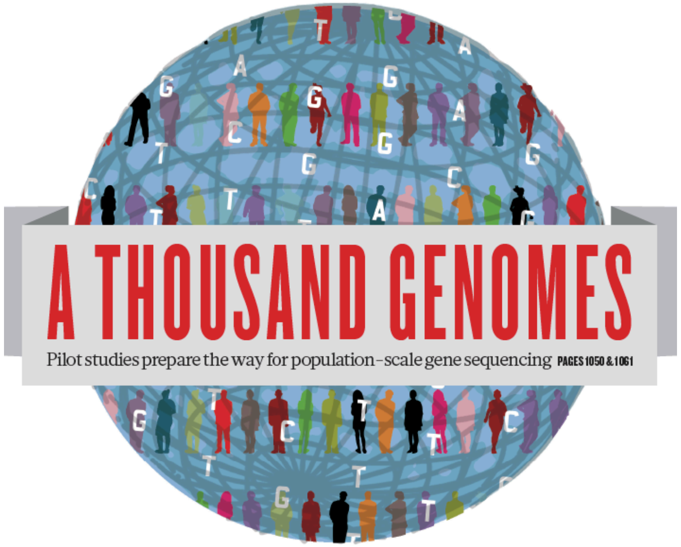

A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project: A Tutorial

Command Line Exercises

EMBL-EBI

Command Line Exercises

These exercises should guide the user through the basics of using several tools which are useful when interacting with 1000 genomes data. For more help please look at these resources

info@1000genomes.org

http://samtools.sourceforge.net/

samtools-help@lists.sourceforge.net

http://vcftools.sourceforge.net/

vcftools-help@lists.sourceforge.net

http://www.ensembl.org

dev@ensembl.org

http://biostar.stackexchange.com/

http://seqanswers.com/

Command Line Exercises

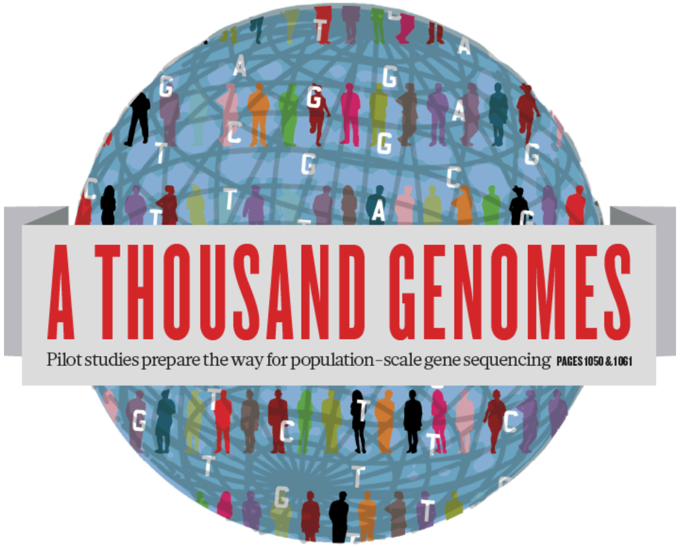
Please note it is important to have read the 1000 Genomes Tutorial Information slides before doing these exercises otherwise you may struggle to start. The answers are provided on a slide after the questions.

The slides are available here:

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

Command Line Tools

- These programs are all required to be able to do these exercises.
- Samtools http://samtools.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
- Vcftools http://vcftools.sourceforge.net/
- The ensembl variation and core apis http://www.ensembl.org/index.html
- The variant effect predictor ftp://ftp.ensembl.org/pub/misc-scripts/Variant_effect_predictor/
- The variation pattern finder ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/variation_pattern_finder/version_1.0
- VCF to PED Converter ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/vcf_to_ped_converter/version_1.0/
- Haploview http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/downloads


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project:

Exercise 1: Finding Data and viewing data on the 1000 genomes ftp site

Exercise: Finding and Viewing Data

- Finding data can use either ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree or http://www.1000genomes.org/ftpsearch
- Find a omni vcf file
- View the data it contains for the region 6:31831625-31834704
- Find the mapped low coverage bam file for HG01375
- View the data it contains for the same region

Answers

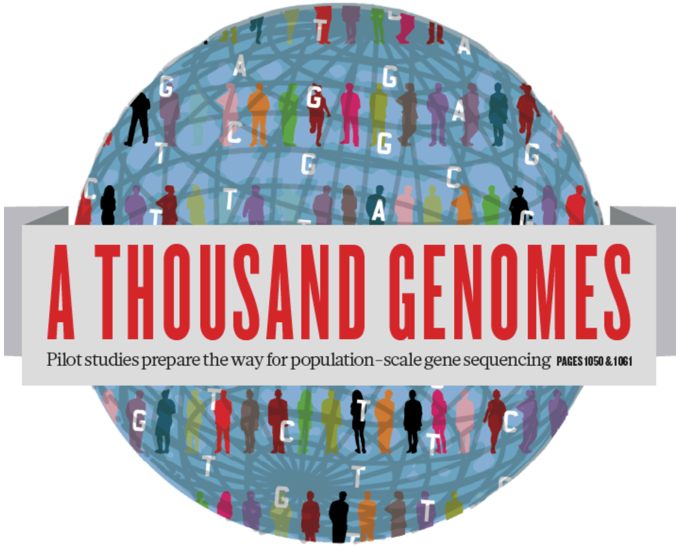
-
  ```
  wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree
  ```

-
  ```
  grep omni current.tree | grep vcf | grep 2012 | grep -v tbi | cut –f1
  ```

-
  ```
  tabix -h ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31831625-31834704
  ```

-
  ```
  grep HG01375 current.tree  | grep low_coverage | grep mapped | grep -v bai | grep -v bas | grep -v unmapped | cut -f1
  ```

-
  ```
  samtools view http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam 6:31831625-31834704
  ```


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project:

Exercise 2: Using 1000 Genomes Tools

Exercise 3 Using 1000 Genomes Tools

- Get a 6:31831625-31834704 slice of our chr6 release vcf file using tabix.
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_release_v2.20101123.snps_indels_svs.vcf.gz
  - Use vcf subset to get just the genotype info for HG01375
  - Direct this into a file
- Use the retrieved vcf file with the variant effect predictor script
- Which variants have deleterious sift/polyphen consequences?
- Use the variant pattern finder to look at the pattern of variation in the same region using the remote chr6 vcf file
- Produce ped and locus information files using the vcf to ped tool for the CEU population
- Have a look at the ped file using Haploview

Answers

- **tabix -h** ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_release_v2.20101123.snps_indels_svs.vcf.gz 6:31831625-31834704 > 6_381831625_3184704.vcf
- **perl variant_effect_predictor.pl** -input 6_381831625_3184704.vcf -sift p -polyphen p -check_existing
  - 6_31832657_C/T, rs141954433 and rs149841290 all have deleterious sift and polyphen results
- **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 6:31831625-31834704 -expand
- **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
- **java –j Haploview.jar**

Any questions?

- Please email info@1000genomes.org if you have any questions or feedback about this resource.

Thanks

- The 1000 Genomes Project Consortium
- Paul Flicek, Laura Clarke
- Richard Smith, Holly Zheng Bradley and Ian Streeter
- Giulietta Spudich and Bert Overduin

A THOUSAND GENOMES

EMBL-EBI