H3ABioNet

Pan African Bioinformatics Network for H3Africa

Downstream analysis of data

Data interpretation

- Data processing requires technical skills and compute resources
- Downstream analysis may require some bioinformatics skills but mostly biological knowledge
- This is the part where you answer your biological question!

H3Africa projects

- Microbiome data –usually correlation of OTU distribution with a phenotype or measurement
- For projects looking for genetic association with diseases:
  - What SNP(s) are potentially associated with my disease?
  - What is their minor allele frequency in my population?
  - Where is the SNP located –in a gene?
  - What is the functional effect of the SNP?
  - Are the SNPs or genes connected?

Variant prioritisation

- After your GWAS your associated SNPs may be already known to be associated with the disease or may be novel
- You may have very few significant SNPs, or none, or too many!
- Can look further at functional effect of a SNP or connected SNPs, or at the gene level: pathway, expression, interactions, known phenotype associations

Variant prioritisation

- Synonymous exonic variants considered ‘silent’ & can be discarded
- Non-synonymous (missense) variants may affect protein function
  - amino acid change does not automatically imply deleteriousness
- Can look at potential impact on function (HUMA)
- Large indels affect function
  - frameshift indels almost always functional
- Splice sites are sensitive to mutation
- Stop-gain/loss, frameshift and splice-site variants are automatically ‘interesting’
- UTR variants don’t affect the protein sequence
  - only have an effect if mutation is in regulatory element
  - there are often thousands to evaluate

Variant prioritisation

- Conservation is important- Variants in regions that are highly conserved across species are likely to be in genes that serve important biological functions
- MAF is important –for a rare or Mendelian disease 1% is good
- Can filter on MAF if you know what MAF you want
- Consider MAF in your population not all populations

Finding SNP info in public data

- dbSNP:
  - Summary of allele frequency across datasets, e.g. 1000 genomes, HAPMAP, HGP, ExAC, ESP6500
  - Info on genome location, if in a gene, if cause of amino acid change
- ExAC (http://exac.broadinstitute.org/)
  - exome aggregation consortium
- SNPedia (https://www.snpedia.com/index.php/SNPedia)
  - Info on effects of variants
- RegulomeDB (http://www.regulomedb.org/)
  - Non-coding variants, data from ENCODE

