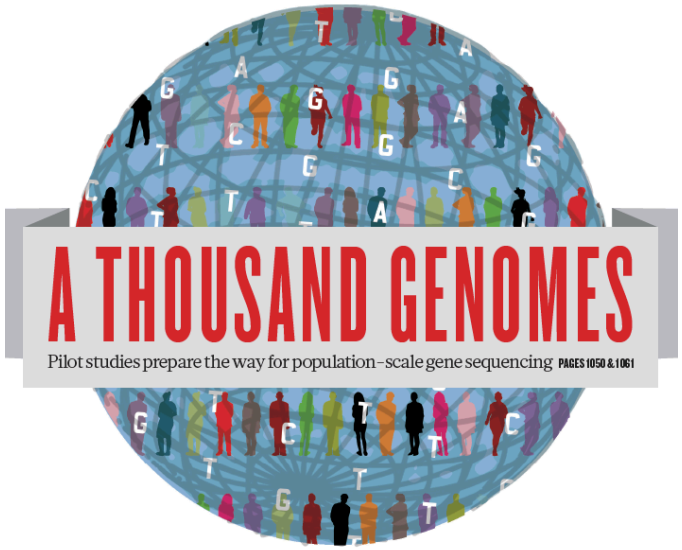


The 1000 Genomes Project

Advanced Information

Laura Clarke

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

EMBL-EBI

Sequence Data

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

EMBL-EBI

Alignment Data

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

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

Alignment data: Extended Cigar Strings

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

Operations include

M - match or mismatch

I - insertion

D - deletion

SAM extends these to include

S - soft clip

H - hard clip

N - skipped bases

P - padding

E.g. Read: ACGCA-TGCAGTtagacgt

Ref: ACTCAGTG----GT

Cigar: 5M1D2M2I2M7S

EMBL-EBI

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">

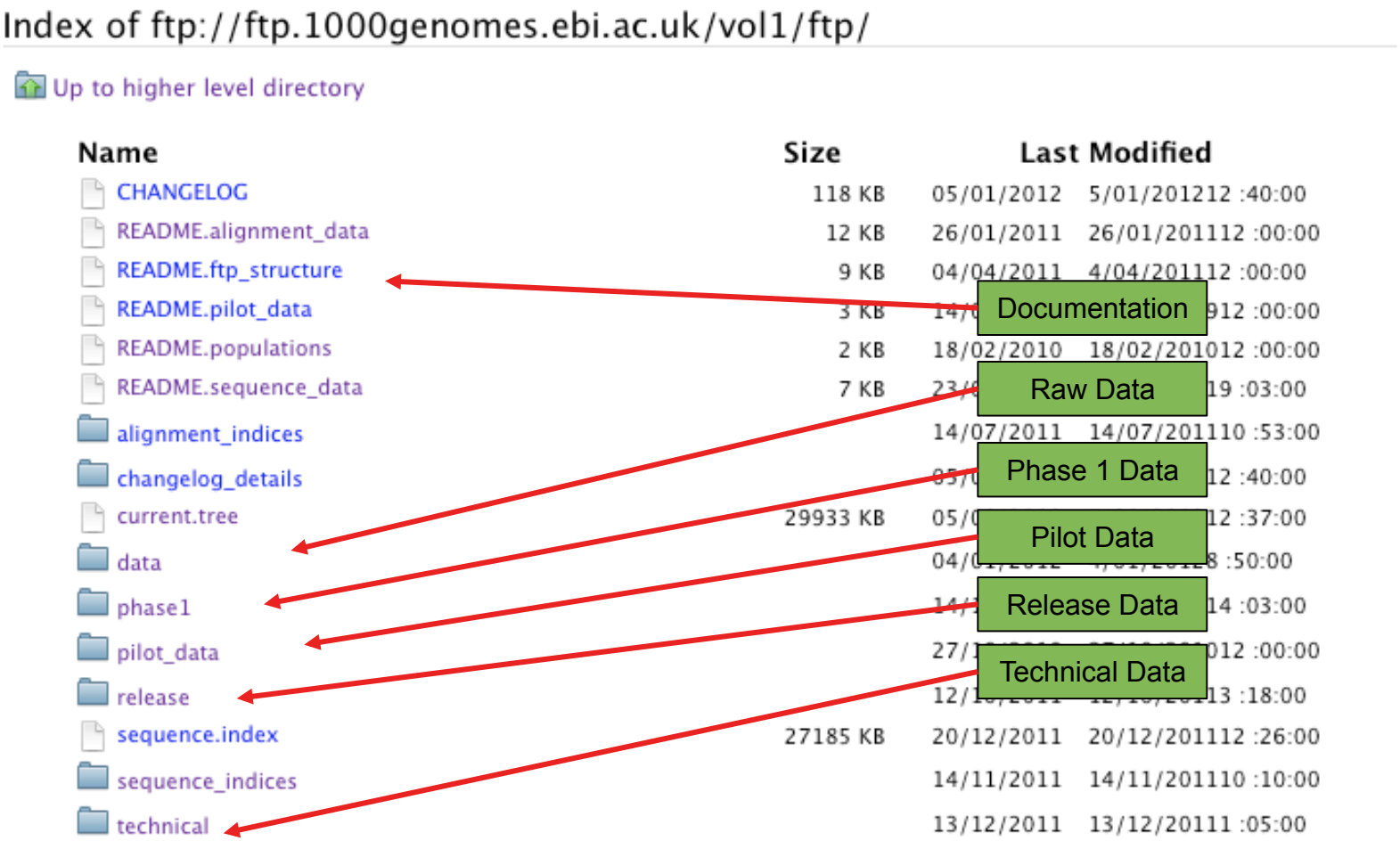
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihoods">

