Genome Annotation and BLAST searches

Marine Genomics

June 1st, 2021

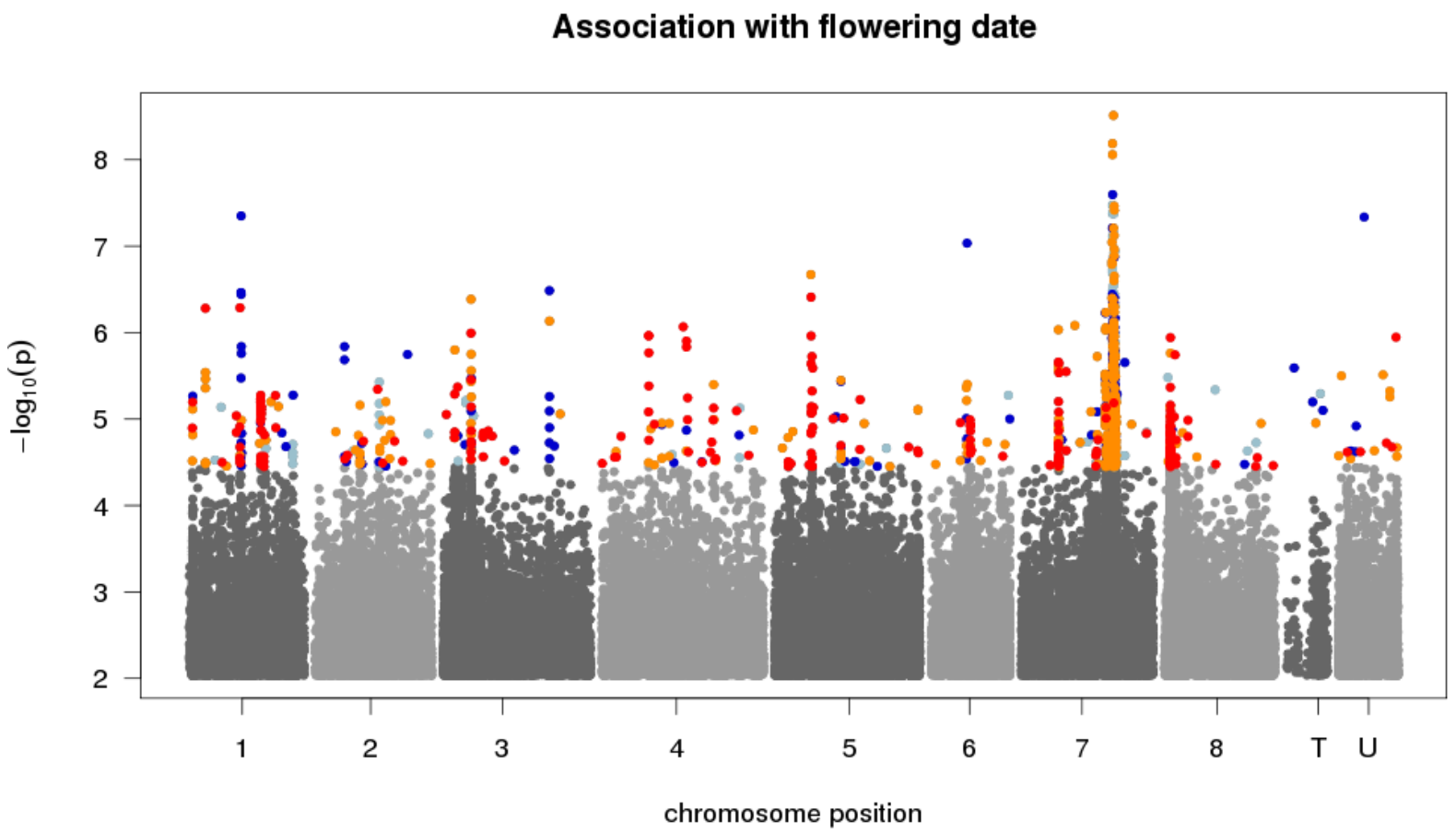
Fst outliers and GWAS

Find SNPs associated with:

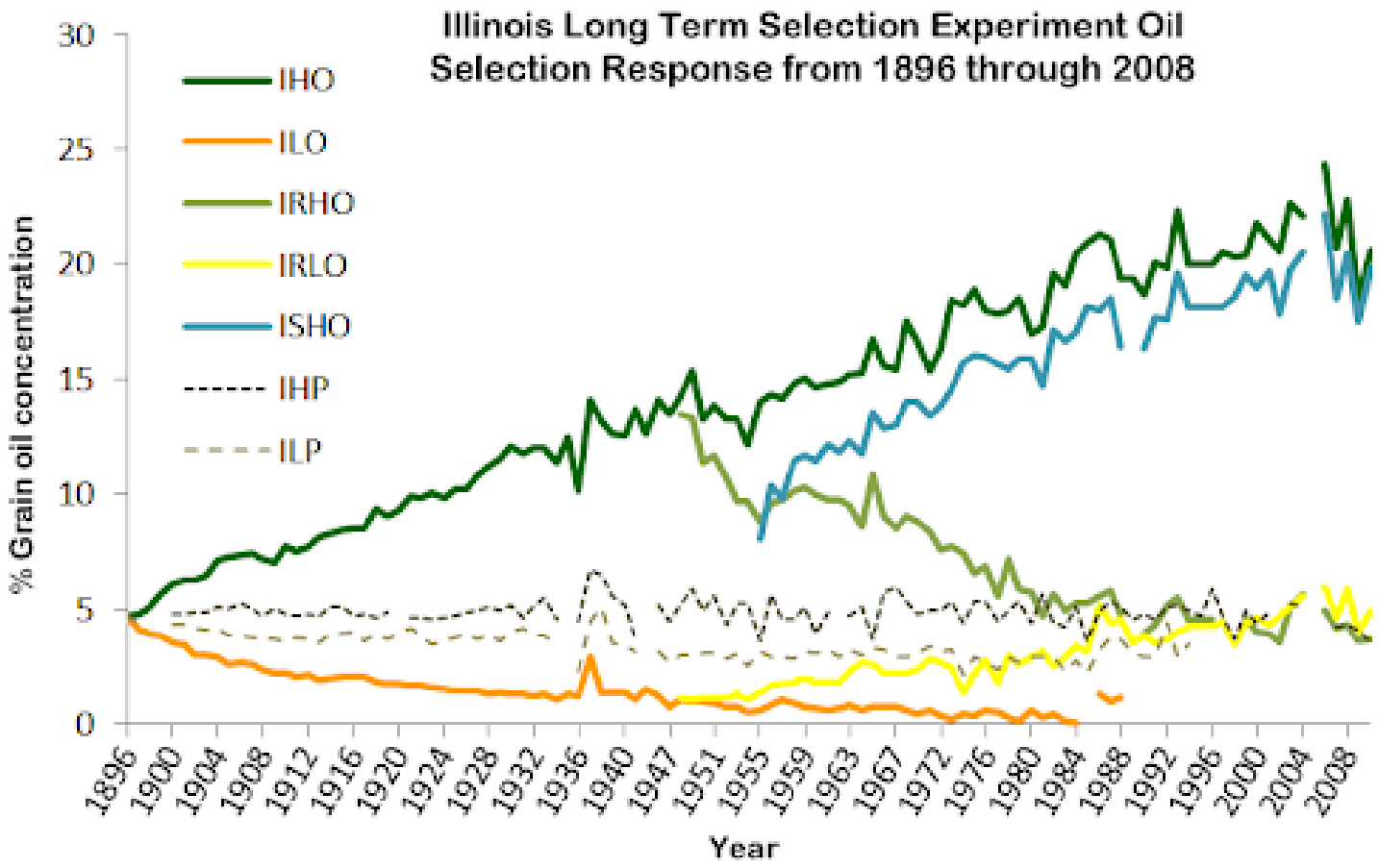
Population divergence (Fst outliers)

A phenotype (GWAS)

An Environment (GWEA)



Selection experiments



How to you find out what a “gene” does?

