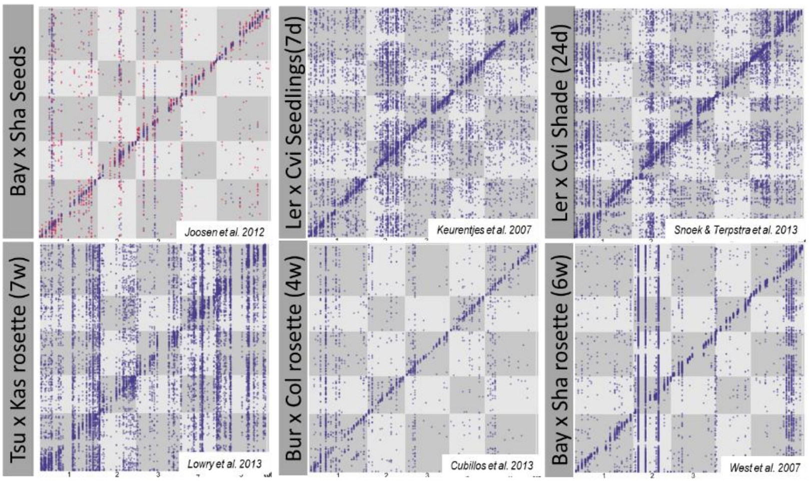
Thesis projects, Systems Genetics & Network Biology 2019-2020

Dr. Basten L. Snoek (l.b.snoek@uu.l)

Theoretical Biology and Bioinformatics

1) Multi population QTL mapping
2) System genetics of plant metabolism
3) Network visualisation of Microbial community dynamics
4) Dynamic gene expression networks
5) Transcription factor binding sites in co-expressed genes
6) Natural variation in expression dynamics underlying germination speed
7) Natural variation explained: Allele mining from GWAS populations
8) Machine learning models for eQTL mapping
9) Machine vision for QTL mapping
10) Gene function enrichment in microbial co-response clusters

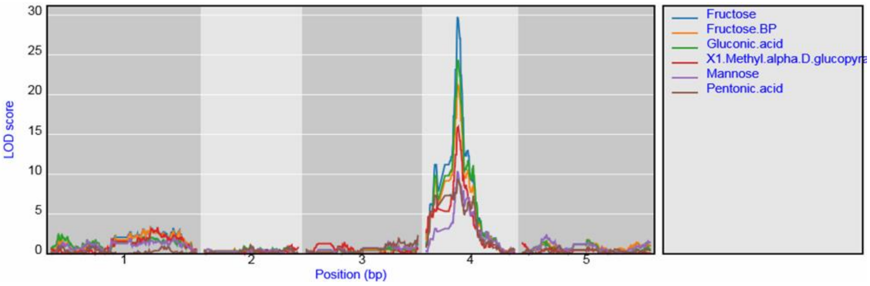
Multi population QTL mapping *(Arabidopsis, possibly C. elegans)*



One of the major goals of quantitative genetics is to unravel the relation between genetic variation and phenotypic variation. Recently, breakthroughs in -omic techniques have enabled advances in finding quantitative trait loci (QTL) for molecular phenotypes, such as gene expression levels (eQTLs). Through these eQTL studies polymorphic regulatory loci can be found. Moreover high quality genomes of individual members of a species are available, making the combined study of different eQTL studies feasible. The increased statistical power from the combined number of samples and increased number of polymorphisms promises to increase the resolution by which the regulatory genes and causal polymorphism can be identified. In this project on Arabidopsis we will combine the polymorphisms generated by the 1001 genomes consortium and the data on genome wide expression variation in recombinant inbred line populations as well as genome wide association data.

Starting skills: Basic R, Basic understanding of quantitative genetics.
Learned skills: eQTL mapping, GWAS, big data handling, big data visualization.

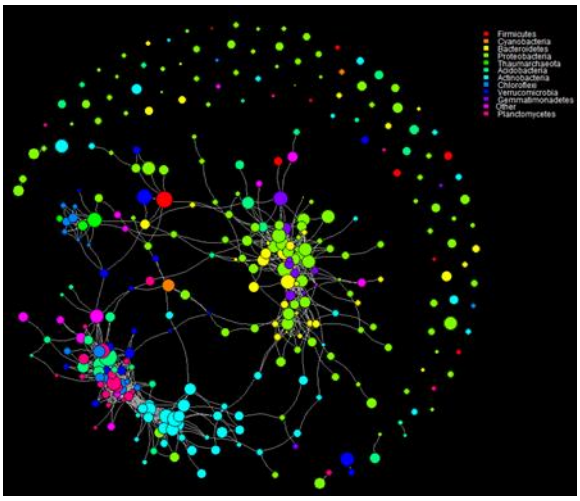
Systems genetics of plant metabolism *(Arabidopsis or Tomato)*



One of the major goals of quantitative genetics is to unravel the relation between genetic variation and phenotypic variation. Recently, breakthroughs in -omic techniques have enabled advances in finding quantitative trait loci (QTL) for molecular phenotypes, such as gene expression (eQTLs) and metabolite (mQTL) levels. Through these studies polymorphic regulatory loci for these molecular phenotypes can be found. Co-location of QTLs indicates a shared regulator between genes, metabolites and genes and metabolites. These can be used to generate a network of variation in molecular phenotype and phenotypic variation. In this project we'll combine variation in the transcriptome and metabolome in the dry Arabidopsis seed to find the shared regulatory loci and predict seedling performance.

