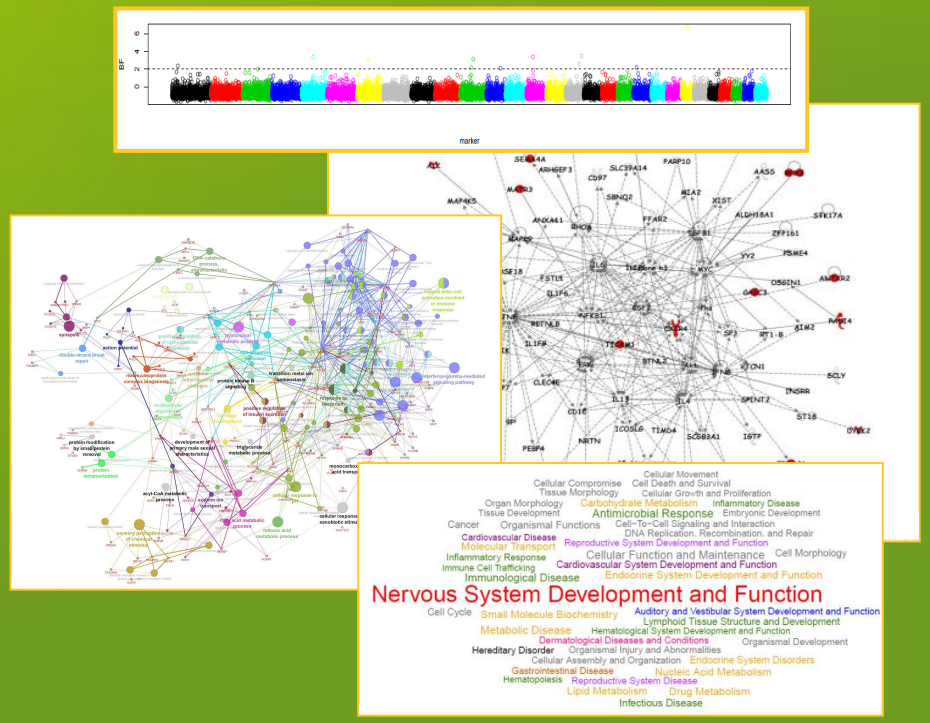


Biological interpretation of selection footprints

FLORI Laurence
GABI, INRA, Jouy-en-josas
INTERTRYP, CIRAD, Montpellier

Overview

SNP chip
NGS (Pool and Ind-Seq)

Identification of footprints of selection in the genome

Ecological variables
e.g. WorldClim DB

Phenotype-free approaches

Association with population ecological variables or phenotypes

Candidate regions

Phenotypes

Overview

SNP chip
NGS (Pool and Ind-Seq)

Identification of footprints of selection in the genome

Ecological variables
e.g. WorldClim DB

Phenotype-free approaches

Association with population ecological variables or phenotypes

Candidate regions

Reverse ecology

Phenotypes

Candidate genes

Systems biology Tools
(e.g. IPA, Cytoscape)

Phenotyping
RNA-Seq

Functional annotation

Principal selective pressures

Whole Genome mutation screening (Seq)

Functional annotation of variants

Candidate causal mutations for adaptive traits

Functional annotation

- **Main goals**

– Identify the main functions and biological pathways in which the genes are involved

– Confirm the biological importance of the genes that have been identified as being positively selected

– Strengthen the credibility of the positive selection model by the development of a sound scenario

Course outline

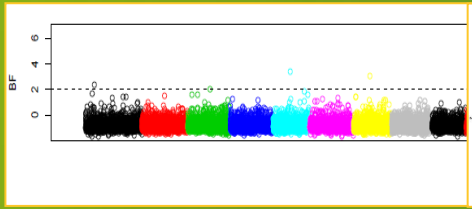
1. From regions under selection to biological interpretation

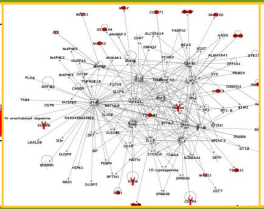
- Annotation of candidate genes using systems biology tools
  - Functional enrichment analysis: Gene ontology, pathway analysis
  - Gene networks analysis
- Inferring the main selective pressures
- Interpretation and story-telling
- Including phenotypes and environmental covariates
- Identifying candidate mutations and prioritizing candidate genes

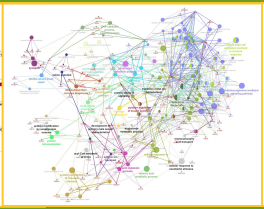
2. Interpretation of selection footprints: some examples

- Phenotype-free approaches
  - Manual functional annotation: Senepol cattle breed
  - Functional annotation using systems biology tools: french dairy cattle breeds, West-African cattle breeds, European bison/cattle
- Association with covariates
  - Phenotypes: dairy traits in French cattle breeds

1. From regions under selection to biological interpretation







INRA
SCIENCE & IMPACT

Interpretation of results obtained with phenotype-free approaches

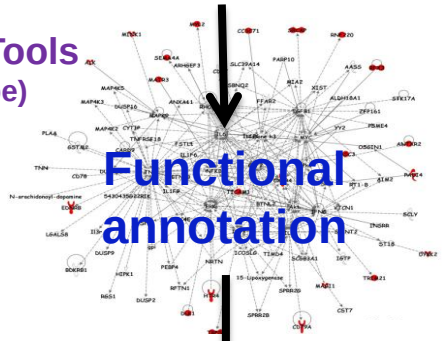
SNP chip
NGS (Pool and Ind-Seq)

Identification of footprints of selection in the genome

Phenotype-free approaches

Candidate regions and genes under selection

Systems biology Tools
(e.g. IPA, Cytoscape)



Functional annotation

Principal selective pressures

Candidate causal mutations for adaptive traits

From regions to candidate genes

- **Identification of chromosomal regions under selection**

⇒ **some arbitrary criteria**

  - Sliding windows (e.g. 1Mb, 500kb overlap)
  - Nb of SNPs with score>significant threshold
  - If several overlapping regions under selection: merging regions or choosing region with the highest peak and the highest proportion of significant SNPs

- **Mapping of regions on corresponding genome assembly on ucsc or ensembl (if available)**

- **List of all genes within regions (Refseq and Other species Refseq)**
- **Criteria to identify one or a few candidate genes per region**
  - e.g. proximity to the peak (e.g. 15-100kb from gene boundaries)

From candidate genes to biological interpretation

- **List of candidate genes with/without associated score**
  - Short list
  - Long list

- **Misleading gene names**

- **Possibly hundreds of papers describing gene functions**

How to find biological sense?

From candidate genes to biological interpretation

List of candidate genes with/without associated score

(one or few per region)

Few regions and candidate genes (e.g. <10)

More candidate genes

Manual functional annotation (e.g. Pubmed, OMIM/OMIA IPA)

Systems biology tools (e.g. DAVID, IPA, Cytoscape)

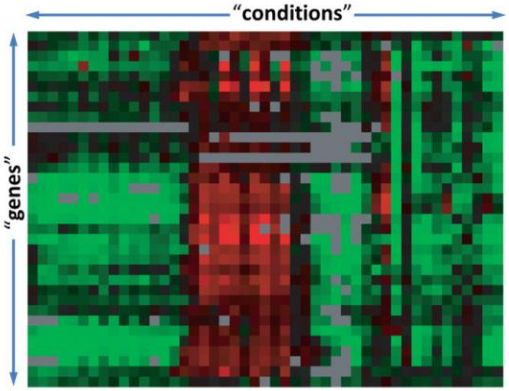
Functions

Biological Pathways

Gene Networks

Annotation using systems biology tools

- **Systems biology**
- **Tools used to interpret results of transcriptomic and proteomic experiments**
- **e.g.: list of differentially expressed genes between different experimental conditions, different tissues**



Annotation using systems biology tools: Functional enrichment analysis

- **Is the list of candidate genes statistically enriched for some functions or some biological pathways?**

- **Needs:**
  - Shared vocabulary: Gene Ontology, Biological pathways
  - Annotation of genes : association between terms and genes or gene products.
  - Statistical tests of enrichment

Annotation using systems biology tools

Functional enrichment analysis: Gene Ontology

- **Gene ontology (GO):** hierarchical vocabulary of terms describing genes and protein
- **Three GO represented by a root ontology term**
  - **cellular component** referring to the place in the cell (e.g. nucleus, ribosome)
  - **biological process** referring to a biological objective (e.g. signal transduction, alpha-glucoside transport)
  - **molecular function** describing activities (e.g. catalytic activity, Toll receptor binding)

- **Developed by a consortium** (geneontology.org)
  => collaborative effort
- **Widely used biological ressource**

- **41 species**
- **Generic terms / Some species-specific terms**

Annotation using systems biology tools

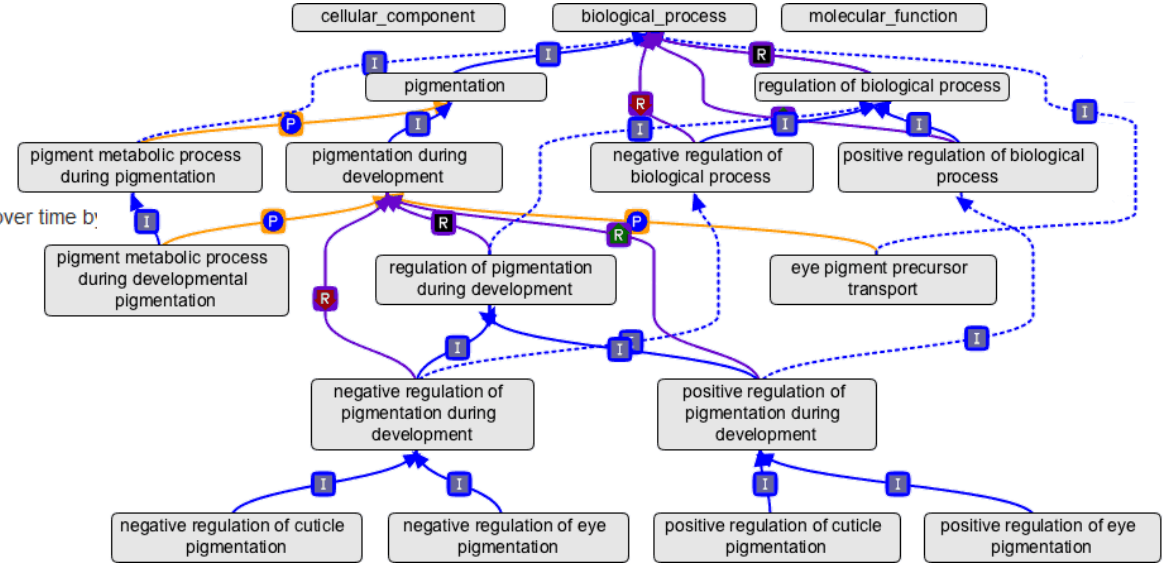
Functional enrichment analysis: Gene Ontology

- **GO structure:**
  - directed acyclic graph
  - types of relationships of child to parent: “is a” or “part of”
  - each term has defined relationships to one or more other terms.

Sample GO Term

The following is a GO term taken from the OBO format file.

- id: GO:0016049
- name: cell growth
- namespace: biological_process
- def: "The process in which a cell irreversibly increases in size over time b
- subset: goslim_generic
- subset: goslim_plant
- subset: gosubset_prok
- synonym: "cell expansion" RELATED []
- synonym: "cellular growth" EXACT []
- synonym: "growth of cell" EXACT []
- is_a: GO:0009987 ! cellular process
- is_a: GO:0040007 ! growth
- relationship: part_of GO:0008361 ! regulation of cell size



geneontology.org

Annotation using systems biology tools

Functional enrichment analysis: Gene Ontology

- **However some (commercial) tools developped their own ontology**

e.g.: Ingenuity Pathway Analysis (IPA) => IPA ontology
three main categories of function

INGENUITY®
PATHWAY ANALYSIS

- Diseases and disorders
- Molecular and cellular functions
- Physiological system development and function

Annotation using systems biology tools

Functional enrichment analysis: Biological pathway

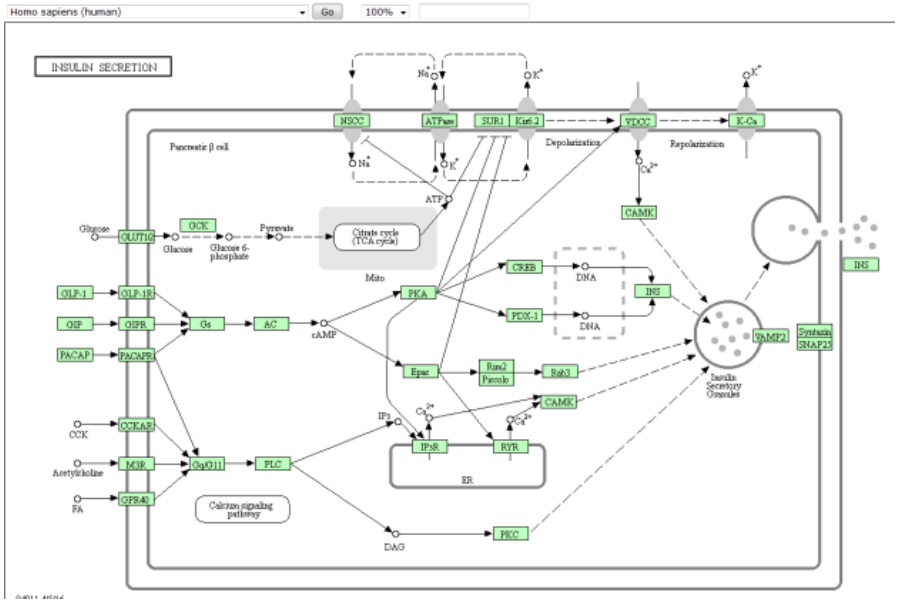
- **Biological pathway**
  - Biochemical engines responsible for the transduction of signals into output responses
  - A series of actions among molecules in a cell that leads to a certain product or a change in the cell.

- **Pathway building**
  - Process of identifying and integrating the molecules entities, interactions, and associated annotations
  - Contribute to the knowledgebase.
  - Can have either a data-driven or a knowledge-driven objective

Annotation using systems biology tools

Functional enrichment analysis: Biological pathway

- **Pathway databases**
  - e.g. consortium effort



  - Integration of several data sources
    - E.g.