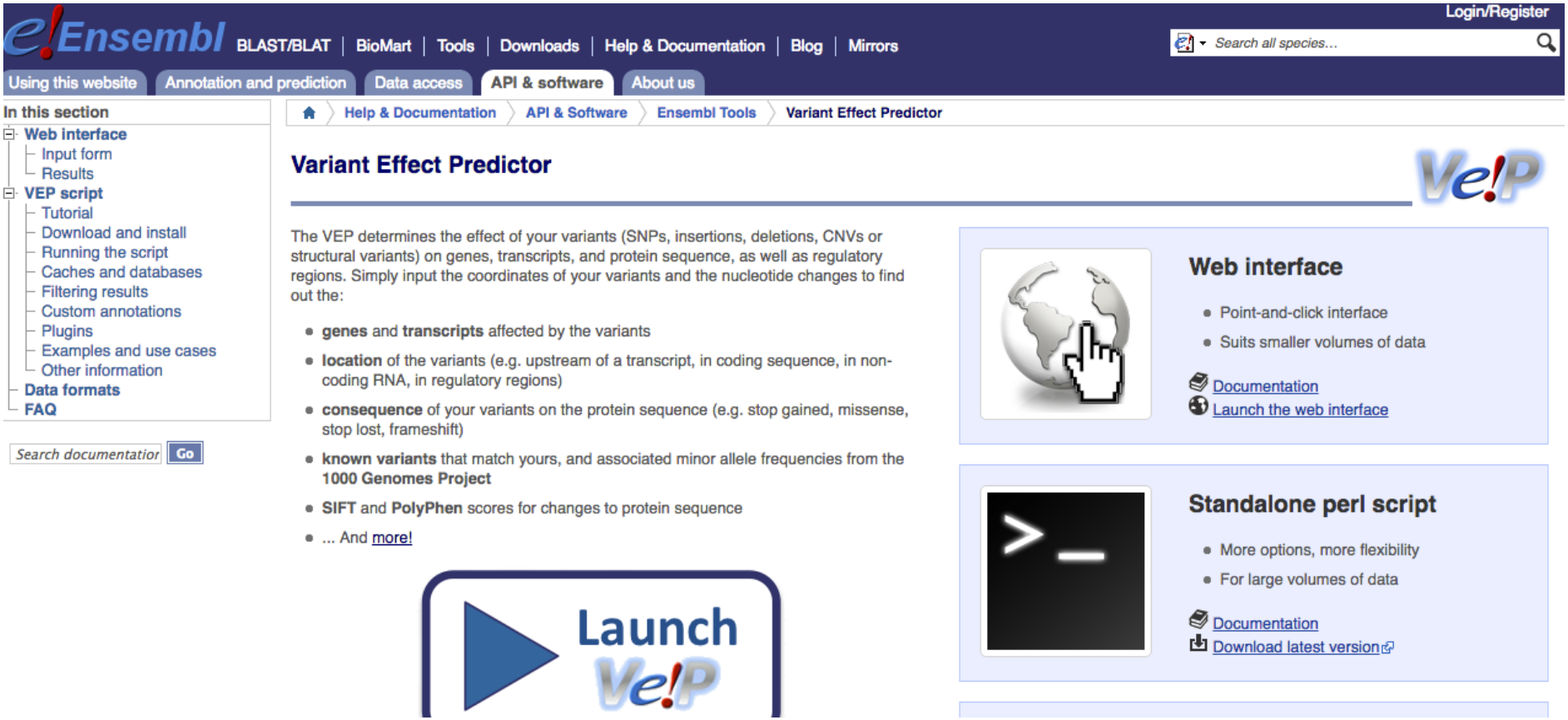
Finding SNP info in public data

- OMIM (http://www.omim.org/)
  - Gene level information on genotype to phenotype
- ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/)
  - Clinical consequence of variants
  - Links variant info to phenotypes
- COSMIC (http://cancer.sanger.ac.uk/cosmic)
  - Catalogue Of Somatic Mutations In Cancer
- PharmGKB (https://www.pharmgkb.org/)
  - Pharmacogenomic information for variants

