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

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

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

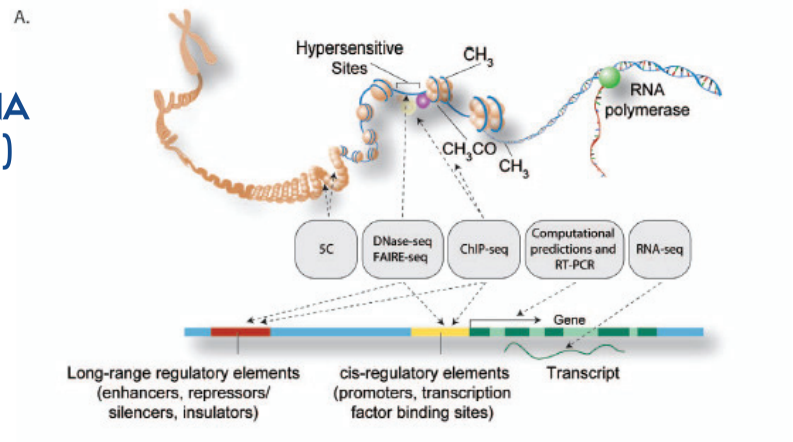
- **disentangle genetic variation**
- **disentangle genetic correlation**

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 their 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?

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
  - Transcriptional Active Genomic Regions
  - Protein-Protein interactions …….

Genomic Feature Models

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

Two genomic feature modeling approaches:

1. one component GBLUP approach (**GBLUP1**)
2. two component GBLUP approach (**GBLUP2**)

- **Test procedures and test statistics used for** identifying genomic features of interest

One Component GBLUP

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

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

$$\mathbf{M_1}: \quad \mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$$

where **g** is the genomic values based on all genetic markers.

Assumptions:

$\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ where $\mathbf{G} = (\mathbf{WW}')/N$

$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$

g and **e** are uncorrelated

One Component GBLUP

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

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

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

Step 3: For each genomic feature compute a summary statistic based on single marker tests within feature such as:

Sum of all single marker test statistics

$$\mathbf{T}_{\mathbf{sum}} = \sum_{i=1}^{n_F} t_i$$

Two Component GBLUP

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

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

$$M_2: \quad \mathbf{y} = \mathbf{Xb} + \mathbf{Zg_f} + \mathbf{Zg} + \mathbf{e}$$

where $\mathbf{g_f}$ is the **genomic values for the feature of interest** and $\mathbf{g}$ is the genomic values based on all genetic markers.

Assumptions:

$\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$ $\quad \mathbf{G} = (\mathbf{WW}')/N$

$\mathbf{g_f} \sim N(0, \mathbf{G_f}\sigma_{g_f}^2)$ $\quad \mathbf{G_f} = (\mathbf{W_f W_f}')/N_f$

$\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$

$\mathbf{g_f}$, $\mathbf{g}$, and $\mathbf{e}$ are uncorrelated

Two Component GBLUP

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

Step 2: For each genomic feature test significance using:

Likelihood ratio test:

$$\mathbf{T_{LRT}} = -2[\mathrm{l}(\widehat{\boldsymbol{\theta}}_1|\mathrm{y}) - \mathrm{l}(\widehat{\boldsymbol{\theta}}_2|\mathrm{y})] \qquad (M_1 <-> M_2)$$

Score test:

$$\mathbf{T_{Score}} = 0.5(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_1)'\widehat{\mathbf{V}}_1^{-1}\mathbf{Z}\mathbf{G}_\mathbf{f}\mathbf{Z}'\widehat{\mathbf{V}}_1^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_1) \qquad (M_1)$$

Goeman et al. 2004

Two Component GBLUP

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

Cross validation procedure for each genomic feature:

- 80% training / 20% validation
- 20 randomly chosen data sets

Step 1: **Fit two component GBLUP** and compute the total genomic values, $\mathbf{g}_t = \mathbf{g}_f + \mathbf{g}$, for the individuals in validation data set

Step 2: **Compute correlation** between the total genomic values obtained from step 1 and from the analysis using all data

Step 3: **Compute predictive ability (PA)** of the model as the average of the correlations across all validation subsets.