  - commercial database

Functional enrichment analysis
Statistical tests

- **Which functional category or biological pathway is more prevalent in the gene list than expected by chance?**

- **Tests**
  - Fisher's exact test (one-tailed, right)
  - Gene Set Enrichment Analysis
  - Correction for multiple testing

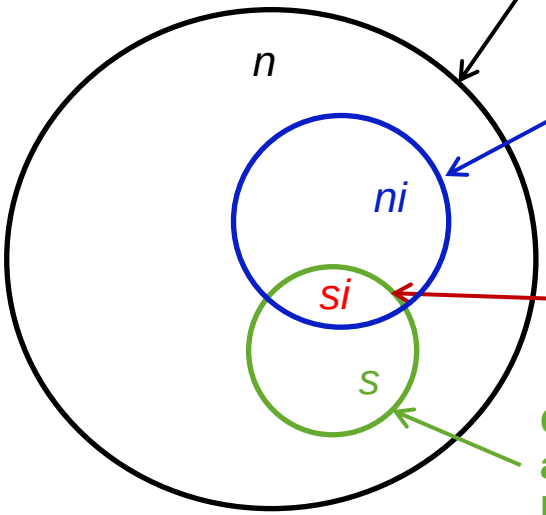
Functional enrichment analysis
Statistical tests

- **Fisher's exact test (one-tailed, right)**

2x2 contingency table

	User genes	All genes tested or genome
In pathway	s_i	ni
Not in pathway	s-s_i	n-ni

Is s_i/s more than random chance comparing to the background of ni/n?



All genes annotated with GO terms (or biological pathways)

Genes annotated with GO term (or biological pathway) *i*

Genes under selection annotated with GO term (or biological pathway) *i*

Genes under selection annotated with GO terms or biological pathways

- **Modified Fisher's exact test (one-tailed, right): e.g. EASE score**

More conservative: uses s_i**-1** instead of s_i

Functional enrichment analysis

Statistical tests

- **Gene Set Enrichment analysis (GSEA)** Subramanian et al, PNAS, 2005
  - Determines if an *a priori* defined set of genes are statistically significant (presumably concordantly different) between two biological states.
    - Sets of genes can be those within a pathway, biological process, etc.
    - Statistical significance determined by permutation (shuffling of the data)
  - Strategy
    - The genes are ordered on the basis of the parameter from the statistical test
    - For each gene set compute an enrichment score (ES) is computed (i.e a measure of how relevant or associated a biological process is for discerning the difference between the two biological states)
    - Create a running sum of a normalized Kolmogorov-Smirnov (non-parametric test) statistic.
    - Permute the class labels a large nb of times, each time recording the maximum ES over all gene sets.
    - Compare the observed ES score to the distribution of the ES scores from the permuted data.
    - Test the hypothesis that no gene set is associated with the class distinction

Functional enrichment analysis
Statistical tests

- **Gene Set Enrichment analysis (GSEA)**
  - A user-supplied ranked list of genes could also be used.
  - It determines whether *a priori* defined sets of genes show statistically significant enrichment at either end of the ranking.
  - A statistically significant enrichment indicates that the biological activity (e.g., biomolecular pathway) characterized by the gene set is correlated with the user-supplied ranking

Functional enrichment analysis
Statistical tests

- **Which functional category or biological pathway is more prevalent in the gene list than expected by chance?**

- **Fisher's exact test** (parametric)

- **Gene Set Enrichment Analysis** (non-parametric)

- **Multiple testing correction:**
  - Bonferroni correction or false discovery rate (FDR)

Functional enrichment analysis

Tools

- DAVID
- EASE
- AmiGO
- GeneGo MetaCore
- GOMiner
- BiNGO & ClueGO integrated with Cytoscape
- sigPathway & GOStat (R/Bioconductor based)
- FuncAssociate
- FatiGO
- GOEAST
- TopGO
- Gene Set Analysis (GSA)
- GSEAPreranked

- KEGG
- Wikipathway
- Reactome

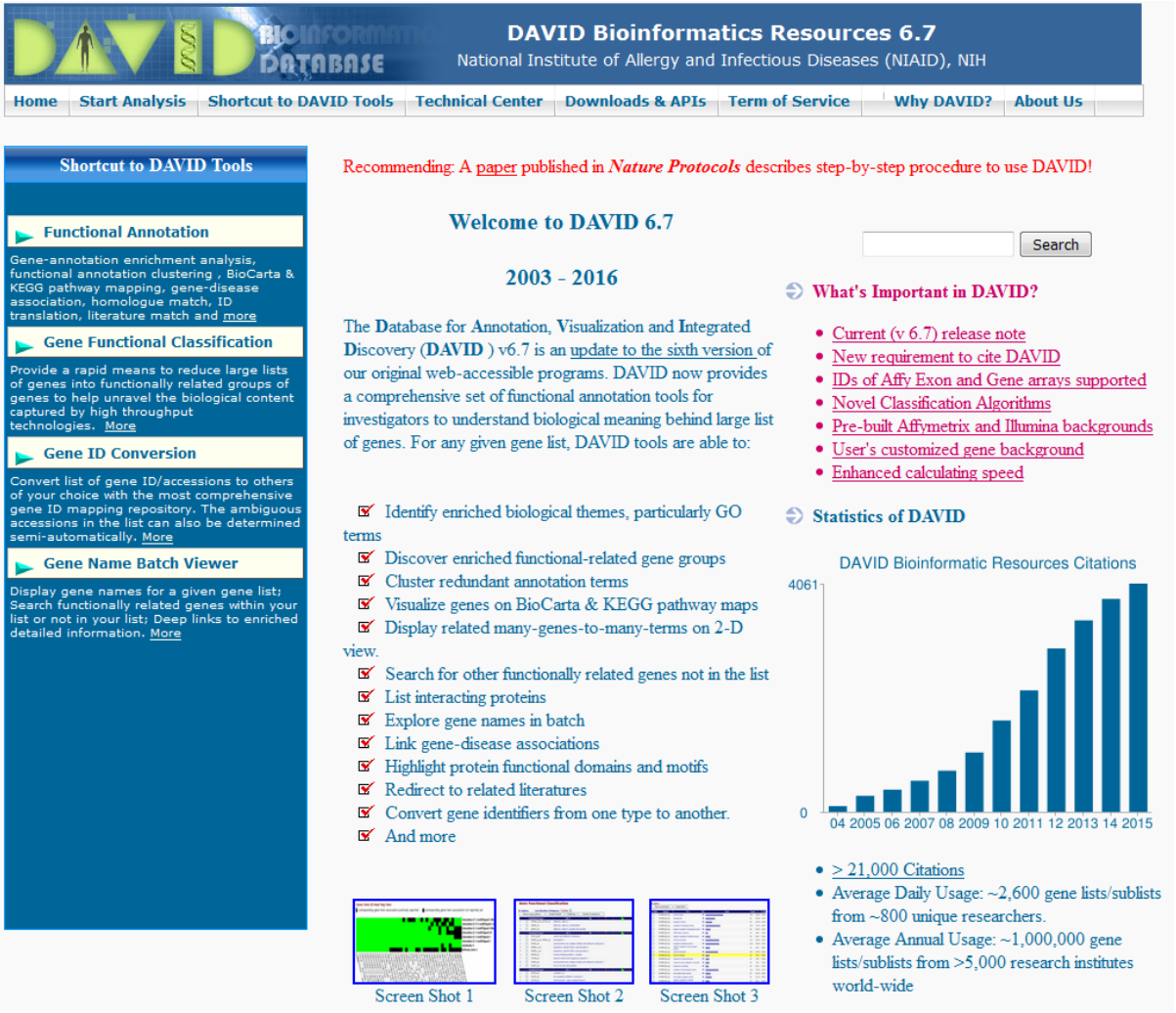
- Genemania in Cytoscape
- PANTHER
- InnateDB

- STRING

- Ingenuity Pathway Analysis

Functional enrichment analysis

Tools



Functional enrichment analysis
Limits

- **Function enrichment analysis (GO)**
  - Structure of GO: difficult to determine which level of hierarchy is most responsible for statistical enrichment
  - The most enriched terms are often broad functional categories, not very informative

- **Pathway analysis**
  - The majority of genes have not been assigned to a pathway
  - Bias towards well-studied signalling pathways
  - Informative about information already known

- **Statistical tests**
  - Choice of the reference gene set

- **Species of interest:** some tools are species specific

Gene network analysis

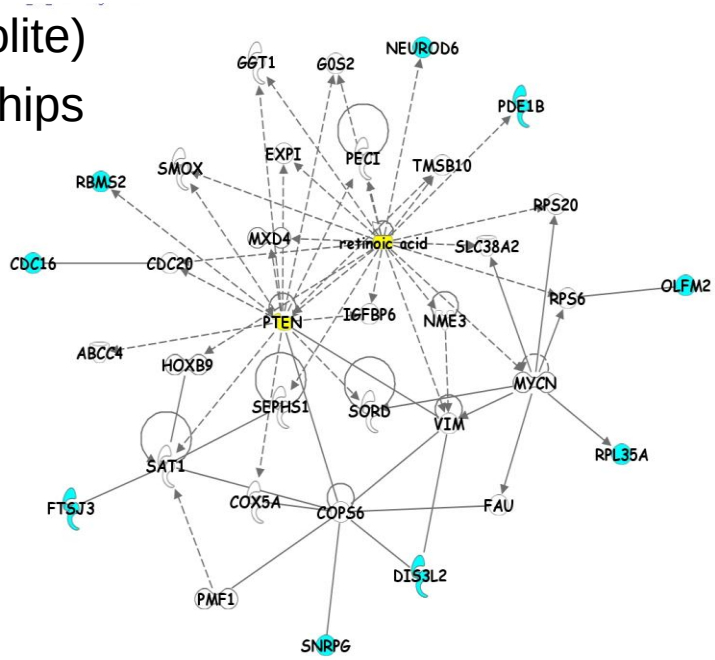
- **Biological processes are mainly controlled by complex networks of molecular interactions**

- **Network**

  – graph in which an entity (molecule or metabolite) is represented by a **node** and entities' relationships by **edges** between nodes

  – Not restricted to one type of nodes of edges

  – Based on publicly available data such as experimentally-supported interactions (e.g. protein-protein interactions: PPI)



Gene network analysis

Molecular interactions databases

- **Different type of interactions:**
  - physical (e.g. PPI, protein-DNA)
  - regulatory (e.g. miRNA-mRNA)
  - biochemicals interactions (e.g.phosphorylation)

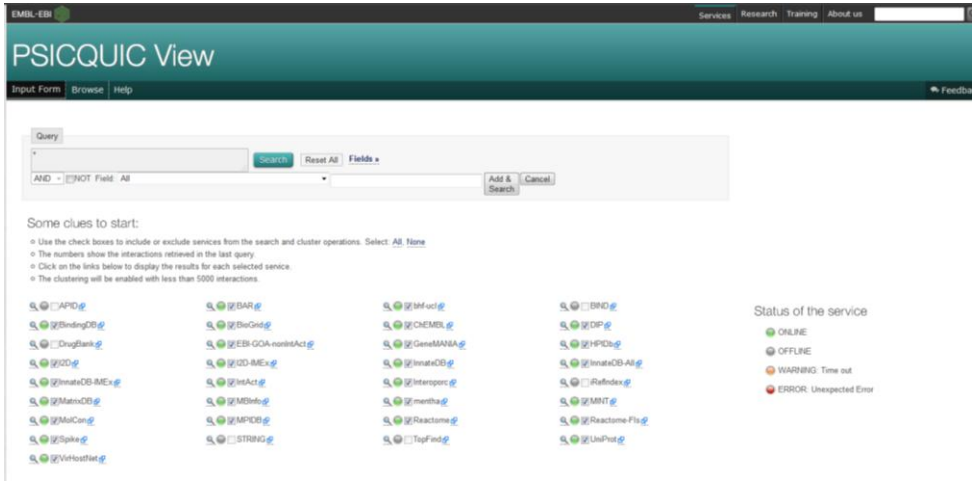
- **Experimentally-validated interation data obtained by:**
  - Primary database: Text-mining from peer-reviewed litterature
    - Manual curation from peer-reviewed litterature
  - Meta-databases :integration of different sources
- **Some databases included *in silico*-predicted interactions**

- **Limited overlap between interaction information of primary databases**
- **Need integartion of information provided by several primary databases**

Gene network analysis

Molecular interactions databases

- **Need integration of information provided by several primary databases:** e.g.
  - Web service PSICQUICK (integrated in Cytoscape and in Bioconductor/R)



  - Commercial: IPA, manual curation by experts

- **The meaning of an edge could vary as it integrates several different types of interaction**

Gene network analysis

A complementary approach to enrichment analysis methods

- **More data-driven**
- **Interactome** (molecular interactions within a biological system) **available for several species**
- **Less biased towards well-studied pathways**
- **Additional possible integration of information associated with nodes or edges**
- **Software development for network visualization**

Gene network analysis

Limits

- **Integration of different type of interactions**
  - The meaning of an edge could vary.
- **Level of confidence associated with interactions**
  - Some technics used to predict PPI in large-scale studies are associated with a high false positive rate (e.g. Yeast-2 Hybrid)
  - For more focused studies, biases toward well-studied biological processes
- **Interactions are context-specific**
- **Species**

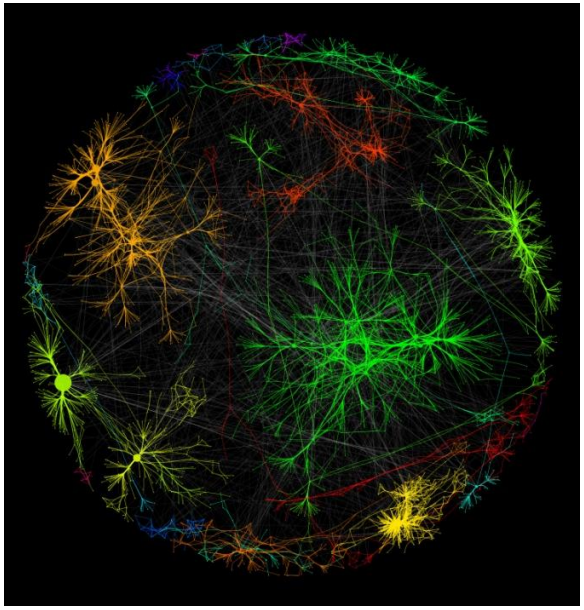
Gene network analysis

Tools

- **Cytoscape (open source):**
  - open source software platform for visualizing complex networks and integrating these with any type of attribute data.
  - a lot of Apps (plugins) available (most of them freely)

