Cornell Probability Summer School 2006

Simon Tavaré

Lecture 3

Ancestral Recombination Graph

Why recombination?

In the era of genomic polymorphism data, the need for models that include recombination is transparently obvious

Many questions are directly focused on recombination:

- linkage disequilibrium (association) mapping
- basic questions about the distribution and nature of recombination events
- novel multilocus summary statistics (e.g., based on haplotype sharing) are likely to become important

The ancestral recombination graph

Generalization of coalescent to allow for recombination

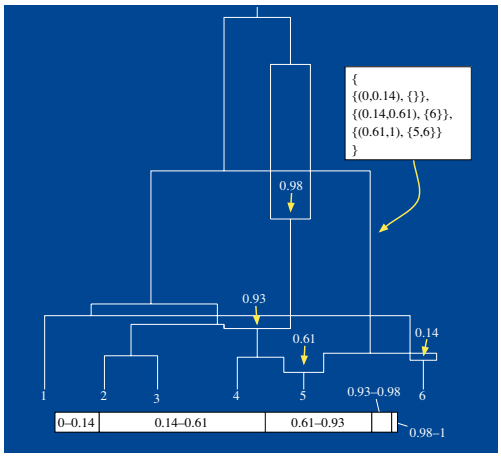
Hudson (1983)

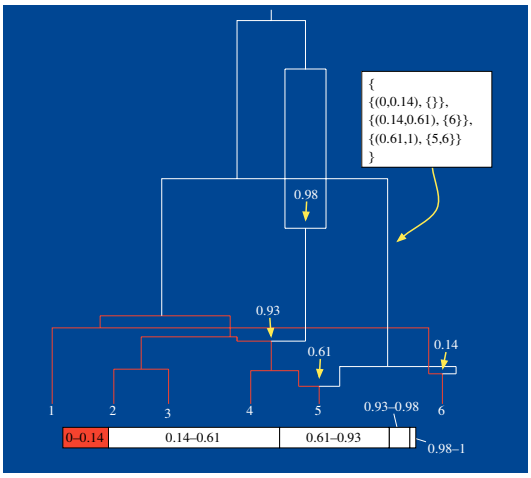
Griffiths & Marjoram (1996, 1997)

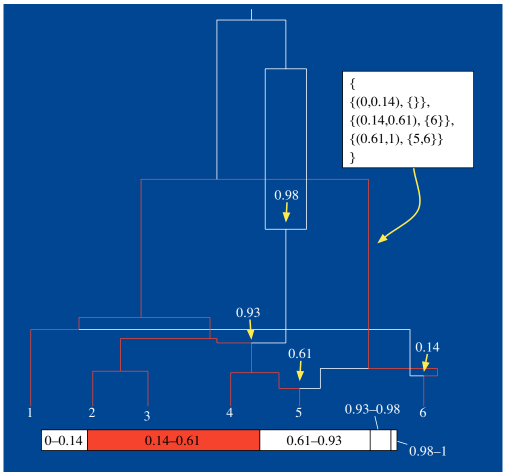
Scaled recombination rate ρ

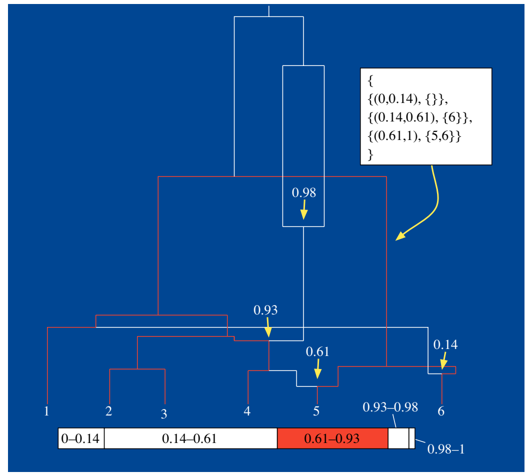
When there are k lineages, split rate is $k\rho/2$, coalescence rate is $k(k-1)/2$

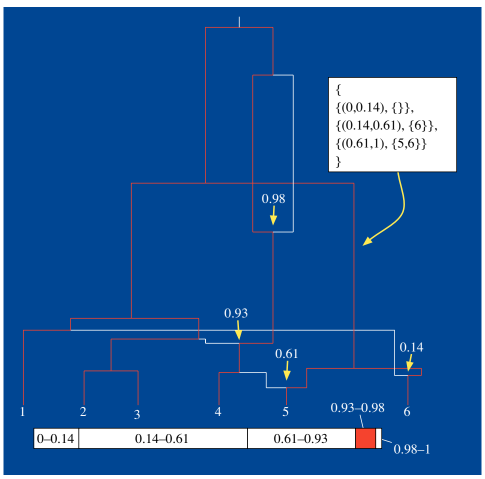
An ARG in action

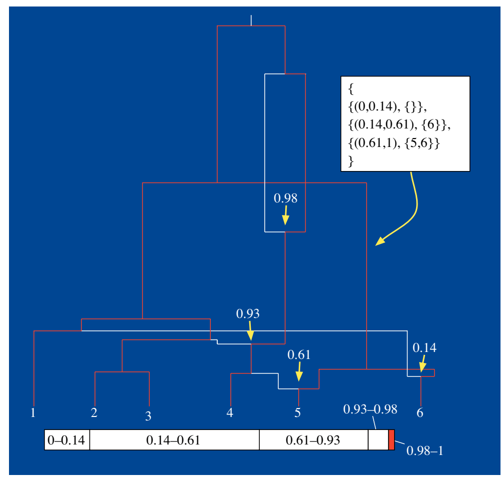




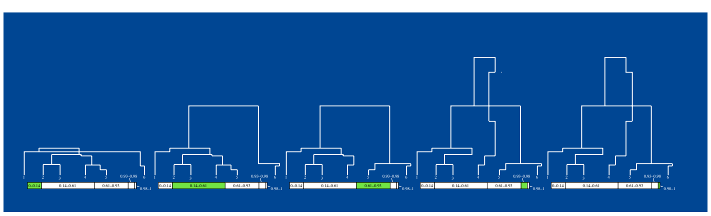








A walk through tree space



As we walk along the chromosome, the trees change — but only gradually.