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

EMBL-EBI

ftp://ftp.1000genomes.ebi.ac.uk
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp



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

EMBL-EBI

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

EMBL-EBI

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

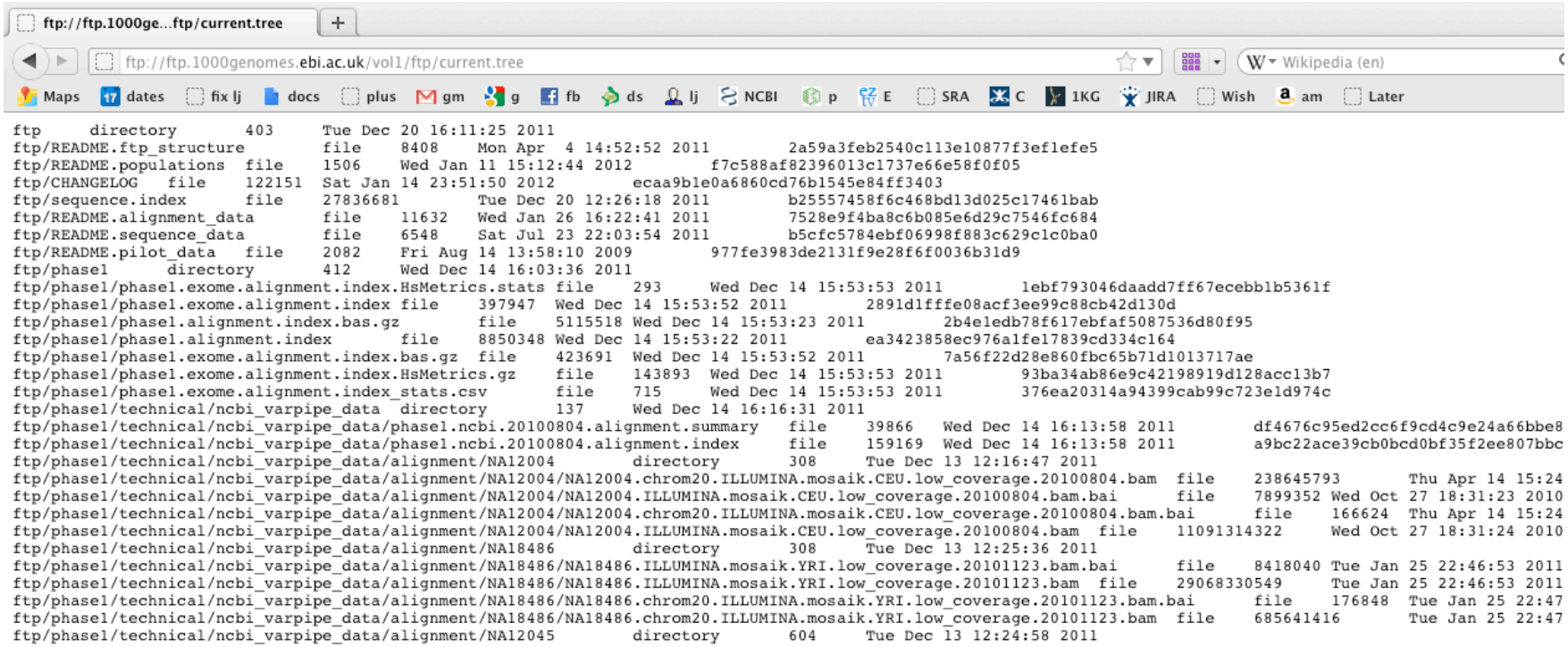
EMBL-EBI

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



A THOUSAND GENOMES

EMBL-EBI

Data Slicing

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

EMBL-EBI

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

EMBL-EBI

Exercise, Finding Data

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

2. Which exome sample from 20110521 has the highest percentage of targets covered at 20x or greater. You need to look at the 20110521.alignment.index.HsMetrics.gz file to find this

3. Find the exome bam file for this sample

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

EMBL-EBI

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? EEBEGEDEGFHHFGHHHHGHHFHHGHHDHHHHHHGHHD HHGGGHHHHHHHHHHHHHHHHGFHHHHGHHHHH

EMBL-EBI

Command Line Tools

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

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

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

- **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

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

- **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