Cytoscape Apps store:

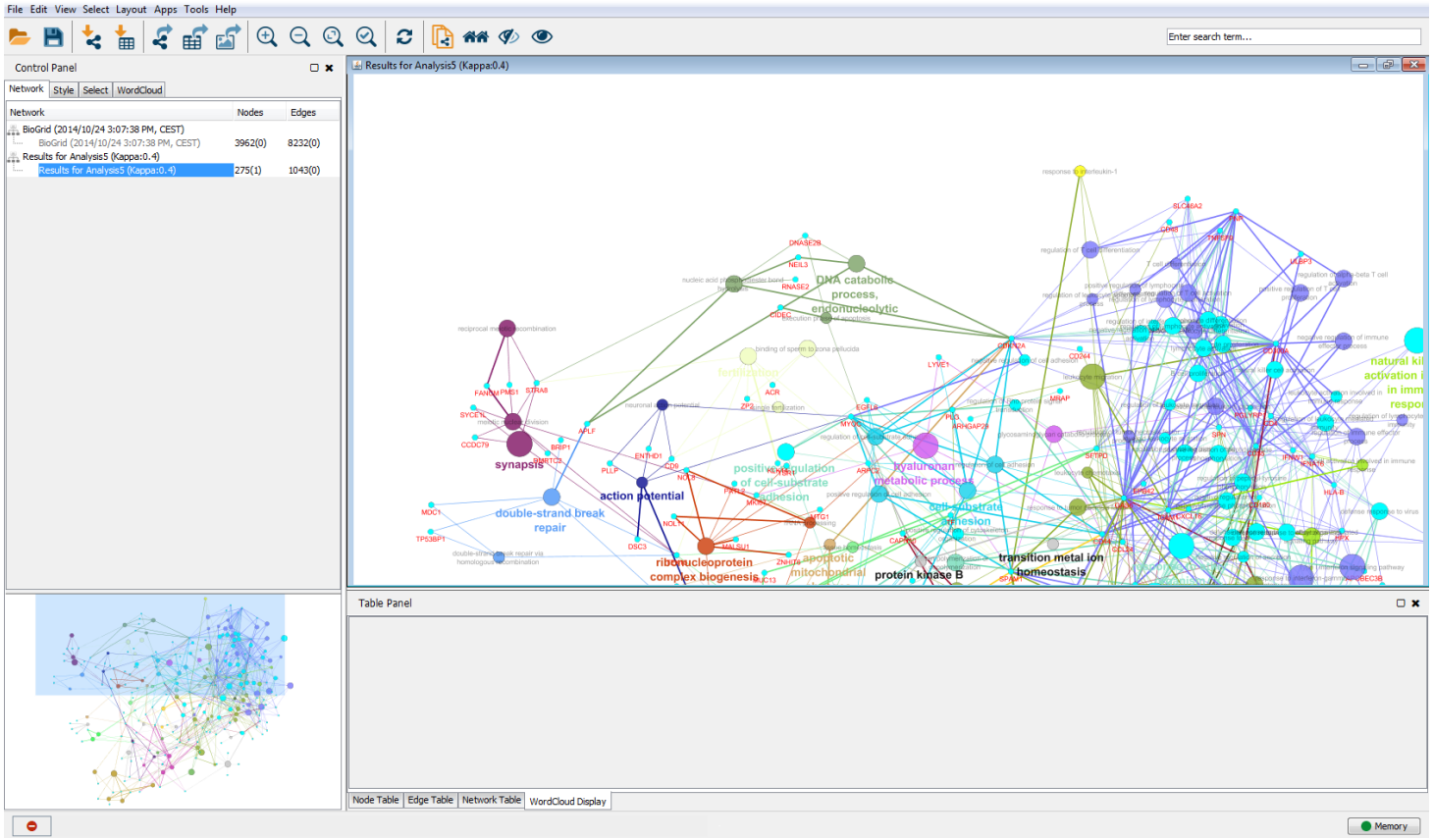
http://apps.cytoscape.org/apps/all



Gene network analysis

Tools

- **Cytoscape interface**

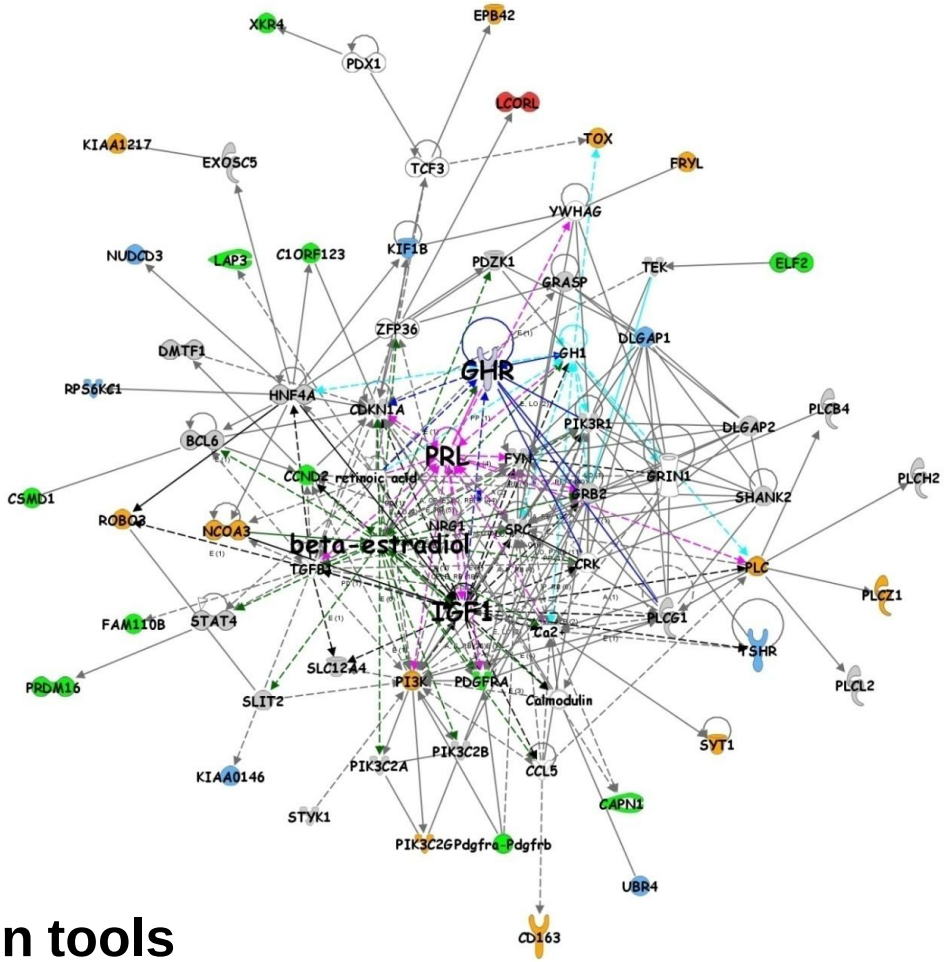


Gene network analysis

Tools

- **Ingenuity Pathway Analysis (commercial)**

INGENUITY®
PATHWAY ANALYSIS

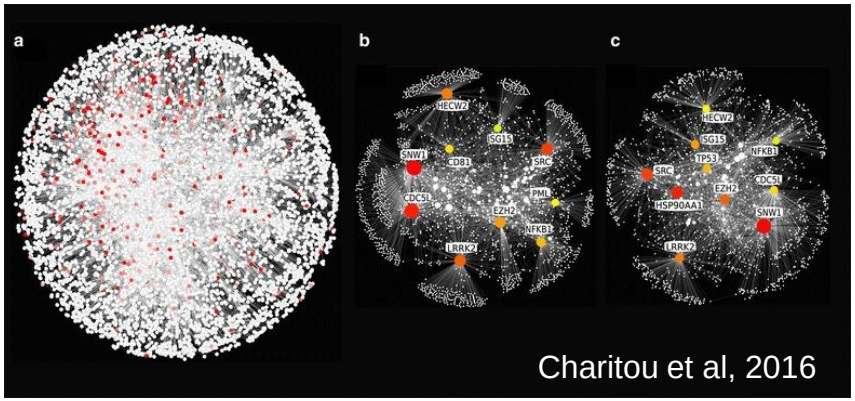


- **Links between tools**

Gene network analysis

Properties and features of gene networks

- **Hubs**
  - **i.e. High degree nodes**
    - Node degree: number of interactions/edges that a node has
  - Important for the network structure
  - Central
  - Less targeted by selection



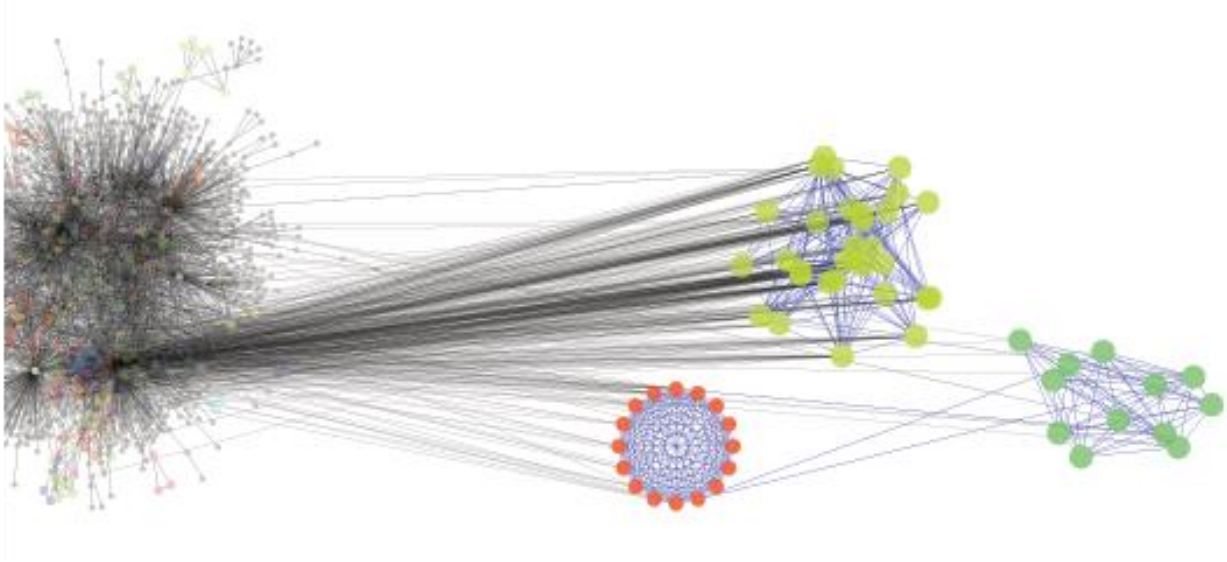
Charitou et al, 2016

- **Bottlenecks**
  - **i.e. Nodes with a high betweenness centrality**
    - nodes which are the crossroads of many shortest paths
    - Distance between nodes: minimum number of steps between them
  - **i.e. Major bridges**

Gene network analysis

Properties and features of gene networks

- **Modules**
  - i.e. group of molecules that preferentially interact which each other (sub-networks)
  - Tend to be enriched for common biological functions or diseases



Gene network analysis

Properties and features of gene networks

- **Tools to identify gene networks properties**
  - **Network Analyst**
  - **Cytoscape Apps:**
    - **CytoHubba**
    - **jActiveModules**

- **Cannot be done with some tools (e.g. IPA)**
  - **Export network in Cytoscape format (need sometimes a specific licence)**

Gene network and selection

- **Position of selected genes selected within PPI network**
  - Kim et al, 2007
    - Hubs tend to be constrained and under negative selection
    - Observed positive selection at the network periphery
  - Qian et al, 2014
    - Signal of selection tend to be underrepresented for genes with fewer neighbors
    - Tend to enrich at subcentral position of the PPI network
    - Molecules of high centrality could be under strong evolutionary constraints
    - Molecules at periphery may not contribute enough to phenotypic effect

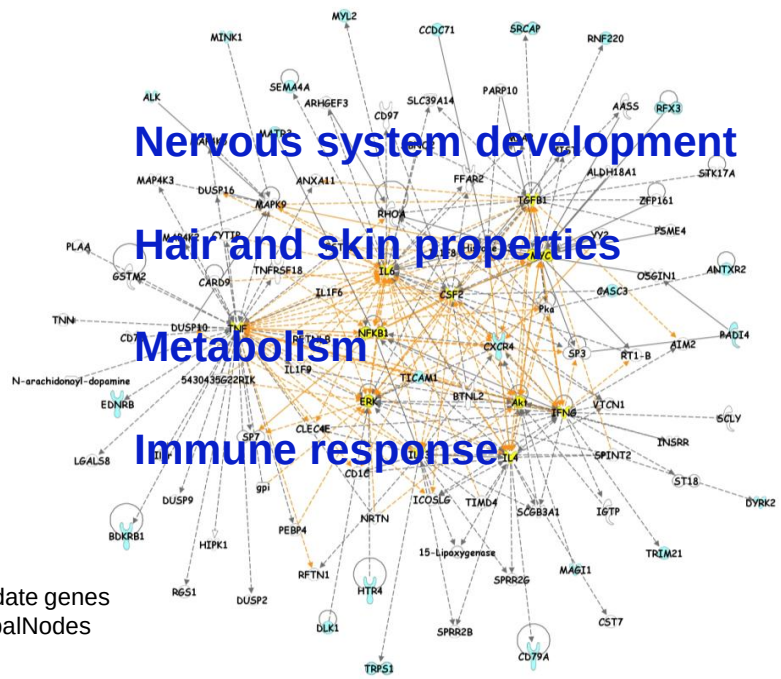
- **Interactions between recently selected genes** (Qian et al, 2014)
  - Closer interaction among genes under selection consistent with the effect of coselection

