General Introduction

Christian Werner — christ_raw

(Computer biologist, and quantitative geneticist) **Roslin Institute**, Edinburgh (Scotland)

Filippo Biscarini — HerrFalloppio

(Biostatistician, bioinformatician, and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

Oscar González-Recio — ogrecio

(Computer biologist, and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

Christian- in a slide

- **University of Gießen** (*BSc Agricultural Science)*
- **University of Gießen** (*MSc Crop Production*)
- **University of Gießen** (*PhD Plant Breeding*)
- **University of Edinburgh** (*Visiting Student*)
- **University of Edinburgh** (*Post-doc*)

Oscar- in a slide

- **Agricultural Engineer** (*Madrid*)
- **Animal Science** (*MSc degree, Madrid*)
- **Animal Genetics & Quantitative Genetics** (*UPM-Madrid & UW-Madison*)
- **University of Wisconsin-Madison** (*Post-doc*)
- **Dept. Environment and Primary Industries-Melbourne, Australia** (*Senior Research Scientist*)
- **INIA - Madrid** (*Senior Research Scientist*)

Filippo - in a slide

- **Roma** (*born*)
- **Perugia** (*MSc degree*)
- **Cork, ICBF** (*Web-design & Database*)
- **Cremona, ANAFI** (*Quantitative Genetics*)
- **Guelph, CGIL** (*Visiting Scientist*)
- **Wageningen, WUR** (*PhD*)
- **Göttingen University** (*post-doctoral researcher*)
- **Lodi, PTP** (*'omics in animals, plants, humans*)
- **Milan** - **CNR** (*tenured researcher*)
- **Cardiff University** (*biostatistician*)
- **Milan** - **CNR** (*senior researcher*)
- **Bruxelles** - **ERC** (*seconded national expert*)
- **Milan - CNR** (*senior researcher, group leader*)

now you – round of introduction

- who
- where from
- type of research / work
- what's your interest on GWAS / motivation

outline of this lecture

1. Laying out the topic
2. Overview and intuition on GWAS

course - in a slide

Introduction to GWAS (Genome-Wide Association Studies)

day	time	activity	topic	hours
Monday	9:30	Lecture 0	General Introduction / Overview of the course	0.5
	10:00	Lecture 1	GWAS: case studies / examples from literature	1
	11:00		coffee break	0.5
	11:30	Lecture 2	basic Linux	1
	12:30	Lecture 3	Introduction to GWAS: Linkage Disequilibrium and Regression Models	0.5
	13:00		lunch break	1
	14:00	Lecture 4	Introduction to GWAS: Linkage Disequilibrium and Regression Models	1.5
	15:30		coffee break	0.5
	16:00	Lab 4	Basic models: linear and logistic regression (demonstration)	1
	17:00	Lecture 5	GWAS, the full model (all SNPs)	1
Tuesday	9:30	Lecture 6	GWAS: The multiple testing issue	1
	10:30		coffee break	0.5
	11:00	Lecture 7	GWAS: Statistical power, Population stratification, Experimental design	2
	13:00		lunch break	1
	14:00	Lecture 9	Practicalities and set-up (server, github repo, etc) and description of datasets	1
	15:00	Lecture 10	data preprocessing: theory	0.5
	15:30		coffee break	0.5
	16:00	Lab 10	data preprocessing: practice	1.5
Wednesday	9:30	Lecture 11	Imputation of missing genotypes: theory	1.5
	11:00		coffee break	0.5
	11:30	Lab 11	practical session on imputation (Beagle)	0.5
	12:00	Lab 12	KNNI imputation (optional)	1
	13:00		lunch break	1
	14:00	Lab 13	GWAS (rrblup): the stand-alone script(s) for the full model	1
	15:00		coffee break	0.5
	15:30	Lab 14	PostAnalysis -In silico study (FUMA)	2