NEED A WELL ANNOTATED GENOME

AND GENE FUNCTIONALIZATION STUDIES

Genome annotation

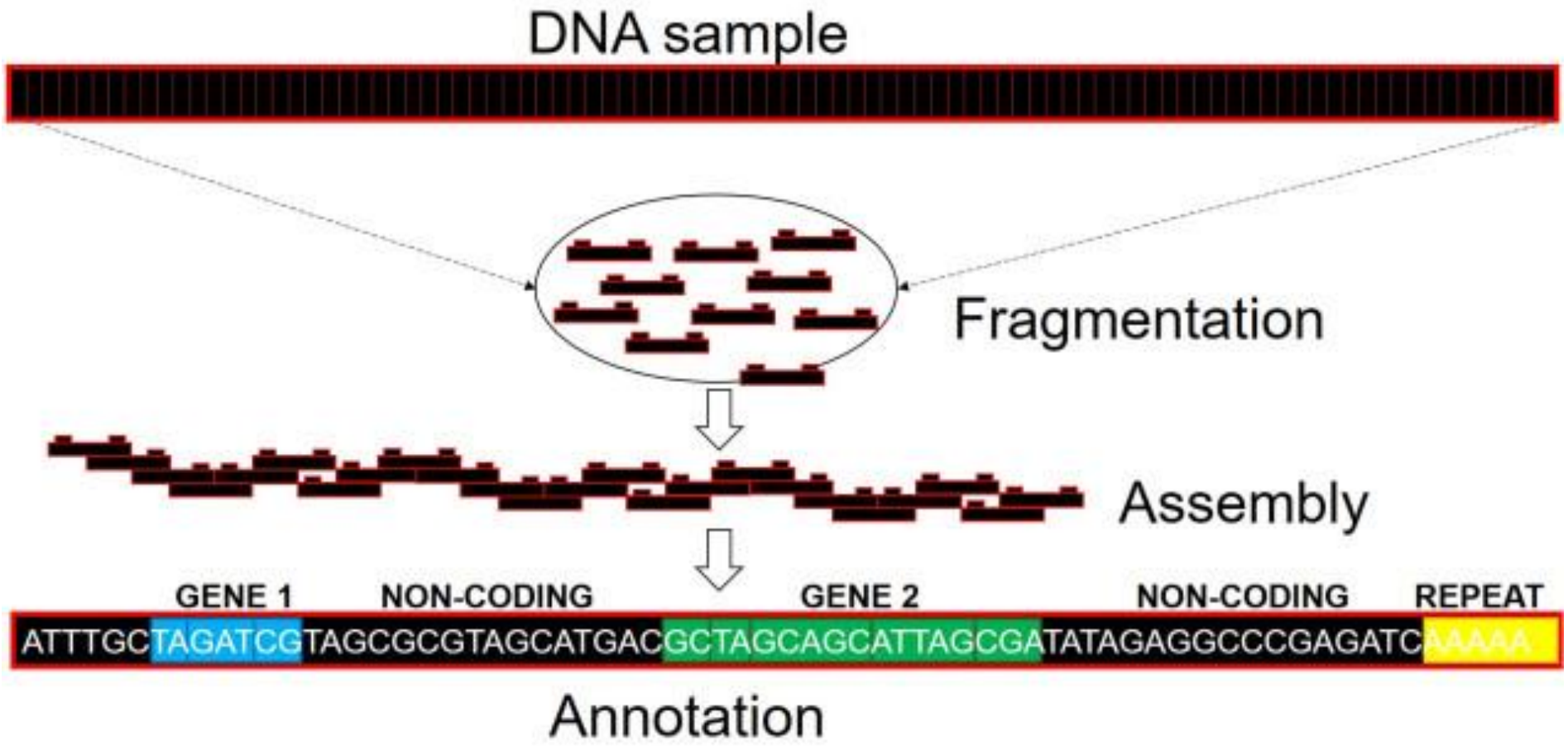
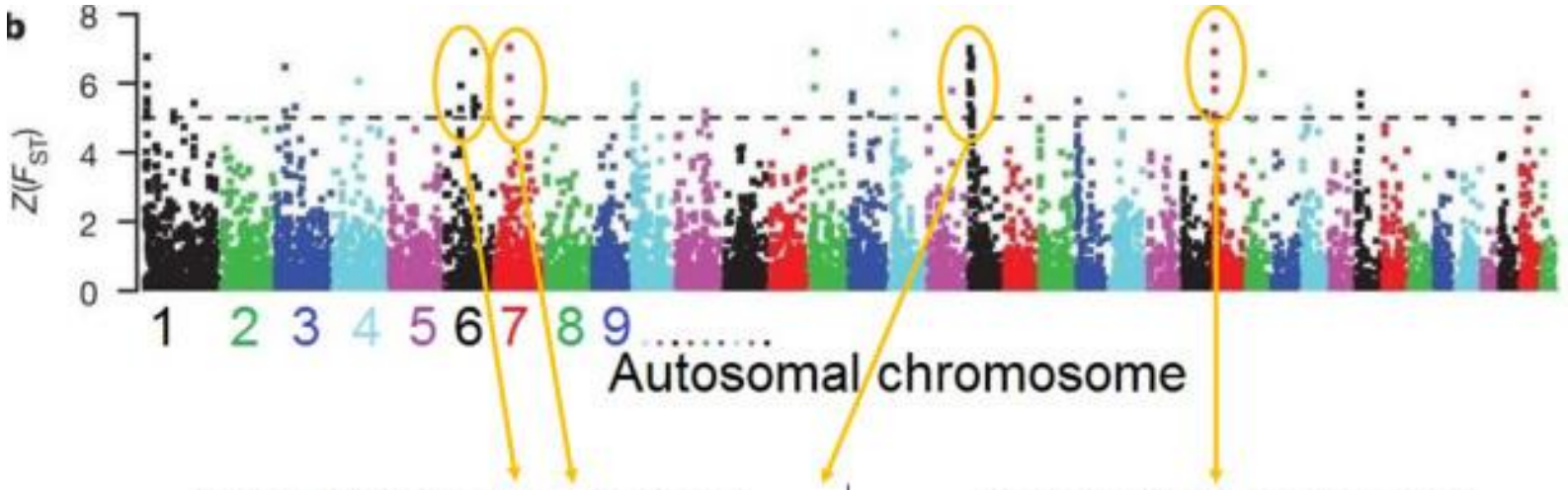