Inferring the main selective pressure

- **Gene network reconstruction and identification of the main functions targeted by selection** (IPA, Cytoscape)

- **Exemple in West-African cattle breeds**

Gautier *et al*, 2009, BMC Genomics
Flori *et al*, 2009, PLoS One
Flori *et al*, 2014, Mol. Ecol



Selective pressures

Climatic conditions

Drought and food shortage

Breeders'choices

Pathogens

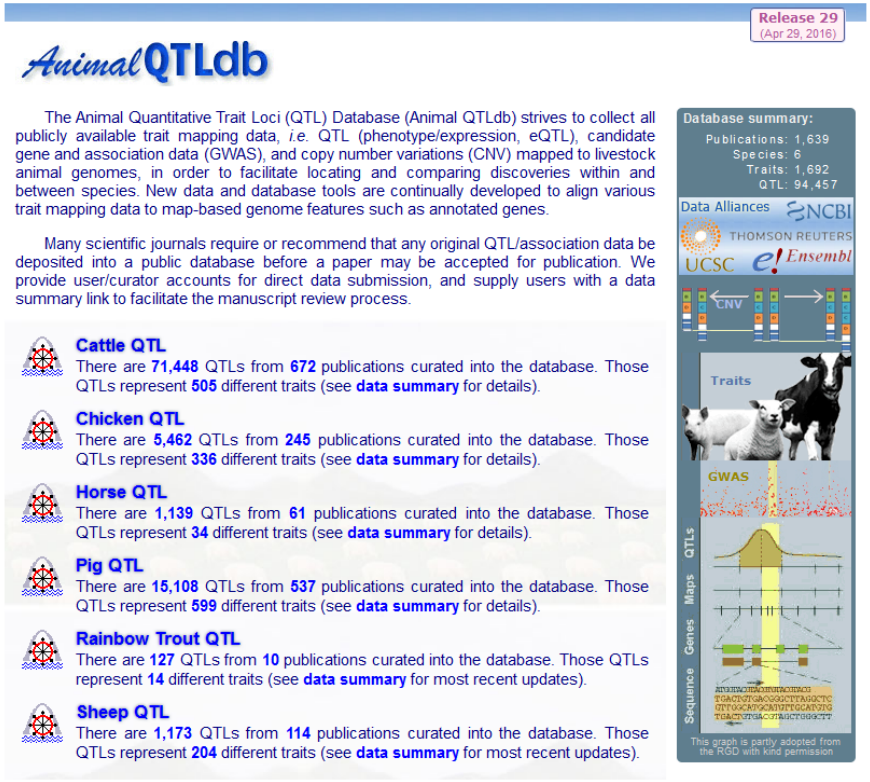
Interpretation and over-interpretation

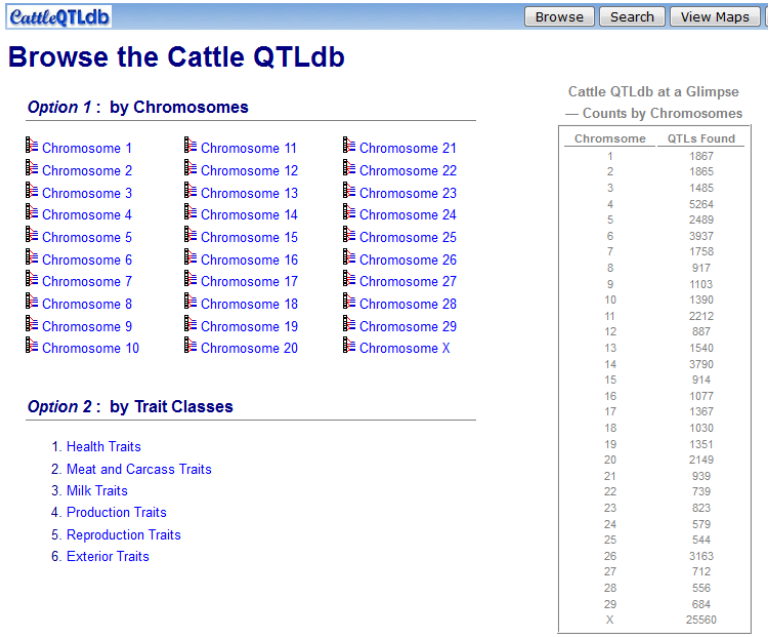
- **Goals of functional annotation of selection footprints**
  - Confirm the biological importance of the genes that have been identified as being positively selected .
  - Stengthen the credibility of the positive selection model by a development of a sound scenario

- **Sometimes a misleading approach prone to storytelling**
  - The majority of genes in a genome have important biological functions
  - easy to find connections between key genes using GO annotation and litterature mining tools
  - Ex: Pavlidis et al, MBE, 2012

Integrating additional information:
QTL location

- **QTL databases**





Association with phenotypes and environmental covariates

SNP chip
NGS (Pool and Ind-Seq)

Identification of footprints of selection in the genome

Ecological variables
e.g. WorldClim DB

Phenotype-free approaches

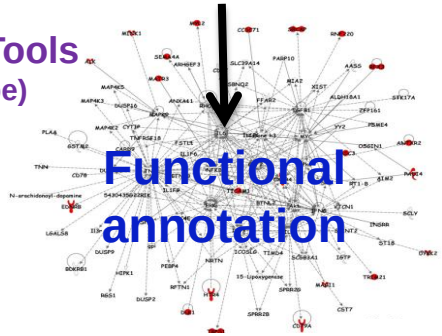
Association with population ecological variables or phenotypes

Reverse ecology

Phenotypes

Candidate regions and genes under selection

Systems biology Tools
(e.g. IPA, Cytoscape)



Functional annotation

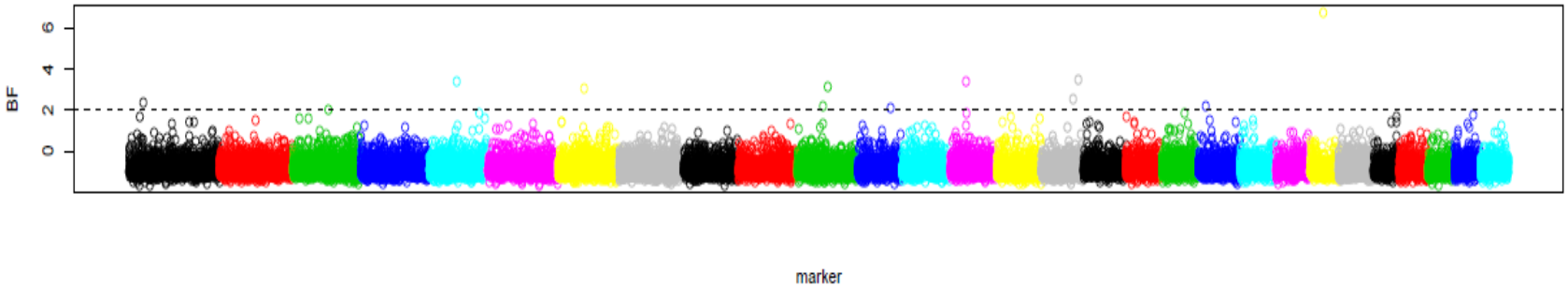
Phenotyping RNA-Seq

Principal selective pressures

Association with phenotypes and environmental covariates

- **Identifying loci underlying local adaptation using correlations between allele frequencies and ecological population variables or phenotypes**
  - Multi-dimensional methods **(PCA, Laloë et al, in prep)**
  - Bayesian model-based approaches **(e.g. Baypass; Gautier, 2015)**
  - LFMM **(Frichot et al, MBE, 2013)**

e.g Mediterranean cattle breeds (GALIMED project): SNPs contribution to the genetic variability according to mean temperature

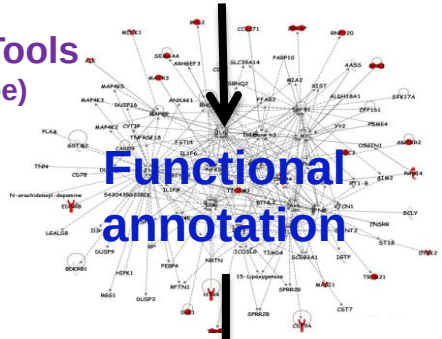


Identifying and annotating candidate mutations and prioritizing candidate genes

Candidate regions and genes under selection

Systems biology Tools
(e.g. IPA, Cytoscape)

Principal selective pressures



Functional annotation

Whole Genome mutation screening (Seq)

Functional annotation of variants

Candidate causal mutations for adaptive traits

Identifying candidate mutations

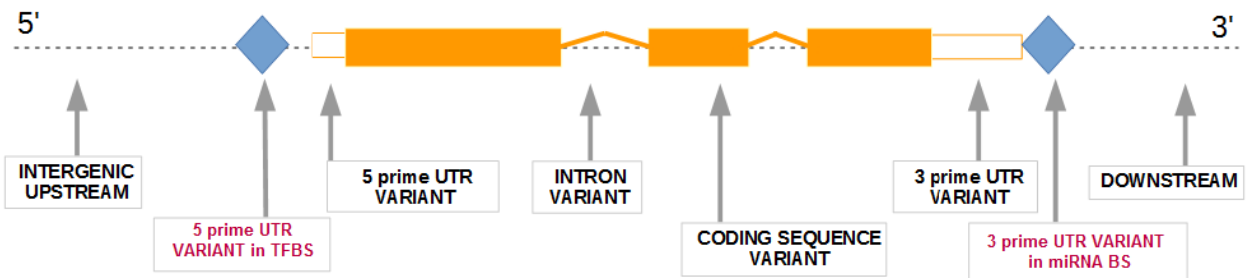
- **Individual whole genome sequencing**
  - **Alignment on reference genome**
  - **Variant calling**
  - **Calculation of allelic frequencies**
  - **Variant annotation**

Comprehensive list of variants in the genome with estimation of their frequency

Annotating candidate mutations

- **Exhaustive list of variants (SNP and indels)**

- **Location of the variant / genes position in reference sequence**



- **Estimation of variant consequences on transcription and protein structure**

Annotating candidate mutations

- **Tools**

e.g.

SnpEff
Genetic variant annotation and effect prediction toolbox.

Ve!P

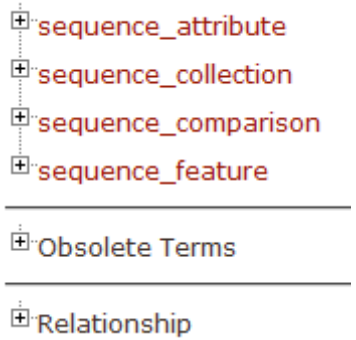
- **Use sequence ontology**
  - i.e. set of terms and relationships used to describe the features and attributes of biological sequence
  - Initially developed by the GO consortium

SO The Sequence Ontology Project
Home Browser Wiki GFF3 Resources About Request A Term Site Map

- **Integrate different information available on variants, e.g.**
  - SIFT
  - GERP scores

Annotating candidate mutations

- **Use sequence ontology**
  - i.e. set of terms and relationships used to describe the features and attributes of biological sequence
  - Initially developed by the GO consortium



Annotating candidate mutations

- **Sequence ontology**

- sequence_attribute
- sequence_collection
- sequence_comparison
- sequence_feature

- Obsolete Terms

- Relationship

- sequence_attribute
  - feature_attribute
  - polymer_attribute
    - nucleic_acid
      - peptidyl
    - synthetic_sequence
    - topology_attribute
  - sequence_location
    - organelle_sequence
    - plasmid_location
    - proviral_location
  - variant_quality
    - coding_variant_quality
    - variant_frequency
    - variant_origin
    - variant_phenotype

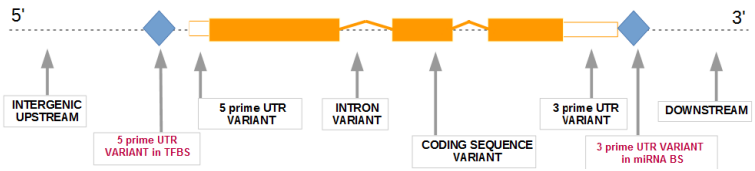
- sequence_collection
  - contig_collection
  - genome
  - peptide_collection
  - variant_collection