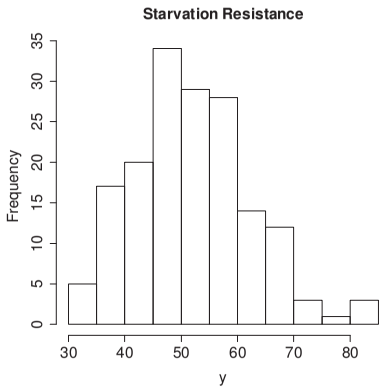
DGRP Data

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

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

- **Whole Genome Sequence data**
  - 2.5M SNPs
  - 20.89 SNP pr Kb

- **Starvation Resistance**: a measure of how long time it takes before a fly dies due to food deprivation



dgrp.gnets.ncsu.edu/data

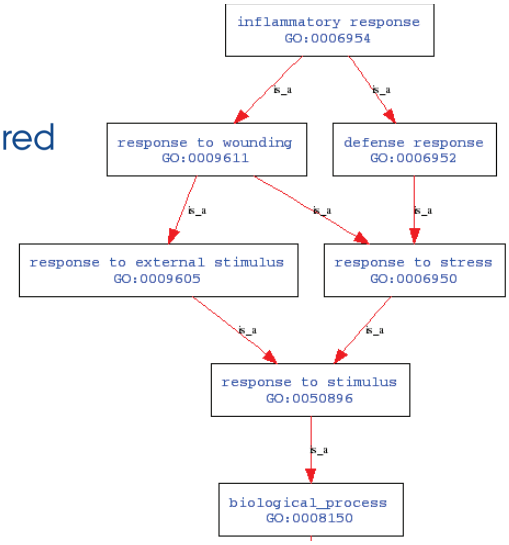
  - 17,324 phenotypic observations
  - on average 104 observation pr line (47-216)
  - high between line variation ($h^2 = 0.39$)

Genomic Features

From sequence … *…to consequence*

Genomic Features defined by **Gene Ontology (GO)**

Gene grouped according to biological processes such as mitosis or immune response, that are accomplished by ordered assemblies of molecular functions

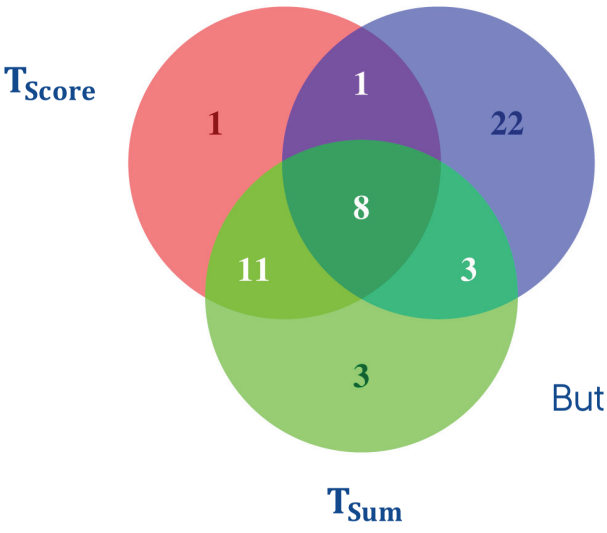


- map SNP to Gene to GO
- 5kb up-/down-stream of gene
- 2066 SNP sets (genomic features)

Results

From sequence … *…to consequence*

Large overlap among the significant genomic features based on summary statistics from **GBLUP1** and **GBLUP2**