Thursday	9:30	Lecture 12	Bioinformatics pipelines: a super-elementary introduction	1
	10:30	Lab 15	Building a pipeline with Snakemake	0.5
	11:00		coffee break	0.5
	11:30	Lab 16 - part I	The GWAS pipeline for continuous and binary phenotypes	1.5
	13:00		lunch break	1
	14:00	Lab 15	Introducing the assignment (+ light touch on RMarkdown)	0.5
	14:30	Assignment (groups)	Build your own GWAS pipeline on new data	1
	15:30		coffee break	0.5
	16:00	Assignment (groups)	Build your own GWAS pipeline on new data	1.5
Friday	9:30	Assignment (groups)	Build your own GWAS pipeline on new data	1.5
	11		coffee break	0.5
	11:30	Assignment (groups)	Build your own GWAS pipeline on new data	1
	12:30	Assignment (groups)	Presenting and discussing results	0.5
	13:00		lunch break	1
	14:00	Assignment (groups)	Presenting and discussing results	1
	15:00		coffee break	0.5
	15:30	Lecture 13	GWAS review: potential limitations	0.5
	16:00	Lecture 14	A glimpse on ROH-based alternatives and resampling methods	0.5

Flexibility!

laying out the topic

Day 1

- **GWAS**: background and intuition
- Linkage disequilibrium
- Linear ‘or’ logistic regression
- **Practicalities** and set up
- Linux
- R

laying out the topic

Day 2

- **Experimental design**
- Statistical Power
- Population stratification
- **Exploratory Data Analysis**
- Preprocessing

laying out the topic

Day 3

- **Imputation**
- **The full model**
- **Functional analysis**

laying out the topic

Day 4

- **Bioinformatic pipelines**
- Introduction
- Snakemake
- Continuous
- Binary

laying out the topic

Day 5

- **Practice – own work**

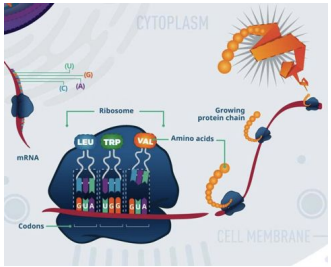
A quick overview of GWAS

Genetic basis of GWAS-GWP

Phenotypic variation

Genetic variants

- Allele substitution effect



- Infinitesimal model

Genetic basis of GWAS-GWP

Phenotypic variation

Genetic variants

$$P = G + E$$

- Continuous trait
- Categorical trait (threshold model)

Genetic basis of GWAS-GWP

Phenotypic variation

Genetic variants

Linkage disequilibrium

$$P = G + E$$

Genotype-Phenotype association → causal or **functional** link

Genetic basis of GWAS-GWP

- A study that agnostically tests hundreds of thousands of single nucleotide polymorphisms (SNPs) densely spaced across the genome for association with a given disease or trait.

- Rationale:
  - Not limited by a priori knowledge of disease process or the results of linkage studies
  - Preferable approach (so far) for diseases of complex etiology
  - Evidence for functional SNPs outside of coding regions

"*First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another*." ENCODE, Nature June 2007

