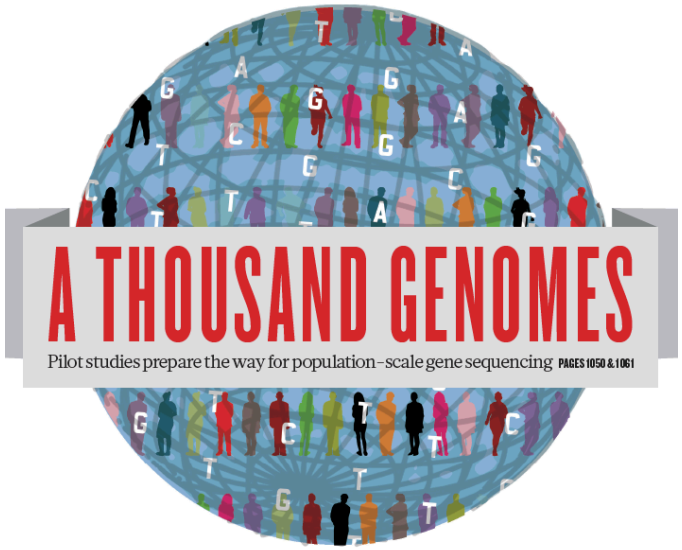


The 1000 Genomes Tutorial Tools for Data Handling and Processing

Laura Clarke

17th February 2012

EMBL-EBI

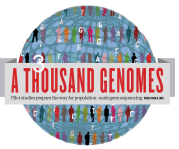
Introduction

This Presentation should give the user an overview of the tools and command lines users can user to gain access to 1000 genomes data sets.