Linked trees are correlated, and the degree of correlation decreases with genetic distance.

One manifestation of this is *linkage disequilibrium* . . .

How common is recombination?

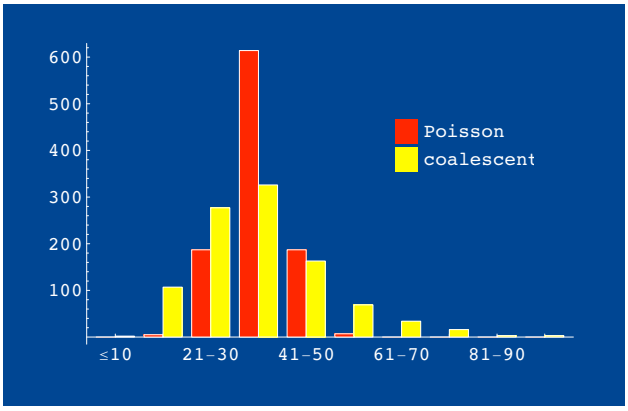
- If one 1 cM $\sim$ 1 Mb, then $r \approx 10^{-8}$ per site
- The per-generation mutation probability u is estimated to be the same or lower
- Thus we should have $\rho \geq \theta$

It follows that there will typically be at least as many recombination events in the history of a sample as there are segregating sites!

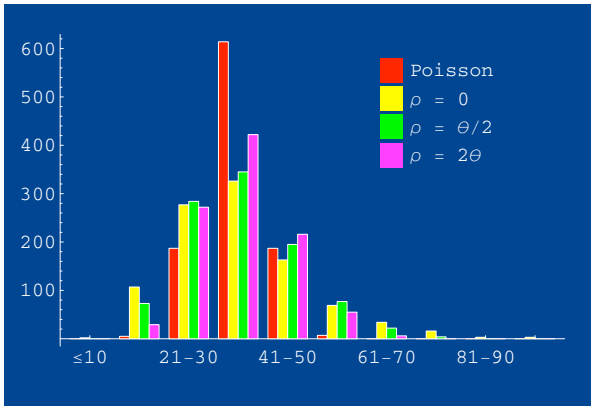
Overcoming the evolutionary variance

- Recombination makes different parts of the genome increasingly independent
- This reduces the evolutionary variance that is due the coalescent
- Finally a sample size greater than 1!

The number of SNPs in 20 copies of 10 kb (1000 runs):



Same thing, with recombination:



Linkage Disequilibrium

Linkage disequilibrium (LD)

Linkage disequilibrium refers to non-random association of alleles at different loci

For definiteness think of loci each with two alleles. Marker locus B, disease locus A

Let p_{ij} denote the probability of the haplotype $A_i\,B_j$, with marginal frequencies p_{i+}, p_{+j} respectively

Define

$$D = p_{11} - p_{1+}\,p_{+1}$$

Measures of LD: D, r^2, d^2

$$D = p_{11}p_{22} - p_{12}p_{21}$$

There are many other measures in the literature, including

$$\begin{aligned} r^2 &= \text{squared correlation between } A \text{ and } B \text{ loci} \\ &= D^2/(p_{1+}p_{2+}p_{+1}p_{+2}) \end{aligned}$$

$$\begin{aligned} d^2 &= (\mathbb{P}(B_1 \mid A_2) - \mathbb{P}(B_1 \mid A_1))^2 \\ &= \left(\frac{p_{A_2B_1}}{p_{A_2}} - \frac{p_{A_1B_1}}{p_{A_1}}\right)^2 \end{aligned}$$

Why linkage disequilbrium?

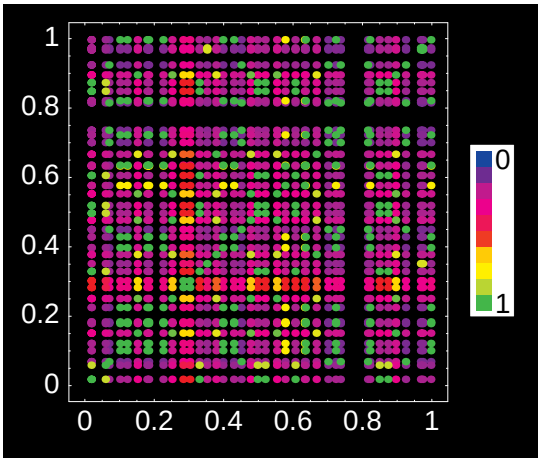
The phrase *linkage disequilibrium* is one of the most misleading in population genetics. First of all

- unlinked genes can be in LD
- linked genes are not necessarily in LD

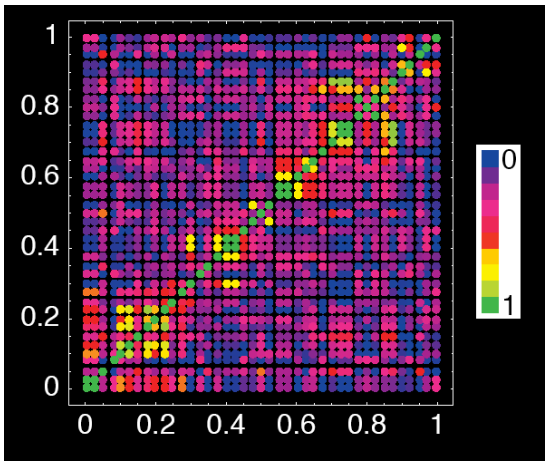
Thus linkage disequilibrium is only indirectly associated with linkage

Useful reference: M. Nordborg & ST (2002)
Linkage disequilibrium: what history has to tell us
Trends in Genetics **18**: 83-90

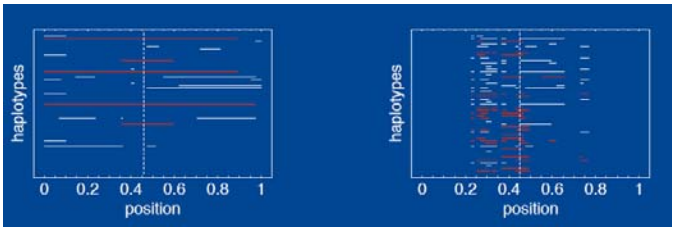
LD with no recombination: $|r|$



LD in theory



Haplotype sharing



The standard ancestral recombination graph with infinite sites mutations and $\theta = \rho = 100$ was used to simulate 50 chromosomes