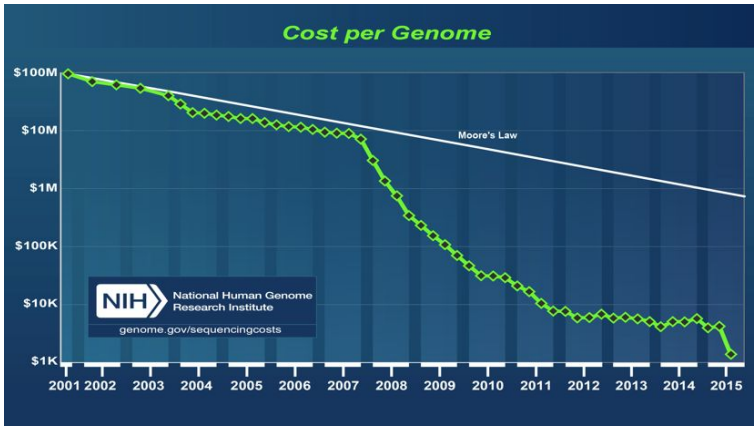
Genetic basis of GWAS-GWP

- Exposure:
  - SNPs
  - Indels
  - *Might* - copy number variants (CNV), haplotypes, other type of markers
- Outcome: easily adaptable to either dichotomous or quantitative outcomes
- Study Populations: Case-control, cross-sectional, or cohort. Family or population based.
  - Need to have large sample sizes!
- Analytic methods: logistic/linear regression
  - Flexibility for different genetic models - additive, dominant, recessive.

Genetic basis of GWAS-GWP

- Genotyping – Since early 2000 (25-100€)
- Genome sequencing - (aprox 1000 €)

- Quantitative genetics – many genes – need many markers
- Use LD between gene and marker
  - We assume that all genes are in LD with at least one marker (we need a high coverage throughout the genome)



Costs matter - approximate current SNP array prices

SNP chip	n. SNP	approximate price
Cow BovineLD	7,900	**$32**
Cow Bovine SNP50	53,000	**$50**
Cow BovineHD	777,000	**$150**
Dog CanineHD	172,000	**$125**

Commercial SNP chips

MaizeSNP50 BeadChip, GeneChip® Rice 44K SNP $\rightarrow\sim$ 100$ per sample

Genome-wide association studies (GWAS)

1. Genotyping costs have decreased dramatically over the last 10 years, in most species
2. Large number of genotyping experiments
   → explosion of GWAS experiments!

Powerful (*supposedly*) approach to the identification of genes**/genomic regions involved in plant, animal and human phenotypes**

Genome-wide association studies (GWAS)

association is one type of statistical problem

- discovery of interesting relationships among variables in large data sets (i.e., **association**);
- division of data sets into several discrete groups (i.e.,**clustering**);
- assignment of observations to groups (i.e., **classification**);
- Extrapolate quantitative outputs based on attributes of observational units (i.e.,**prediction**);
- etc.

