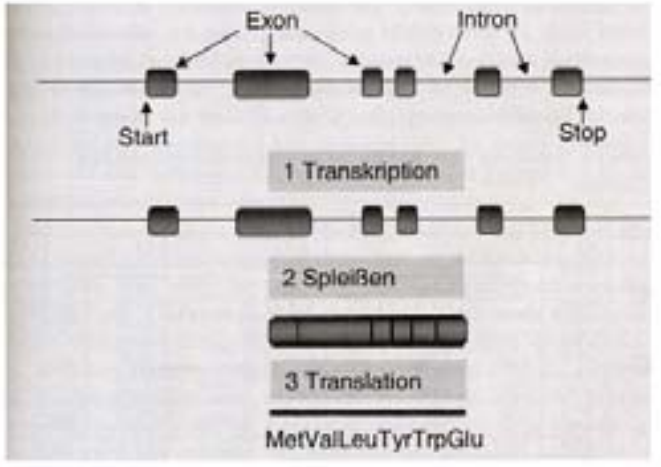
Detecting Signals in Genomewide Association Studies

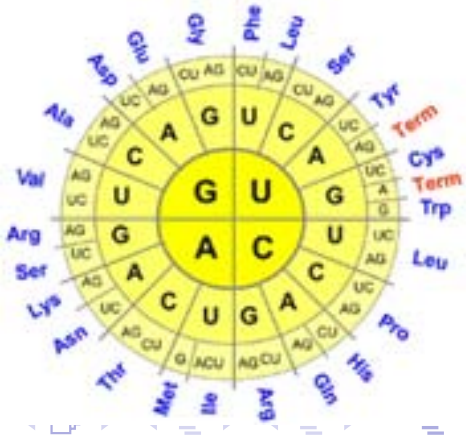
Hansjörg Baurecht

7th July 2010

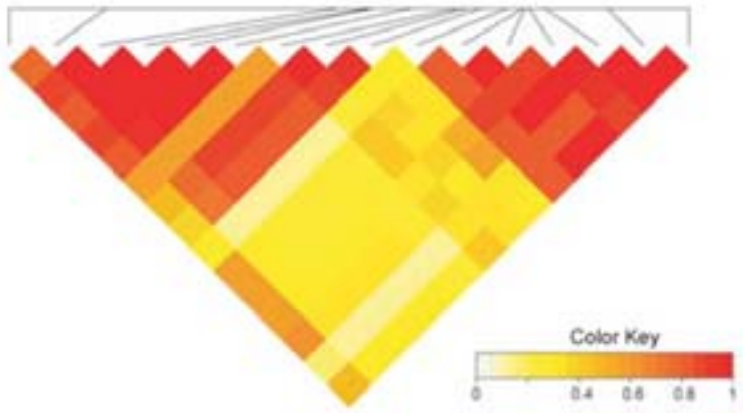
Introduction into Genetics



- basepairs: A-T, C-G
- single nucleotide polymorphism (SNP)
- gene sequence: promoter, exons, introns, terminator
- synonymous/ missense mutation
- regulatory sequences in introns may lead to alternate splicing
- deletion/ insertion lead to frameshift



LD-structure



- LD refers non-independent inheritance of alleles
- assumption: genome organized in blocks whith strong LD
- exploiting LD structer enables high coverage of genes
- association between SNP and phenotype might not indicate functional variant
- GWAS useful for further dissection of complex traits
- Affymetrix chip uses unbiased selection of polymorphisms

Multiple Test Problem

- Bonferroni correction
- assume const. LD, adjust effective number of tests (Zondervan,Cardon 2007)
- weaker local significance levels e.g. 10^{-6}, 10^{-5} (Arking et al. 2006)
- region interesting if 2 or more SNPs in modest/strong LD have nominal p-value below an even weaker threshold