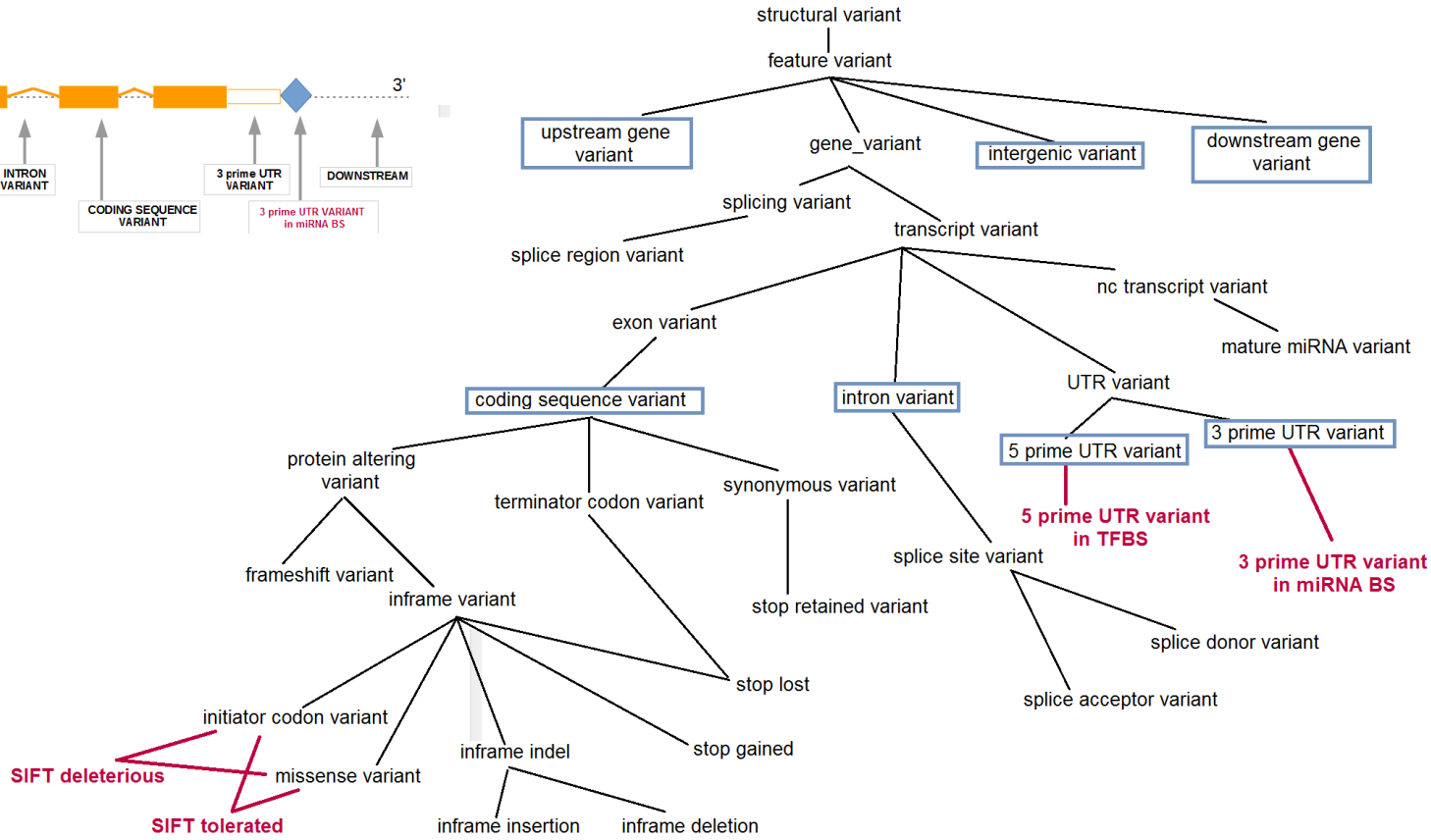
- sequence_comparison
  - no_variation
  - sequence_variant
    - functional_variant
      - dominant_negative_variant
      - gain_of_function_variant
      - lethal_variant
      - loss_of_function_variant
      - loss_of_heterozygosity
      - null_mutation
      - transcript_function_variant
        - level_of_transcript_variar
        - transcript_processing_va
        - transcript_stability_variar
        - transcription_variant
      - translational_product_functi
    - structural_variant
      - copy_number_change
      - feature_ablation
      - feature_amplification
      - feature_fusion
      - feature_translocation
      - feature_variant

- sequence_feature
  - junction
    - chromosome_breakpoint
      - clone_insert_end
      - clone_insert_start
      - deletion_junction
      - exon_junction
    - insertion_site
      - polyA_site
    - restriction_enzyme_cleavage_
      - splice_junction
      - trans_splice_junction
  - region
    - biological_region
    - biomaterial_region
    - experimental_feature
    - topologically_defined_region
  - sequence_alteration
    - UPD
    - copy_number_variation
    - deletion
      - indel
    - insertion
    - inversion
    - structural_alteration
    - substitution

Annotating candidate mutations

- **Sequence ontology in VEP (ensembl)**





Annotating candidate mutations

- **Prediction of regulatory variant effect on transcription/expression**
  - e.g. variant within TFBS
  - Integrated in VEP for some species
  - Partial annotation in regulatory regions for other species
    need other tools: e.g. Lasagna, mrSNP

- **Prediction of variant effect on protein structure:**
  - **SIFT:** applies on non-synonymous polymorphism and predicts whether an amino acid substitution affects protein function.
    - based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences.
    - Tolerated/Deleterious (high, low confidence)

Annotating candidate mutations

- **Prediction of variant effect in the whole sequence**

  - **Genomic Evolutionary Rate Profiling (GERP)** :
    a method for producing position-specific estimates of evolutionary constraint using maximum likelihood evolutionary rate estimation and discovering "constrained elements“ that is indicative of a putative functional element.

    - Based on multiple sequence alignments and phylogenic tree
    - Constraint intensity quantified in terms of a "rejected substitutions" (RS) score, i.e. the number of substitutions expected under neutrality minus the number of substitutions "observed" at the position.
    - Positive scores represent a substitution deficit and thus indicate that a site may be under evolutionary constraint. Negative scores indicate that a site is probably evolving neutrally

Variants prioritization

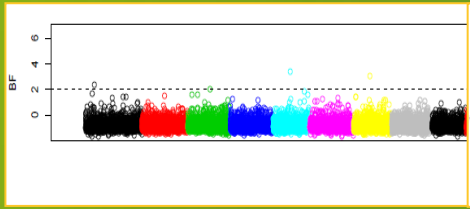
Different categories based on impact on protein level or protein structure, e.g.:

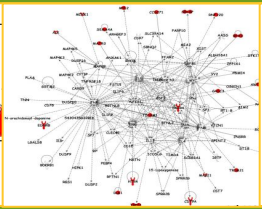
- Strong impact:
  - Non synonymous mutation deleterious (SIFT)
  - 5'UTR within TFBS
  - 3'UTR within miRNA binding site
  - In splice region
  - With positive GERP score
- Weakest impact
  - 5'UTR not in TFBS
  - 3'UTR not in miRNA binding site
  - Downstream
  - Upstream
  - Intronic
  - Non synonymous mutation tolerated

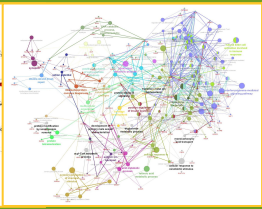
Candidate genes prioritization

- **Criteria**
  - Proximity to the position with the highest score (peak)
  - Number of variants in categories with strong effect (after variant annotation)
  - Allelic frequencies
    - mutation fixed (freq=1)
    - Mutation with high frequency
- **Functional annotation of these new list of candidate genes using systems biology tools**

2. Some examples of biological interpretation







INRA
SCIENCE & IMPACT

Course outline

1. **From regions under selection to biological interpretation**

   - Annotation of candidate genes using systems biology tools
     - Functional enrichment analysis: Gene ontology, pathway analysis
     - Gene networks analysis
   - Inferring the main selective pressures
   - Interpretation and story-telling
   - Including phenotypes and environmental covariates
   - Identifying candidate mutations and prioritizing candidate genes

2. **Interpretation of selection footprints: some examples**

   - Phenotype-free approaches
     - Manual functional annotation: Senepol cattle breed
     - Functional annotation using systems biology tools: french dairy cattle breeds, West-African cattle breeds, European bison/cattle
   - Association with covariates
     - Phenotypes: dairy traits in French cattle breeds

Manual functional annotation

The exemple of the Senepol cattle breed

- **Senepol: a European taurine breed with a small proportion of zebu ancestry.**

- **Living in tropical area** (Caribbean, St Croix island)

Flori *et al*, 2012, PLoS One

- **Identification of selection footprints**

  - 153 individuals genotyped on 47,365 SNPs
  - Tests based on the extent of haplotype homozygosity: iHS (Voight et al, 2007) and Rsb (Tang et al, 2007)

    Package rehh, Gautier & Vitalis, 2012
  - 1Mb sliding windows (0.5Mb overlap)
  - Selection of regions with at least 2 SNPs exceeding the significant threshold (P<0.0001)

Manual functional annotation

The exemple of the Senepol cattle breed

- **Only four regions under selection**

Region	BTA	Position (Mb)	Peak position (Mb)	iHS_{SEN}	$Rsb_{EUT/SEN}$	$Rsb_{ZEB/SEN}$	Nb of significant SNPs	Gene closest to the maximum
#1	1	52.6–53.6	52.9	3.984	NS	3.984	2	No gene found
#2	1	2.4–3.4	3.2	NS	4.161	NS	6	TIAM1
#3	1	4.7–5.8	5.5	NS	4.538	NS	8	GRIK1
#4	20	38.6–39.6	39.5	4.055	4.576	4.055	3–6	RAI14

BTA1 =>*polled* locus

BTA20 =>*slick* locus involved in a short and sleek hair coat and in thermotolerance/adaptation to tropical conditions

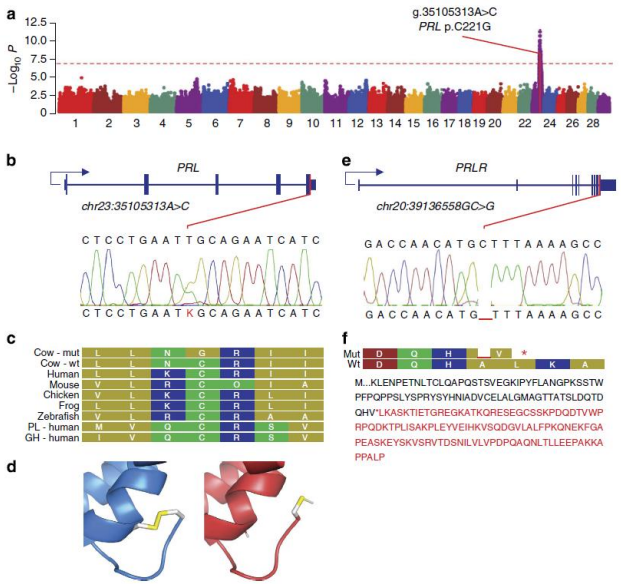
– Identificarion of candidate gene: boundaries located less than 25kb from the peak position

Manual functional annotation

The exemple of the Senepol cattle breed

- **Identification of the mutation responsible of the slick phenotype within the PRLR (Littlejohn et al, 2014)**

- **PRLR is located in the region #4 found under selection in Senepol**



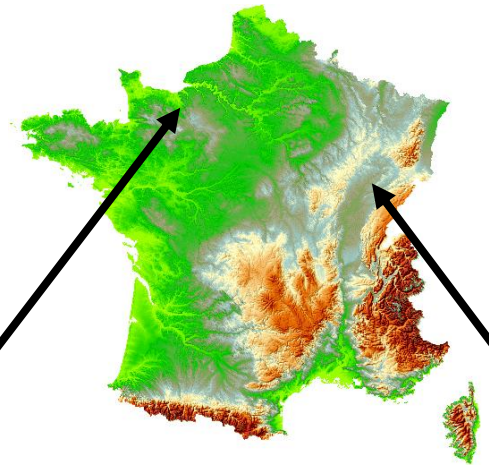
PRLP p.Leu462 WT

Annotation using systems biology tools

The example of the French dairy cattle breeds

- **The three main French dairy cattle breeds =>artificial selection**

Flori *et al*, 2009, PLoS One



HOLSTEIN

70%

NORMANDE

15%

MONTBELIARDE

~15%

Breed	Herd-Book creation	Population size *(2002)*	Ne *(généalogie)*	Milk production *(kg)*	Content *(°/oo fat/Protein)*
MON	1872	1,799,200	34	7,441	38.8 / 32.5
NOR	1883	2,106,000	61	6,595	44.2 / 36.0
HOL	1922	11,535,378	42	8,628	40.9 / 31.6

- **Increase in milk production but decline in reproductive performances**

Annotation using systems biology tools
The example of the French dairy cattle breeds

- **Identification of selection footprints**
  - 2,803 bulls genotyped on 42,486 SNPs
  - Differentiation (Fst) between breeds (Nicholson et al, 2002)
  - Empirical approach
    - combining information between closely related SNPs
    - Smoothed individual SNP Fst values over each chromosome for observed and simulated datasets
    - Calculation of local q-values

- **13 regions under selection** (qvalue<0.05)

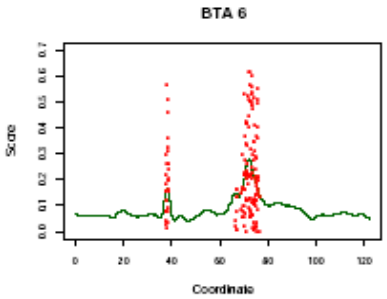
- **Some regions contain genes with causal variants with a strong effect on milk production trait (GHR) or coloration (MC1R)**

Annotation using systems biology tools

The example of the French dairy cattle breeds

- **Regions under selection**

#	BTA	Start-End (peak position) in Mb	F_{ST} at the peak position (qvalue)	candidate gene	Breeds within which region is also significant
1	3	57.084–58.505 (58.343)	0.375 (0.0298)	CCCBL2	
2	4	78.833–80.43 (79.701)	0.667 (0.0298)	NUDCD3	
3	5	20.301–23.091 (21.02)	0.483 (0.0298)	na	NOR, HOL
4	5	97.803–100.826 (98.26)	0.557 (0.0298)	PIK3C2G	NOR, HOL
5	5	108.461–109.236 (109.182)	0.403 (0.0401)	CD163	
6	5	110.286–111.861 (111.552)	0.46 (0.0435)	ANO2	
7	6	37.433–38.756 (37.963)	0.566 (0.0298)	LAP3/LCORL	MON
8	6	66.599–66.935 (66.809)	0.165 (0.0435)	na	
9	6	68.938–76.32 (72.024)	0.616 (0)	PDGFRA	NOR
10	14	22.02–25.567 (22.634)	0.591 (0)	na	MON, NOR
11	18	12.987–14.058 (13.36)	0.632 (0)	MC1R	MON, HOL
12	20	31.964–33.757 (32.277)	0.523 (0.0298)	GHR	
13	26	22.137–23.191 (22.983)	0.509 (0.0298)	C10ORF76	



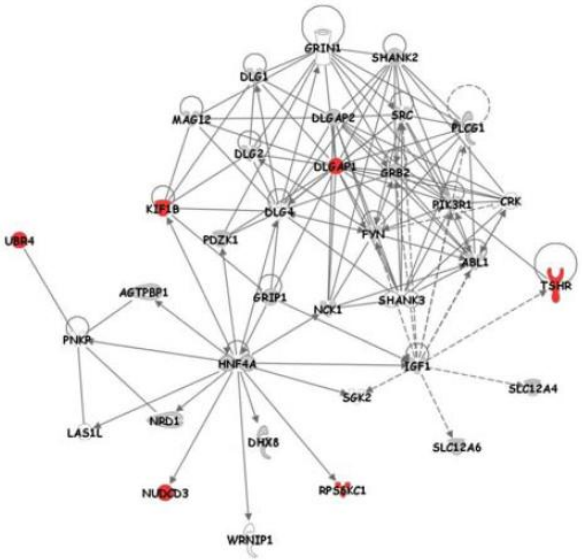
- Some regions contain genes with causal variants with a strong effect on milk production trait (GHR) or coloration (MC1R)
- Some regions contain loci involved in morphological traits

