Scientific Focus Area 1

Genetic Architecture of Complex Traits

Peter Sørensen

Center for Quantitative Genetics and Genomics (QGG)

Data

From sequence ... *...to consequence*

- **Whole-genome sequences** and **multiple novel trait phenotypes** from large numbers of individuals from **multiple populations**

Enable us to **better understand genetic architecture** of complex traits

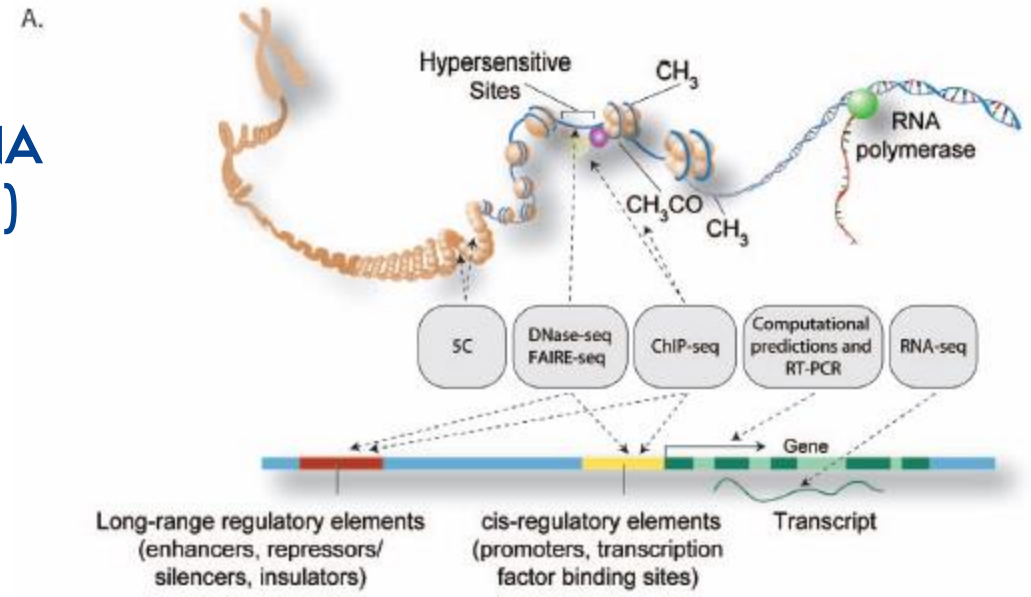
- **disentangle genetic variation**
- **disentangle genetic correlation**

Two important parameters for **prediction of trait phenotypes** and **consequences of selection decisions** in breeding programs

More data

From sequence … *…to consequence*

Encyclopedia of DNA Elements (ENCODE)



- **Molecular phenotypes** (transcriptome, proteome or metabolome) associated to the traits/diseases of interest

- **Molecular-interaction maps** that provide insight into the structural and functional organization of the genomes

What to do?

From sequence ... *...to consequence*

- How do we translate the massive collection of data at different levels of biology into useful biological knowledge?

- Can we use these different layers of data to improve predictive models of complex traits and diseases?

- Which statistical modeling approaches should we use?

Outline

From sequence ... *...to consequence*

1. Genomic feature model analyses of complex traits
2. Genomic Heritability: What Is It?
3. Using whole-genome sequence data to study genetic control of environmental variance in inbred lines of Drosophila Melanogaster
4. Using biological pathways for decomposing genomic variance for complex traits in dairy cattle and Drosophila Melanogaster
5. Genomic feature model analyses of growth trait in pure breed Duroc pigs
6. Current plans

Genomic Feature Model Analysis of Complex Traits

Peter Sørensen

Center for Quantitative Genetics and Genomics (QGG)

Department of Molecular Biology and Genetics
Aarhus University
Denmark

Genomic feature models

From sequence ... *...to consequence*

- **Statistical modeling approach** that evaluates the collective action of sets of genetic variants defined by a genomic feature: "**...model a feature of the genome**"

- **Genomic features are defined by grouping** genetic variants according to a certain classification scheme such as:

  - Chromosomes/Genes
  - Biological Pathways
  - Gene or Sequence Ontologies
  - Prior QTL regions
  - Expression or Methylation patterns
  - Protein-Protein or Protein-Metabolite interactions .....