Genomewide Association Studies

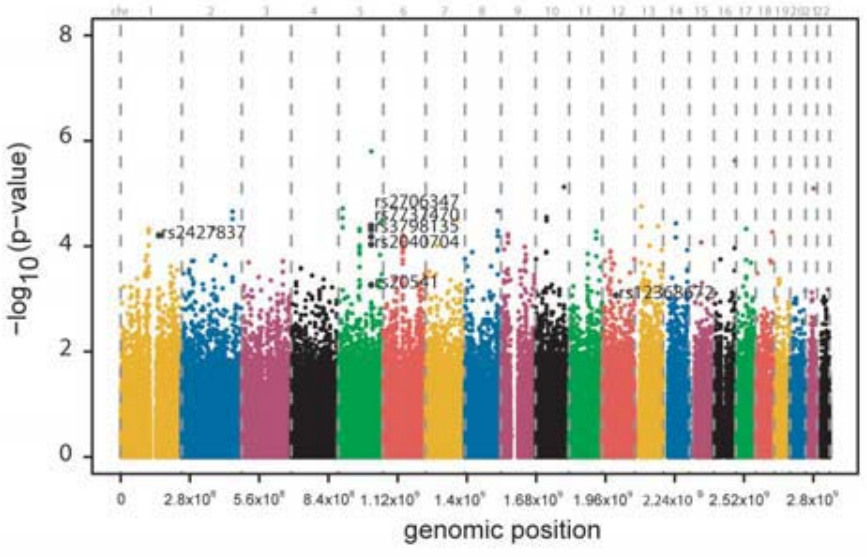
SNP1	
AA	$\bar{y}_{wt}$
AG	$\bar{y}_{het}$
GG	$\bar{y}_{hom}$

SNP2	
CC	$\bar{y}_{wt}$
GC	$\bar{y}_{het}$
GG	$\bar{y}_{hom}$

r^2_{01}

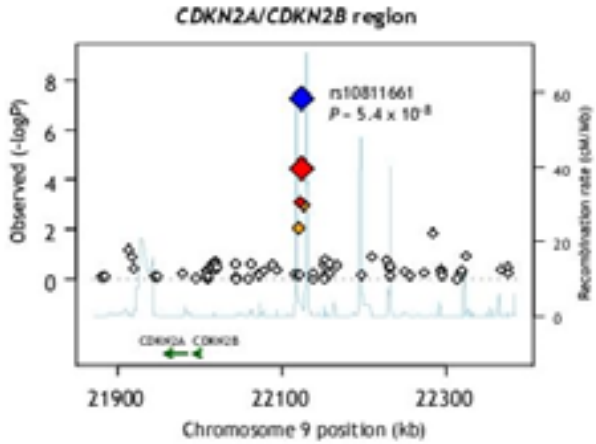
x_{01}

... 500 000 times



Inspriation

- clumping: group SNP-based results based on empirical LD estimates (PLINK v1.07) (2)
- de Bakker's regional plots (3)
- adopt idea of kernel weights