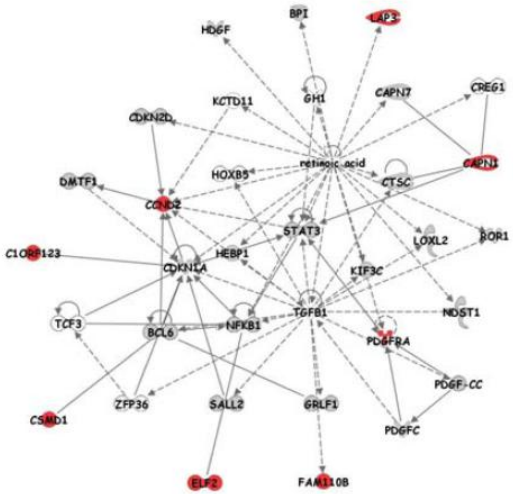
Annotation using systems biology tools

The example of the French dairy cattle breeds

Holstein

7 eligible genes/8



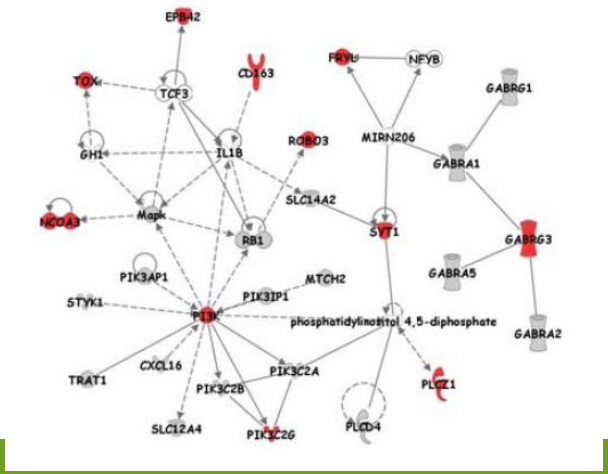


Montbeliarde

13 eligible genes /16

Normande

14 eligible genes/19



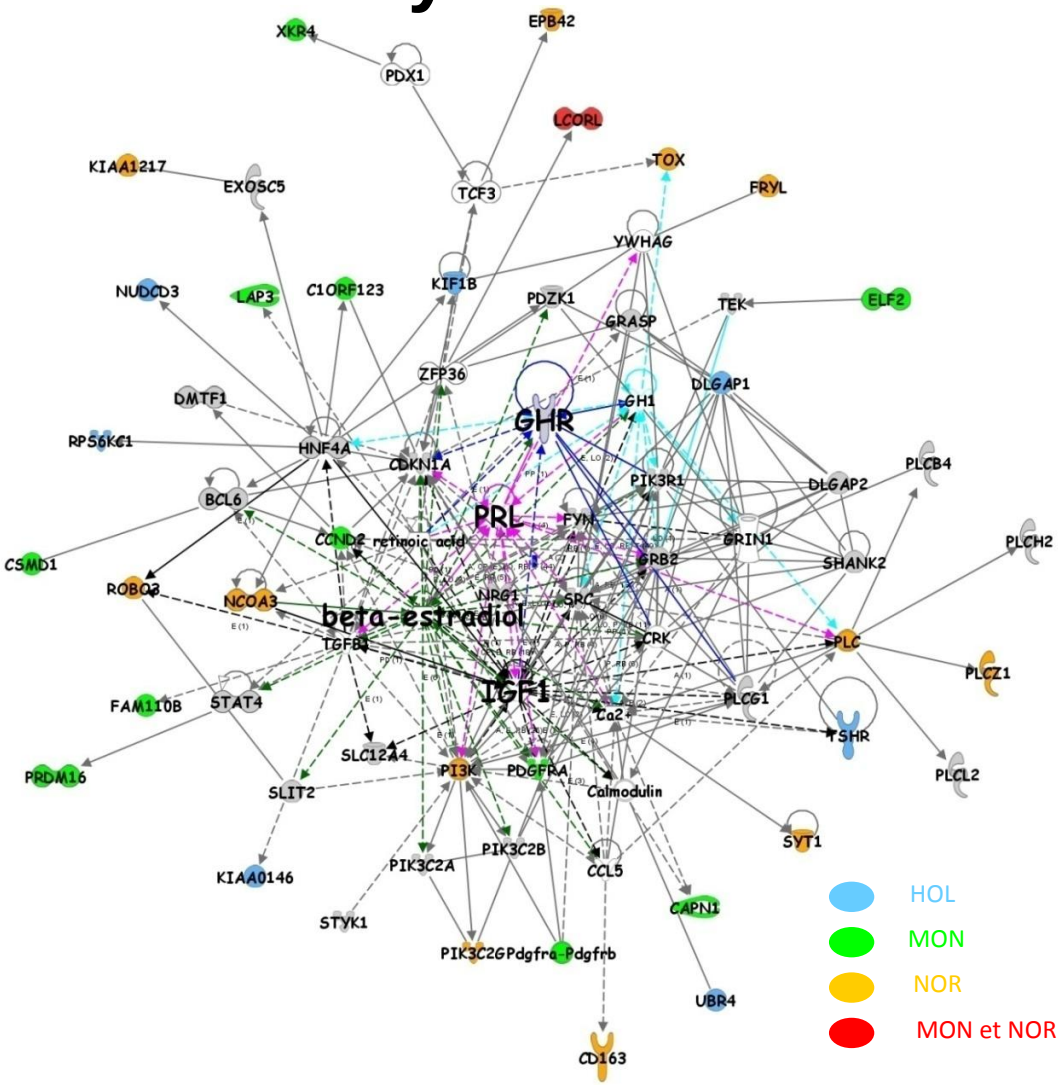
- **Significant networks contain different genes under selection in the three breeds**
- **But genes under selection in the each breed are involved in the same biological pathway**
- **Genome Plasticity of genome response to the same selective pressure**

INGENUITY PATHWAY ANALYSIS

Annotation using systems biology tools

The example of the French dairy cattle breeds

- **Global gene network**
  - contains genes under selection in at least one breed
- **Central role of somatotropic and gonadotropic axes in response to artificial selection**
- **Illustrates the antagonism between milk production and reproduction**



Course outline

1. **From regions under selection to biological interpretation**

   - Annotation of candidate genes using systems biology tools
     - Functional enrichment analysis: Gene ontology, pathway analysis
     - Gene networks analysis
   - Inferring the main selective pressures
   - Interpretation and story-telling
   - Including phenotypes and environmental covariates
   - Identifying candidate mutations and prioritizing candidate genes

2. **Interpretation of selection footprints: some examples**

   - Phenotype-free approaches
     - Manual functional annotation: Senepol cattle breed
     - Functional annotation using systems biology tools: French dairy cattle breeds, West-African cattle breeds, European bison/cattle
   - Association with covariates
     - Phenotypes: dairy traits in French cattle breeds

Annotation using systems biology tools

The exemple of the West-African cattle breeds

- **West-African cattle: models of adaptation to tropical conditions**

Gautier et *al*, 2009, BMC Genomics

Exemple of the West-African cattle breeds :
annotation using systems biology tools

- **Identification of selection footprints**

  - 342 individuals genotyped on 36320 SNPs
  - Differentiation (Fst, Bayesian model) between 9 West-African populations
  - Decision rule to identifiy no-neutral loci based on BF, expressed in deciban unit ($dB_i=10\log_{10}(BF_i)$)
  - Smoothed individual SNP BF values over each chromosome
  - Permutations to estimate local p-values
  - Correction of the local p-values by computing q-value

- **53 regions under selection** (at the 5% local FDR)

Exemple of the West-African cattle breeds :
annotation using systems biology tools

- **SNP annotation using Transmap Refseq**
  - 46,598 Refseq ID anchored to the bovine genome assembly
  - A SNP is considered representative of a gene if localized within gene boundaries extended by 15kb upstream and downstream
  - Annotation of the Refseq using IPA => 7,177 different genes

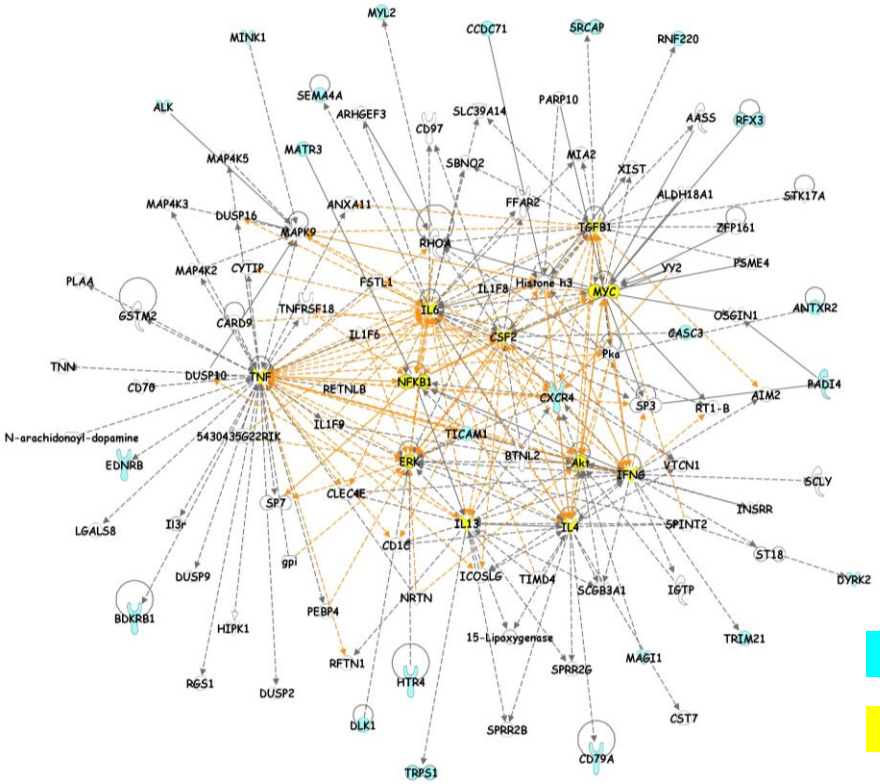
- **Identification of 42 candidate genes**

- **Functional analysis using IPA**
- **Network analysis using IPA and functional annotation of each significant network**
  - Each network contains at most 35 molecules

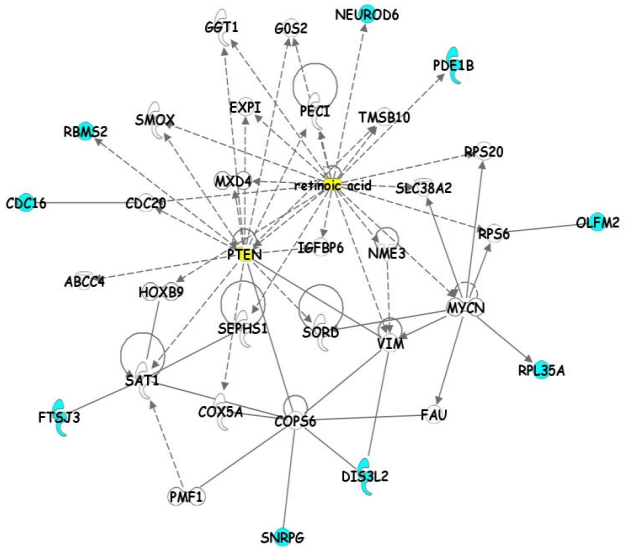
Exemple of the West-African cattle breeds :

annotation using systems biology tools

A. Network N (obtained merging N1, N2 et N4) 22 genes under selection



B. Network N3 9 genes under selection



Genes under selection

Network nodes

=> Nervous system diseases and disorders

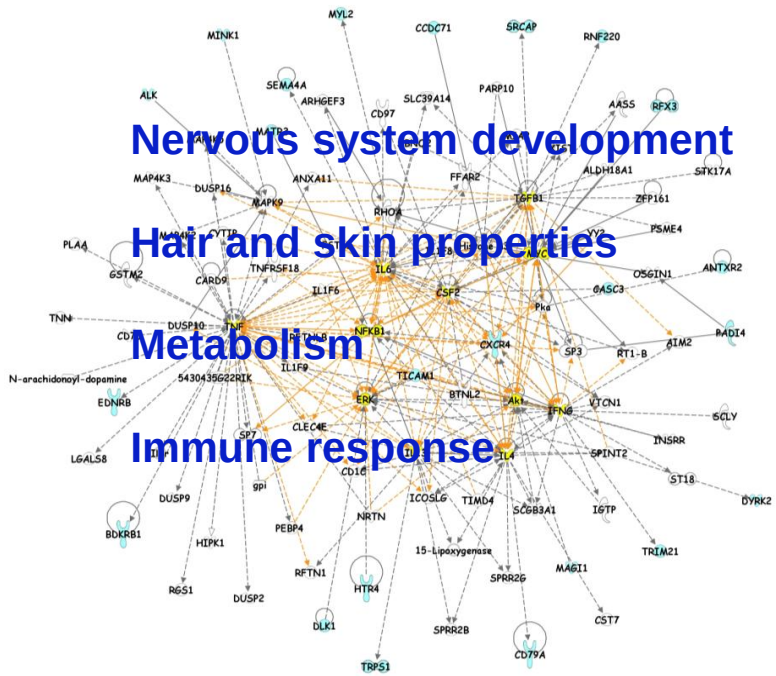
=> Innate and adaptive immune response

INGENUITY PATHWAY ANALYSIS

Exemple of the West-African cattle breeds :

annotation using systems biology tools

Gautier *et al*, 2009, BMC Genomics
Flori *et al*, 2009, PLoS One
Flori *et al*, 2014, Mol. Ecol



Candidate genes
PrincipalNodes

Selective pressures

Climatic conditions

Drought and food shortage

Breeders'choices (coat color, horn)