Genomic feature models

From sequence ... *...to consequence*

- **Estimate genetic parameters** (e.g. enrichment, variance or correlation) associated with each feature
  - Bayesian approaches
  - GBLUP approaches
  - many others.....

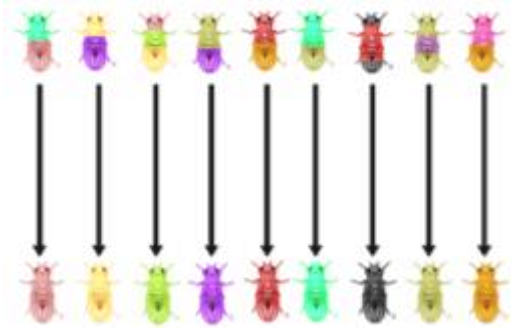
- **Novel insight into the genetic architecture of complex traits** by identifying genomic features that:

  - are enriched for associated variants
  - explain high proportions of genetic (co)variances
  - provide better model fits
  - better predict phenotypes

DGRP

From sequence ... *...to consequence*

- **205 inbred lines** from the DGRP population derived from 20 generations of full sib mating



- **Whole Genome Sequence data**
  ~ 1.8M SNPs
  ~ 20 SNP pr Kb

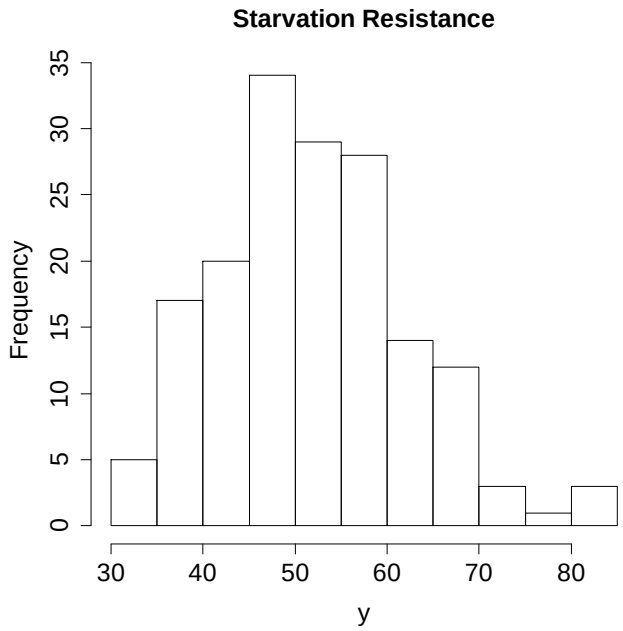
dgrp.gnets.ncsu.edu/data

- A **range of complex trait phenotypes** (e.g. starvation resistance, startle response, chill coma recovery)
- **Access to a wealth of annotation data** that can be used to link the SNPs to different types of genomic features (e.g. genes, pathways, QTL regions, MAFs, gene and sequence ontologies, and so on)

Phenotypes

From sequence ... *...to consequence*

Starvation Resistance: a measure of how long time (hours) it takes before a fly dies due to food deprivation



Precise phenotyping:

n = 19,361 phenotypic observations

- males and females
- 50 observation pr line/sex
- 10 flies/replicate/sex

One Component GBLUP

From sequence ... *...to consequence*

Step 1: Fit a single one component linear mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$$

where $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ is the genomic values based on all genetic markers (e.g. $\mathbf{G} = (\mathbf{WW}')/m$)

$h^2 = 0.39$ (narrow sense)

$r_{m,f} = 0.67$ (males – females)