The horizontal axis, representing chromosomal position, corresponds to ~ 100 kb

Focal mutations and haplotype sharing

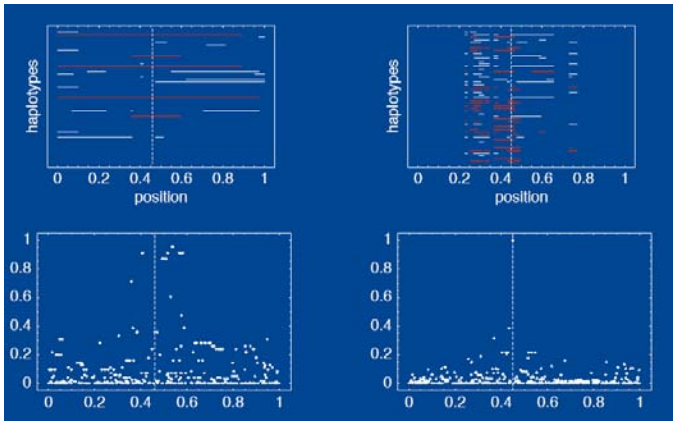
The chromosomal positions of the focal mutations are indicated by the vertical lines

- The previous plot shows the extent of haplotype sharing with respect to the most recent common ancestor (MRCA) of the focal mutation among the 50 haplotypes
- The horizontal lines indicate segments that descend from the MRCA of the focal mutation.
- Red indicates that the current haplotype also carries the focal mutation; black that it does not

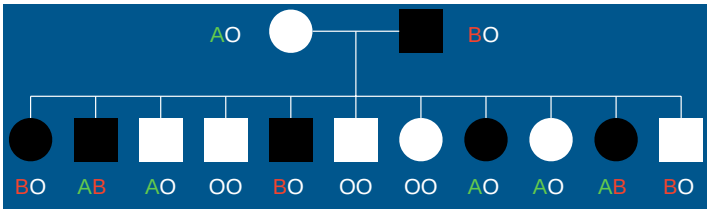
Focal mutations and haplotype sharing

- Note that the red segments necessarily overlap the position of the focal mutation
- For clarity, segments that do not descend from the MRCA of the focal mutation are excluded, and haplotypes that do not carry segments descended from the MRCA of the focal mutation are therefore invisible
- Next plot shows behaviour of d^2

Haplotype sharing and LD (d^2)



Traditional mapping methods: pedigrees

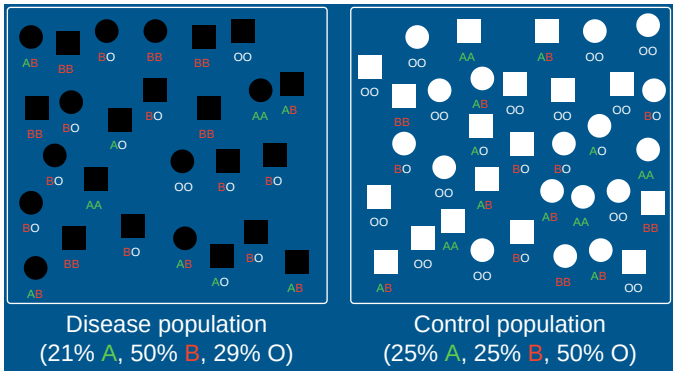


In human genetics, pedigrees are increasingly insufficient:

- They contain too few recombination events to allow fine-scale mapping
- They contain too few individuals to handle complex traits where each allele has only a small effect

Linkage disequilibrium (LD) mapping

Solution: look for population associations



Why LD mapping?

Advantages:

- No pedigrees or crosses needed
- Utilizes historical recombination events

Disadvantages:

- Statistical nightmare

Association Studies

Regulatory Element Variants

Project: *Identify and characterize functionally variable regions that are likely to contribute to complex phenotypes and disorders in human populations through effects on regulation of gene expression*

- Gene expression level as phenotype
- Survey gene expression in multiple individuals
- Survey nucleotide polymorphism in the same individuals
- Map responsible genetic factors

Outline

- Gene expression
- HapMap and ENCODE projects
- Illumina BeadArray technology
- Results
- Statistical aspects

Control of Gene Expression

cis-acting regulatory elements of genes

- Near the coding regions on the same DNA molecule. They act only on the coding regions on this molecule

— promoters, operators, enhancers

trans-acting regulatory elements of genes

- Encode proteins that can diffuse to any molecule of DNA in the cell and regulate it by binding to cis regulatory regions or to other proteins bound there

— repressors, transcription factors

HapMap `www.hapmap.org`

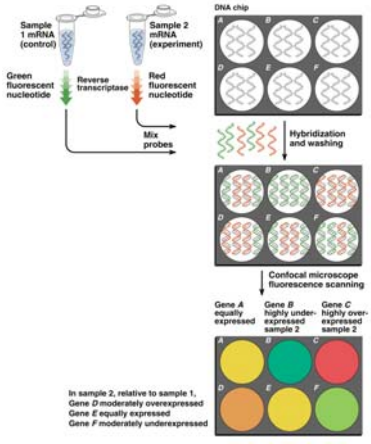
ENCODE `http://www.genome.gov/10005107`

The ENCODE Project: ENCyclopedia Of DNA Elements

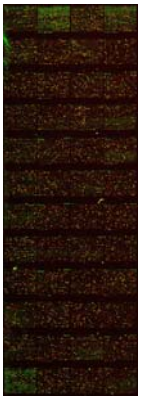
- Overview
- Consortium Membership
- Data Release Policy
- Accessing ENCODE Data
- Common Consortium Resources
- Target Selection Process and Target Regions
- Comparative Sequence Analysis
- Coordination with HapMap
- Meeting Reports
- Request for Application (RFA)
- Press Releases and Publications
- Program Staff