Pathogens

Course outline

1. **From regions under selection to biological interpretation**

   - Annotation of candidate genes using systems biology tools
     - Functional enrichment analysis: Gene ontology, pathway analysis
     - Gene networks analysis
   - Inferring the main selective pressures
   - Interpretation and story-telling
   - Including phenotypes and environmental covariates
   - Identifying candidate mutations and prioritizing candidate genes

2. **Interpretation of selection footprints: some examples**

   - Phenotype-free approaches
     - Manual functional annotation: Senepol cattle breed
     - Functional annotation using systems biology tools: French dairy cattle breeds, West-African cattle breeds, European bison/cattle
   - Association with covariates
     - Phenotypes: dairy traits in French cattle breeds

Annotation using systems biology tools
The example of the European bison

Gautier et al, under revision

- **Wisent: the largest European herbivore, emblematic of the continent wildlife**

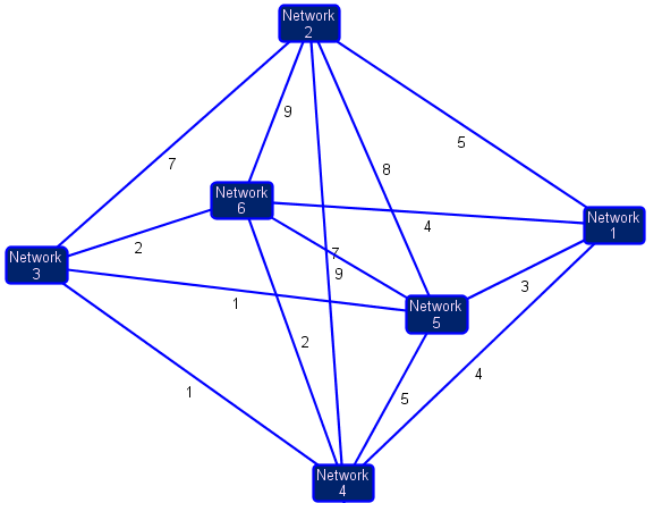
- **Identification of selection footprints between the European bison and cattle**

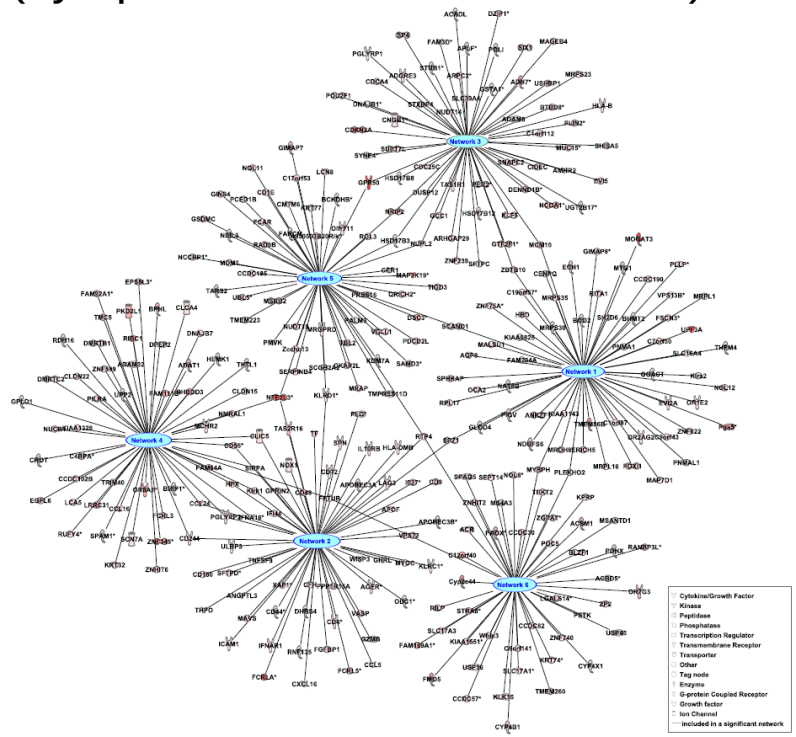
  - Using individual whole-genome sequences (10X, 100nt paired-end)
  - Calculating K_a/K_s (Kimura, 1983; *KaKs calculator*, Zhang et al, 2006):
    - the K_a/K_s ratio is an indicator of selective pressure acting on a protein-coding gene.
    - i.e. ratio of the number of non-synonymous substitutions per non-synonymous site (K_a) to the number of synonymous substitution per synonymous site (K_s), in a given period of time.
  - Genes with a K_a/K_s ratio above 1 are evolving under positive selection

Annotation using systems biology tools

The example of the European bison

- **873 transcripts under selection** => 450 genes ready for functional and network analyses
- **70% of the genes participated to a global network** corresponding to six significant interconnected networks (by up to 12 common molecules)





INGENUITY PATHWAY ANALYSIS

Annotation using systems biology tools

The example of the European bison

- **Functional analysis**

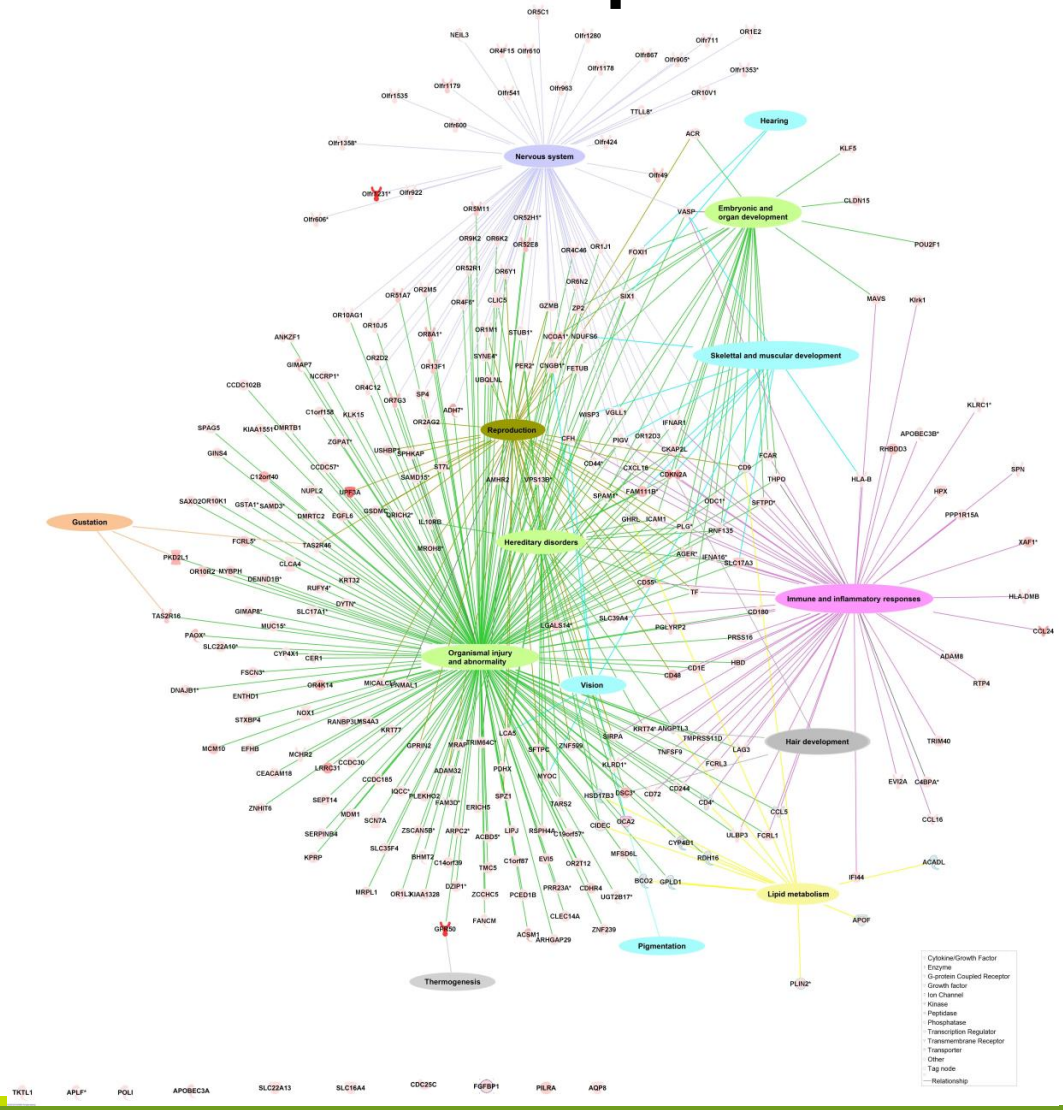
Main functional categories	Diseases and Biological Functions	p-value	Number of molecules
Physiological System Development and Function	Nervous System Development and Function	$1.26x10^{-16}$ - $3.05x10^{-02}$	54
	Hematological System Development and Function	$1.88x10^{-04}$ - $3.05x10^{-02}$	39
	Immune Cell Trafficking	$1.88x10^{-04}$ - $3.05x10^{-02}$	34
	Embryonic development	$2.07x10^{-03}$ - $3.05x10^{-02}$	15
	Organ development	$2.07x10^{-03}$ - $3.05x10^{-02}$	11
Diseases and Disorders	Inflammatory Response	$1.88x10^{-04}$ - $3.05x10^{-02}$	35
	Infectious Disease	$2.70x10^{-04}$ - $3.05x10^{-02}$	14
	Connective Tissue Disordes	$3.29x10^{-04}$ - $3.05x10^{-02}$	12
	Organismal Injury and Abnormalities	$3.29x10^{-04}$ - $3.05x10^{-02}$	208
	Skeletal and Muscular Disorders	$3.29x10^{-04}$ - $3.05x10^{-02}$	11
Molecular and Cellular Functions	Cell Death and Survival	$2.96x10^{-05}$ - $3.05x10^{-02}$	33
	Cellular Compromise	$2.96x10^{-05}$ - $3.05x10^{-02}$	24
	Cell Morphology	$1.10x10^{-04}$ - $3.05x10^{-02}$	22
	Cellular Assembly and Organization	$1.10x10^{-04}$ - $3.05x10^{-02}$	17
	Cell-To-Cell Signaling and Interaction	$1.17x10^{-04}$ - $3.05x10^{-02}$	94

  - Most significant category: Nervous System Development and Function

INGENUITY PATHWAY ANALYSIS

Annotation using systems biology tools

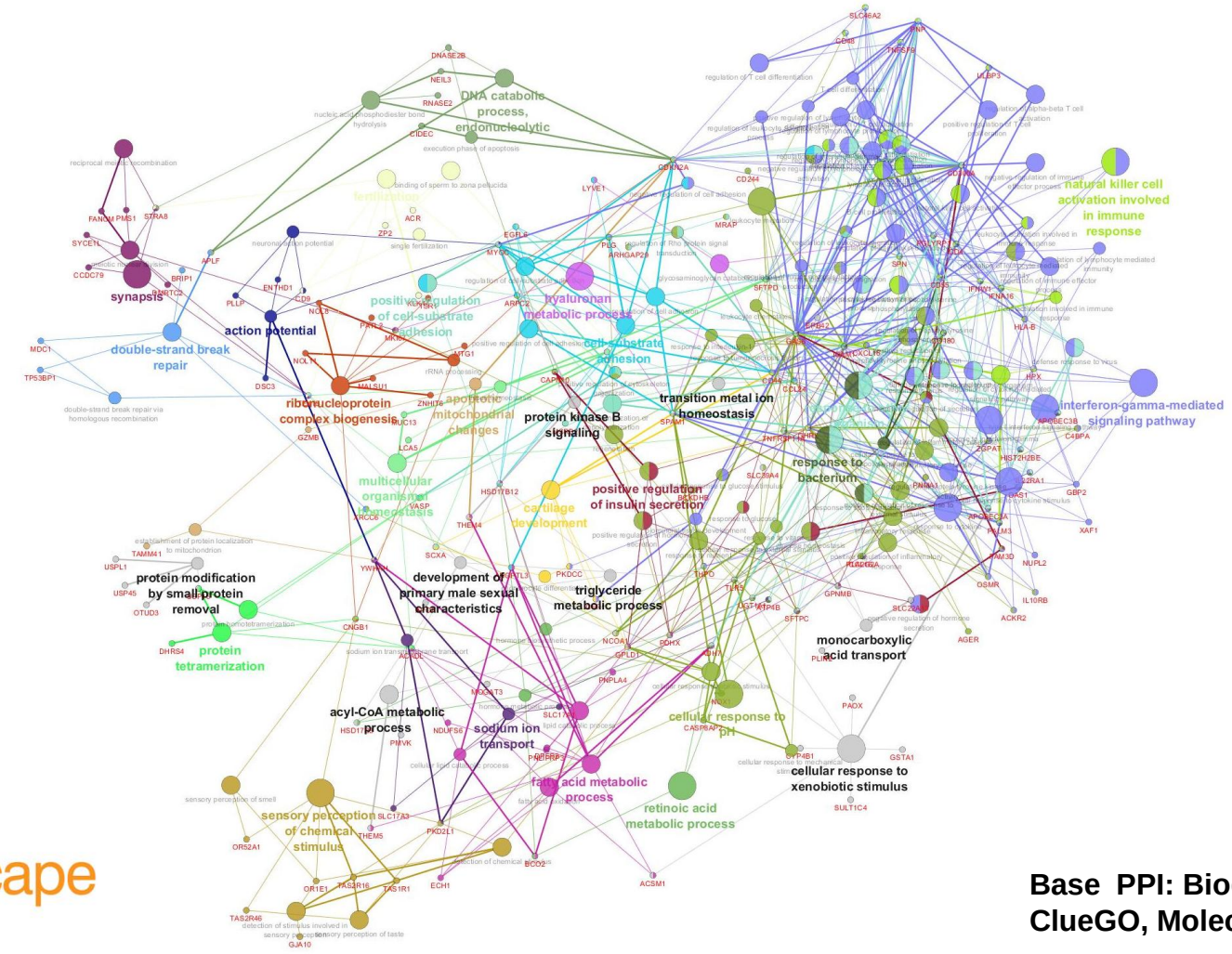
The example of the European bison



- **Many genes are related to several functions**
  =>pleiotropic role
- **Some genes underlie obvious distinctive features between wisent and cattle**