Starting skills: Basic R, Basic understanding of quantitative genetics.
Learned skills: eQTL mapping, big data handling, big data visualization.

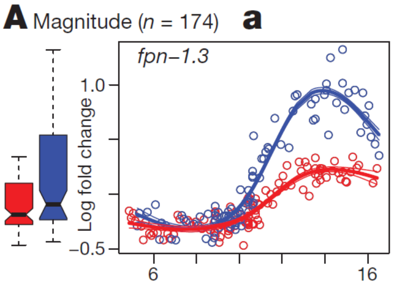
Network visualization of Microbial community dynamics



Network structure and visualizations are increasingly used to study microbial communities and their function. As communities and function are dynamic this needs to be integrated in these network studies. For example, we have data on the bacterial and fungal soil community before, during and after drought treatments. By generating networks within and between the different time points, we can uncover which community clusters are affected at what time. Moreover we can link this to the plants which have been growing on these soils and discover the finer details of the role of the microbial community throughout the response to drought.

Starting skills: Basic R
Learned skills: Cytoscape or iGraph, custom network generation, big data handling, big data visualization.

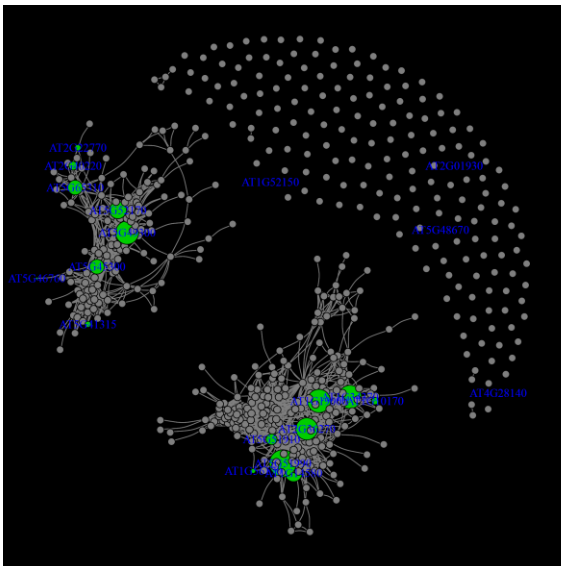
Dynamic gene expression networks *(C. elegans)*



One of the major goals of quantitative genetics is to unravel the relation between genetic variation and phenotypic variation. Recently, breakthroughs in -omic techniques have enabled advances in finding quantitative trait loci (QTL) for molecular phenotypes, such as gene expression (eQTLs) levels. Through these eQTL studies polymorphic regulatory loci can be found. However these studies do not include the dynamic nature of gene expression into account. Now we have highly detailed time series on the heat stress response of *C. elegans*. This data can be used to determine the speed at which genetically different individuals react to heat stress. This response information can be included in the QTL mapping model and used to identify the regulatory loci involved in the dynamics of gene expression.

Starting skills: Basic R, Basic understanding of quantitative genetics.
Learned skills: eQTL mapping, random forest, big data handling, big data visualization.

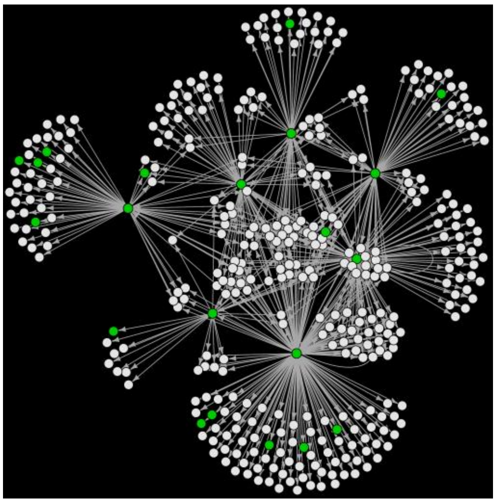
Transcription factor binding sites in co-expressed genes *(Arabidopsis / C.elegans)*



Co-expressed genes are likely to share a regulator. More specifically genes that share an expression quantitative trait locus (eQTL) are likely to share a regulator physically present on the QTL. This could be a transcription factor or other dna or rna binding element. To investigate if transcription factors leave a recognisable motif in genes affected by the same trans-regulatory locus, we use MEME or other motif search software to identify those motifs. These motifs will then be used to find the link with the candidate regulator. Moreover the recently generated data on the binding site of many specific transcription factors enables a direct comparison or enrichment to elude their regulatory role in eQTL hotspots.

Starting skills: Basic R, Basic understanding of quantitative genetics.
Learned skills: eQTL mapping, enrichments, network generation, big data handling, big data visualization.