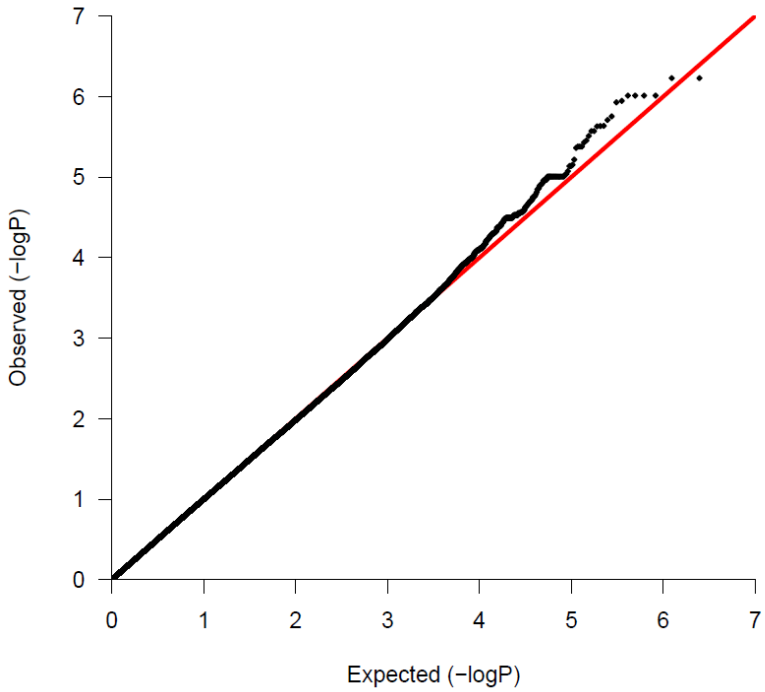
One Component GBLUP

From sequence ... *...to consequence*

Step 2: Backsolve to get single markers effects and test statistics:

$$\hat{\mathbf{s}} = \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\hat{\mathbf{g}}$$

$$\mathbf{t}_{\hat{s}_i} = \frac{\hat{s}_i^2}{\mathbf{Var}(\hat{s}_i)}$$



- many weak signals

Are the associated variants randomly spread over the genome?

Genomic Features

From sequence ... *...to consequence*

Genomic features defined by grouping genetic variants according to **genome-wide expression patterns**

Expression data

- A subset (38 lines) of the DRGP population
- RNA from a pool of 3- to 5-d-old flies (25 flies/sex/line)
- Affymetrix Drosophila 2.0

(Ayroles et a. 2009)

Association analyses

- starvation resistance line means
- expression levels of 10000 genes
- Simple linear regression

Rank gene according to the degree of association between its expression levels and the phenotype (starvation resistance)

Summary Statistic

From sequence ... *...to consequence*

Step 3: For each genomic feature compute a **summary statistic** that **quantify the degree of association**:

$$\mathbf{T_{count}} = \sum_{i=1}^{m_F} I(t_i > t_0)$$

$$\mathbf{T_{sum}} = \sum_{i=1}^{m_F} t_i$$

$$\mathbf{T_{cvs}} = \hat{\mathbf{g}}'\hat{\mathbf{g}}_\mathbf{f} \qquad \hat{\mathbf{g}}_\mathbf{f} = \sum_{i=1}^{m_F} \mathbf{W}_i\hat{S}_i$$

Null Hypothesis

From sequence ... *...to consequence*

Self-contained: Genetic variants linked to genomic feature are not associated to the trait phenotype:

$$\sigma^2_{g_f} = 0$$

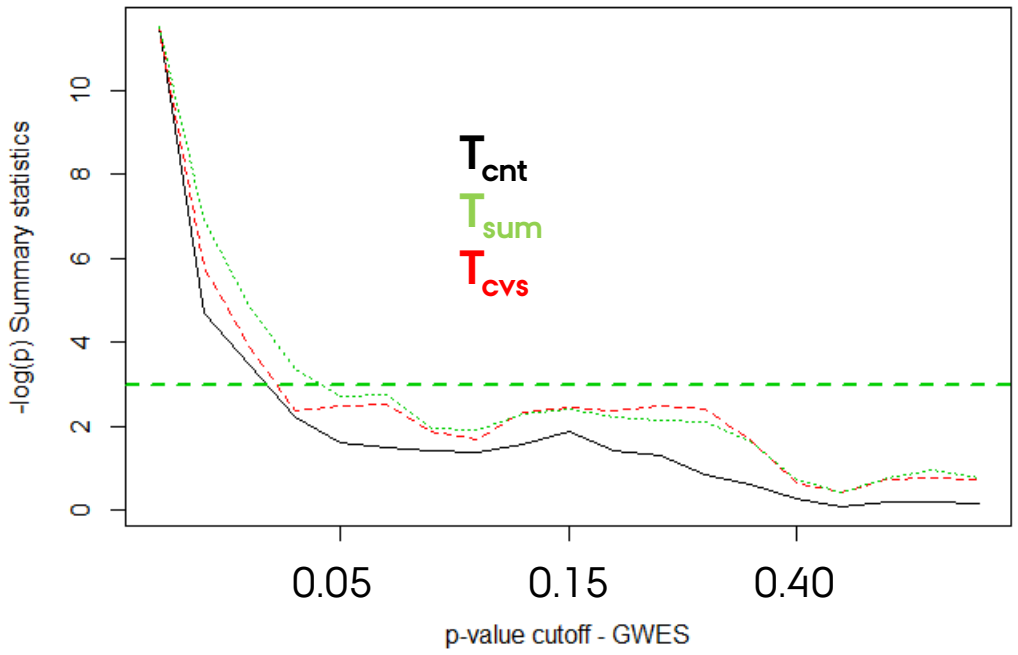
Competitive: Genetic variants associated to the trait phenotype are randomly and evenly distributed over the genome:

$$\sigma^2_{g_f} = \frac{\sigma^2_g}{N} N_f$$

Goeman et al. 2006

Associated Genomic Features

From sequence ... *...to consequence*



Similar patterns for the 3 summary statistics

Two Component GBLUP

Step 1: Fit a two component linear mixed model for each genomic feature:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg_f} + \mathbf{Zg_{nf}} + \mathbf{e}$$

where $\mathbf{g_f}$ is the **genomic values for the feature of interest** and $\mathbf{g_{nf}}$ is the genomic values based on the remaining set of genetic markers.

$$\mathbf{g_f} \sim N(0, \mathbf{G_f}\sigma^2_{g_f}) \quad \mathbf{G_f} = (\mathbf{W_f}\mathbf{W_f}')/m_f$$
$$\mathbf{g_{nf}} \sim N(0, \mathbf{G_{nf}}\sigma^2_{g_{nf}}) \quad \mathbf{G_{nf}} = (\mathbf{W_{nf}}\mathbf{W_{nf}}')/m_{nf}$$

Predictive Ability

From sequence ... *...to consequence*

Step 3: For each genomic feature compute a **summary statistic** that quantify the **predictive ability**.

Cross validation scheme: 80% training - 20% validation

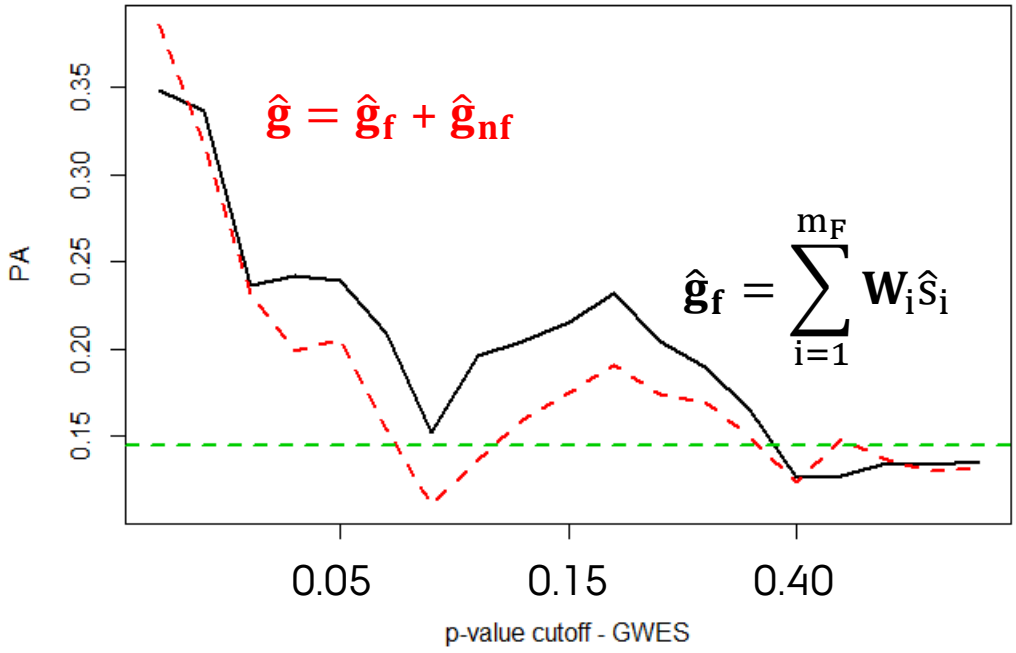
Compute predictive ability **(PA)** of the model as the average of the correlations across all validation subsets.