Novel SNP annotation

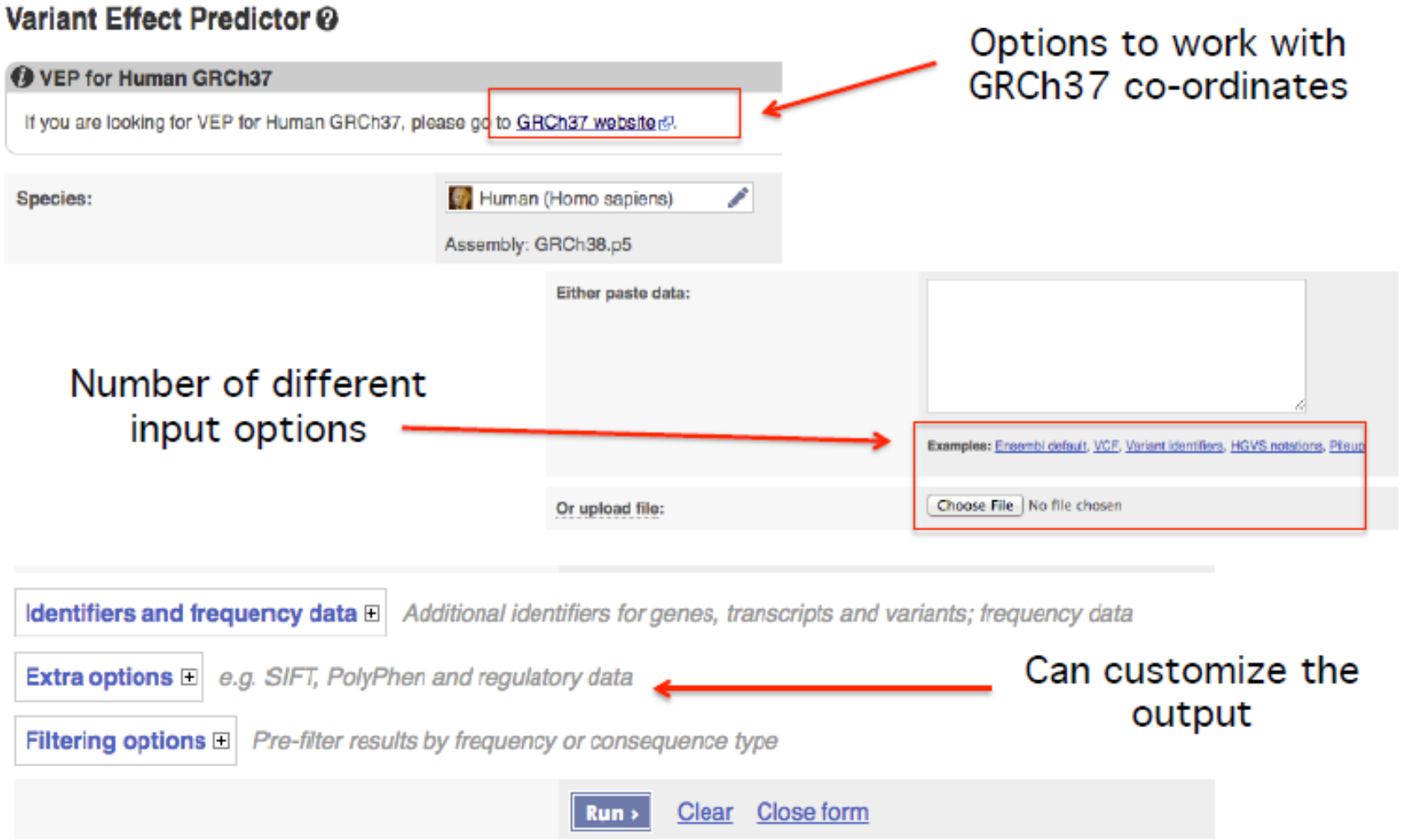
- SNP function prediction tools –ANNOVAR (Annotation Of VARiants) –command line and web version, includes:
  - SIFT (http://si:.jcvi.org/www/SIFT_chr_coords_submit.htm) looks at coding variants
  - PolyPhen (http://gene,cs.bwh.harvard.edu/pph2/) –coding and nsSNPs
- FATHMM (http://fathmm.biocompute.org.uk/) functional analysis through HMMs, coding and on-coding variants
- GATK Variant annotator (https://www.broadins=tute.org/gatk/index.php)
- Ensembl Variant effect predictor tool (http://www.ensembl.org/info/docs/tools/vep/index.html)

An example - Ensembl VEP



Images Collen Saunders' slides

An example - Ensembl VEP



Images Collen Saunders' slides

An example - Ensembl VEP

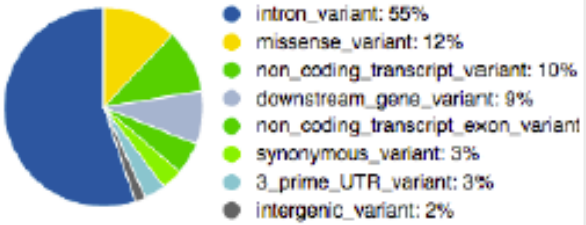
Variant Effect Predictor results

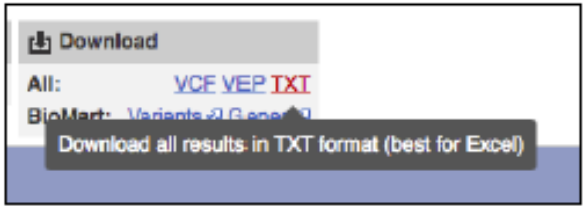
Job details

Summary statistics

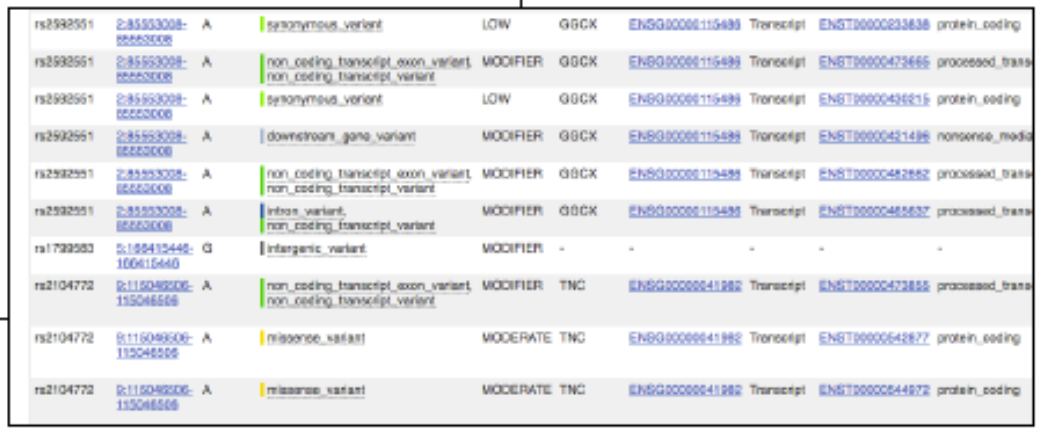
Category	Count
Variants processed	6
Variants remaining after filtering	6
Novel / existing variants	0 (0.0%) / 6 (100.0%)
Overlapped genes	5
Overlapped transcripts	51
Overlapped regulatory features	-

Consequences (all)





Coding consequences



Images Collen Saunders' slides

Gene-level analysis

- Top down or bottom up approach
- Mapping SNPs to genes and doing analysis at the gene level
- How to merge p-values
  - Determine what is extent of gene –no. bp up- and downstream
  - average, lowest, correct for gene length
- Can look at expression of genes in relevant cells
- Look at known gene-phenotype associations
- Look at pathways genes are associated with

Looking at phenotype info

- **Phenomizer** (http://compbio.charite.de/phenomizer/) – useful for clinical studies

- **Phenolyzer** (http://phenolyzer.usc.edu/)
  - Looks for links between a gene list and phenotypes
  - Looks for links between a genomic region and phenotypes

Phenolyzer result

Submission information

Phenotypes are interpretated.

4 genes are entered within the genelist.

At most 2000 genes will be found in details, for the complete list, please download the report here.

1 disease terms have been entered, among which, 1 terms have corresponding records in our database.

They are: ehlers_danlos_syndrome

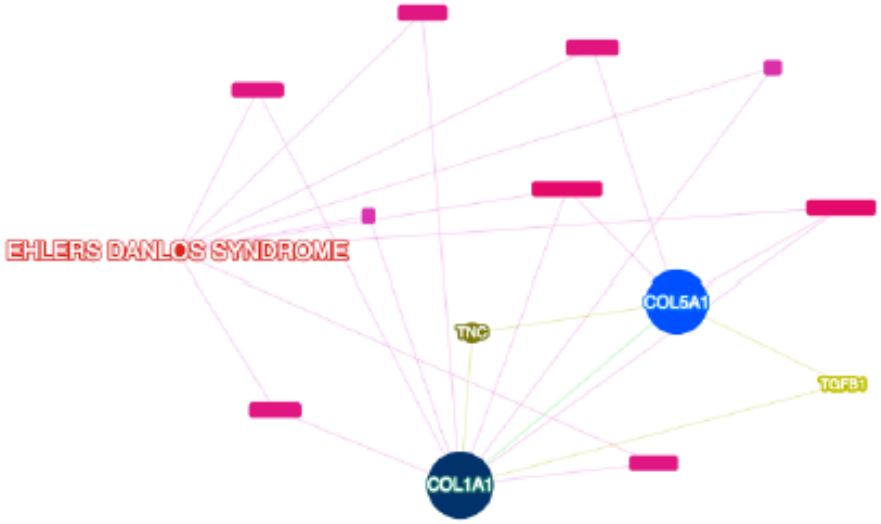
The GENELIST/REGION SPECIFIC REPORT could be found Here(4 genes).

The GENELIST/REGION SPECIFIC GENE LIST could be found Here(4 genes).

The WHOLE REPORT could be found Here(7635 genes).

The FINAL GENE LIST could be found Here(7635 genes).

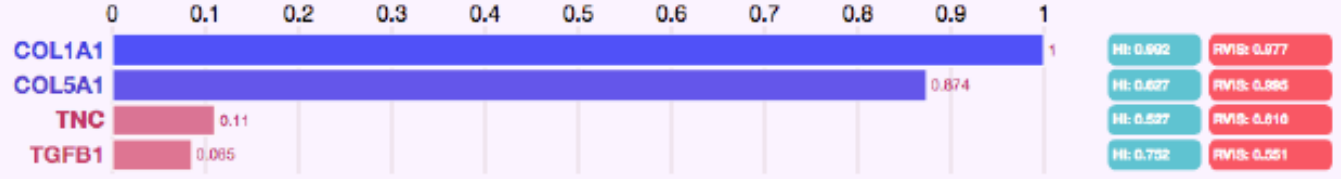
The SEED GENE REPORT could be found Here(19 genes).



Summary | Network | Barplot | Details

Barplot

View source data



http://phenolyzer.usc.edu/

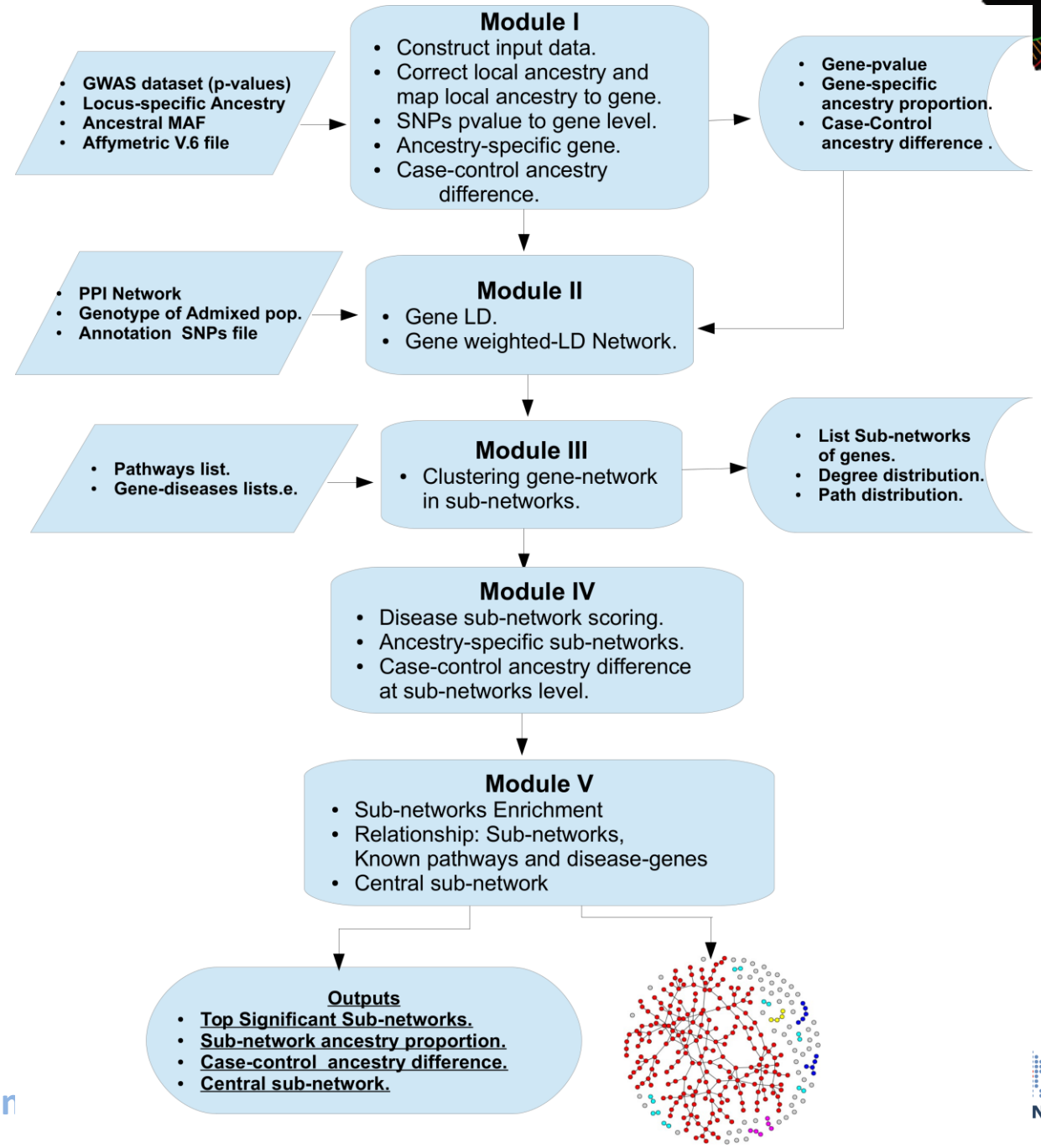
Pathway analysis

- Many SNPs found in GWAS have moderate effect sizes, could be combination of SNPs
- Genes work together in interaction networks and pathways
- Try to find enrichment of pathways in SNP list
- Can do either:
  - Candidate pathway analysis
  - Genome-wide pathway analysis
- Tools tend to collapse SNPs to a gene, methods should correct for this in p-values

