

# Collaborative **exercise**

Christian Werner

*(Quantitative geneticist and biostatistician)* **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

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



HerrFalloppio

Oscar González-Recio

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



OscarGenomics



# GWAS workflow/pipeline - collaborative exercise


- Build your own workflow!
  - a. Download the data
  - b. Prepare the data (look at the phenotypes and genotypes!)
  - c. Filter the data
  - d. Impute missing genotypes
  - e. Run the GWAS
- Data on stump tail sperm defect of Swiss Large White boars

(<https://zenodo.org/record/4081475#.YKPfmnUzZhE>)




from: <https://petpigworld.com/large-white-pigs-the-essential-guide/>

# GWAS workflow/pipeline - collaborative exercise

- Build your own workflow!
  - a. Download the data
  -  **b. Prepare the data (look at the phenotypes and genotypes!)**
  - c. Filter the data
  - d. Impute missing genotypes
  - e. Run the GWAS
- Data on stump tail sperm defect of Swiss Large White boars



 (<https://zenodo.org/record/4081475#.YKPfmnUzZhE>)



!! beware !! no missing SNP  
genotypes (data already filtered  
and imputed)

get data with missing SNP  
genotypes from here:  
[http://www.jackdellequerce.com/data/GWAS\\_cohort.tar.gz](http://www.jackdellequerce.com/data/GWAS_cohort.tar.gz)

# GWAS workflow/pipeline - collaborative exercise

- Build your own workflow!
    - a. Download the data
    -  **b. Prepare the data (look at the phenotypes and genotypes!)**
    - c. Filter the data
    - d. Impute missing genotypes
    - e. Run the GWAS
  - Data on stump tail sperm defect of Swiss Large White boars
-  (<https://zenodo.org/record/4081475#.YKPfmnUzZhE>)

take a **subset of the SNPs** (to make calculations -and trial/errors- quicker), **e.g. chromosomes 10-14** (the mutation is on chromosome 12)



# GWAS workflow/pipeline - collaborative exercise

- Build your own workflow!
  - a. Download the data [from [http://www.jackdellequerce.com/data/GWAS\\_cohort.tar.gz](http://www.jackdellequerce.com/data/GWAS_cohort.tar.gz)]
  - b. Prepare the data (look at the phenotypes and genotypes!)
  - c. Filter the data
  - d. Impute missing genotypes
  - e. Run the GWAS



the most difficult part will probably be preparing the data files before filtering etc. (look at the column names of the phenotype file, their order etc.)



the IDs of pigs have underscores (eg. ctrl\_1): check the `-vcf-iid` and `-double-id` options in Plink

We will break the exercise in 4 steps:

- 1) download & prepare
- 2) EDA & filtering
- 3) imputation
- 4) GWAS

} collaborative revision  
after each step!



# GWAS workflow/pipeline - collaborative exercise

## - TIPS

- download the data: e.g. wget
- uncompress the data (tar.gz file: e.g. gunzip, tar)
- look at the data: phenotypes? genotypes?
- the genotypes are in the Plink binary format!
- you'll need the `--bfile` option to read them with Plink
- you can check the Plink online manual: <https://www.cog-genomics.org/plink/>
- (it's always a good thing to learn to use the internet to look for code syntax help)
- subset the genotype files (and maybe convert to ped/map)
- anything to be done with the phenotypes?



# GWAS workflow/pipeline - collaborative exercise

- Build your own workflow!
  - a. Download the data [from [http://www.jackdellequerce.com/data/GWAS\\_cohort.tar.gz](http://www.jackdellequerce.com/data/GWAS_cohort.tar.gz)]
  - b. Prepare the data (look at the phenotypes and genotypes!)
  - c. Filter the data
  - d. Impute missing genotypes
  - e. Run the GWAS



the most difficult part will probably be preparing the data files before filtering etc. (look at the column names of the phenotype file, their order etc.)



the IDs of pigs have underscores (eg. ctrl\_1): check the `-vcf-id` and `-double-id` options in Plink

1. First, try on your own (individually, groups)
2. Then, let's do it all together!





# GWAS - bonus assignment

## MOLECULAR ECOLOGY

Molecular Ecology (2013) 22, 3949–3962

doi: 10.1111/mec.12376

### Genomic dissection of variation in clutch size and egg mass in a wild great tit (*Parus major*) population

ANNA W. SANTURE,<sup>\*1</sup> ISABELLE DE CAUWER,<sup>\*†1</sup> MATTHEW R. ROBINSON,<sup>\*</sup>  
JOCELYN POISSANT,<sup>\*</sup> BEN C. SHELDON<sup>‡</sup> and JON SLATE<sup>\*</sup>

<sup>\*</sup>Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK, <sup>†</sup>Laboratoire de Génétique et Evolution des Populations Végétales, UMR CNRS 8198, Bâtiment SN2, Université des Sciences et Technologies de Lille - Lille 1, F-59655, Villeneuve d'Ascq Cedex, France, <sup>‡</sup>Department of Zoology, Edward Grey Institute, University of Oxford, Oxford, OX1 3PS, UK

data from a paper  
on genetic analysis  
of **clutch size** and  
**egg mass** in *Parus  
major*





# GWAS - **bonus assignment**



## Phenotypes

- egg numbers (clutch size)
- egg mass



# GWAS - **bonus assignment**

- repository: <https://datadryad.org/resource/doi:10.5061/dryad.ck1rq>
- article: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12376>

focus on:

1. understand the data and the problem/project at hand
2. manipulate the data to get them in the same format as the dogs and rice data before the filtering/imputation steps

challenges:

- multiple phenotypes per individual (over time)
- errors/missing values in the genotype data (! WARNING !)