EMBL-EBI

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

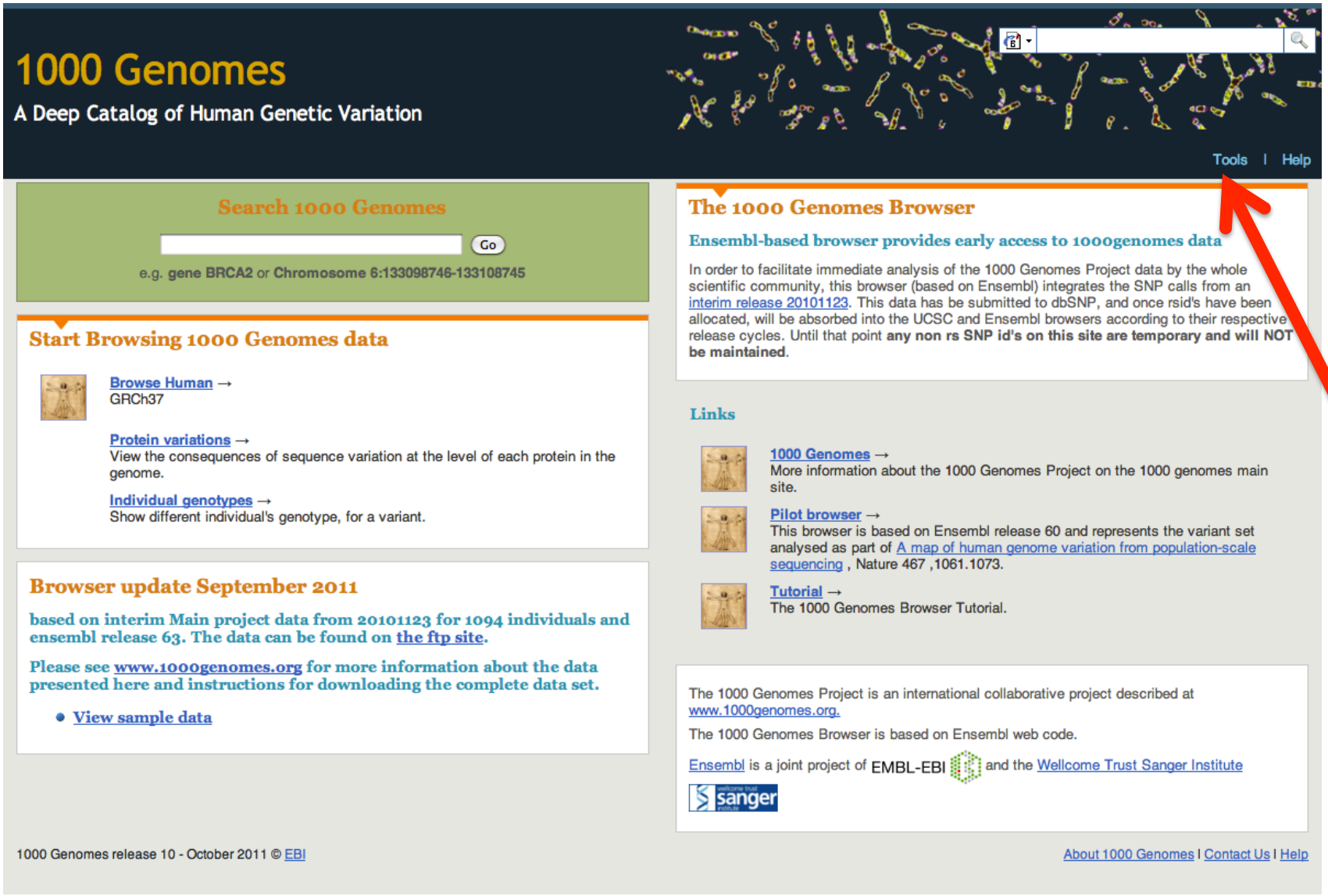


EMBL-EBI

Summary

- Where to find the Tools
- Variant Effect Predictor
- Data Slicer
- Variation Pattern Finder
- VCF to PED
- API and Database access
- Data Availability
- Announcements

EMBL-EBI



http://browser.1000genomes.org

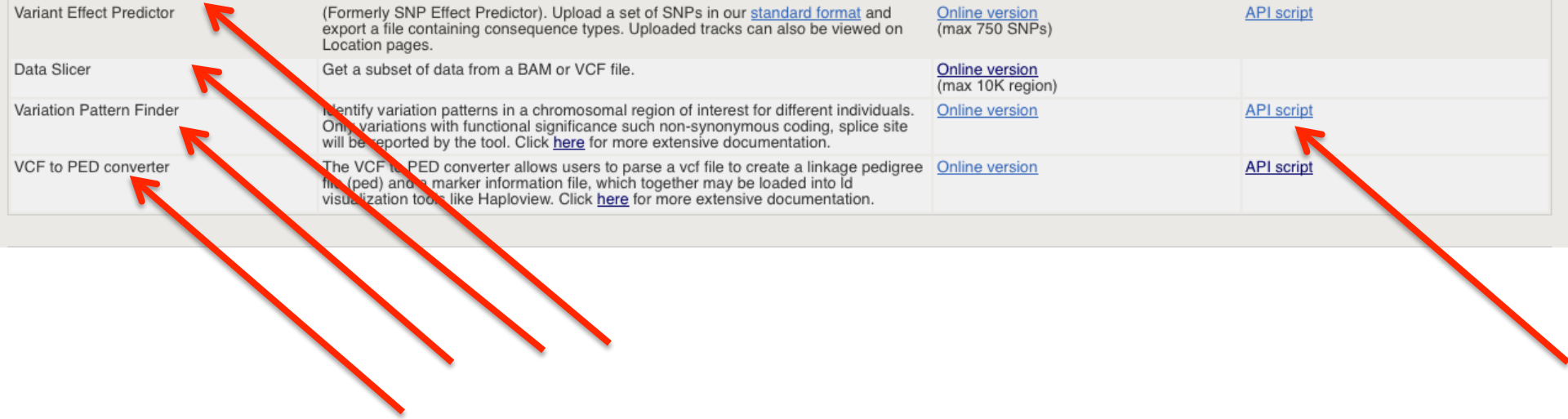
Tools page

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:

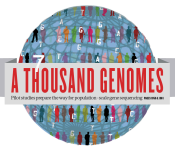
Tool	Description		
Assembly converter	Map your data to the current assembly. Accepted file formats: GFF, GTF, BED, PSL N.B. Export is currently in GFF only	Online version	API script
ID History converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	Online version (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



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



EMBL-EBI

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 >

Variant Effect Predictor

- **perl variant_effect_predictor.pl** -input 6_381831625_3184704.vcf -sift p -polyphen p –check_existing
- less variant_effect_output.txt