Image credit: https://theg-cat.com/tag/genome-annotation/

Annotation to find functional differences



Gene ontology category	Total number of genes
Nervous system development	89
Sperm-egg recognition	6
Regulation of molecular function	24
Digestion (esp. starch)	10
Other	3

Axelsson et al. 2013: https://www.nature.com/articles/nature11837 and https://theg-cat.com/tag/genome-annotation/

Gene functionalization is an entire field

A lot of papers use homology to infer or suggest function, which is fine, but doesn’t provide proof of what a gene does.

New tools to enable gene functionalization:

CRISPR-Cas9

Morpholinos

It is very easy to tell a story…..

A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans

Pavlos Pavlidis,*[1] Jeffrey D. Jensen,[2] Wolfgang Stephan,[3] and Alexandros Stamatakis[1]

[1]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany

[2]Ecole Polytechnique Fédérale de Lausanne, School of Life Sciences, Lausanne, Switzerland

[3]Section of Evolutionary Biology, Biocenter, University of Munich, Planegg-Martinsried, Germany

***Corresponding author:** E-mail: pavlidisp@gmail.com.

Associate editor: Arndt von Haeseler

Abstract

In the age of whole-genome population genetics, so-called genomic scan studies often conclude with a long list of putatively selected loci. These lists are then further scrutinized to annotate these regions by gene function, corresponding biological processes, expression levels, or gene networks. Such annotations are often used to assess and/or verify the validity of the genome scan and the statistical methods that have been used to perform the analyses. Furthermore, these results are frequently considered to validate "true-positives" if the identified regions make biological sense a posteriori. Here, we show that this approach can be potentially misleading. By simulating neutral evolutionary histories, we demonstrate that it is possible not only to obtain an extremely high false-positive rate but also to make biological sense out of the false-positives and construct a sensible biological narrative. Results are compared with a recent polymorphism data set from *Drosophila melanogaster*.

