Admixture and stratification

Paul McKeigue
Centre for Population Health Sciences
Schoolof Molecular, Genetic and Population Health Sciences
College of Medicine and Veterinary Medicine
University of Edinburgh

Statistical definition of genetic stratification

- Population is stratified if there are allelic associations between unlinked loci
  - Allelic associations generated by stratification are independent of map distance
  - associations generated by admixture or drift decay with map distance
- Stratification is not eliminated by a single generation of random mating
  - unlike Hardy-Weinberg equilibrium which holds after one generation of random mating

Statistical model-based definition of genetic admixture

- Admixture between *K* subpopulations with different allele freqs generates gametes that are a mosaic of segments with ancestry from each subpopulation
  - allelic associations decay with map distance
  - these allelic associations are explained by a model based on a mosaic of segments from K subpopulations
  - contrast with stratification which gives rise to allelic associations that are independent of map distance

Why stratification and admixture usually co-exist

- Where there is stratification, some gene flow between subpopulations is likely to occur unless social / geographic barriers are very strong
- Stratification within an admixed population is often maintained by continuing gene flow, or by social stratification
- Possible scenario for admixture without stratification: pulse of admixture followed by random mating in an island population

Quantifying genetic stratification

- F_{st}: fixation index subpopulation-total
  - Ratio of average allelic variance in subpopulations to allelic variance in ancestral total population
  - Can be computed for 2 or more subpopulations, or conditional on a continuous variable
  - F_{st} between inbred strains is 1.
- F_{st} between human continental groups ~ 0.05 to 0.15
- F_{st} between European subpopulations ~ 0.005

How population stratification may confound genetic associations with phenotype

- If C lies on a pathway from which causal information flows to exposure X and disease D, C is a confounder of the association between X and D
- If subpopulations differ in disease risk for environmental or genetic reasons, any genetic variant X that has different frequency in subpopulations will be associated with disease D
- Extreme example: association of lactase gene polymorphism with height in European-Americans

How important is it to control for confounding by population stratification?

- In populations formed by gene flow from different continental groups (e.g. Mexicans), stratification can be strong
  - Confounding effects are usually weak unless marker shows large variation between subpopulations
- When trying to detect small genetic effects (< 1% of variance), even weak stratification may be enough to confound associations
  - almost all genetic association studies now focus on small effects

Controlling for confounding by population stratification

- *Classical approaches*:
  - restrict to homogeneous population, stratify by demographic variables
- *Family-based designs*
  - study transmission of gene copies from heterozygous parents
- *Measure the confounder and adjust for it*
  - model stratification as a mixture of subpopulations ("structured association")
  - *principal components analysis* to infer latent variables that generate associations between unlinked loci

Statistical modelling of stratification

- Applications: investigate genetic structure of population, or control for stratification as a confounder
- Two possible approaches
  - Use principal components analysis, retaining K-1 principal components that summarize allelic associations between unlinked loci
  - Fit standard model of admixture and stratification with K subpopulations (ADMIXMAP, STRUCTURE, ANCESTRYMAP)
    - Compare models with different values of K

Principal components analysis

- *M* variables measured on *N* individuals
- Rotate the *M* axes to define *M* latent variables (principal components) that are linear combinations of the original variables
  - 1st component is axis that maximizes proportion of variance explained
  - 2nd component is axis that maximizes proportion of residual variance explained, and so on
- Principal components are evaluated as eigenvalues and eigenvectors of covariance matrix