Natural variation in expression dynamics underlying germination speed *(Arabidopsis)*

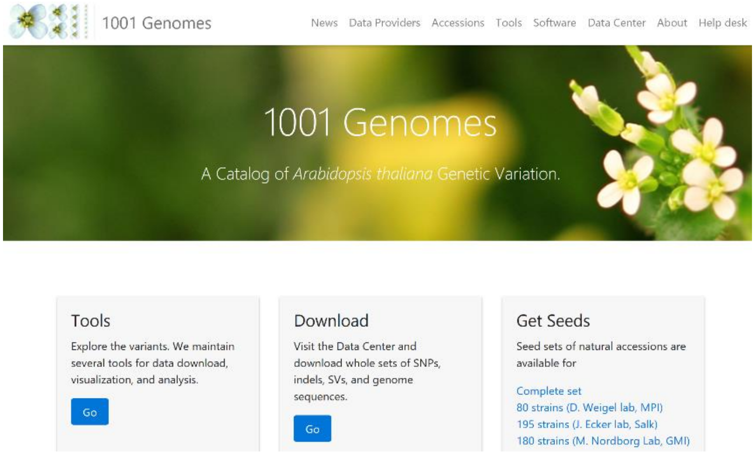


One of the major goals of quantitative genetics is to unravel the relation between genetic variation and phenotypic variation. Recently, breakthroughs in -omic techniques have enabled advances in finding quantitative trait loci (QTL) for molecular phenotypes, such as gene expression (eQTLs) levels. Through these eQTL studies polymorphic regulatory loci can be found. However these studies do not include the dynamic nature of gene expression into account. Now we have highly detailed time series on the germination process of Arabidopsis. This data can be used to determine the germination speed of genetically different individuals and learn about their development. This developmental information can be included in the QTL mapping model and used to identify the regulatory loci involved in the dynamics of gene expression.

Starting skills: Basic R, Basic understanding of quantitative genetics.
Learned skills: eQTL mapping, big data handling, big data visualization.

Natural variation explained: Allele mining from GWAS populations (Arabidopsis)

The increased availability of individual specific genome wide genetic differences in Arabidopsis has enabled the mining of functionally different alleles and regulatory motifs from sequence data. By combining phenotypic differences found after mutations in specific genes and natural occurring allelic differences a list with candidate alleles can be made. For these candidates the phenotypic difference can be predicted and tested by a complementation test. In this way part of the mechanism underlying the link between the genotype and phenotype can be uncovered.



1001 Genomes

A Catalog of *Arabidopsis thaliana* Genetic Variation.

Tools

Explore the variants. We maintain several tools for data download, visualization, and analysis.

Download

Visit the Data Center and download whole sets of SNPs, indels, SVs, and genome sequences.

Get Seeds

Seed sets of natural accessions are available for

Complete set
80 strains (D. Weigel lab, MPI)
195 strains (J. Ecker lab, Salk)
180 strains (M. Nordborg Lab, GMI)

Starting skills: Basic R, Basic understanding of quantitative genetics.
Learned skills: Sequence comparison, big data handling, big data visualization.

Machine learning models for eQTL mapping *(Arabidopsis, Tomato, C. elegans)*

One of the major goals of quantitative genetics is to unravel the relation between genetic variation and phenotypic variation. Recently, breakthroughs in -omic techniques have enabled advances in finding quantitative trait loci (QTL) for molecular phenotypes, such as gene expression (eQTLs) levels. Through these eQTL studies polymorphic regulatory loci can be found. A standard linear model is most frequently used to identify the QTLs. However not all genetic variation results linear behaving phenotypic variation. Here we propose to use machine learning models, such as random forest to identify eQTLs and compare the performance to the linear models used normally.

Starting skills: Basic R, Basic understanding of quantitative genetics.
Learned skills: eQTL mapping, machine learning, big data handling, big data visualization.

Machine vision for QTL mapping *(Tomato)*

High-throughput phenotyping and image analysis have a great potential in genetics and plant breeding, due to increased precision of the measurement or by detection of new previously overlooked phenotypes. Here we have set of images and 3d data of a RIL population of Tomato seedlings. The genotypes of this population are known enabling QTL mapping and detection of regulatory loci for the scored traits. In this project we want to automatically score phenotypes from hundreds of images and link this with the genetic variation.

Starting skills: Basic Python and R, Basic understanding of quantitative genetics.

Learned skills: QTL mapping, image analysis, big data handling, big data visualization.

Gene function enrichment in microbial co-response clusters

Many microbes co-respond to biotic and abiotic variation in the environment and therefor form co-response clusters. These co-response clusters could interact and possibly determine the future dynamics of the microbial community or even the community as a whole, including plants and animals. By searching for enriched gene functions in microbial co-response clusters we could predict future dynamics or explain the overserved changes during the co-response.

Starting skills: Basic Python and R, Basic understanding of genomics.
Learned skills: Network generation, clustering, enrichment. big data handling, big data visualization.