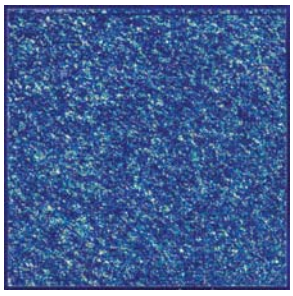
Special Announcement: The modENCODE Project new

Expression Measured by DNA Chips

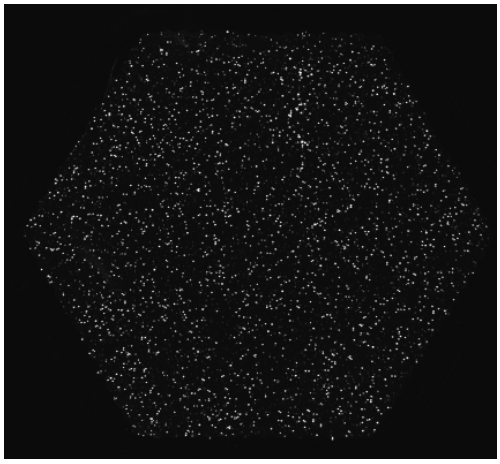


Array Technology

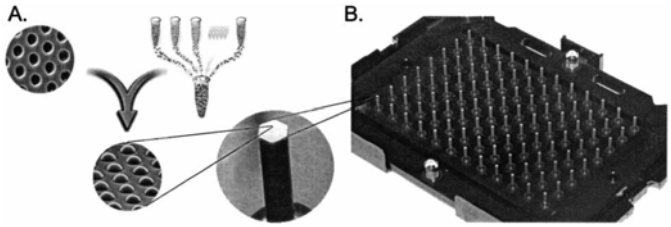




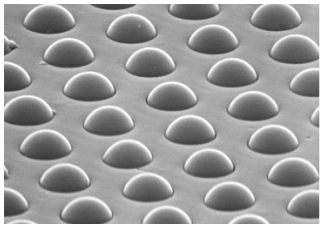
Illumina BeadArray

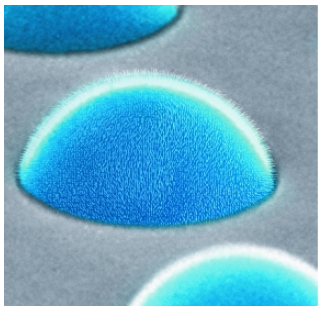


Beads



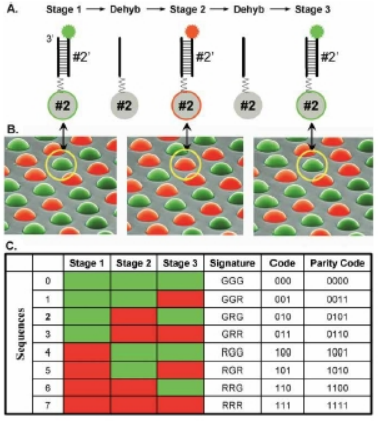
Illumina BeadArray





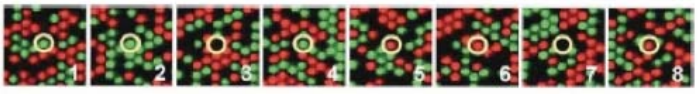
One bundle is 1.4mm wide
50,000 3-micron beads per bundle
6 micron spacing

Decoding



		Stage 1	Stage 2	Stage 3	Signature	Code	Parity Code
Sequences	0				GGG	000	0000
	1				GGR	001	0011
	2				GRG	010	0101
	3				GRR	011	0110
	4				RGG	100	1001
	5				RGR	101	1010
	6				RRG	110	1100
	7				RRR	111	1111

Decoding



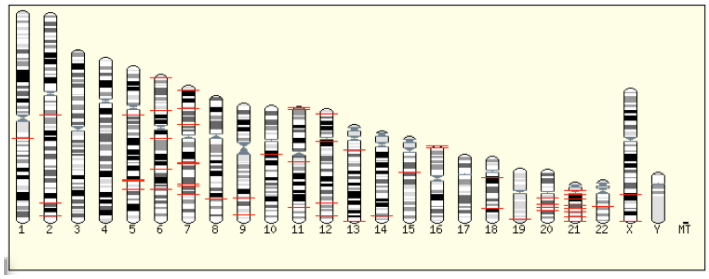
Galinsky VL. Automatic registration of microarray images. I. Rectangular grid. Bioinformatics 2003, 19: 1824-1831.

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG et al. Decoding randomly ordered DNA arrays. Genome Research 2004, 14: 870-877.

Which genes?

- 700 transcripts
  - 350 from ENCODE regions
  - 250 from Human chr21 (Down Syndrome)
  - 100 from Human chr20 (10 Mb 20q12-13.13, type II diabetes and obesity)
- Samples: lymphoblastoid cell lines from unrelated HapMap individuals.
- 60 European (CEU), 60 Yoruban (YRI), 45 Japanese (JPT), 45 Han Chinese (HCB)

Where are our genes?

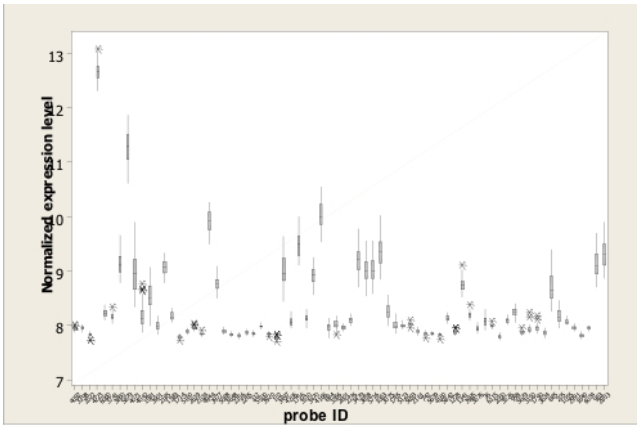


Analyzing Bead Data

- The experiment
  - RNA → 2 IVTs → 3 replicates of each
- Low-level analysis of bundle data
  - Image analysis and segmentation
  - Background correction
  - Normalization
- Using bead-level data

Results

Phenotypic variation: expression in sample



Linear Models

i – individual ($i = 1, 2, \ldots, 60$)

genotype$_i$ – one of (eg) A/A, A/T, T/T, N/N

score$_i$ – number of A in genotype

Additive model:

$$Y_i = \alpha + \beta\,\text{score}_i + \text{error}_i$$

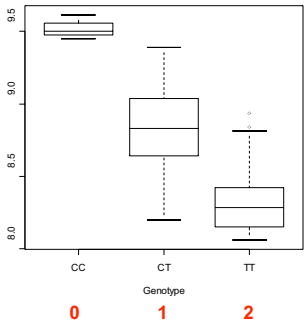
Non-additive model:

$$Y_i = \alpha + \sum_g \beta_g\, I(\text{genotype}_i = g) + \text{error}_i$$

Is $\beta = 0$? Are $\beta_g = 0$?