Inference vs Prediction

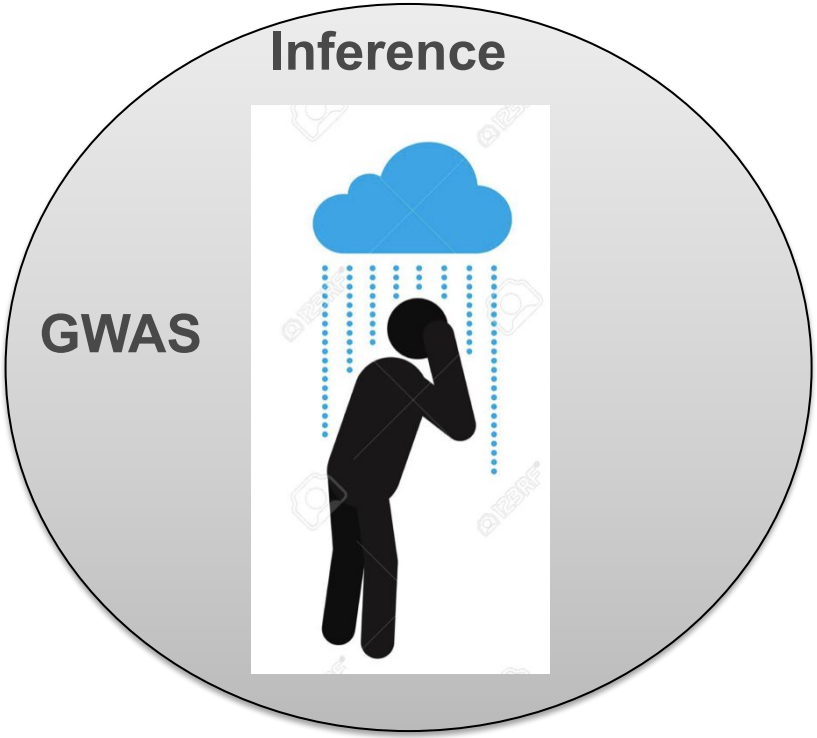
Inference

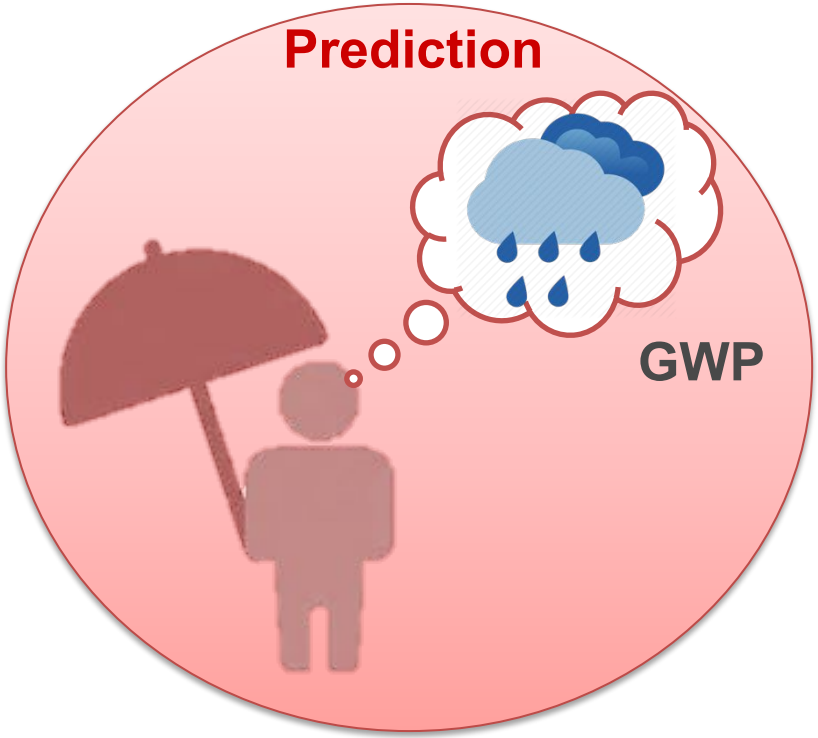
- Determine the effect of a covariate on the response
- Determine the causal relationship between a covariate and the response
- More difficult (in general)

Prediction

- Educated guess of the outcome
- Expected behaviour in the future
- Based on proxies/markers

Inference vs Prediction





Inference vs Prediction

- Know the past
- Predict the future
- Act consequently

INFERENCE

Genetic basis of GWAS-GWP

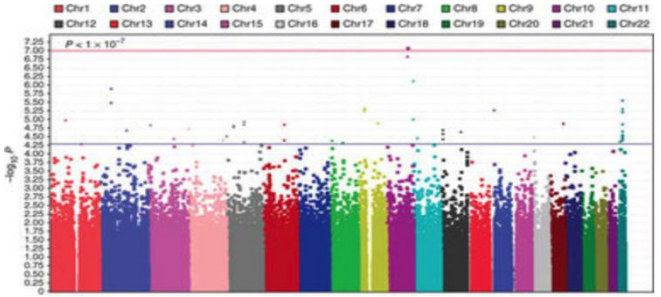
GWAS goal

- Detect genomic markers/regions associated to phenotypes (traits) of interest
- Find biological pathways of interest
- Interaction between treatments/drugs and genes

GWP goal

- Calculate the genomic risk score /predisposition of a disease or trait
- Calculate the genomic merit of individuals
- Predict future performance
- Sire/dam selection in animal breeding

