



**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](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.broadinstitute.org/gatk/index.php>)
- Ensembl Variant effect predictor tool (<http://www.ensembl.org/info/docs/tools/vep/index.html>)



# An example - Ensembl VEP

The screenshot shows the Ensembl VEP website interface. At the top, there is a navigation bar with links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. A search bar is located on the right. Below the navigation bar, there are tabs for 'Using this website', 'Annotation and prediction', 'Data access', 'API & software', and 'About us'. The 'API & software' tab is selected, and the breadcrumb trail shows 'Help & Documentation > API & Software > Ensembl Tools > Variant Effect Predictor'. The main content area is titled 'Variant Effect Predictor' and includes a description of the tool's function: 'The VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. Simply input the coordinates of your variants and the nucleotide changes to find out the:'. Below this, a list of features is provided: genes and transcripts affected, location of variants, consequence on protein sequence, known variants from the 1000 Genomes Project, SIFT and PolyPhen scores, and more. A 'Launch Ve!P' button is prominently displayed. To the right, there are two panels: 'Web interface' with a globe icon and 'Standalone perl script' with a terminal icon. Both panels list their respective features and provide links to documentation and the latest version.

**Variant Effect Predictor**

The VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. Simply input the coordinates of your variants and the nucleotide changes to find out the:

- **genes** and **transcripts** affected by the variants
- **location** of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- **consequence** of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- **known variants** that match yours, and associated minor allele frequencies from the 1000 Genomes Project
- SIFT and PolyPhen scores for changes to protein sequence
- ... And [more!](#)

**Web interface**

- Point-and-click interface
- Suits smaller volumes of data

[Documentation](#)  
[Launch the web interface](#)

**Standalone perl script**

- More options, more flexibility
- For large volumes of data

[Documentation](#)  
[Download latest version](#)

**Launch Ve!P**

Images Collen Saunders' slides

H3Africa Data management workshop – 12<sup>th</sup> May 2016, Senegal



# An example - Ensembl VEP

## Variant Effect Predictor

### VEP for Human GRCh37

If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Options to work with GRCh37 co-ordinates

Species:

Human (Homo sapiens)

Assembly: GRCh38.p5

Either paste data:

Number of different input options

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#), [Pileup](#)

Or upload file:

Choose File No file chosen

Identifiers and frequency data

Additional identifiers for genes, transcripts and variants; frequency data

Extra options

e.g. SIFT, PolyPhen and regulatory data

Can customize the output

Filtering options

Pre-filter results by frequency or consequence type

Run >

Clear

Close form

Images Collen Saunders' slides

H3Africa Data management workshop – 12<sup>th</sup> May 2016, Senegal





# An example - Ensembl VEP

### Variant Effect Predictor results

Job details [B]  
Summary statistics [B]

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)

- intron\_variant: 55%
- missense\_variant: 12%
- non\_coding\_transcript\_variant: 10%
- downstream\_gene\_variant: 9%
- non\_coding\_transcript\_exon\_variant: 3%
- synonymous\_variant: 3%
- 3\_prime\_UTR\_variant: 3%
- intergenic\_variant: 2%

#### Coding consequences

- missense\_variant: 78%
- synonymous\_variant: 22%

Variant ID	Ref	Alt	Consequence	Impact	Gene	Transcript	Effect
rs2592551	C	A	synonymous_variant	LOW	GGCX	ENSG00000115489	Transcript
rs2592551	C	A	non_coding_transcript_exon_variant, non_coding_transcript_variant	MODIFIER	GGCX	ENSG00000115489	Transcript
rs2592551	C	A	synonymous_variant	LOW	GGCX	ENSG00000115489	Transcript
rs2592551	C	A	downstream_gene_variant	MODIFIER	GGCX	ENSG00000115489	Transcript
rs2592551	C	A	non_coding_transcript_exon_variant, non_coding_transcript_variant	MODIFIER	GGCX	ENSG00000115489	Transcript
rs2592551	C	A	intron_variant, non_coding_transcript_variant	MODIFIER	GGCX	ENSG00000115489	Transcript
rs1739283	G	C	intergenic_variant	MODIFIER	-	-	-
rs2104772	A	G	non_coding_transcript_exon_variant, non_coding_transcript_variant	MODIFIER	TNC	ENSG00000411982	Transcript
rs2104772	A	G	missense_variant	MODERATE	TNC	ENSG00000411982	Transcript
rs2104772	A	G	missense_variant	MODERATE	TNC	ENSG00000411982	Transcript

Download

All: [VCF VEP TXT](#)

BioMart: [Variants 2 Gene 1](#)

Download all results in TXT format (best for Excel)

Images Collen Saunders' slides

H3Africa Data management workshop – 12<sup>th</sup> May 2016, Senegal





# 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 interpreted.

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](#)

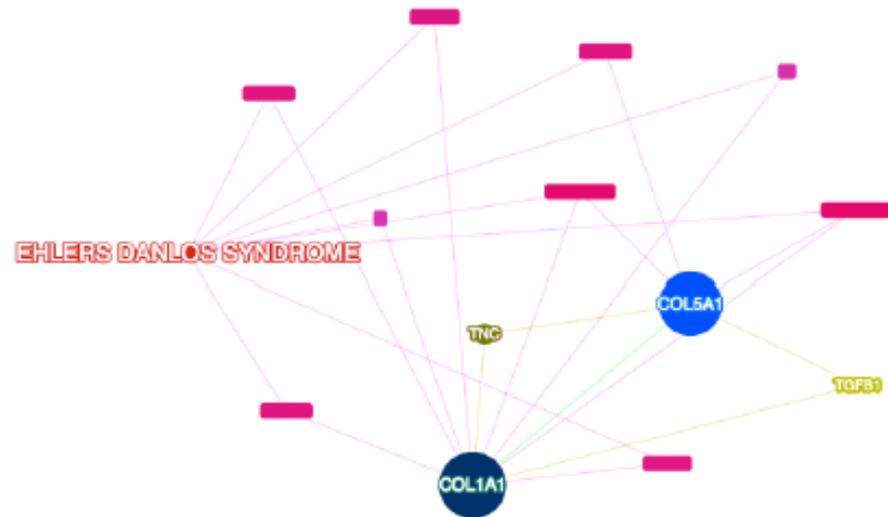
The GENELIST/REGION SPECIFIC REPORT could be found [Here\(4 genes\)](#).

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\(7635 genes\)](#).

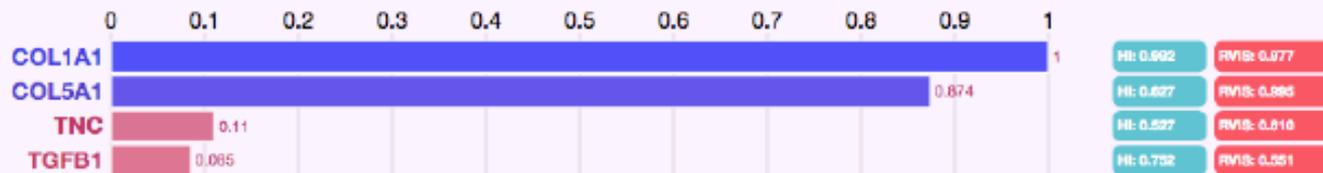
The SEED GENE REPORT could be found [Here\(19 genes\)](#).



- Summary
- Network
- Barplot
- Details

## Barplot

View source data



<http://phenolyzer.usc.edu/>

H3Africa Data management workshop – 12<sup>th</sup> May 2016, Senegal





# 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

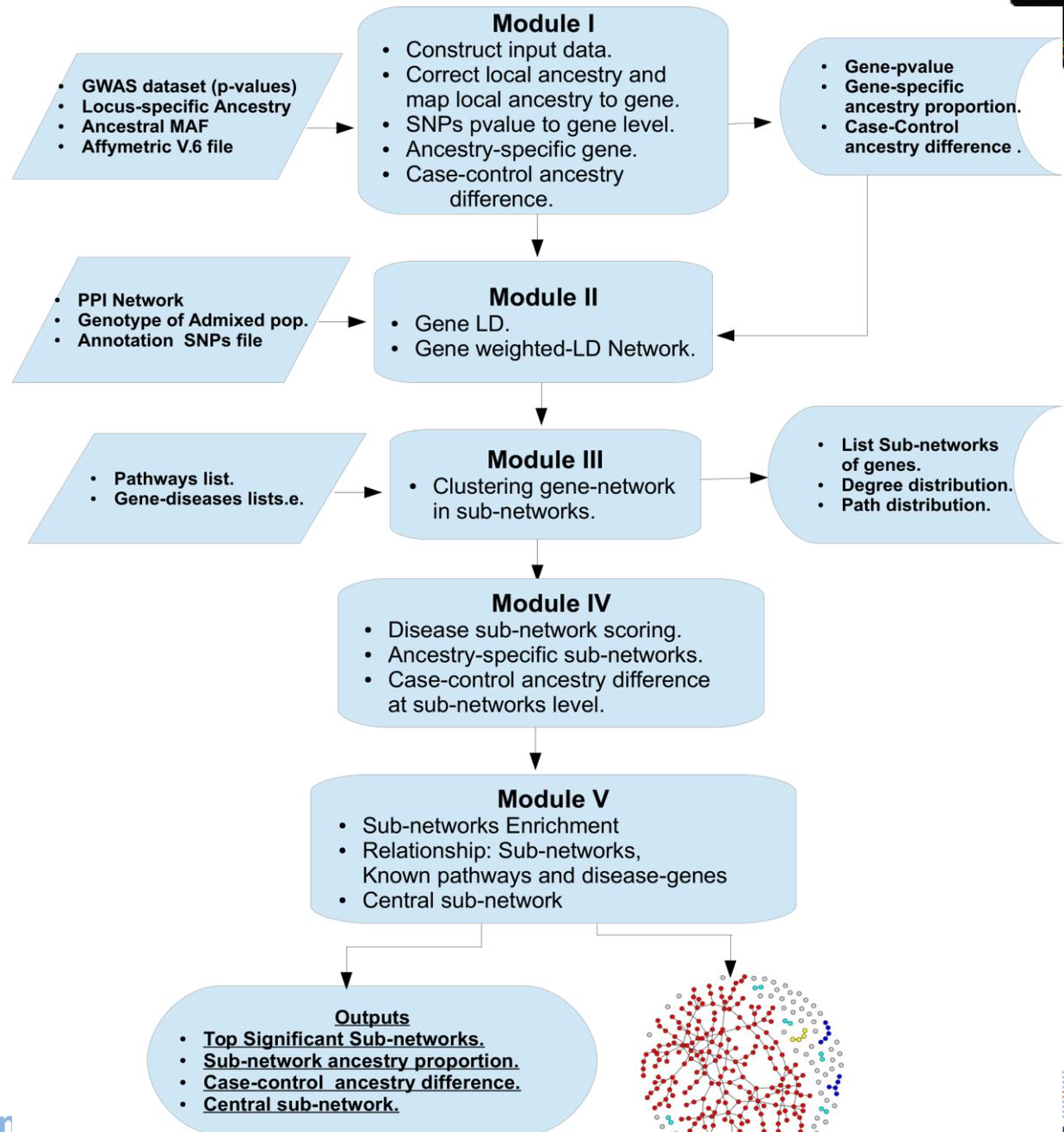


# An example -ancGWAS

- 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>

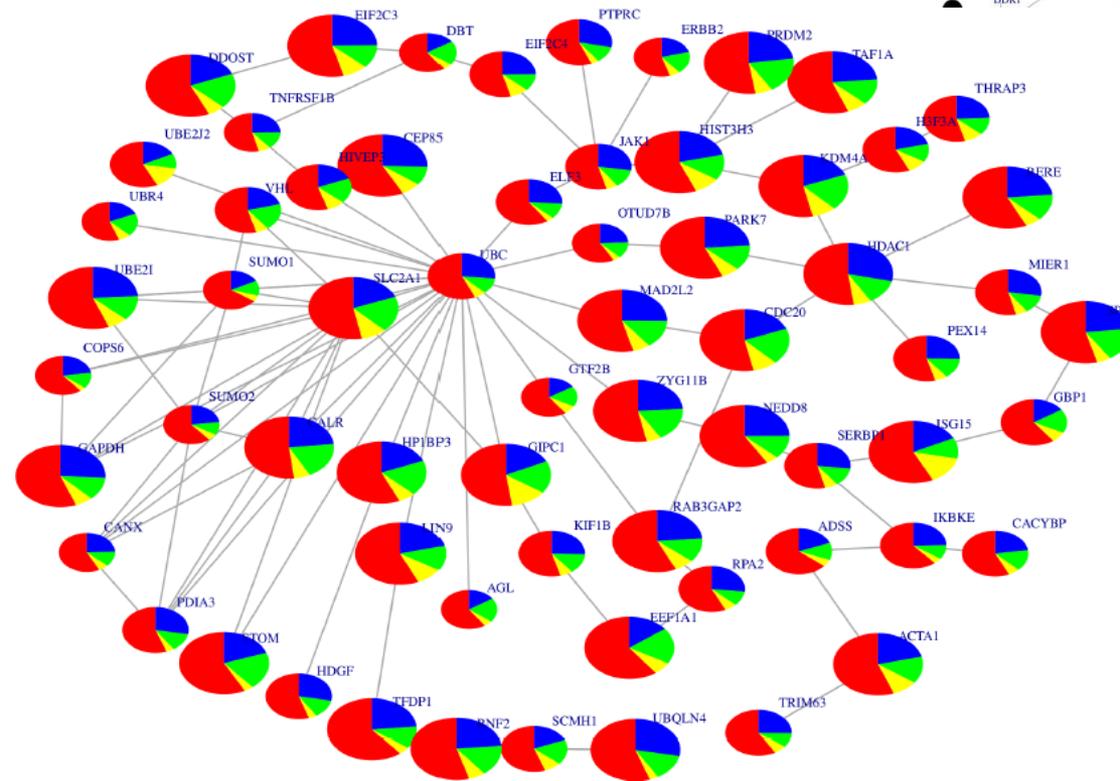
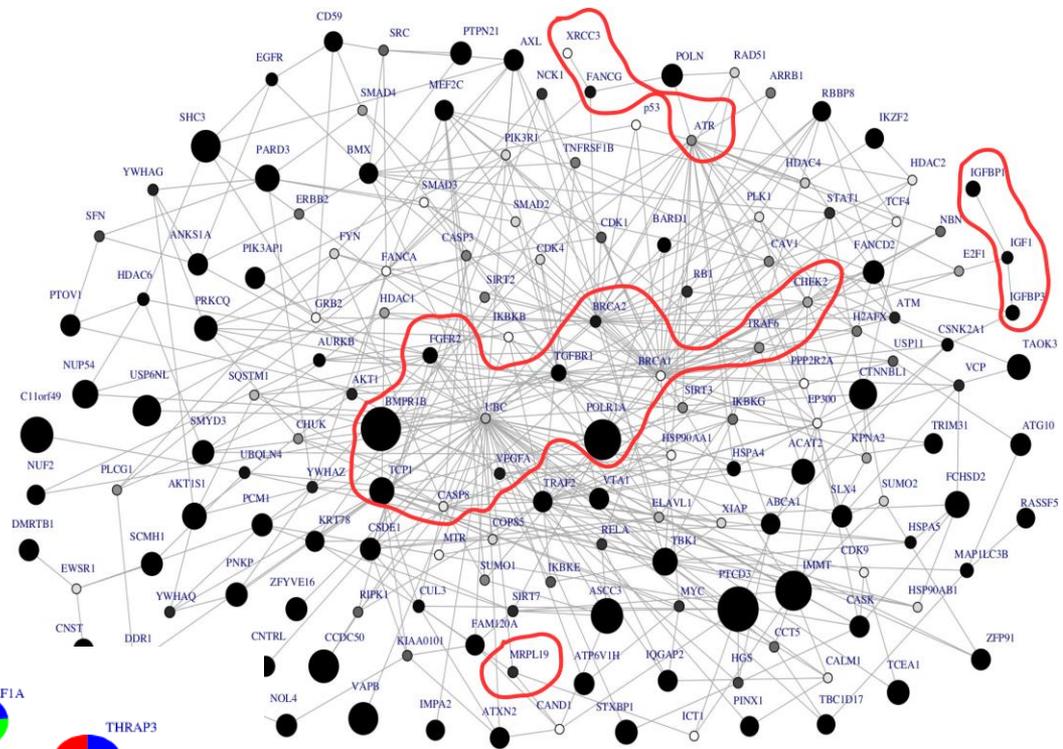


# ancGWAS pipeline





# ancGWAS result



12<sup>th</sup> May 2016, Senegal



# 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

**Galaxy** Analyze Data Workflow Shared Data Visualization Help User Using 0 bytes

We have updated Galaxy to the forthcoming 16.04 release and made some configuration and deployment changes, so you may encounter a few problems. We are working

**Tools**

- NGS: BamTools**
- NGS: Picard**
- NGS: VCF Manipulation**
- NGS: Peak Calling**
- NGS: Variant Analysis**
  - [ANNOVAR](#) Annotate VCF with functional information using ANNOVAR
  - [SnEff](#) Variant effect and annotation
  - [SnEff Available Databases](#)
  - [SnEff Download](#) Download a new database
  - [Mutate Codons](#) with SNPs
  - [Variant Annotator](#) process variant counts
  - [FreeBayes](#) - updated bayesian genetic variant detector
  - [Naive Variant Caller](#) - tabulate variable sites from BAM datasets

**Galaxy** is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#). You can install your own Galaxy by following the [tutorial](#) and choose from thousands of tools from the [Tool Shed](#).

**080+**

**Public Galaxy Servers**  
and *still* counting

**History**

search datasets

**Unnamed history**

0 b

This history is empty. You can [load your own data](#) or [get data from an external source](#)



# Data manipulation & visualisation

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with 'Galaxy' and 'Analyze Data' tabs. A notification banner states: 'We have updated Galaxy to the forthcoming 16.04 release and made some configuration and deployment changes, so you may...'. On the left, a 'Tools' sidebar lists various categories like 'NGS: RNA Structure', 'NGS: Du Novo', 'NGS: Gemini', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'CloudMap', 'Phenotype Association', 'BEDTools', 'Genome Diversity', 'EMBOSS', 'Regional Variation', 'FASTA manipulation', 'Multiple Alignments', 'Metagenomic Analysis', 'Multiple regression', 'Multivariate Analysis', 'Motif Tools', 'STR-FM: Microsatellite Analysis', and 'NCBI SRA Tools'. The 'phyloP interspecies conservation scores' tool is selected, showing its description: 'phyloP interspecies conservation scores (Galaxy Version 1.0.0)'. The 'Dataset' section has a dropdown menu with 'No interval dataset available.' and an 'Execute' button. A warning message below states: 'This currently works only for builds hg18 and hg19.' The 'Dataset formats' section explains: 'The input can be any interval format dataset. The output is also in interval format. (Dataset missing?)'. The 'What it does' section describes: 'This tool adds a column that measures interspecies conservation at each SNP position, using conservation scores for primates pre-computed by the phyloP program. PhyloP performs an exact P-value computation under a continuous Markov substitution model. The chromosome and start position are used to look up the scores, so if a larger interval is in the input, only the score for the first nucleotide is returned.'



# 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