#Uploaded_variation Location Allele Gene Feature Feature_type Consequence cDNA_position CDS_position Protein_position Amino_acids Codons Exi sting_variation Extra

rs138094825 6:31831667 A ENSG00000204385 ENST00000414427 Transcript DOWNSTREAM - - - - - rs138094825 -

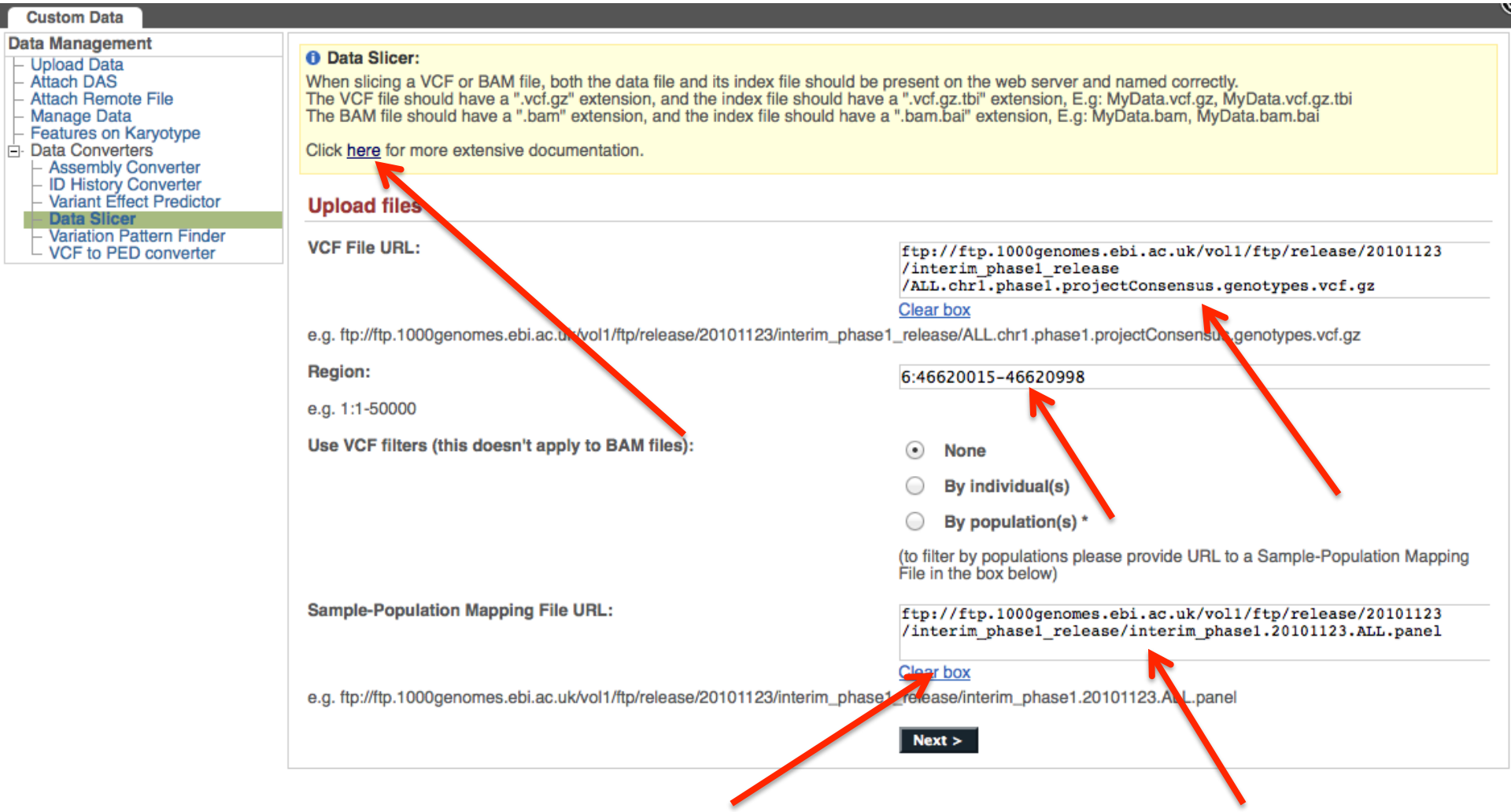
rs138094825 6:31831667 A ENSG00000204385 ENST00000229729 Transcript INTRONIC - - - - - rs138094825 -