Some examples



GWAS –Rheumatoid arthritis

BMC Proceedings

Home About Articles Submission Guidelines

Volume 3 Supplement 7

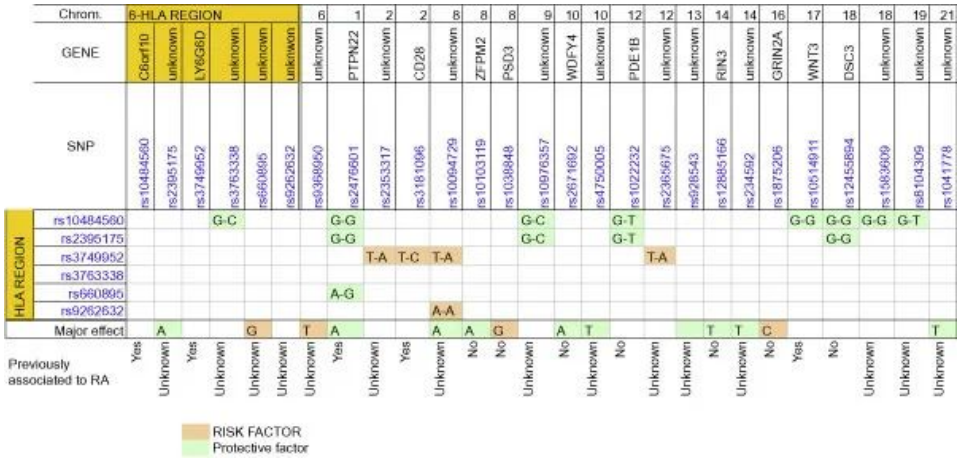
Genetic Analysis Workshop 16

Proceedings | Open Access | Published: 15 December 2009

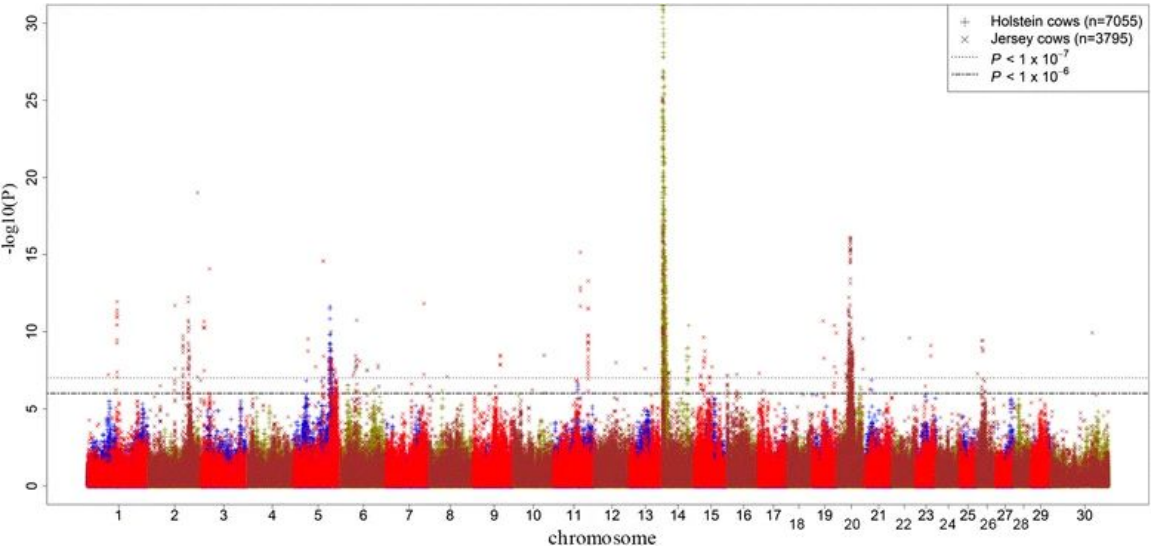
Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model

Oscar González-Recio, Evangelina López de Maturana, Andrés T Vega, Corinne D Engelman ✉ & Karl W Broman

BMC Proceedings 3, Article number: S63 (2009) | Cite this article



GWAS – Milk yield dairy cattle



- **DGAT1** gene on BTA 14 in dairy cows (HOL and JER)
- milk fat content

BMC Genetics

Home About Articles Submission Guidelines

Research article | Open Access | Published: 22 July 2015

Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows

Hassan Aliloo, Jennie E. Pryce, Oscar González-Recio, Benjamin G. Cocks & Ben J. Hayes

BMC Genetics 16, Article number: 89 (2015) | Cite this article

4154 Accesses | 8 Citations | 3 Altmetric | Metrics

GWAS – Fertility dairy cattle

A number of **fertility-related** traits (additive and dominance)

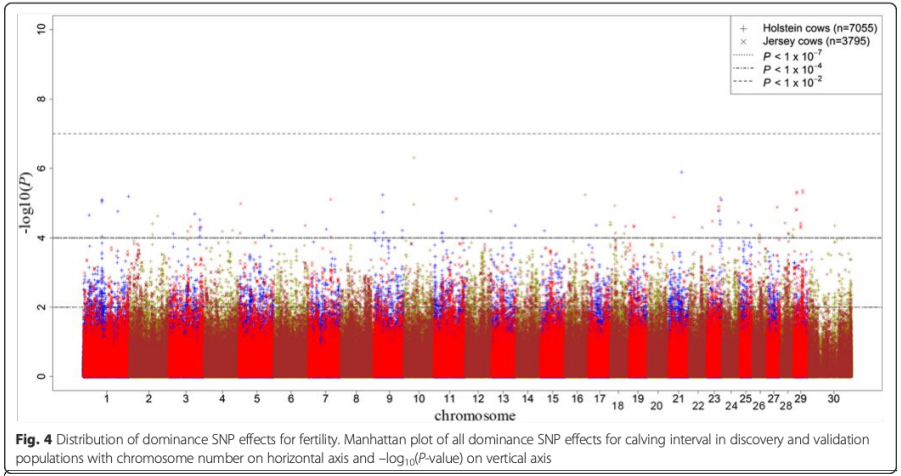


Fig. 4 Distribution of dominance SNP effects for fertility. Manhattan plot of all dominance SNP effects for calving interval in discovery and validation populations with chromosome number on horizontal axis and $-\log_{10}(P\text{-value})$ on vertical axis

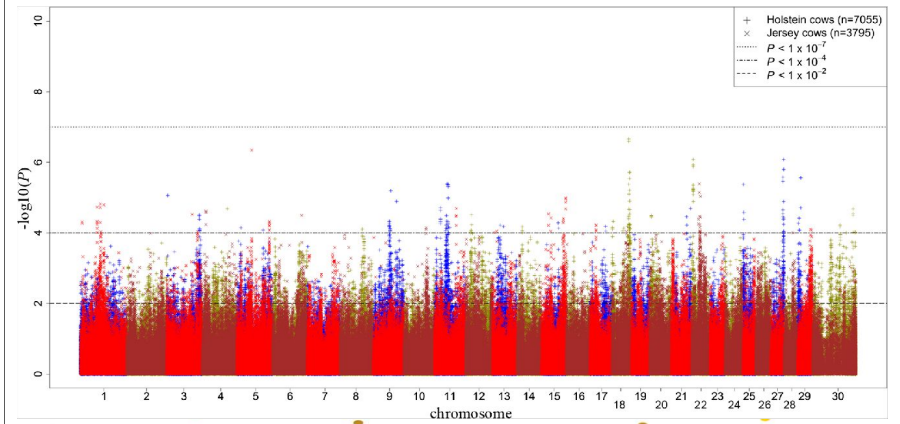


Fig. 2 Distribution of additive SNP effects for fertility. Manhattan plot of all additive SNP effects for calving interval in discovery and validation populations with chromosome number on horizontal axis and $-\log_{10}(P\text{-value})$ on vertical axis