EMBL-EBI

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

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

EMBL-EBI

Exercises, Command Line Tools

5. Get a slice of HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam for 7:114304000-114305000 (FoxP2 exon)

6. Get the equivalent section of the 20110521 release chr 7 genotypes file

7. Use vcftools vcf-subset to get the genotypes for HG00737, does HG00737 have any variant sites in this location?

8. Use this piece with tools, the variant effect predictor, the vcf pattern finder

9. Are there any snps with deleterious sift/polyphen consequences?

10. What is the most common pattern of variation in this region?

11. Use the vcf to ped script with 6:31830700-31840700 and population CEU

12. How many different haplotype blocks does the section contain?

EMBL-EBI

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 GTTTGCTGCAAGGACGATTGTTTATATTTTCACATCGCACTTAATTTCCTTGCATCTCTGCCACAAG TAGCCAGTT S=??DDBGE@CGGAE@BABIACB? A@ACCCGCGCBH=GCGEBAEBCDHHECIHBBGDHEIHHCGABIAAIHHCGBR RG:Z:SRR099984 NM:i:0 OQ:Z:HHHHEHHHHHHHHFHHHHHHHHHGEHGHHHHHHHHHHBHFHGFHHHHHHHFHHHEHHFHHH HHHHDBGFGGHHFEHF

SRR099984.344934 163 7 114304134 59 76M = 114304429 370 TTTTCACATCGCACTTAATTTCCTTGCATCTCTGCCACAAGGAGCCAGTTAGGAATTTTTTTTCAATA CATTTTCT 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

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

> 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 | vcf-subset –c HG00737 > HG00737.vcf

> vcf-stats HG00737.vcf

'hom_RR_count' => 16,

Exercise Answers, Command Line Tools

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

> grep SIFT variant_effect_output.txt

> rs182138317 7:114304331 A ENSG00000128573 ENST00000393489 Transcript NON_SYNONYMOUS_CODING 1949 1567 523 A/T Gcc/Acc - PolyPhen=possibly_damaging;SIFT=deleterious

Exercise Answers, Command Line 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)

EMBL-EBI

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

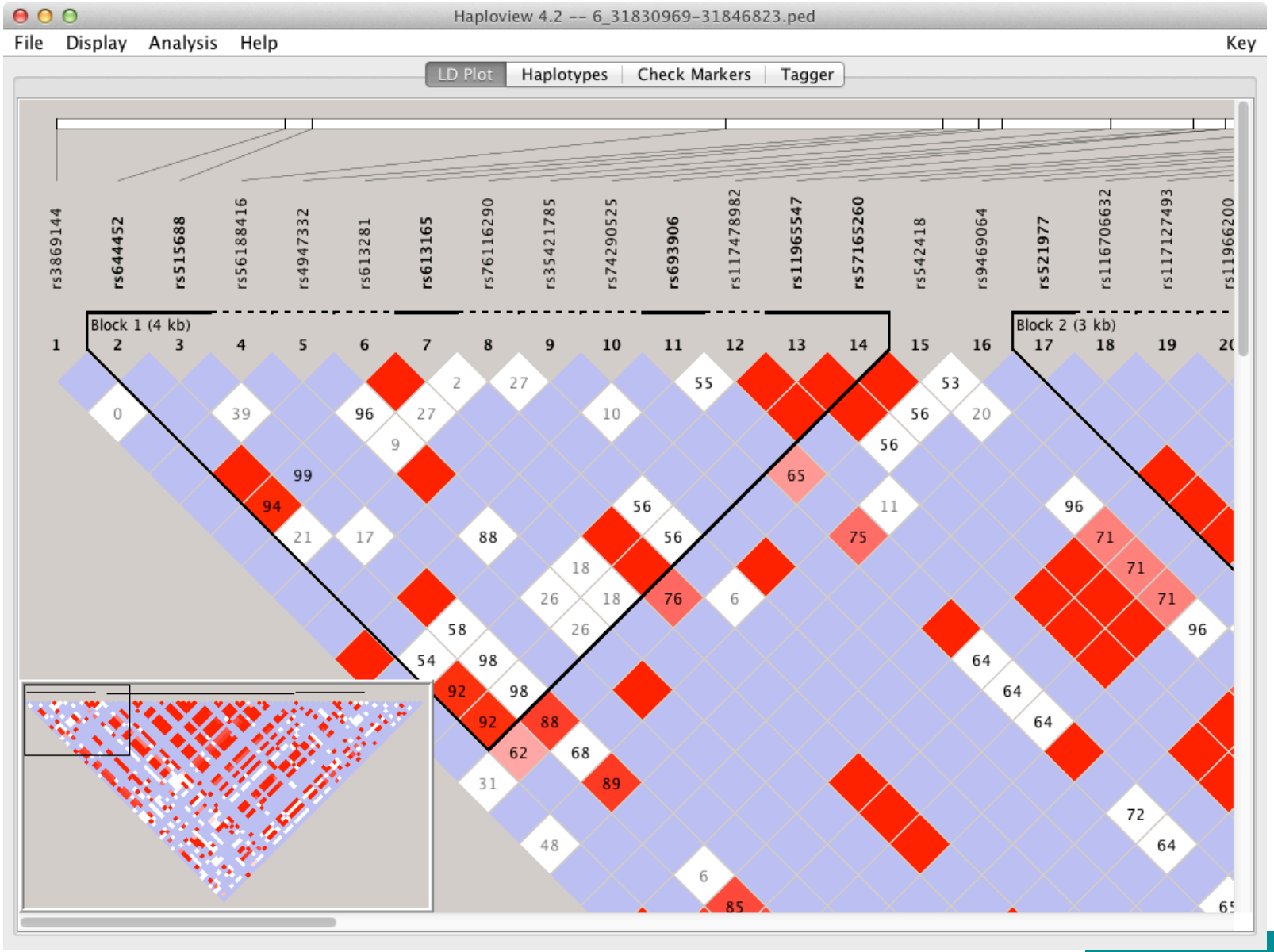
> ls ./

> 6_31830700-31840700.info

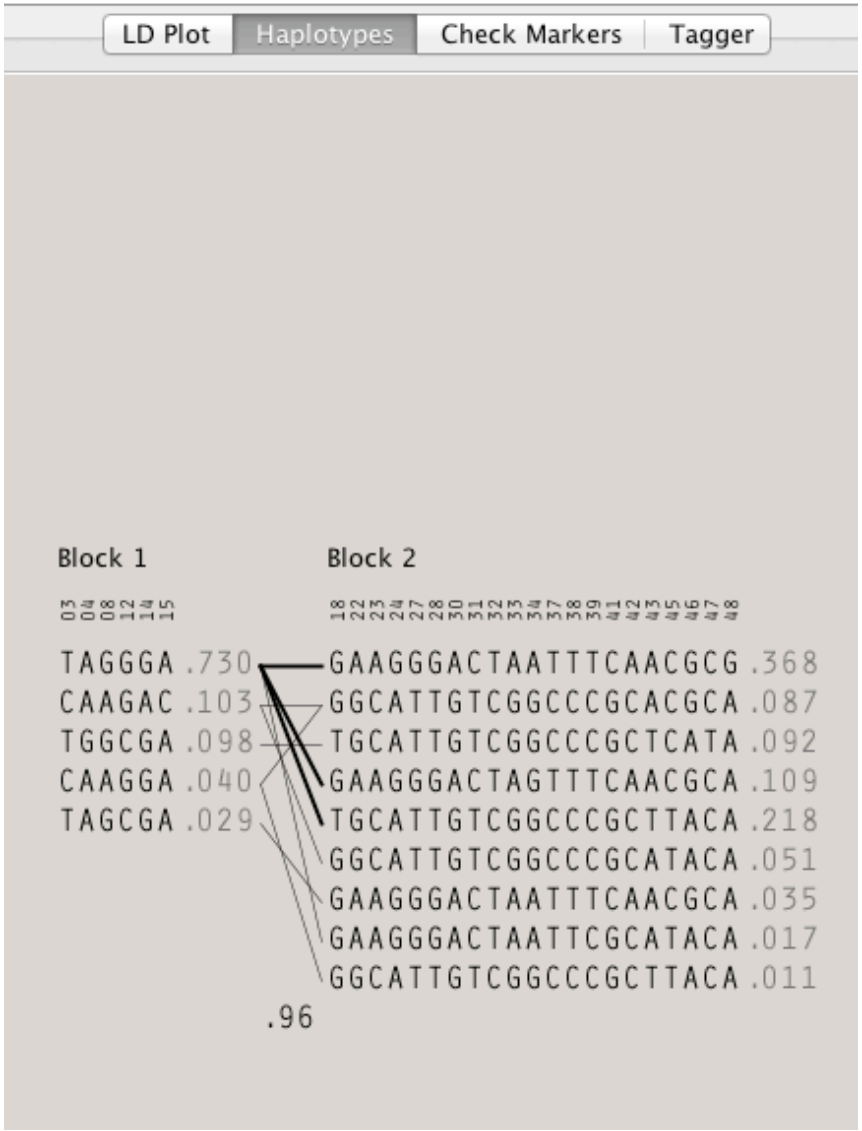
> 6_31830700-31840700.ped

EMBL-EBI

Exercise Answers



Exercise Answers



Questions

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

EMBL-EBI

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