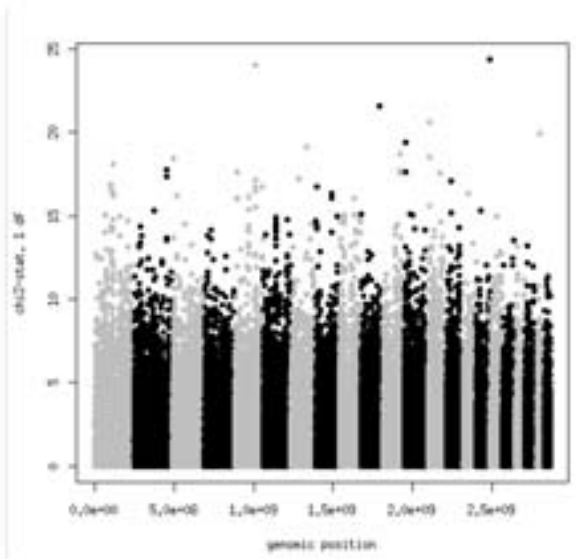


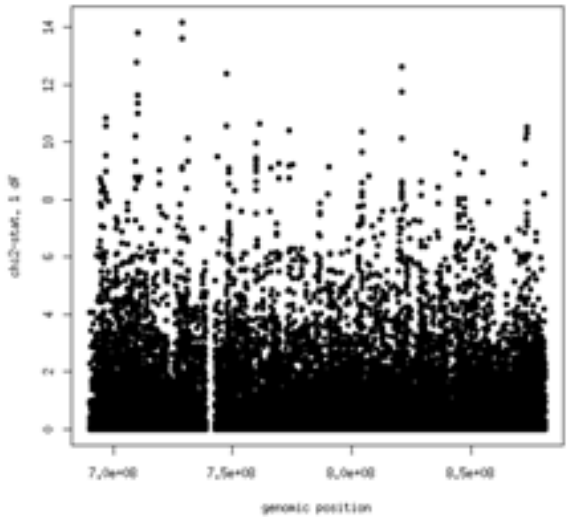
- given are univariate χ^2 statistics of p SNPs
- weighted composite statistic $\psi_0 = \frac{\sum_{j=1}^{p} K_\lambda(x_0,x_j) r_{0j}^2 \chi_j^2}{\sum_{j=1}^{p} K_\lambda(x_0,x_j) r_{0j}^2}$
- $K_\lambda(x_0, x_j) = D\left(\frac{|x-x_0|}{\lambda}\right)$, $\lambda = 100kb$, $x_j =$ gen. Pos. des SNPs

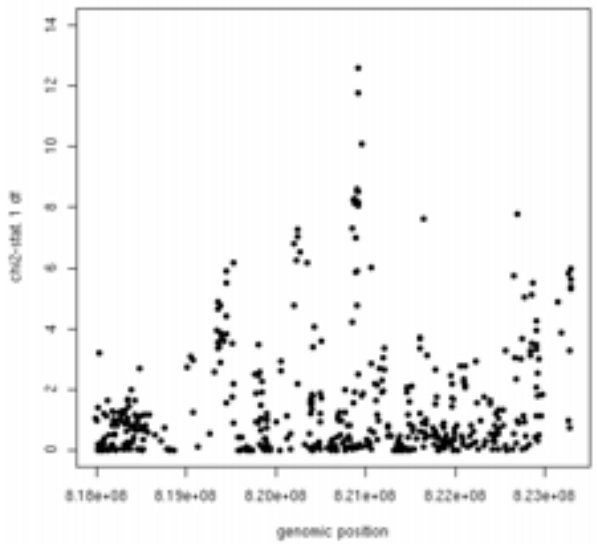
$$D(t) = \begin{cases} \frac{3}{4}(1-t^2) & \text{if } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