Annotation using systems biology tools

The example of the European bison



Cytoscape

Base PPI: Biogrid
ClueGO, Molecular Process

Annotation using systems biology tools

The example of the European bison

- **Some genes underlie obvious distinctive features between wisent and cattle**

Key functions	Presumed adaptation
➢ Hair development and thermogenesis	=> to cold climatic conditions (wisent)
➢ Olfactory and taste receptor genes	=> to food ressources (Forest-habitat, wisent)
➢ Genes involved in immune reponse	=> to different pathogen exposures (wildlife/domestication)
➢ Genes involved in lipid metabolism and mammary gland development	=> artificial selection in cattle (dairy traits)
➢ Genes involved in key processes of nervous system and in several by-product phenotypes of domestication	=>to domestication/wildlife

Annotation using systems biology tools

The example of the European bison

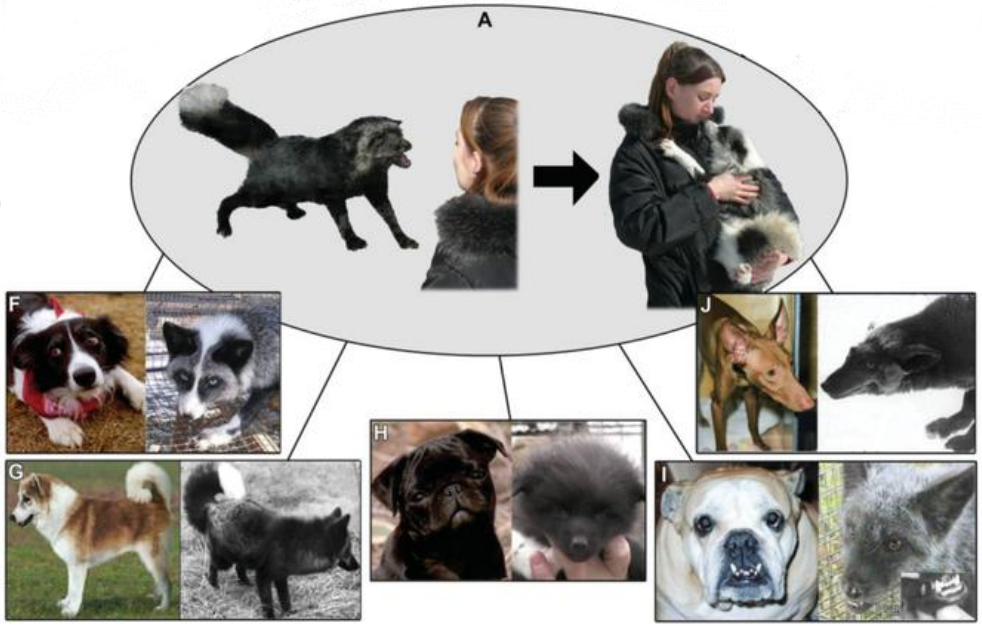
- **Numerous genes involved in the domestication syndrom**
- **Main modified traits in domestication syndrom**

Selection

- Docility
- Juvenile behavior

By-products phenotypes of domestication

- Depigmentation (white patches)
- Floppy and reduced ears
- Reduced muzzle and jaws
- Smaller teeth
- Shape of the tail (curly)
- More frequent oestrus

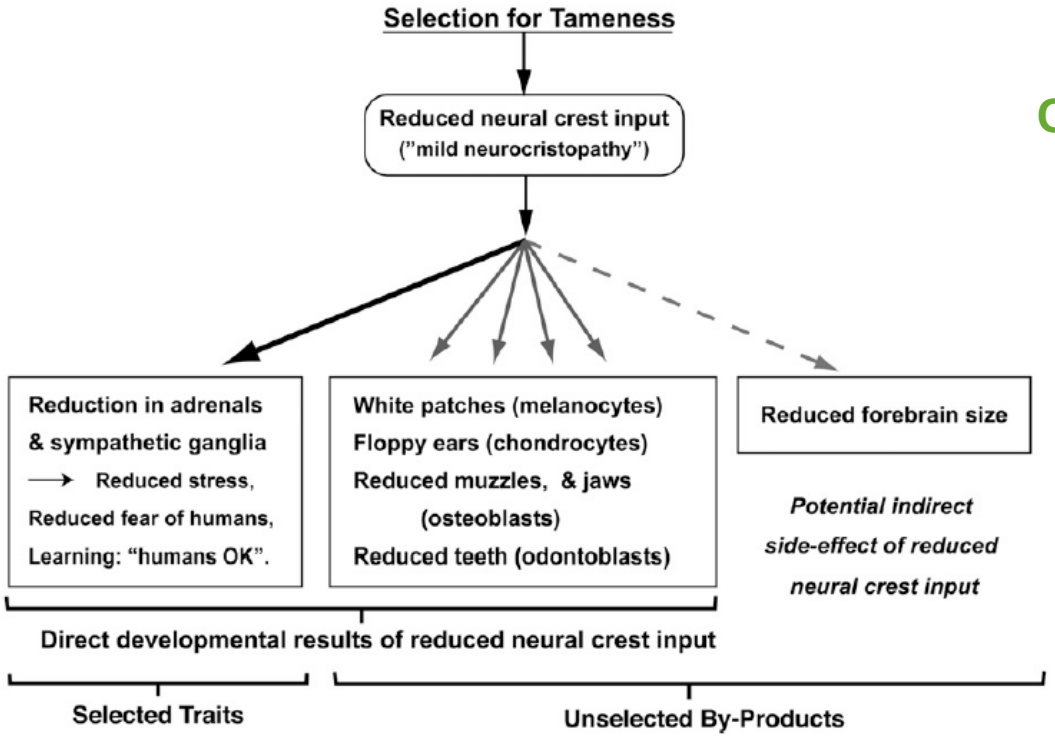


Belyaev and Trut , 1989; Trut et al, 2009

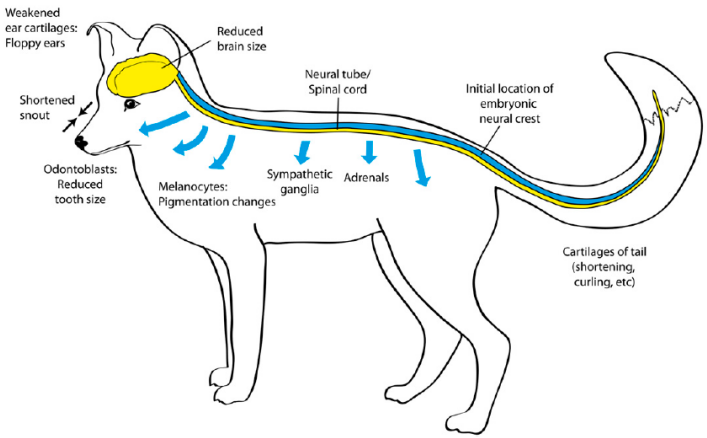
Annotation using systems biology tools

The example of the European bison

- **Our analysis give an empirical support to the unified explanation of the domestication syndrome in mammals proposed by Wilkins et al, 2014**



Central role of the neural crest



Course outline

1. **From regions under selection to biological interpretation**

   - Annotation of candidate genes using systems biology tools
     - Functional enrichment analysis: Gene ontology, pathway analysis
     - Gene networks analysis
   - Inferring the main selective pressures
   - Interpretation and story-telling
   - Including phenotypes and environmental covariates
   - Identifying candidate mutations and prioritizing candidate genes

2. **Interpretation of selection footprints: some examples**

   - Phenotype-free approaches
     - Manual functional annotation: Senepol cattle breed
     - Functional annotation using systems biology tools: French dairy cattle breeds, West-African cattle breeds, European bison/cattle
   - Association with covariates
     - Phenotypes: dairy traits in French cattle breeds

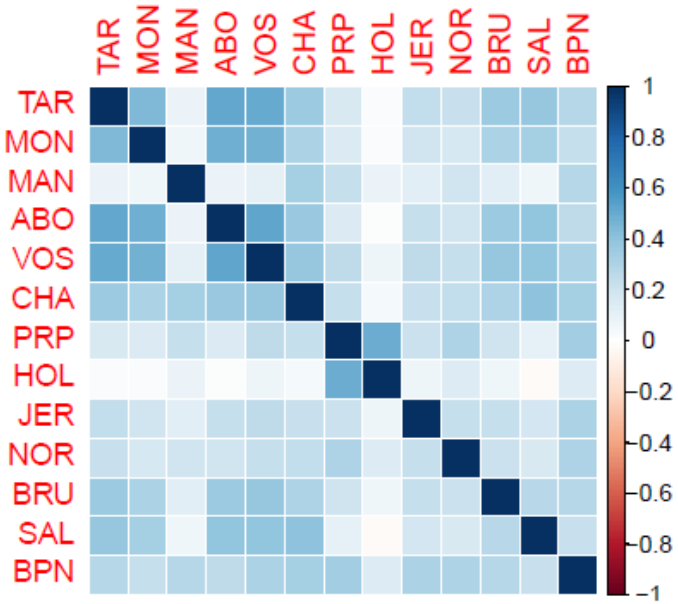
Association with covariates

The example of the French cattle breeds

Gautier and Flori, *in prep*

- **Identification of regions under selection associated with dairy traits**
  - Dairy traits: average milk production, lactation length, fat and protein content
  - 13 breeds
  - 50K SNP chip
  - Baypass (Gautier, 2015):
    - Auxiliary covariate model

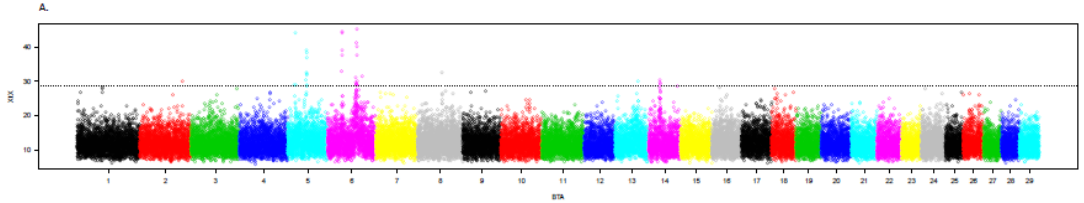
Correlation plot based on Ω



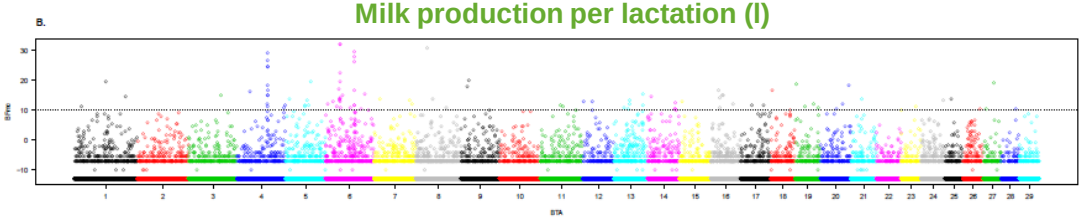
Association with covariates

The example of the French cattle breeds

- **Identification of regions under selection associated with dairy traits**
  - XtX> 28.6 (0.01% PODs)

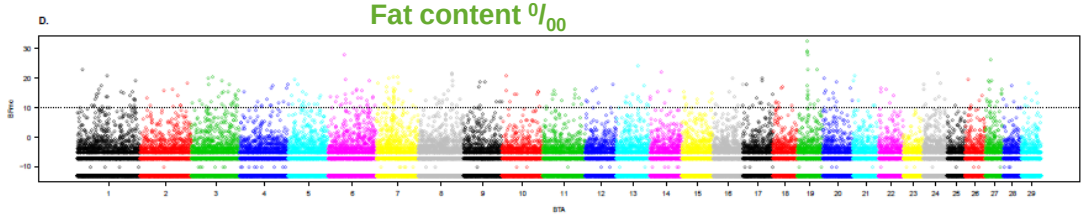


Milk production per lactation (l)

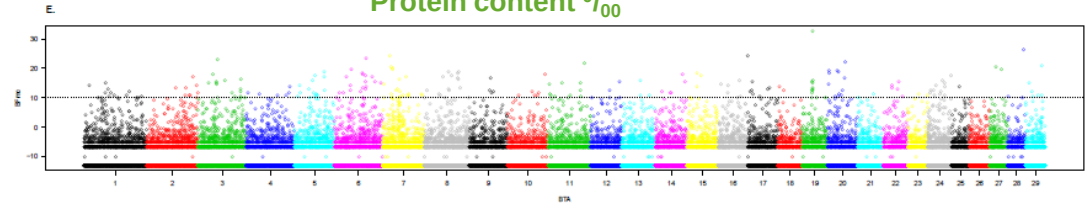


  - BFmc>10=>strong evidence
  - BFmc>20 =>decisive evidence

Fat content $^0/_{00}$



Protein content $^0/_{00}$



Association with covariates

The example of the French cattle breeds

- **Identification of regions under selection associated with dairy traits**

  - Using Refgene file from UMD3.1 assembly on ucsc (14627 Refseq ID)
  - Annotation of SNPs using the Refgene file: a SNP is representative of a gene if it is located within gene boundaries +/- 15kb

    =>94492 single ID

  - BFmc>20 =>decisive evidence
    - » Milk production: NUDC3, LAP3, LCORL
    - » Fat content: 24 RefSeq
    - » Protein content 5 Refseq

  - Choice of a less stringent criteria would be more informative (e.g. BFmc>10)

Conclusion

- **Functional and network analysis might be powerful to find a biological interpretation of footprints of selection**

- **Taking special care to avoid storytelling**

  - Criteria to define a candidate region and candidate genes (e.g. significant threshold, nb of significant SNP/gene or region, distance from peak)
    - » If too stringent criteria => loss of information
  - Integrating other information
    - » Sequencing data=>exhaustive list of variants that can be annotated and allelic frequencies
    - » QTL information
    - » Using association methods with population environmental covariates or phenotypes

Some references

- **Charitou et al, GSE, 2016.** Using biological network to integrate, visualize and analyse genomic data

- **Khatri et al, PLoS Computational biology, 2012.** Ten years of pathway analysis: current approaches and outstanding challenges

- **Pavlidis et al, MBE, 2012.** A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans.