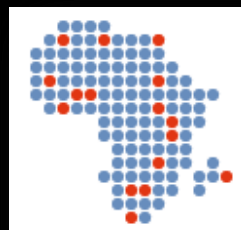
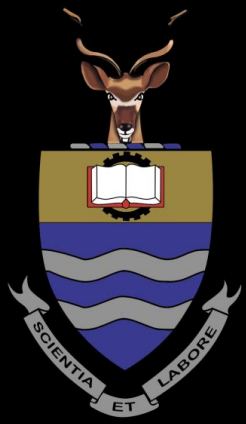


H3ABioNet

Pan African Bioinformatics Network for H3Africa



H3ABioNet Data Management Workshop

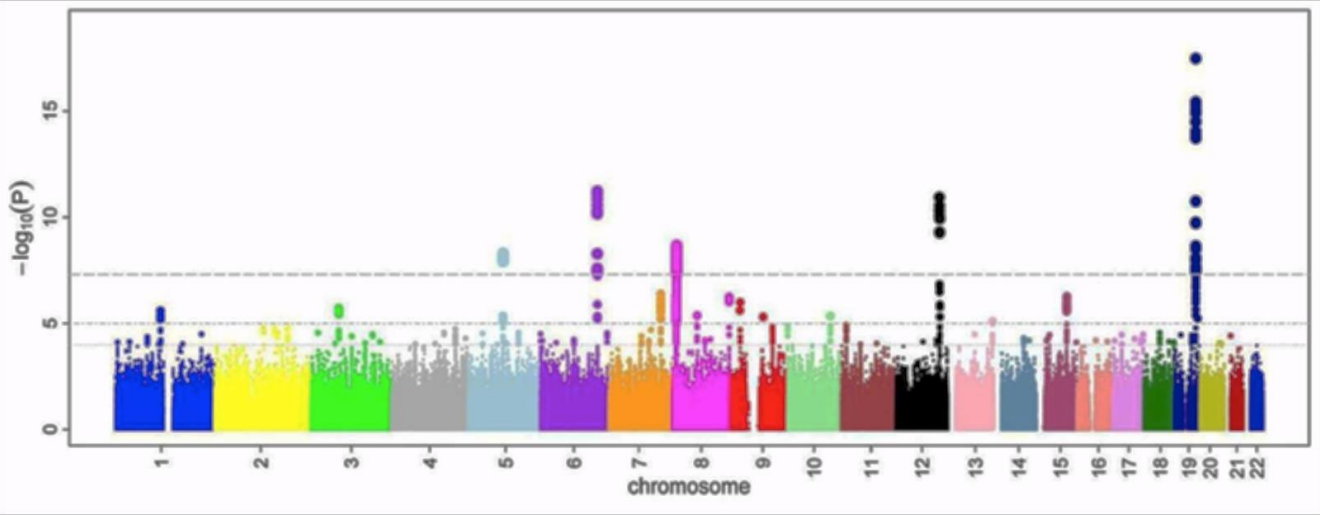
Shaun Aron

June 2014

Post GWAS visualisation and analysis

Post GWAS analysis

- What do you do once you have identified a set of SNPs associated with a particular phenotype of interest?



Post GWAS analysis

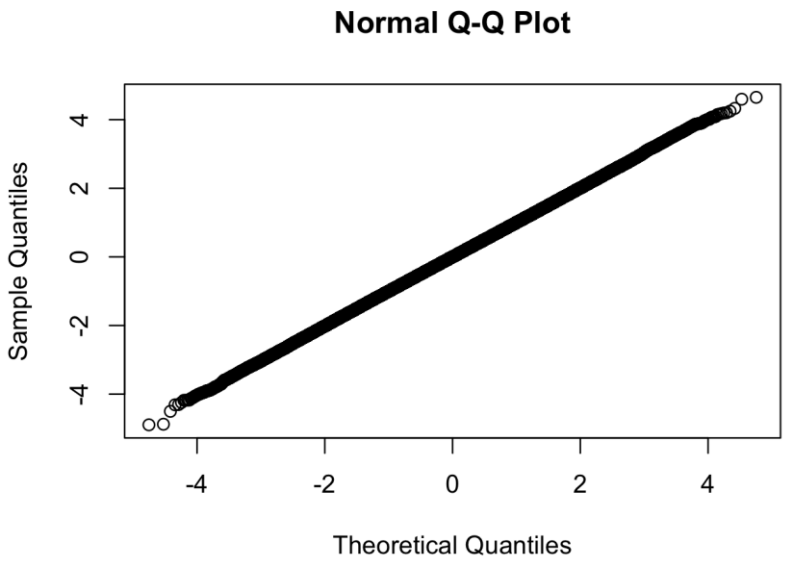
- If you're lucky…
  - Variants are in a gene previously linked to your phenotype/disease
  - Celebrate….not just yet…
  - Explore further implications of the variant
  - Replicate then celebrate!
- If you're not so lucky
  - Look for approaches to interrogate and prioritise associated SNPs for further analysis
  - Look for alternative approaches to analyse GWAS data

Post GWAS visualisation

- Q-Q plots
- Manhattan plots
- Evoker

Q-Q plots

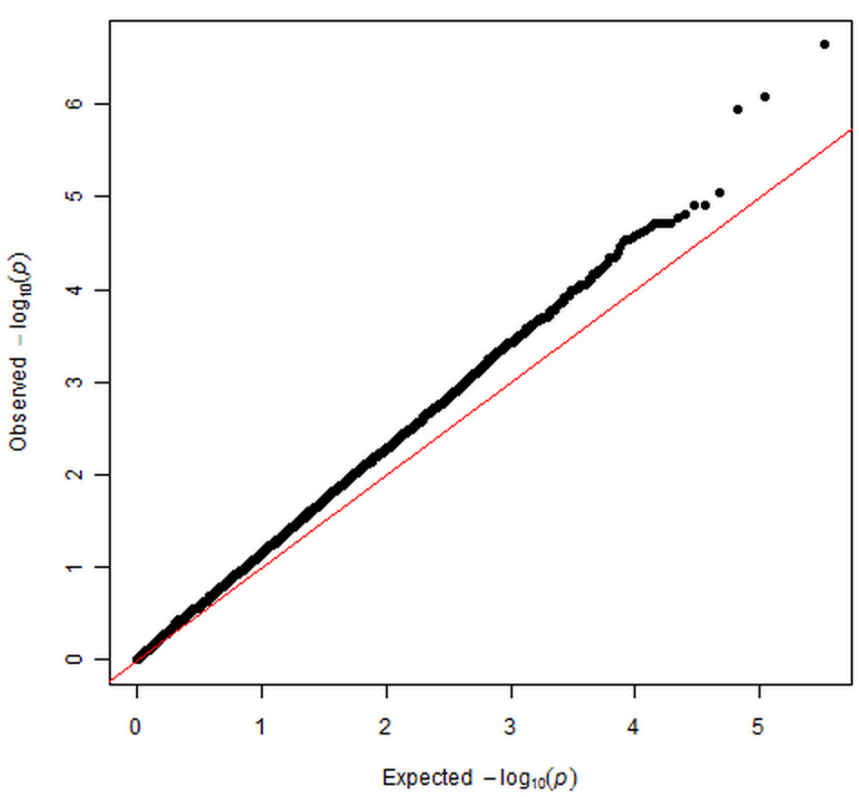
- A Q–Q plot is a **probability plot**, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- Essentially a GWAS QQ plot aims to plot the quantile distribution of observed p-values (y-axis) vs the quantile distribution of expected p-values (x-axis)

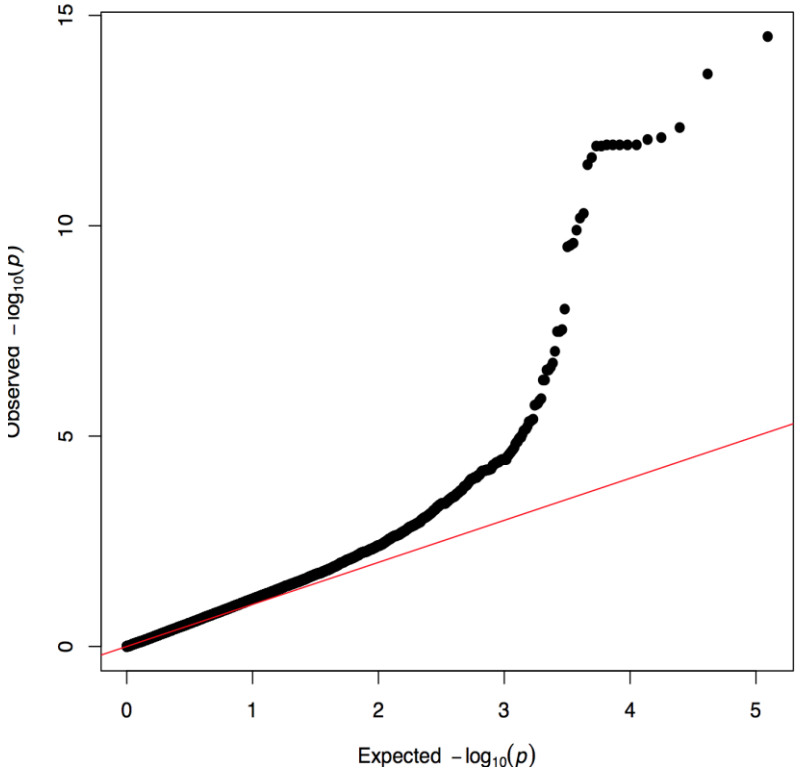


A normal Q–Q plot comparing sample quantile data on the vertical axis to a standard normal distribution on the horizontal axis.

The second distribution is often theoretical

QQ plots in GWAS





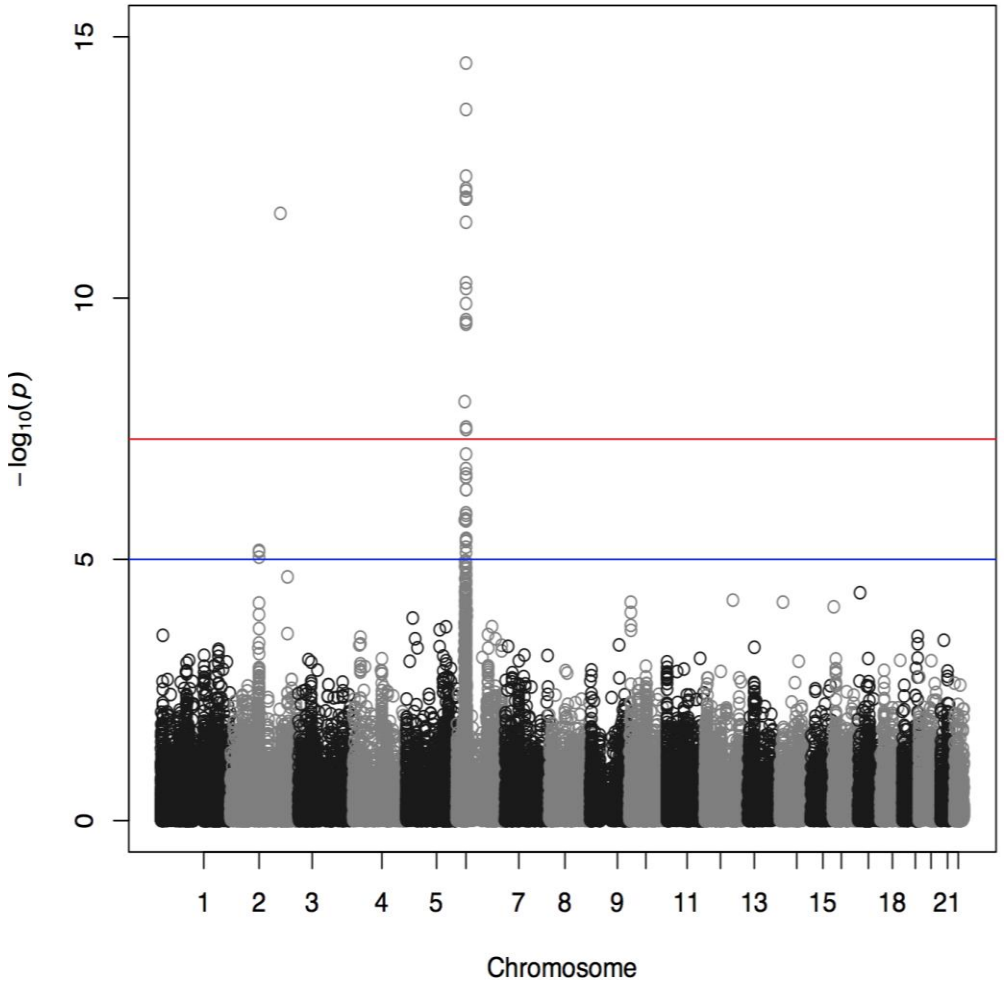
Visualizing Associations -Manhattan plot

To generate Manhattan plot you can use several available R scripts to plot your p-values. E.g qqman in R

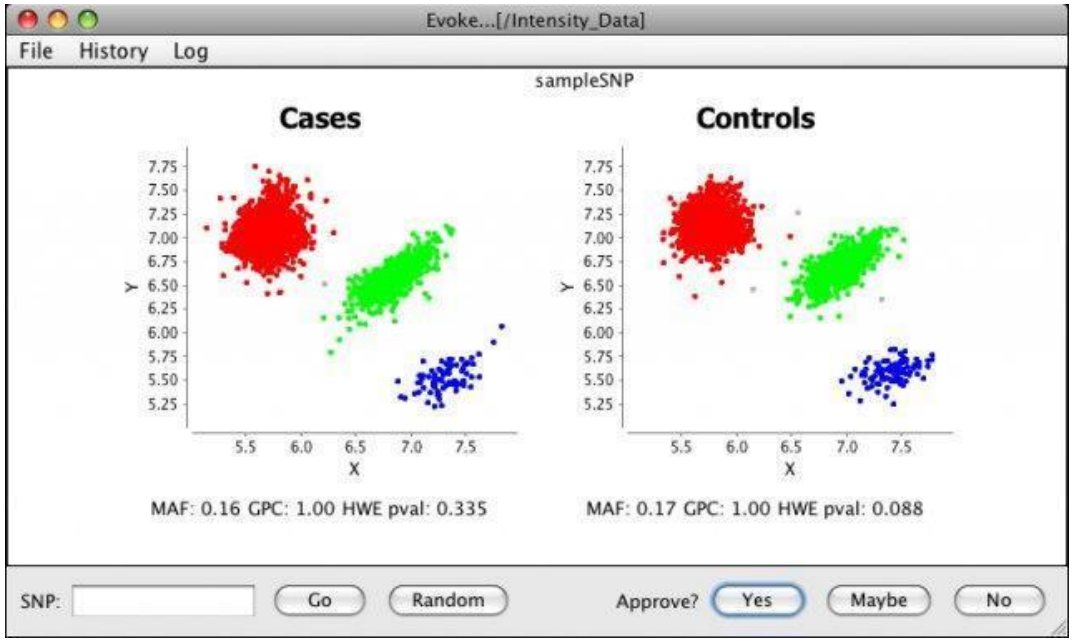
Each point is a SNP laid out across the human chromosomes from left to right, and the heights correspond to the strength of the association to disease.

Red line: Standard P-value cutoff

Blue line : Suggestive P value cutoff



Reviewing Cluster plots for Predicted Associations - EVOKER



- Evoker is a a fast, user-friendly and interactive interface tool for visualizing genotype cluster plots, and provides a solution to the computational and storage problems related to working with large datasets.
- Evoker requires four data file types, which provide information about samples, SNPs, genotype calls and X/Y allelic intensities.
- Plots for particular markers can be quickly called up by searching on the marker name, along with summary statistics on minor allele frequency, genotyping call rate and Hardy–Weinberg equilibrium *P*-value

https://www.sanger.ac.uk/resources/software/evoker/

Evoker: Cluster plot chromosome 2 best SNP

P value= 2.389e-12

PostGWAS SNP prioritisation

Post GWAS approaches

- Functional annotation of SNPs
  - Functional annotation of associated SNPs
- Pathway analysis
  - Pathways associated with genes in which associated SNPs are identified
- Genome-wide complex trait analysis (GCTA)
  - Alternative approach to typical GWAS

Functional annotation of SNPs

- If you have a few associated SNPs, functional annotation can be done manually
  - Extract information from genome browsers such and Ensembl or UCSC
- If you have a large number of associated SNPs there are several tools that do automated annotation and prioritisation

Functional annotation of SNPs

- ANNOVAR – ANNotation of VARiants
- Functional annotation pipeline for variants from various organisms
- Command line version and web service called wANNOVAR

wANNOVAR

- Given a list of SNPs:
  - Annotates the functional effect on genes for non-synonymous SNPs
  - Calculate the predicted functional effect using tools such as SIFT, PolyPhen2, MutationTaster
  - Retrieve allele frequencies in public databases (1000 genomes)
  - Uses a variant reduction process to identify a subset of potentially deleterious variants (command line version)

wANNOVAR

Column Name	Explanation
Func	Variant function (exonic, intronic, intergenic, UTR, etc)
Gene	Gene Name. By default, RefSeq gene definition is used, but users can choose from other gene definition systems.
ExonicFunc	Exonic variant function (non-synonymous, synonymous, etc)
AAChange	Amino acid changes
Conserved	Region-level phastCons LOD scores
SegDup	Sequence identity score for the segmental duplication region where variant is located in
ESP5400 ALL	Alternative allele frequency in all subjects in the NHLBI-ESP project with 5400 exomes
1000g2011may ALL(hg19)	Alternative allele frequency data in 1000 Genomes Project
1000g2010jul(hg18,3datasets: ceu, yri, jptchb)	Same as above
dbSNP	by default, dbSNP135 for hg19, dbSNP132 for hg18. Users can select a different dbSNP version.
AVSIFT	Whole-exome SIFT scores for non-synonymous variants
LJB PhyloP	Whole-exome PhyloP scores
LJB SIFT	Whole-exome LJBSIFT (1-SIFT) scores
LJB PolyPhen2	Whole-exome PolyPhen version 2 scores
LJB LRT	Whole-exome LRT scores
LRT MutationTaster	Whole-exome MutationTaster scores
LJB GERP++	Whole-exome GERP++ scores

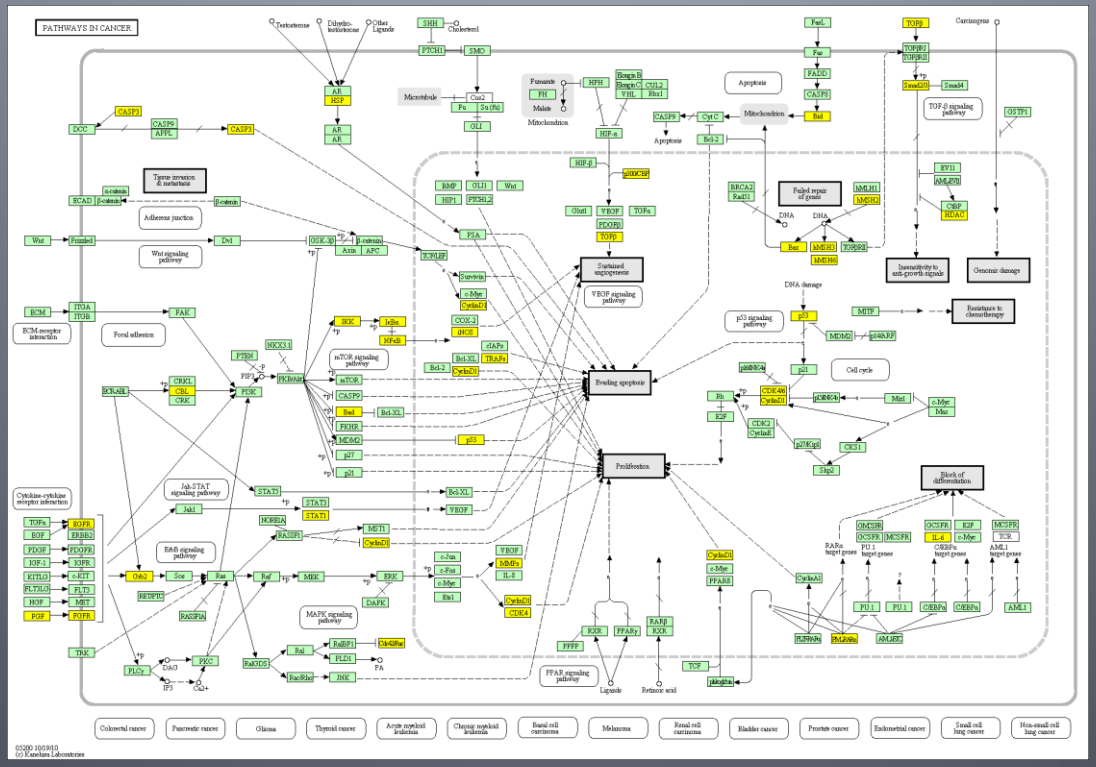
wANNOVAR

Func	Gene	ExonicFunc	AAChange
exonic	NOC2L	synonymous SNV	NM_015658:c.C1654T:p.L552L
exonic	NOC2L	synonymous SNV	NM_015658:c.C657T:p.L219L
intronic	PLEKHN1		
exonic	AGRN	nonsynonymous SNV	NM_198576:c.G5125C:p.G1709R
exonic	TTLL10	synonymous SNV	NM_153254:c.C120T:p.P40P
exonic	TTLL10	nonsynonymous SNV	NM_153254:c.C884G:p.P295R
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.C16T:p.R6W
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.A193G:p.S65G
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.C200T:p.P67L
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.G466A:p.D156N
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.G619C:p.D207H
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.G649A:p.G217S
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.C694T:p.R232C
exonic	B3GALT6	nonsynonymous SNV	NM_080605:c.T925A:p.S309T
exonic	PUSL1	nonsynonymous SNV	NM_153339:c.C517A:p.L173I
splicing	CPSF3L		

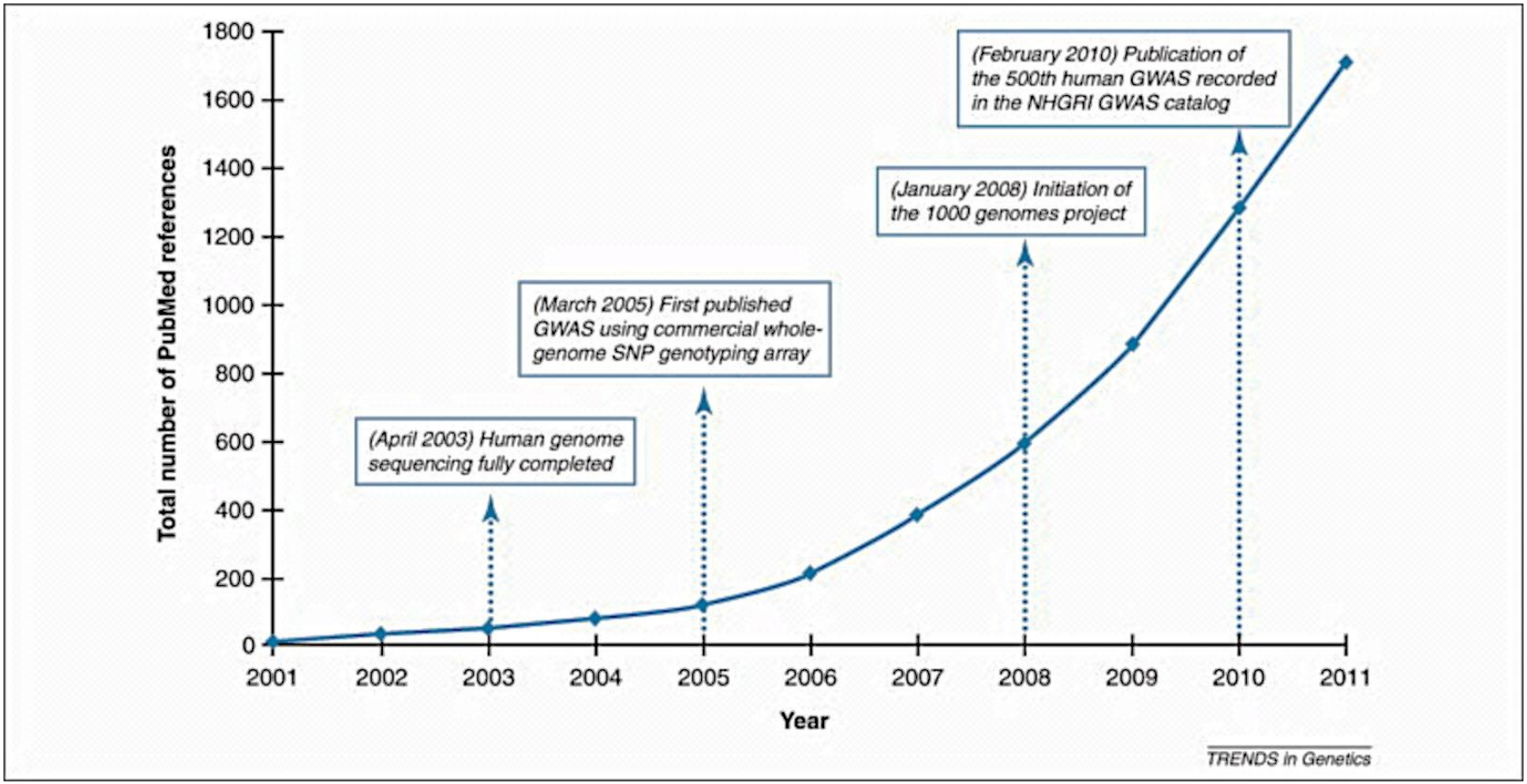
Additional variant annotation tools

- RegulomeDB
  - Include annotations for non-coding regions
  - Includes data from ENCODE project
  - TFBS, DNase I Hypersensitive sites
- Variant Effect Predictor
  - User friendly interface
- Ensembl Perl API or BioMart

Pathway Analysis



Pathway analysis



Pathway analysis

- Based on the premise that:
  - Most GWAS-implicated common alleles exhibit modest effect sizes
  - Genes function within biological pathways and interact with biological networks
  - Jointly may affect the risk of a complex disease
  - Is a particular pathway "enriched" with your associated SNPs/genes compared to all pathways?

Pathway analysis

- Two main approaches
  - Candidate pathway analysis
    - Pre-select pathways linked to phenotype
    - Assess enrichment in these pathways only
  - Genome-wide pathway analysis
    - Interrogate enrichment in all pathways

Pathway analysis

- Earlier analysis methods were based on gene lists and not variants
  - Only assess variants found within genes or within close proximity to a gene
  - Imputation to increase gene coverage
- More recent methods have been developed for GWAS analysis
  - Raw genotype data
  - List of SNPs and associated P-values

Pathway analysis

- Pathway analysis tools adapted for GWAS:
  - Most tools utilise one association signal per gene
  - GWAS – include multiple signals per gene in some cases
  - Newer methods correct for this occurrence based on LD and P-values used as the input data

Pathway analysis

- Active area of development
- Most tools differ in the method used to calculate enrichment and pathway databases used
- Pathway-enrichment tools
  - GRAIL, MAGENTA, DAVID, INRICH, ALIGATOR, WebGestalt, GWAS3D, Ingenuity Pathway Analysis (IPA)

Alternative approach to GWAS

Genome-Wide Complex Trait Analysis GCTA

GWAS story so far…

- GWAS studies have uncovered hundreds of SNPs significantly associated with complex traits
- Yet for any one trait these SNPs account for only a small fraction of the genetic variation
- GWAS tests statistical association of a single SNP with trait/disease
- Can only account for a small percentage of the heritability

Two possible explanations

- Causal variants each explain such a small amount of variation (small effect size) that their effects fail to reach stringent GWAS significance levels

- Causal variants are not in complete LD with the SNPs that have been genotyped

Heritability estimates

- Twin and family studies have investigated the genetic and environmental origins of individual differences in various traits
  - The extent to which genetic variance can account for observed or phenotypic variance.
  - Estimates of genetic heritability for various traits have been determined e.g. height, weight, cognitive ability, BMI etc.

Missing heritability

- Current associated SNPs only explain a small percentage of the estimated heritability in most traits/diseases
- Where is the "missing heritability"?

Alternative approach

- Genome-wide Complex Trait Analysis (GCTA)
  - Provide an estimate of heritability using genome-wide data from unrelated samples
- Premise is based on the concept of traditional heritability studies in families and twins
- Use all SNPs together in unrelated individuals to estimate the amount of genetic variance explained by all the SNPs

GCTA

- GCTA does not identify specific genes or SNPs associated with a trait
  - Uses chance similarity across genome-wide SNP data to predict phenotypic variance on a pairwise basis in a large sample of unrelated individuals
  - This yields a measure of the proportion of phenotypic variance explained by all SNPs in a GWAS dataset for the associated trait

GCTA

- If there is a significantly high amount of variance, then this shows that the phenotype variation of the trait of interest is likely due to common variants in the data with small effect size
- Caveat of GCTA approach – Requires fairly large sample size for an accurately powered result (>5000)

Tools - Links

- wANNOVAR
  - http://wannovar2.usc.edu/
- Evoker
  - https://www.sanger.ac.uk/resources/software/evoker/
- Regulome DB
  - http://regulome.stanford.edu/
- Variant effect predictor
  - http://www.ensembl.org/info/docs/tools/vep/index.html
- DAVID
  - http://david.abcc.ncifcrf.gov/
- INRICH
  - http://atgu.mgh.harvard.edu/inrich/
- WebGestalt
  - http://bioinfo.vanderbilt.edu/webgestalt/
- GWAS3D
  - http://jjwanglab.org/gwas3d
- GCTA
  - http://www.complextraitgenomics.com/software/gcta/
- qqman R package
  - https://github.com/stephenturner/qqman