Key words: genome scanning, positive selection, gene ontology, validation, literature mining.

Research artic

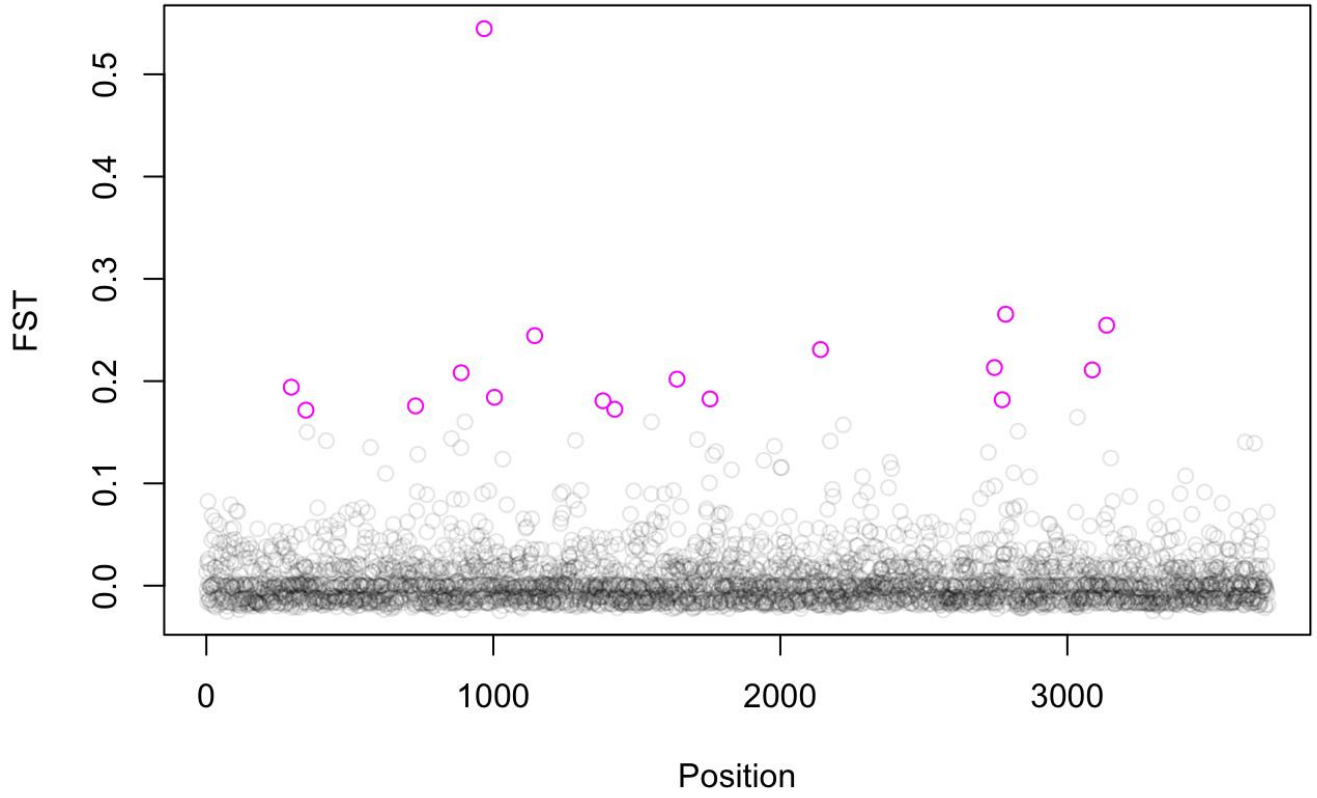
What we learn regardless

The number and effect size of SNPs associated with a trait is still very good information.

The evolution of a trait that is controlled by a few versus many genes is going to be different.

= genetic architecture

Our data for this week



Fst outliers

- Xuereb et al. 2018
- 17 outliers identified between the “north” and “south” populations

- Pulled 20kb region surrounding those SNPs from the reference genome (Parastichupus parvimensis)