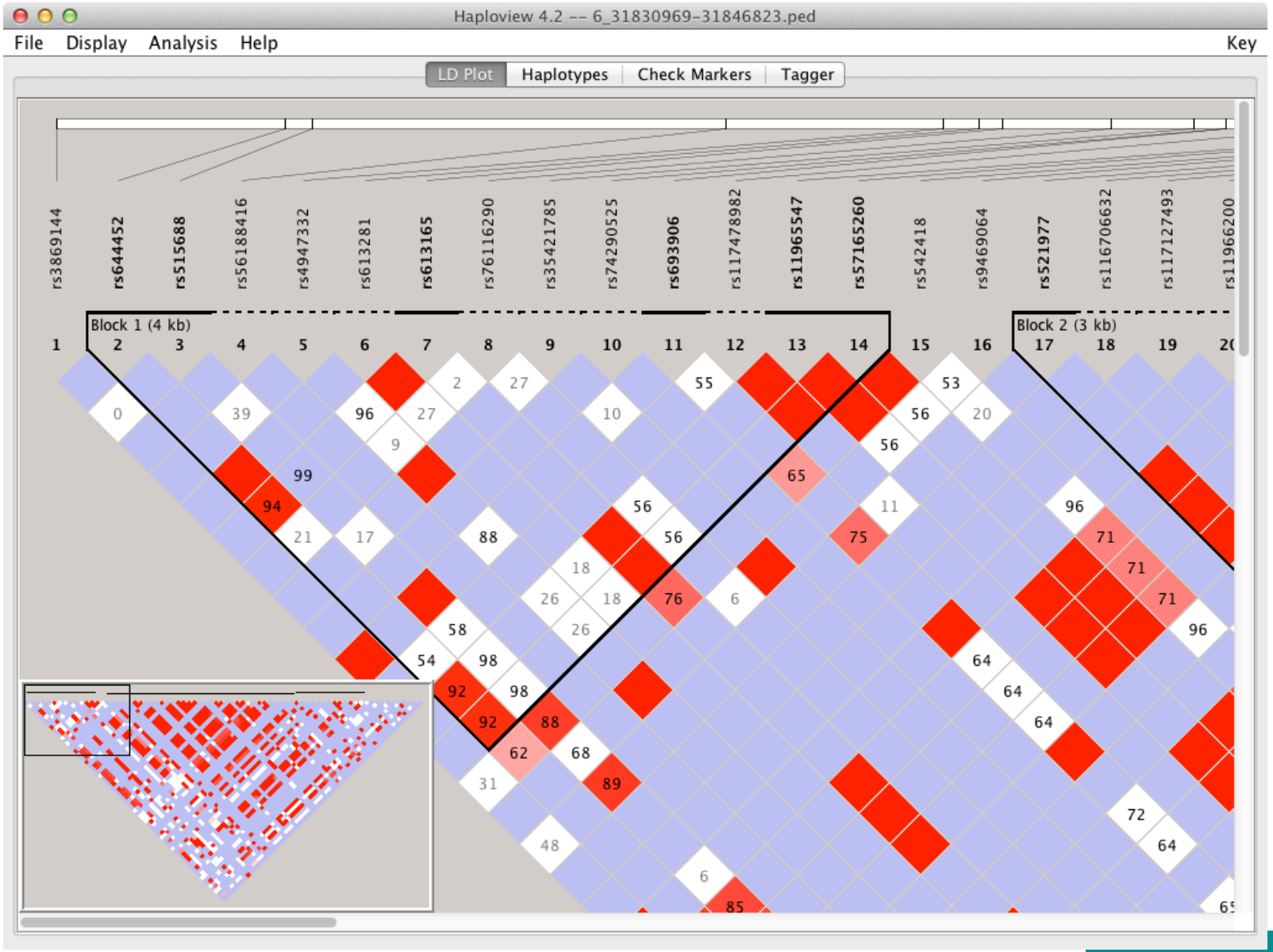
Select one or more populations from the scrollable list:

ASW
CEU
CHB
CHS
CLM
FIN
GBR
IBS
JPT
LWK

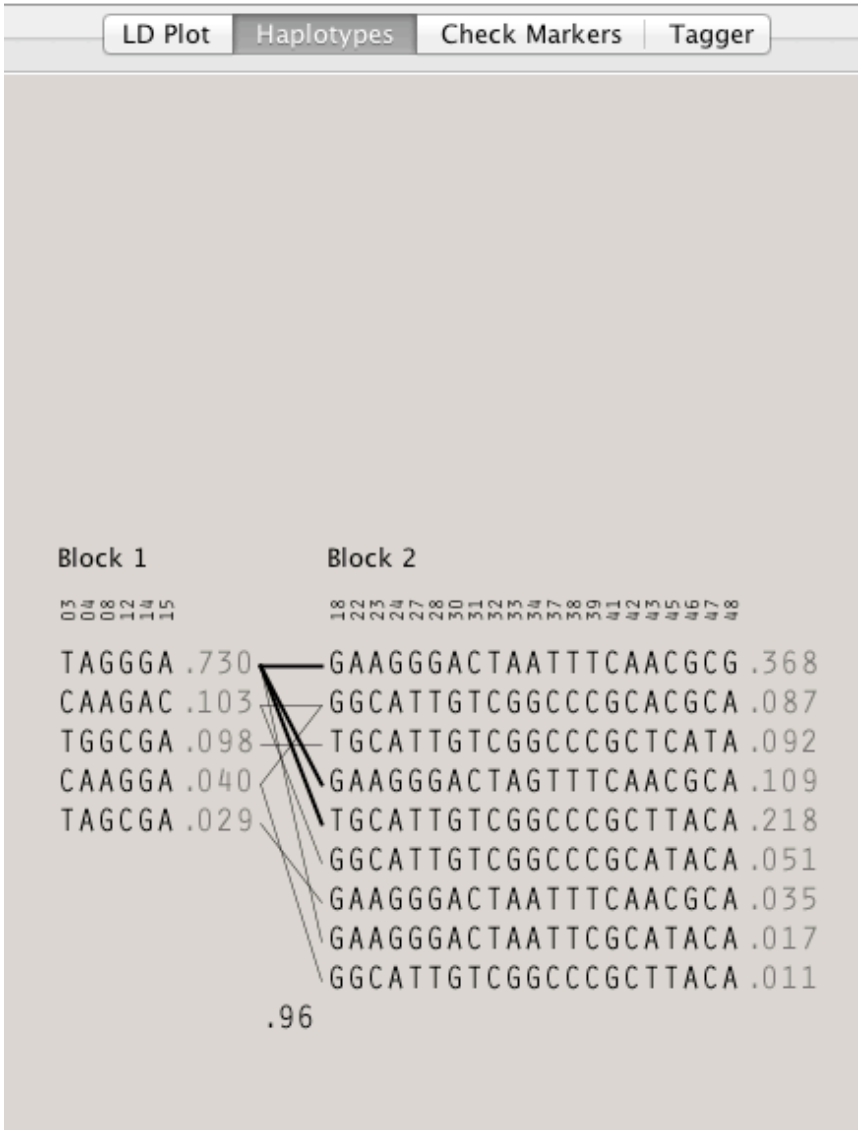
Next >

Your linkage pedigree and marker information files have been generated:
Right click on the file name and choose "Save link as .." from the menu:
Marker Information File  Linkage Pedigree File

# Exercise Answers



# Exercise Answers




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

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


EMBL-EBI

# 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

A THOUSAND GENOMES

EMBL-EBI

# Questionaire

http://goo.gl/AxAR0

# File Formats


EMBL-EBI

# 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

# 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

## 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

**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

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">

# 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

**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

A THOUSAND GENOMES

EMBL-EBI

# 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

# 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

# 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
```

# 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/

# 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

# Exercise, Finding Data

15. How many GRCh37 omni vcf files are in technical/ working

16. Which exome sample from 20110521 has the highest percentage of targets covered at 20x or greater

17. Find the exome bam file for this sample

18. Get a slice of this exome bam file between 7:114173990-114175942

# 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 | grep b37 | wc –l

> 32

> 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

```
>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	GAACCATATTTGGTGTACATAGGCATAAAGAATTTTGCATAAAACCCCCTTGTGGGATTTTATTCATACATAGGTT	SD@GIB>BFDDHDCDBBJCAFHHJBBDDEHDBFFDCHJB<CCC4IIHHIECGCGGGAEEE@AEBH??@H@?CFDBS	RG:Z:SRR099984	NM:i:0	OQ:Z:DE@DEE?EEBEGEDEGFHHFGHHHHGHHFHHGHHDHHHHHGHHDHHGGGHHHHHHHHHHHHHHHHGFHHHHGHHHHH
```

# Command Line Tools

# 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/
- http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations

# 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 Existing_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

# Data Slicing

- 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 6:31833625-31833704
- 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 6:31830969-31846823 | vcf-subset -c HG01375
- http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice

# Variation Pattern Finder

- Remote or local tabix indexed VCF input
- 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

# 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_svs.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[],ENST00000544672-SPLICE_SITE[],ENST00000229729-SPLICE_SITE[],ENST00000375562-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 |

A THOUSAND GENOMES

EMBL-EBI

# 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

# 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_svs.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 -population CEU
- Output should be two files
- 6_31830969-31846823.info
- 6_31830969-31846823.ped

# Haploview

- haploview


http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview

# 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

# 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, Command Line Tools

19. Get a slice of HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam for 7:114304000-114305000 (FoxP2 exon)
20. Get the equivalent section of the 20110521 release chr 7 genotypes file
21. Use vcftools vcf-stats to specify which SNP transition happens most in this section
22. Use this piece with tools, the variant effect predictor, the vcf pattern finder
23. Are there any snps with deleterious sift/polyphen consequences?
24. What is the most common pattern of variation in this region?
25. Use the vcf to ped script with 6:31830700-31840700 and population CEU
26. How many different haplotype blocks does the section contain?

# Exercise Answers, Command Line Tools

> grep HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam /nfs/1000g-archive/vol1/ftp/current.tree | cut -f1 | grep -v bam. | awk '{print "ftp://ftp.1000genomes.ebi.ac.uk/vol1/  "$1}'

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

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

SRR099984.29321596 163 7 114304108 65 76M = 114304379 346 GTTTGCTGCAAGGACGATTGTTTATATTTTCACATCGCACTTAATTTCCTTGCATCTCTGCCACAAGTAGCCAGTT S=??DDBGE@CGGAE@BABIACB?A@ACCCGCGCBH=GCGEBAEBCDHHECIHBBGDHEIHHCGABIAAIHHCGBR RG:Z:SRR099984 NM:i:0 OQ:Z:HHHHEHHHHHHHHFHHHHHHHHGEHGHHHHHHHHHHBHFHGFHHHHHHHFHHHEHHFHHHHHHHDBGFGGHHFEHF

SRR099984.344934 163 7 114304134 59 76M = 114304429 370 TTTTCACATCGCACTTAATTTCCTTGCATCTCTGCCACAAGGAGCCAGTTAGGAATTTTTTTTCAATACATTTTCT S>??D?C??B>A6BBB?C>AAFF1ECCB9FECBDAD=CAEG&AGDDBGAB@GFB@@?>B781<=<?@87>=55>5S RG:Z:SRR099984 NM:i:1 OQ:Z:FHEHHHHGGEFD6EFGGEBDEGG/I@EG8GGDBBBE=FBFE,BEEEDEDEFFEE@FB@F8:3>@?@D==A?77@77

A THOUSAND GENOMES

EMBL-EBI

# 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_svs.genotypes.vcf.gz 7:114304000-114305000 > 20110521.vcf

> vcf-stats 20110521.vcf

> 'G>A' => 5

# 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 Tools

> 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

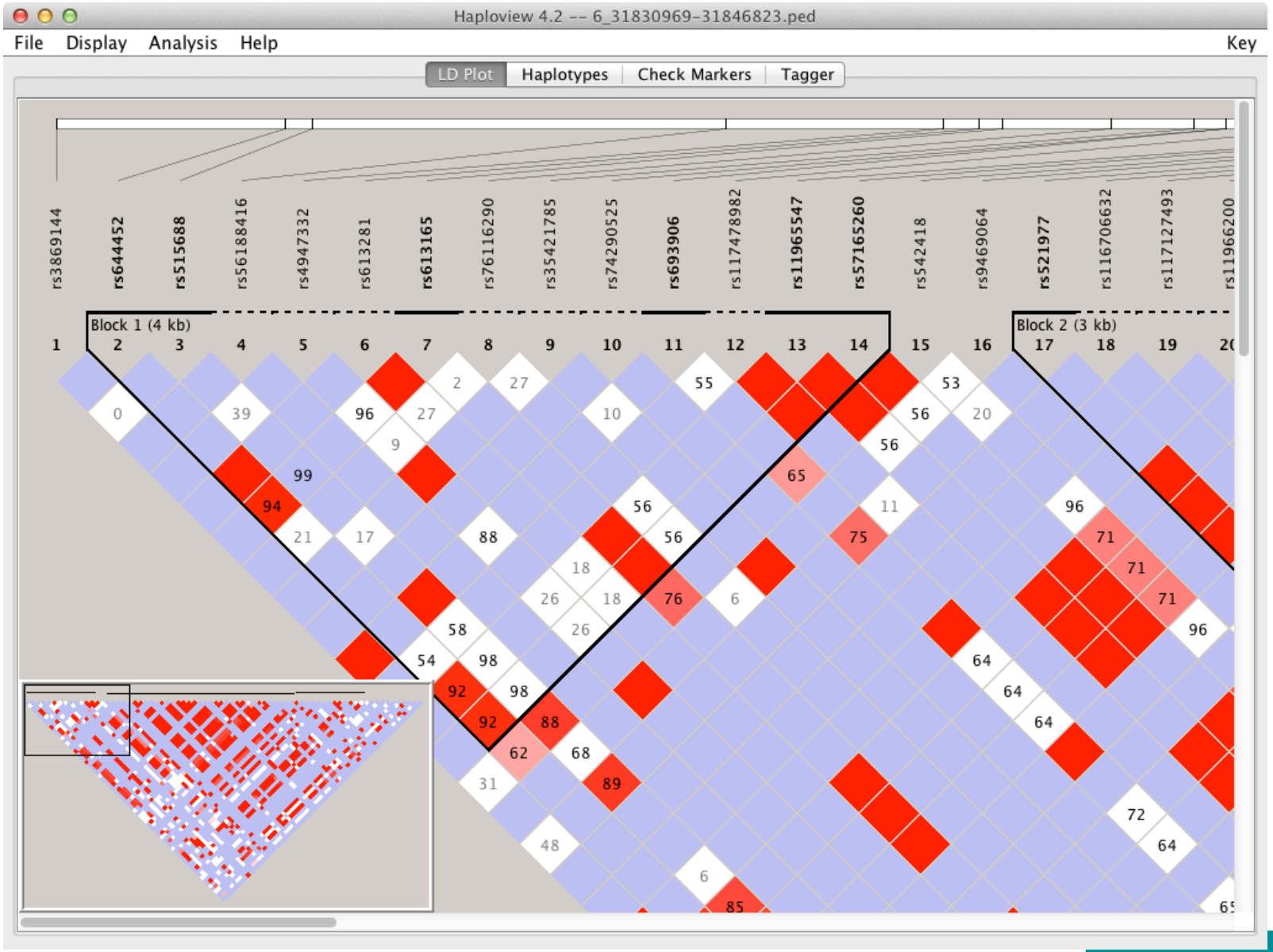
7:114304563_rs1378771[C] 7:114304630_rs1378772[A] 7:114304969_rs2396765[T]

> 14.38 - - C|T - A|T - - - - - T|C TSI(30)

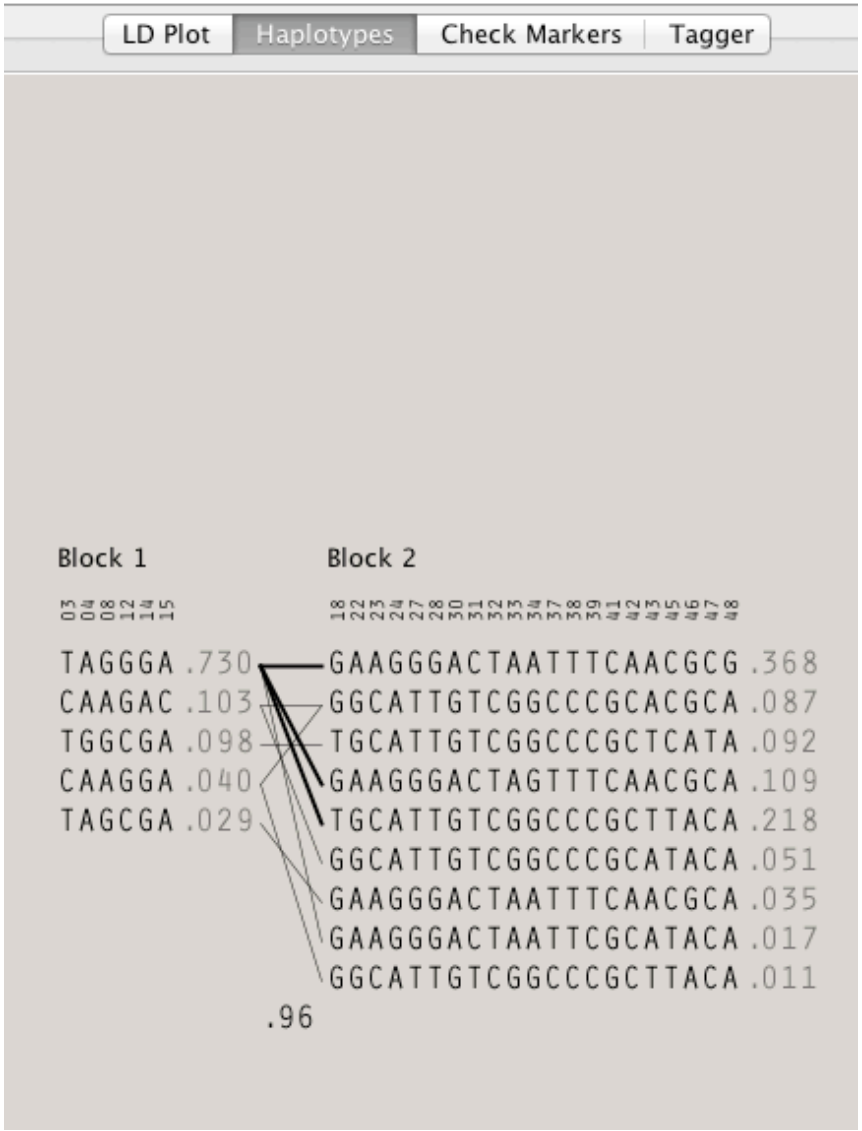
# Exercise Answers: Command Line Tools

```
> 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




A THOUSAND GENOMES

EMBL-EBI

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

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


EMBL-EBI

# Thanks

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

A THOUSAND GENOMES

EMBL-EBI

# Questionaire

http://goo.gl/ud1KM