6_31832657_C/T 6:31832657 T ENSG00000204385 ENST00000229729 Transcript NON_SYNONYMOUS_CODING 1883 1862 621 R/H cGc/cAc - PolyPhen=possibly_damaging;SIFT=deleterious

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

Data Slicing

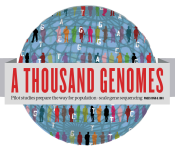


A THOUSAND GENOMES

EMBL-EBI

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



EMBL-EBI

Variation Pattern Finder

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

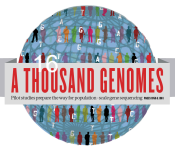
Variation Pattern Finder Output

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

EMBL-EBI

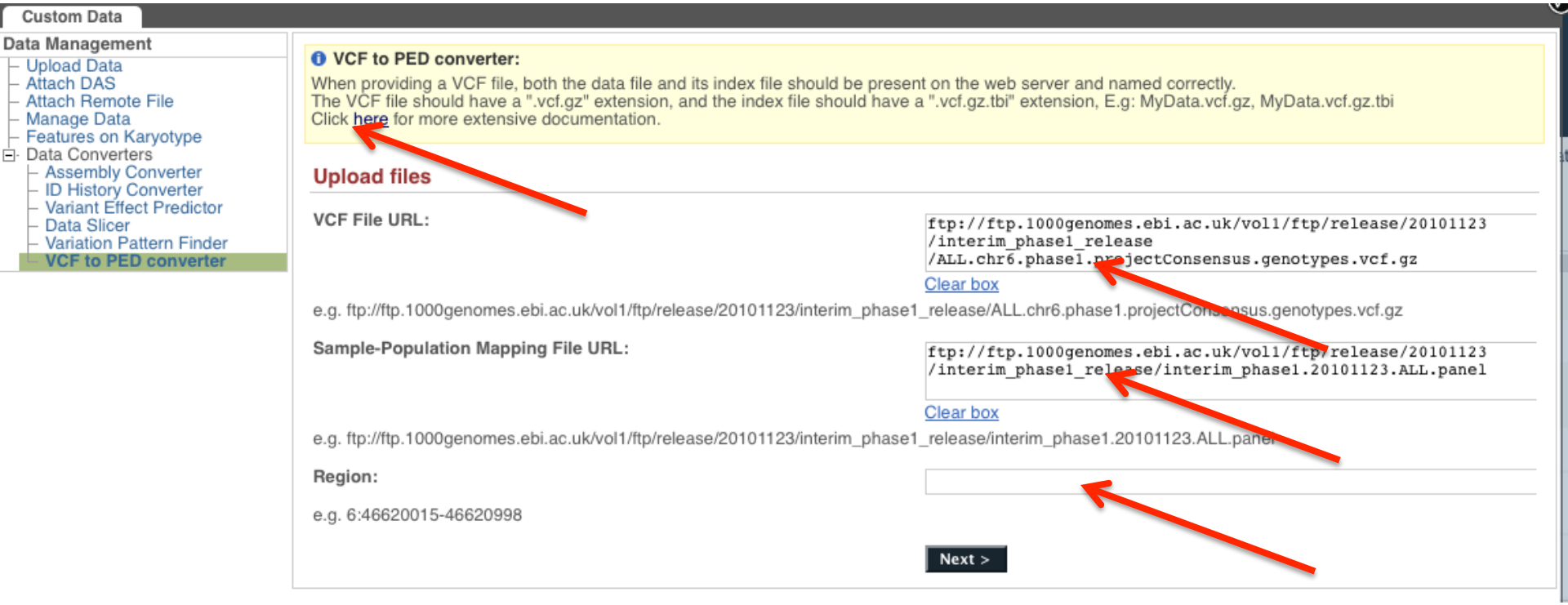
VCF to PED

- 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



EMBL-EBI

VCF to PED

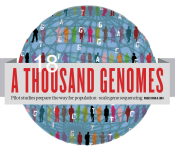


A THOUSAND GENOMES

EMBL-EBI

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



EMBL-EBI

Haploview

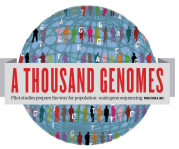
- haploview

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



EMBL-EBI

Data Availability

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

Announcements

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

Questions

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

EMBL-EBI

Thanks

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

A THOUSAND GENOMES

EMBL-EBI