GWAS –curly hair in cattle

nature genetics

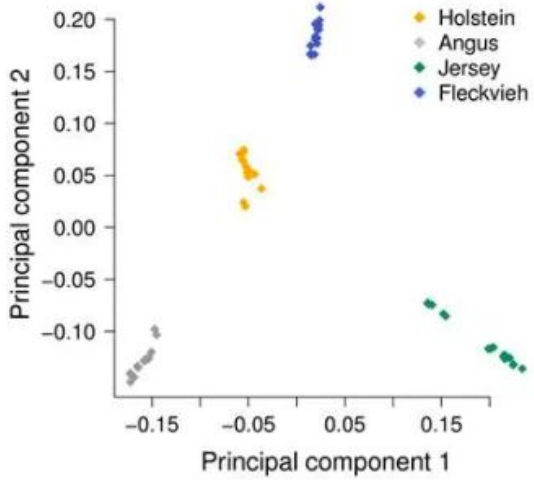
Article | Published: 13 July 2014

Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle

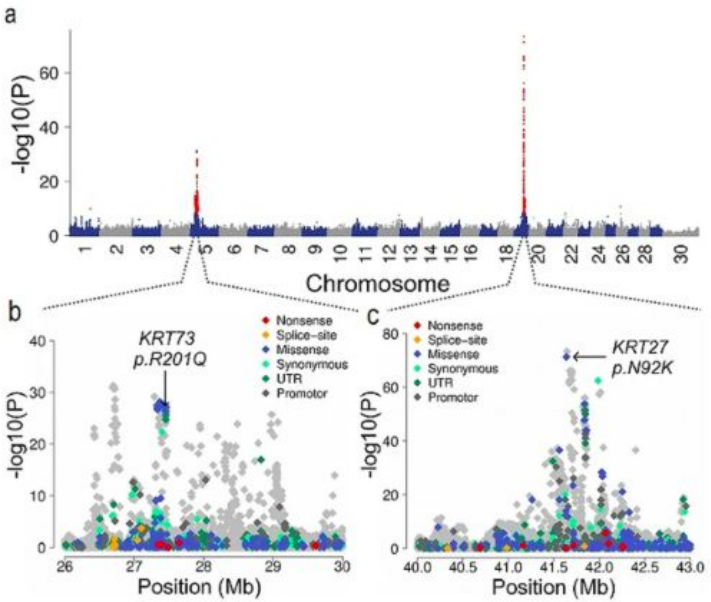
Hans D Daetwyler, Aurélien Capitan, [...] Ben J Hayes

Nature Genetics **46**, 858–865(2014) | Cite this article

1351 Accesses | **351** Citations | **113** Altmetric | Metrics



From: Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle



Manhattan plot showing the association of 17,640,970 imputed variants with the proportion of daughters with curly hair in 3222 Fleckvieh bulls (**a**). Red dots represent variants with $P < 10^{-9}$. Detailed overview of the associated regions on chromosomes 5 (**b**) and 19 (**c**). Variants in the promoter (defined to encompass 1,000 bp upstream of the transcription start), in the untranslated regions (UTR) and in the amino acid coding region are highlighted with different color. The associated region on BTA5 encompasses Krt71, which underlies curly hair in various species. Variant calling yielded four missense mutations in Krt71 (p.R133W, p.F143I, p.N177I, p.P452H); however, none of them was polymorphic in the 43 resequenced Fleckvieh animals. Functional annotation of the variants within the QTL region on BTA5 revealed that 12 closely linked missense mutations in Krt73, Krt2 and Krt76 are highly significantly associated with curly hair in Fleckvieh cattle. Among those, only the p.R201Q mutation in Krt73 (c.G602A, chr. 5: 27,445,800 bp, ss682156288) was predicted to be damaging by PolyPhen-2 and SIFT analysis.

GWAS - Schizophrenia

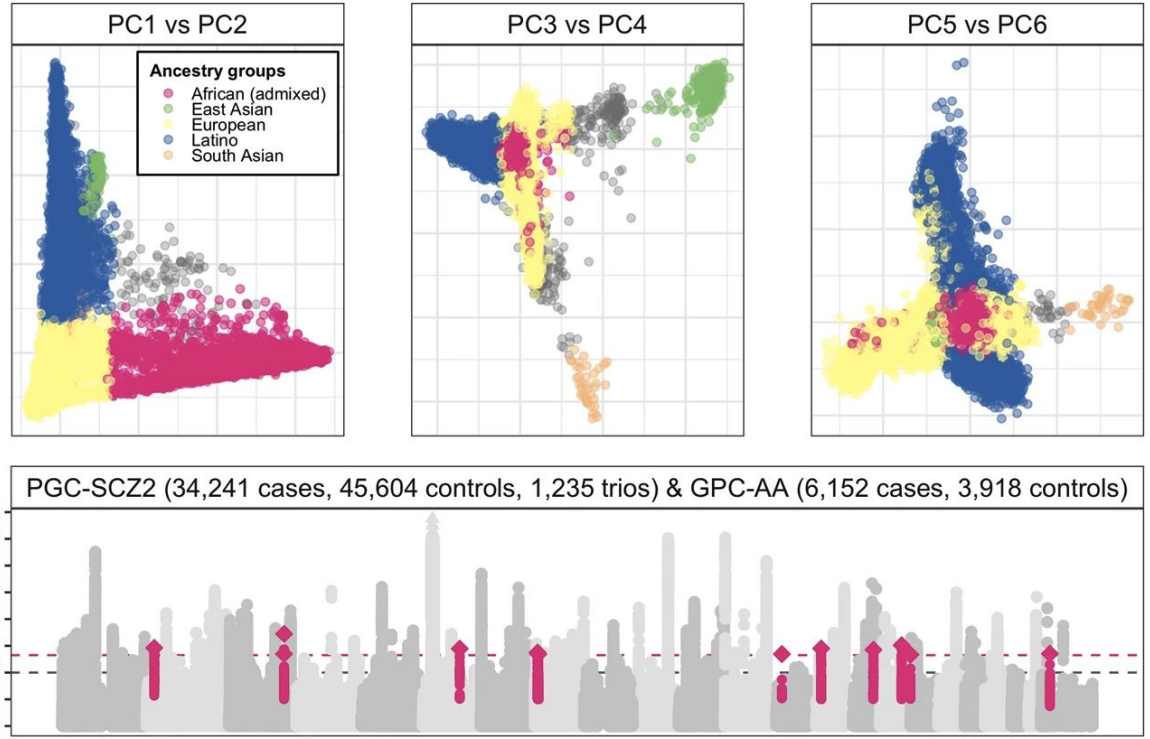
Molecular Psychiatry

Article | Open Access | Published: 07 October 2019

Contributions of common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry

Tim B. Bigdeli ✉, Giulio Genovese, [...] Carlos N. Pato

Molecular Psychiatry (2019) | Cite this article



What will you learn

- How GWAS work
- Use the right type of analyses
- Identify and understand the individual steps involved in a GWAS project
- Understand the limitations of GWAS
- Visualize results of GWAS
- Assemble the different steps into a reproducible pipeline

Regarding GWAS - what we will do

0. getting the data
1. data preprocessing **(EDA)**
2. data preprocessing **(filtering)**
3. **imputation** of missing genotype data
4. **GWAS** basic models
5. single SNP vs many SNP
6. continuous/binary traits
7. population structure
8. Manhattan plots/qq-plots (post-hoc analysis)
9. build the **pipeline**

NEXT LECTURE

Introduction to GWAS: Linkage disequilibrium and Linear Regression