$$\mathbf{PA_2} = \sum_{j=1}^{n_{cv}} \mathrm{corr}(\hat{\mathbf{g}}_{\mathbf{obs}}, \hat{\mathbf{g}}) \qquad \hat{\mathbf{g}} = \hat{\mathbf{g}}_{\mathbf{f}} + \hat{\mathbf{g}}_{\mathbf{nf}}$$

$$\mathbf{PA_1} = \sum_{j=1}^{n_{cv}} \mathrm{corr}(\hat{\mathbf{g}}_{\mathbf{obs}}, \hat{\mathbf{g}}_{\mathbf{f}}) \qquad \hat{\mathbf{g}}_{\mathbf{f}} = \sum_{i=1}^{m_F} \mathbf{W}_i \hat{s}_i$$

Predictive Genomic Features

From sequence ... *...to consequence*



Conclusion

From sequence ... *...to consequence*

Genomic feature model analysis can be used to reveal biological relevant classification schemes that:

- explain higher proportions of genomic variances
- provide better model fit
- increase predictive ability of the statistical model

Implemented using standard linear mixed modeling approach:

- more comparisons to other alternatives is needed
- “best” may depend on the goal of the analyses
- influence of family structure and type of feature
- many classification schemes => fast method

Many more studies needed to determine which genomic feature classification scheme are the “best”

Outline

From sequence ... *...to consequence*

1. Genomic feature model analyses of complex traits
2. Genomic Heritability: What Is It?
3. Using whole-genome sequence data to study genetic control of environmental variance in inbred lines of Drosophila Melanogaster
4. Using biological pathways for decomposing genomic variance for complex traits in dairy cattle and Drosophila Melanogaster
5. Genomic feature model analyses of growth trait in pure breed Duroc pigs
6. Current plans

Current plans

From sequence ... *...to consequence*

- Comparison of alternative GFM modeling approaches
- Multiple trait and multiple genomic feature models
- GFM for disease traits

- Additive and non-additive multi-trait genomic models of micro-environmental sensitivity (**Fabio Morgante**, **Trudy Mackay**)

- GFM for production/disease traits in dairy cattle (**LingZhao Fang**)
- GFM for growth traits in pigs (**Pernille Sarup**)
- GFM for schizophrenia in humans (**Palle Duun Rohde**)

- GFM for gene by environment in ryegrass (**PD**, **Torben Asp**)
- Methods for mapping rare variants in dairy cattle (**Qianqian Zhang**, **Goutam Sahana**)
- Contribution of structural variants to genomic variance (**PhD**)
- GFM using microRNA and long non-coding RNA (**PD**, **Christian Bendixen**)

Acknowledgements

From sequence ... *...to consequence*

- **Aarhus University**
  - Stefan M. Høj-Edwards
  - Palle Jensen
  - Pernille Sarup
  - Daniel Sorensen
  - Per Madsen

- **This work was financially supported by the following research projects:**
  - GenSAP (DK-Strategic Research Council)
  - Quantomics