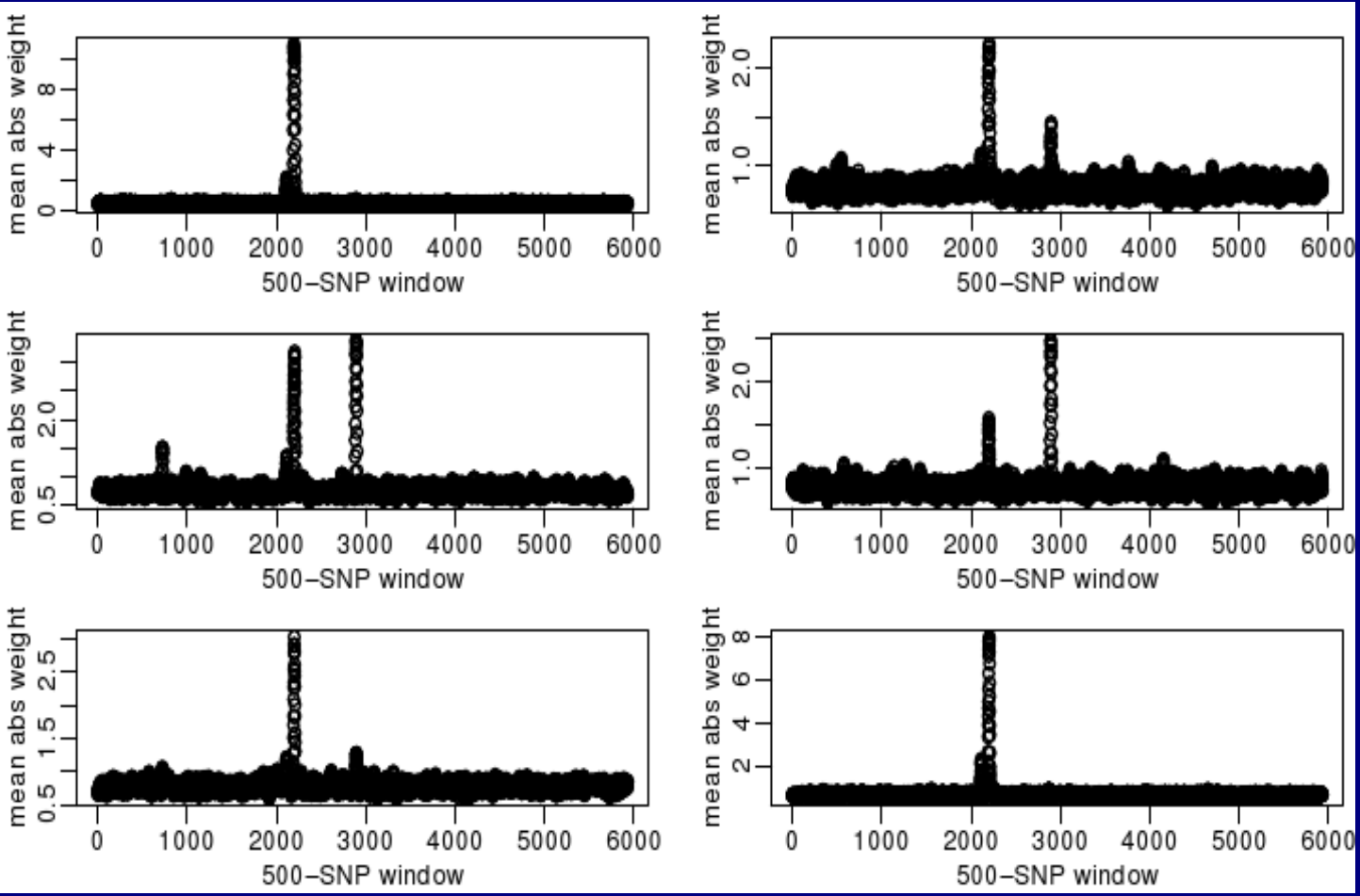
Eigenvalues, eigenvectors and principal components

- For efficient computation, can use the N x N matrix of covariance between persons
- With N individuals typed at M loci (M > N), the covariance matrix has N eigenvalues and N eigenvectors
- Eigenvalues are positive numbers proportional to variance accounted for by each PC.
- Eigenvectors are vectors of weights defining the rotation to new coordinates
  - used to compute PC scores for each individual

Principal components analysis with genome-wide marker panels (Patterson 2006)

- Regress genotype at locus *t* on genotypes at loci *t*, *t-1*, ...
- Use residuals from these regression models to calculate covariance matrix for genotypes
  - adjustment in regression model eliminates most short-range allelic associations generated by haplotype structure
- Use this covariance matrix for a principal components analysis

EIGENSTRAT with 500K tag SNPs: mean average weight of 500-SNP windows on first six principal components



Eliminating artefactual principal components to detect stratification only

- Thin the marker panel to obtain a panel of SNPs that are not in strong LD with each other (PLINK option `--ldpruned`)
- Regress genotype at locus *t* on genotypes at loci *t*, *t-1*, ...
- Drop clumps of markers that contribute strongly to one principal component (e.g. HLA region)
- Exclude related individuals (detected in N x N covariance matrix between persons)