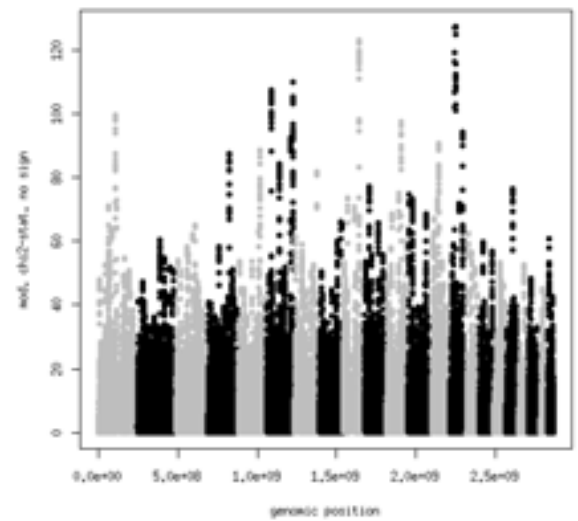
Estimation of Lambda

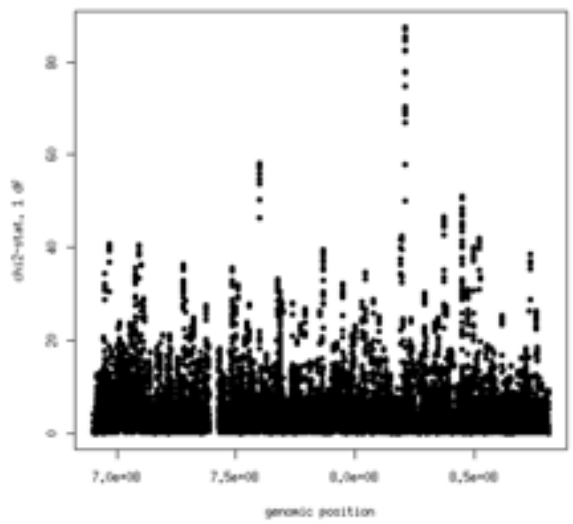
- estimate λ by 5-fold Crossvalidation with $\kappa : \{1, ..., N\} \rightarrow \{1, ..., \mathbb{K}\}$
- $CV(\hat{\lambda}) = \frac{1}{\mathbb{K}} \sum_{\kappa=1}^{\mathbb{K}} \left[\bar{y}^{\kappa} - \frac{\sum_{j=1}^{p} K_{\lambda}(x_0^{-\kappa}, x_j^{-\kappa}) r_{0j}^{2\,-\kappa} \mathbb{E}(y^{-\kappa})}{\sum_{j=1}^{p} K_{\lambda}(x_0^{-\kappa}, x_j^{-\kappa}) r_{0j}^{2\,-\kappa}} \right]^2 \rightarrow \min$
- calculate observed $\bar{y}^{\kappa}$ for each genotype per SNP
- estimate $\mathbb{E}(y^{-\kappa})$ by regression model for each genotype per SNP
- calculate weighted average of $\mathbb{E}(y^{-\kappa})$ of p SNPs within the range of λ
- plug in $\hat{\lambda}$ in formula of composite statistic ψ

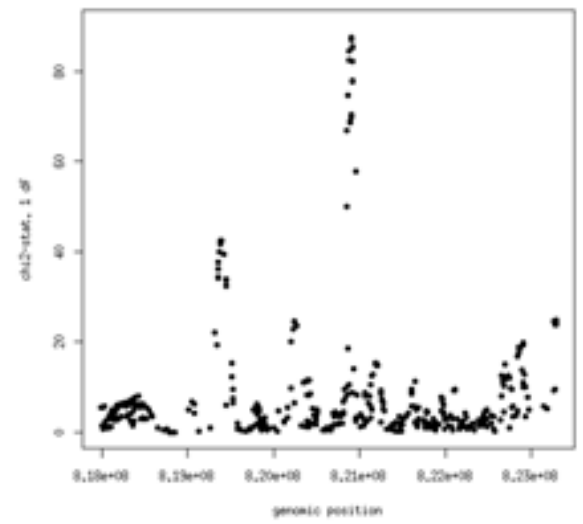

chi2-stat, 1 df
genomic position


chi2-stat, 1 df
genomic position


chi2-stat 1 df
genomic position


genomic position


chi2-stat, 1 df
genomic position


chi2-stat, 1 df
genomic position

Summary and Discussion

- discrimination of interesting and non-interesting regions
- automatically optimizing of the screening method
- incorporates the idea of regional plots and clumping

- Simulation studies: when rejects ψ_0 the null hypothesis?
- how affect $K_{\lambda(x_0,x_j)}$ and r_{0j}^2 the statistic ψ_0?
- how is ψ_0 distributed under the null hypothesis
- computational very intensive

References

[1] Yang J, Benyamin B, et al. (2010): Common SNPs explain a large proportion of the heritability for human height, Nature Genetics, 42(7):565-69.

[2] Purcell S, Neale B, et al. (2007): PLINK: a toolset for whole-genome association and population-based linkage analysis. AJHG, 81.

[3] Broad Institute of Harvard et al. (2007): Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels, Science, 316:1331-36.