Additive Model

Additive association model:
Linear regression e.g. CC = 0, CT = 1, TT = 2.



- slope of line
- p-value
- r^2

Additive Model

Additive association model:
Linear regression e.g. CC = 0, CT = 1, TT = 2.

60 CEU
45 CHB
44 JPT
60 YRI

→ 688 probes; 374 genes

Phase I HapMap; MAF > 0.05

CEU: 762,447 SNPs
CHB: 695,601
JPT: 689,295
YRI: 799,242

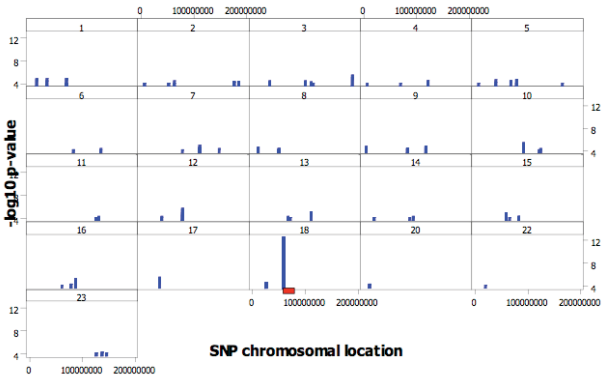
} ~1/5kb

Assessing significance

Three methods used:

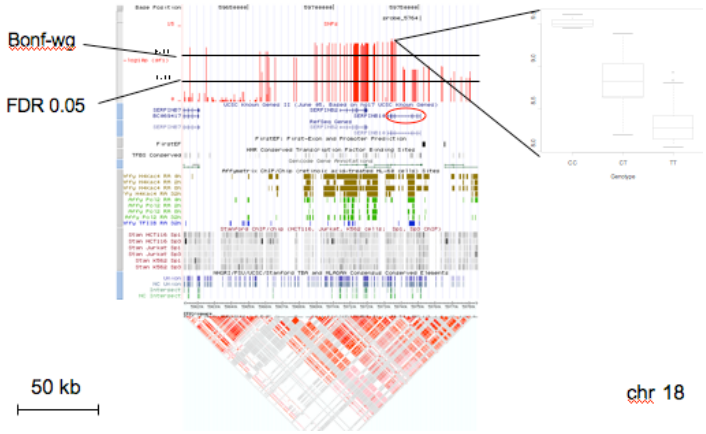
- Bonferroni (genome-wide or local)
  - nominal p-value $\approx 10^{-10}$
- Permutation-based (genome-wide or local)
- FDR (local) $q = 0.05$

Example: *SERPINB10* in CEU



serine proteinase inhibitor B10

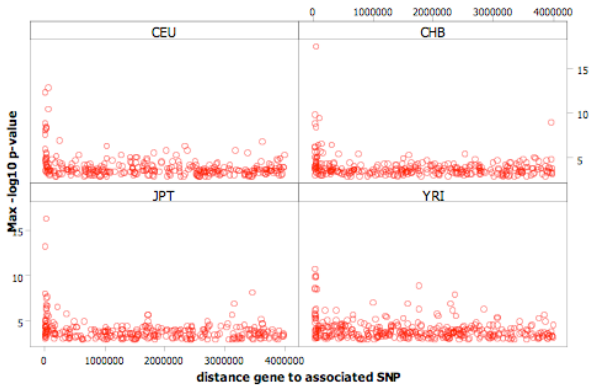
Example: *SERPINB10* in CEU



Comparing the populations

Separate analysis for each group

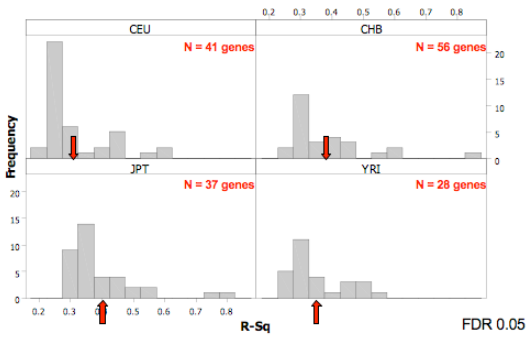
Effect of distance from the gene



distance < 4Mb; $-\log_{10} p > 3$

Magnitude of significant cis- effects (r^2)

How much of phenotypic variation is explained by most significant SNP?



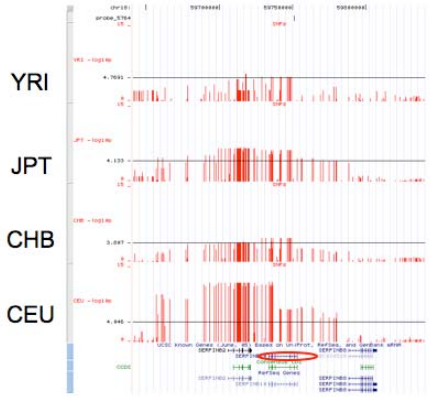
FDR 0.05

Replication: CEU, CHB, YRI, JPT

Overlap of genes with significant cis- signals

FDR 0.05

SERPINB10: expression-SNP association



FDR 0.05

Conclusions

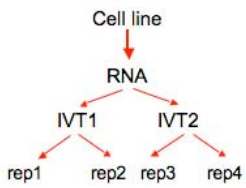
- Large number of genes with significant expression variation in a human population samples
- Strong association between individual genes and specific SNPs
- Replication of significant signals in other populations
- Prospects for identification of functionally variable regulatory regions in the whole genome

Too early for premature optimism?

Experiments and Analysis

- Functional assays to confirm associations
  - Promoter assays, allele specific expression
- Analysis of HapMap trios (QTDT; heritability).
- Development of interaction models
- Association with Copy Number Variants (CNVs)
- Genome-wide expression screening (48,000 transcripts) of all 270 HapMap individuals.

Whole-genome gene expression



illumina Human 6 x 2 gene GEX arrays

~48,000 transcripts interrogated

270 HapMap individuals:
CEU: 30 trios, 90 total
CHB: 45 unrelated
JPT: 45 unrelated
YRI: 30 trios, 90 total

2 IVTs each person
2 replicate hybridizations each IVT

Whole-genome gene expression

Additive association model:
Linear regression e.g. CC = 0, CT = 1, TT = 2.

60 CEU
45 CHB
45 JPT
60 YRI

→ ~15,000 genes

Phase II HapMap; MAF > 0.05

CEU: ~3M SNPs
CHB: ~3M SNPs
JPT: ~3M SNPs
YRI: ~3M SNPs

} ~1/kb

45 billion tests per population!

Statistical Issues

Full employment for statisticians!

A: Bead-level Data

Each hexagonal bundle has $\sim 50,000$ intensities

Illumina now provides

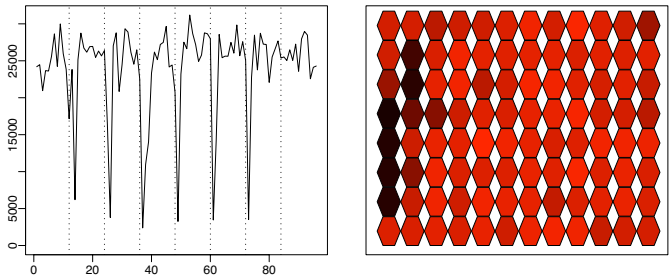
- probe sequences
- coordinates of each bead center
  - bead is 3×3 pixel box
- (identity and) intensity of each bead
- control probe features

Now able to consider further diagnostics about

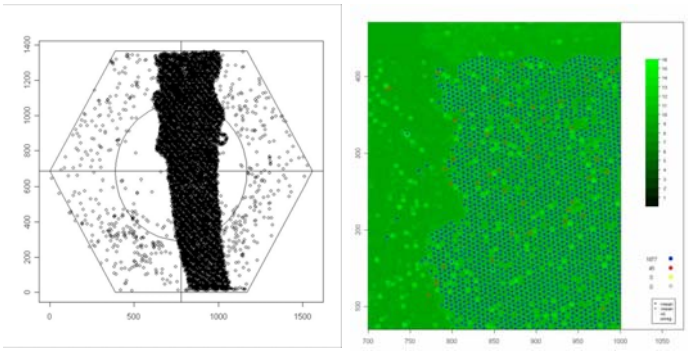
- bead quality
- background signal (hybridization dynamics)
- bundle quality

We have developed an analysis package *beadarray* in R in the latest BioConductor release, 1.8.
(`http://www.bioconductor.org/`)

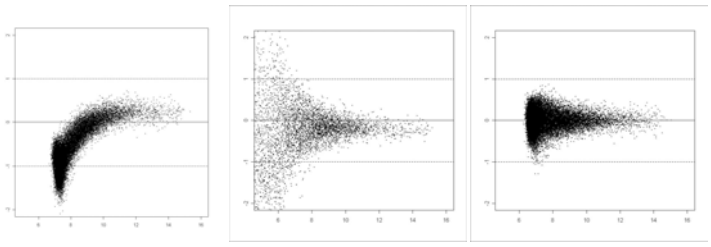
Diagnostic plots



Outlier detection



Normalization – MA plots



Non-normalized | Background normalized | Background corrected

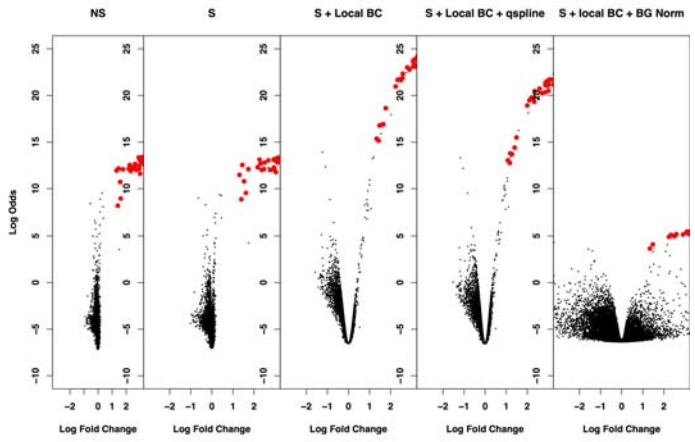
B: Spike-in Experiments

- In the microarray business, *truth* is hard to come by
- Control experiments:
  - Spike-ins
  - Dilution series

B: Spike-in Experiments

- In the microarray business, *truth* is hard to come by
- Control experiments:
  - Spike-ins
  - Dilution series
- Complex mouse background hybridised to array.
- Human genes spiked at known concentrations:
  1000 pM, 300, 100, 30, 10, 3 (2 replicates)
  1.0, 0.3, 0.1, 0.03, 0.01, 0.003 (1 replicate)
- Besides 33 spikes, nothing is differentially expressed

Volcano plots



C: Assessing Significance

- It is the ranking that counts!
  - False discovery rates
- Indirect associations
- Interactions
  - . . . among SNPs
  - . . . among phenotypes
- Data mining

D: Interactions — CART, Clustering

- Phenotypes: clustering, common regulatory elements
- Genotypes: CART
  - should pick out nonlinear interactions
  - significance assessed by predictive error
  - then search for SNPs in high LD with CART SNPs
- Does it work?
  - Sample sizes? Power?
  - Simulating data – how do be generate bigger test samples?