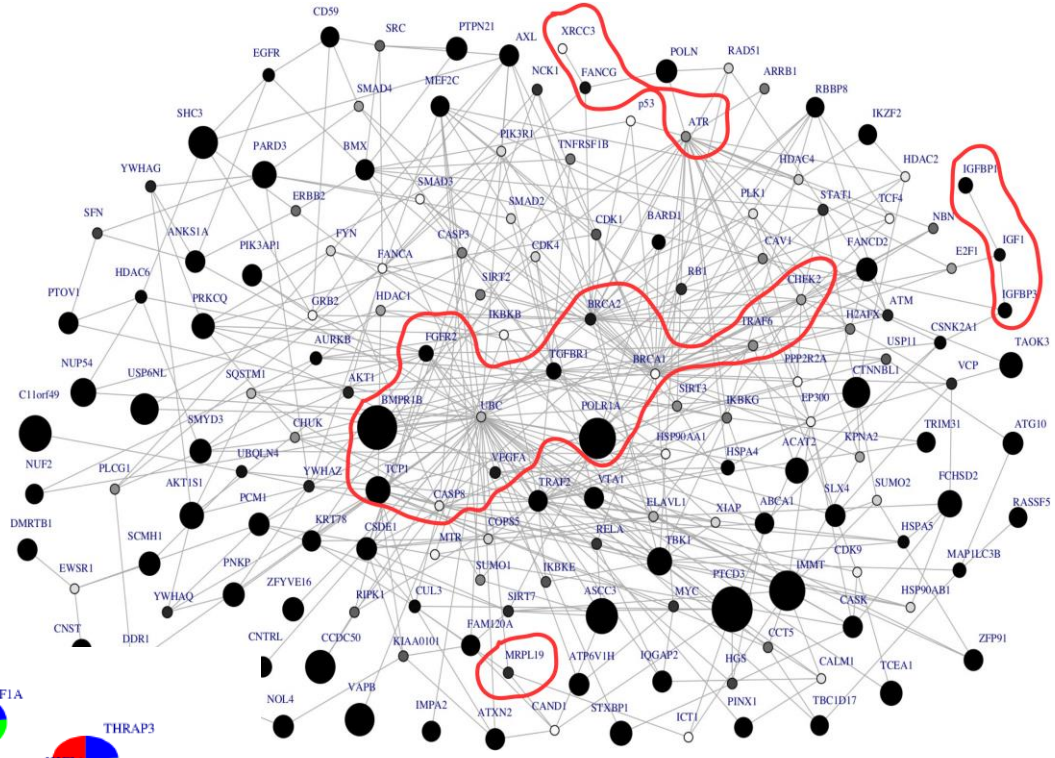
An example -ancGWAS

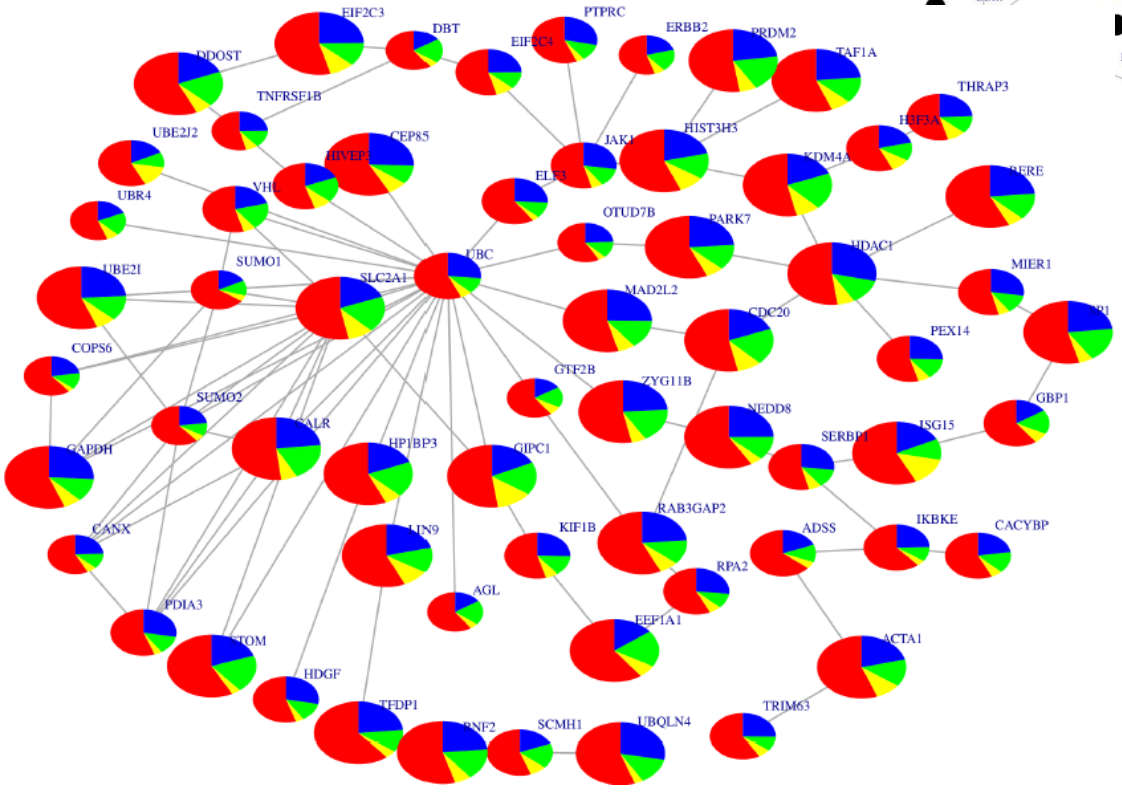
- Algebraic graph-based centrality measure for identifying significant disease sub-networks
- Accounts for linkage disequilibrium
- Integrates the association signal from GWAS data sets into the human protein–protein interaction (PPI) network
- Can look for subnetworks enriched in certain ancestries for admixed populations
- http://www.cbio.uct.ac.za/~emile/software.html

ancGWAS pipeline



ancGWAS result





12th May 2016, Senegal

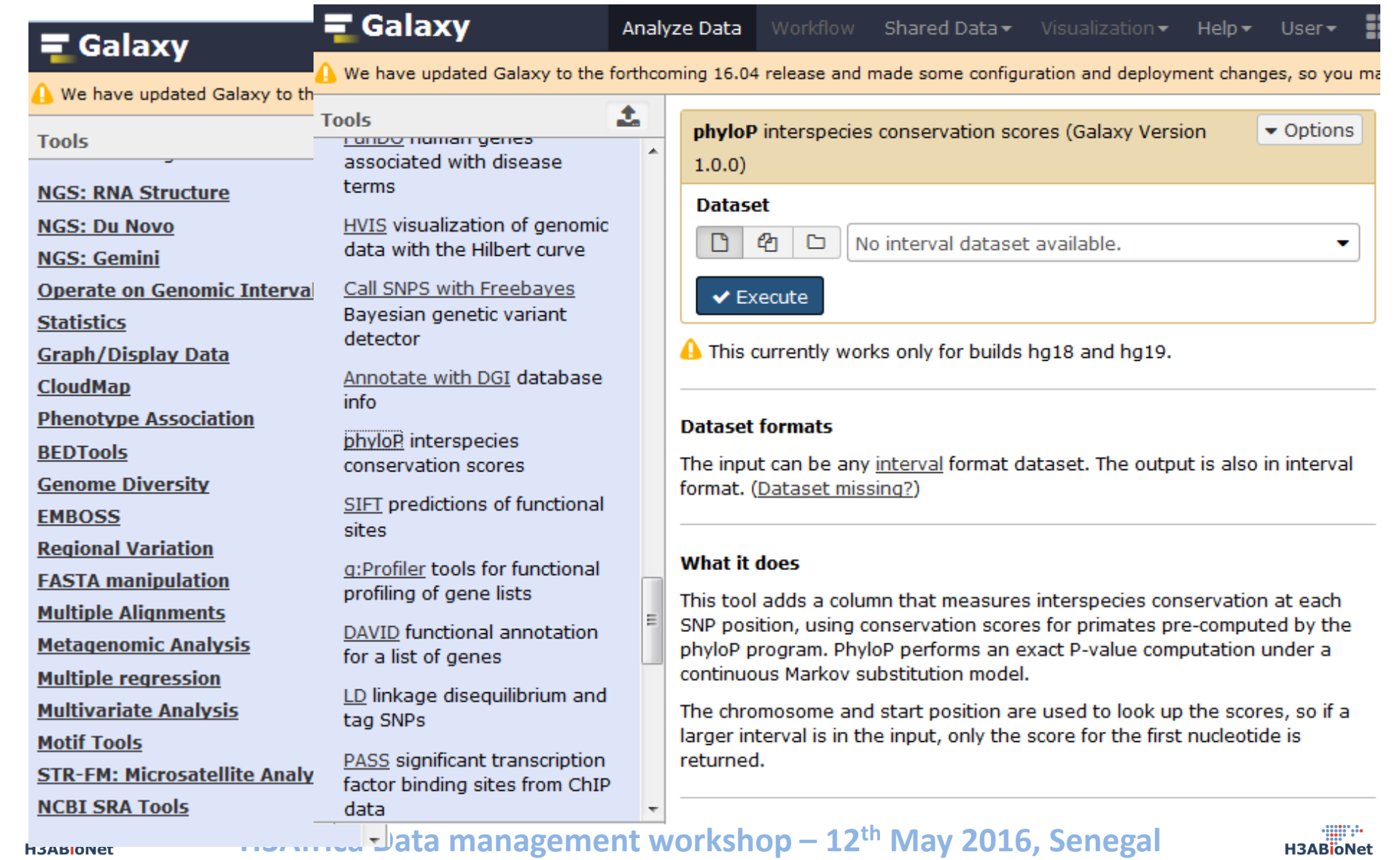
H3ABioNet

Data manipulation & visualisation

- Data visualization is important, especially when dealing with big data
- Genesis –improving STRUCTURE and PCA plots
- However, the more data you are dealing with the more technical skills are required!
- Galaxy tools –easy to use data manipulation tools

Data manipulation & visualisation

Data manipulation & visualisation



Conclusions

- Data generation is only half the project
- Data processing and analysis is a major component
- Biological analysis and interpretation is key, and can take time
- Many tools are available for small-scale or large-scale users
- Make sure you visualise the data appropriately
- Go back to the literature to look at relevance of results in the context of other work