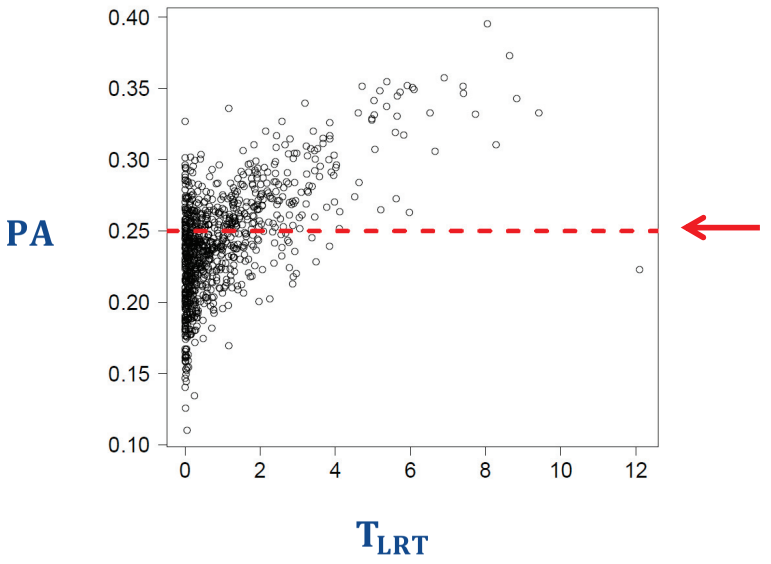


But it is still difficult to decide ….

Results

From sequence … *…to consequence*

Higher LRT is linked to higher PA: $y = Xb + Zg_f + Zg + e$



Predictive ability of one component linear mixed model

$y = Xb + Zg + e$

Results

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

Rank correlation between PA and –logP of summary statistics

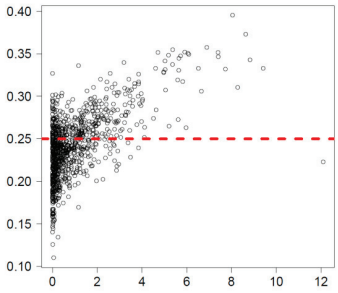
GBLUP1

T_{Sum}: 0.50

GBLUP2

T_{LRT}: 0.52

T_{Score}: 0.37



Results

From sequence … *…to consequence*

High ranking genomic features defined by gene ontologies

GO	PA	LR	H^2	N_f
GO:0007474	0.395	8.05	84.8%	29,018
GO:0042742	0.373	8.64	50.5%	6,225
GO:0007464	0.357	6.90	48.2%	14,263
GO:0048747	0.355	5.38	25.5%	5,256
GO:0048103	0.352	5.92	33.6%	2,468
GO:0007465	0.352	4.72	31.2%	5,541
GO:0045893	0.351	7.40	69.0%	24,874
GO:0007458	0.350	6.05	38.7%	3,739
GO:0008587	0.349	6.10	61.3%	16,124
GO:0010389	0.348	5.19	19.1%	682

explain a **larger proportion of genomic variance**, provide **better model fit** or **predictive ability** for starvation resistance in DRGP

Novel Insights?

From sequence … *…to consequence*

Biological interpretation is difficult … BUT high ranking genomic features:

- **are associated with genes involved in growth and development** (increased developmental time seems to be evolutionary important in populations that have elevated starvation resistance) (Rion and Kawecki, 2007)

- **are associated with genes involved in immune system** (immune response is costly and lines with a down regulated immune response might have more resources available for energy storage, and could thereby resist starvation for longer periods) (Lochmiller and Deerenberg, 2000)

- **are associated with genes whose expression levels are associated to starvation resistance** (Ayroles et al., 2009)

Conclusion

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

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

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

Conclusion

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

A number of genomic feature model approaches are available and can easily be implemented using standard linear mixed modeling methods:

- **One component GBLUP approach** is computational fast and the significant associated genomic features are similar to the ones obtained from two component GBLUP approach

- **Two component GBLUP approach** more computational demanding, but allow us to use the predictive ability associated to the genomic feature as a measure of importance

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