E: NHGRI — ENDGAME Consortium

Enhancing Development of Genome-wide Association Methods

Statistical aspects:

- Assessing the utility of genome-wide association studies to understand human genetic variation and its role in health and disease
- Developing new, analytic and computational strategies and resources for use by the scientific community to find genes associated with disease
- Understanding the role of gene-gene and gene environment interactions in disease susceptibility

References

Dunning M, Thorne NP, Camilier I, Smith ML, Tavaré S (2006) Quality control and low-level statistical analysis of Illumina BeadArrays. *REVSTAT*, **4**, 1–30

Stranger BE, Forrest MS, Clark AG, Minichiello M, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavaré S, Deloukas P, Dermitzakis ET (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet*, **1**, 695–704

Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW et al. High throughput DNA methylation profiling using universal bead arrays. Genome Research 2006, 16: 383-393.

Acknowledgements

Illumina:

- Gary Nunn
- Brenda Kahl
- Semyon Kruglyak

Sanger Institute:

- Barbara Stranger
- Matthew Forrest
- Manolis Dermitzakis
- Panos Deloukas

Cornell

- Andy Clark

UCSD:

- Roman Sasik

USC:

- Peter Calabrese
- Han Wang

Cambridge:

- Natalie Thorne
- Mark Dunning
- John Marioni
- *Isabelle Camilier*
- Mike Smith

Cornell Probability Summer School 2006

Simon Tavaré

Lecture 4

Rejection algorithm

Model depends on parameter θ. Want posterior distribution $\mathbb{P}(\theta \mid \mathcal{D})$

- Select θ' from prior distribution
- Simulate data $\mathcal{D}'$ from model with parameter θ'
- If $\mathcal{D} = \mathcal{D}'$, accept θ'
- Repeat

Accepted θ' have distribution $\mathbb{P}(\theta \mid \mathcal{D})$

Approximate Bayesian Computation

Acceptance rate is too low for rejection algorithm. Instead select summary statistic S of data $\mathcal{D}$, and threshold ϵ.

- Select θ' from prior distribution
- Simulate data $\mathcal{D}'$ from model with parameter θ', compute summary statistic S'
- If $\rho(S, S') < \epsilon$, accept θ'
- Repeat

Accepted θ' have approximately distribution $\mathbb{P}(\theta \mid \mathcal{D})$

How to choose and use summary statistics?

- If the model is too complicated to calculate likelihoods, can't find a *sufficient* statistic
- Local-linear regression (Beaumont et al. Genetics, 2002)
- *Projection pursuit* is a nonlinear regression method

$$\theta = \alpha_0 + \sum_{j=1}^{M} f_j(\boldsymbol{\alpha}_j^T \boldsymbol{S}) + \epsilon$$

New algorithm — Peter Calabrese

- Training set of size n
  - Select $\{\theta_i'\}$ from prior distribution
  - Simulate data $\{\mathcal{D}_i'\}$
  - Compute vector of summary statistics $(S_{i1}', S_{i2}', \ldots, S_{im}')$
- Fit projection pursuit regression function $\hat{f}$ such that $\hat{f}(S_{i1}', S_{i2}', \ldots, S_{im}') \approx \theta_i'$
- Apply $\hat{f}$ to summary statistics $(S_1, S_2, \ldots, S_m)$ from data $\mathcal{D}$
- If $|\hat{f}(S_1, S_2, \ldots, S_m) - \hat{f}(S_{i1}', S_{i2}', \ldots, S_{im}')| < \epsilon$ accept θ_i'

To improve behavior,

- Repeat, substituting accepted θ_i' for prior distribution

Comments:

- Start with $n = 100,000$
- Usually iterate 3 times, accepting a fraction ≈ 0.47 each time
- End with $10,000$ observations
- Repetition moves towards the posterior (iterates use different fs)

A test example

Allozyme frequency data and the ESF

- Urn models, branching processes, coalescents
- θ-biased permutations
- Distance of ESF to independent Poisson components
- Feller coupling

- $K = \#$ of types is sufficient for θ

Summary statistics

Sample of $n = 100$, $\theta = 5$

$\pi(\theta) \sim$ exponential, mean 10

- number of types
- mode $= \#$ individuals with most popular type
- homozygosity
- $\#$ singletons

Results

Statistic	Mean error	% within factor of 2 of truth	Coverage 50% credibility region	Width 50% credibility region
Homozygosity	1.70	90	51	3.26
Number singletons	1.43	81	68	3.00
Mode	2.18	82	53	4.48
Number alleles (sufficient)	1.17	99	57	2.17
ABC all 4	1.21	97	53	2.17
ABC all but number alleles	1.23	98	55	2.31

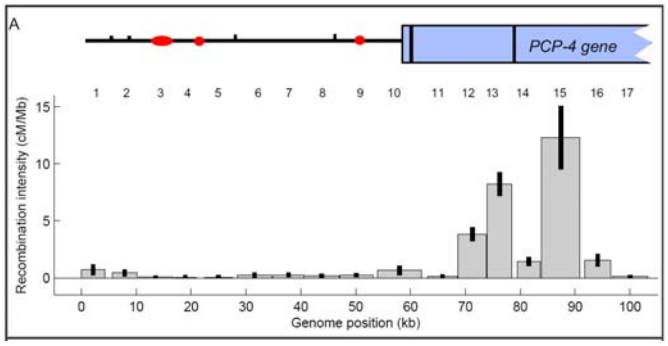
Conclusions

- The results when using the sufficient statistic are similar to when the ABC method is given all four statistics ...
- & when given the three statistics not including K
  - mean error below 1.23, 98% or greater within factor 2, 50% credibility region width less than 2.31
- Results are better than when any of the non-sufficient statistics are used individually
  - mean error above 1.70, 90% or fewer not within factor 2, 50% credibility width greater than 3.00

Recombination hotspots

Sperm-typing data from Arnheim's laboratory

Tiemann-Boege I, Calabrese P, ..., Arnheim N (PloS Genetics 2006)



Motivation

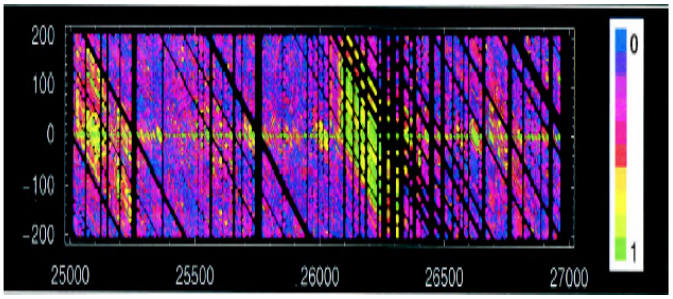
- Want fine-scale recombination rates:
  - Choose tag SNPs
  - Many statistical tests require genetic distances
  - Set parameters for simulations
- Can't sperm-type entire genome
- What can we infer from linkage disequilibrium data?

SNP data

- Perlegen: 24 European-Americans, 23 African-Americans, 24 Chinese
- HapMap: 30 European-American trios, 30 African trios, 45 Chinese, 45 Japanese
- Arnheim's 100kb region: Perlegen 124 SNPs, HapMap 69 SNPs

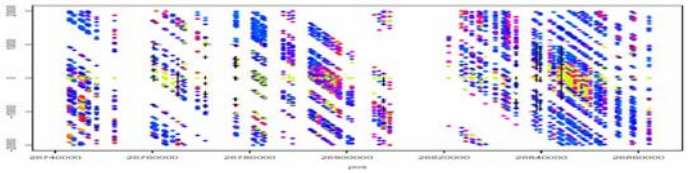
2,000 kb window: 200 kb up/downstream

Innan, Padhukasahasram, Nordborg (Genome Research 2003)

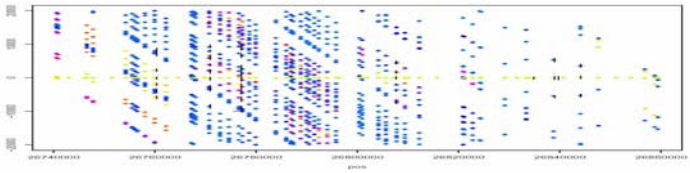


100 kb window: 20 kb up/downstream

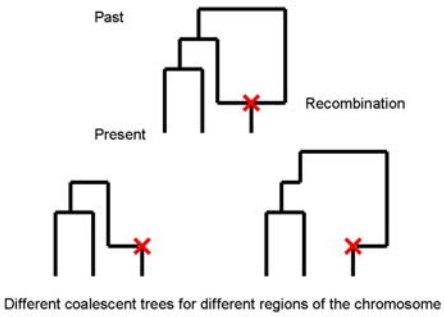
Perlegen



HapMap



Ancestral recombination graph



Allow recombination rate to vary along the chromosome

Simulating ARGs

- ms — Hudson, Bioinformatics 2002
- McVean & Cardin, Phil Trans R. Soc. B 2005
- Marjoram and Wall, BMC Genetics 2006

Getting it done

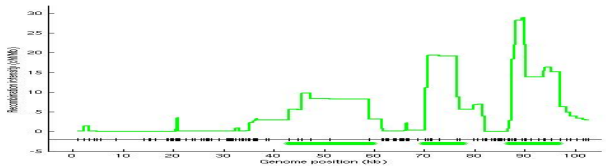
Table 3. Run-times. Average time per simulation, as a fuction of sample size based on 20 trials, assuming $\theta = 10^{-3}$/bp and $\rho = 5 * 10^{-4}$/bp. Simulations were run on a 2.8 GHz Intel Xeon processor. Dashes correspond to simulations that could not be completed because they required too much (> 3 GB RAM) memory.

n	Length (Mb)	SMC	ms
1000	2	0.9	7.2
	5	2.1	62.6
	10	4.3	473.6
	20	8.3	6459.6
	50	20.9	-
	100	41.6	-
	200	83.9	-
4000	2	4.0	10.6
	5	10.4	-
	10	22.2	-
	20	40.7	-
	50	105.8	-
	100	201.5	-
	200	406.1	-

Existing methods: LDHat

McVean, Myers, . . . , Donnelly (Science 2004)

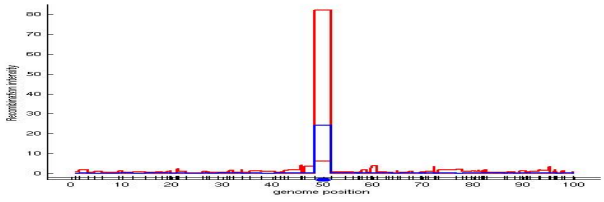
- Recombination rate is piecewise constant
- Reversible Jump Markov Chain Monte Carlo
- Composite likelihood
- Statistical significance



Existing methods: Hotspotter

Li and Stephens (Genetics 2003)

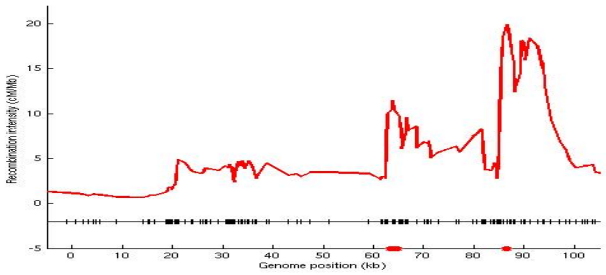
- Different recombination rate between each pair of consecutive SNPs
- New "coalescent-like" model easier to calculate likelihoods



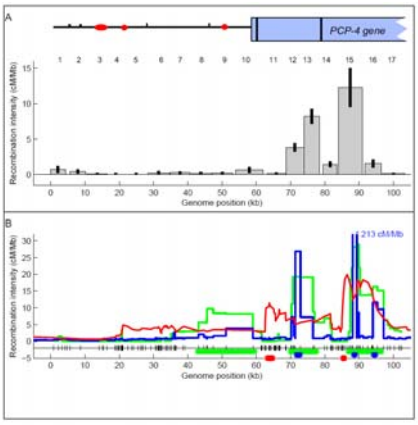
Estimating fine-scale recombination rates

Summary statistics: for sliding windows of 20 SNPs, compute the following 4 statistics on all 20, left 10, right 10 SNPs

- homozygosity
- number of alleles
- D' averaged over all pairs of SNPs
- # of non-overlapping pairs of SNPs that violate four-gamete test



Arnheim's region



European populations