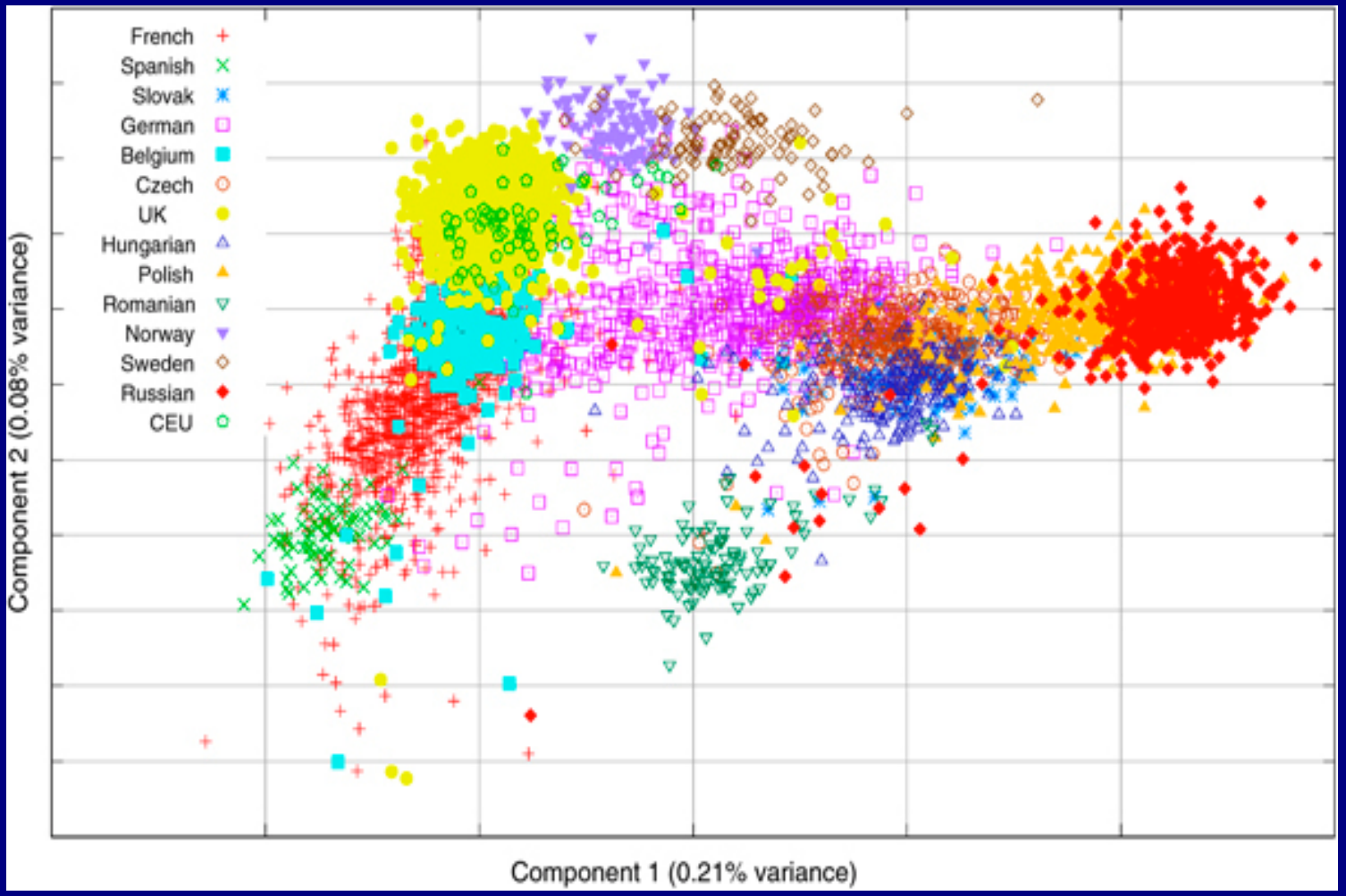
Principal components analysis as a statistical model

- M observed variables are modelled as linear combinations of C independent (unobserved) gaussian variables, plus noise
- Can infer the number C of principal components required to account for the observed covariance matrix
- All rotations of the C axes are equally compatible with the data
  - contours of multivariate gaussian are spherical

Testing how many principal components to retain

- Can calculate test for stratification based on proportion of variance explained by first principal component (largest eigenvalue)
  - repeat for next component until test is not significant
- For *N* individuals typed at *M* loci, stratification can be detected if $F_{ST} > 1/\sqrt{M\,N}$
  - to detect two subpopulations with $F_{ST} \sim 0.001$ requires $M\,N > 10^6$
- Each individual can be assigned coordinates in a *K-1* dimensional space

Scatter plot of scores on first two principal components in 6000 Europeans typed with 300K SNPs (Heath 2008)



"Structured association" or principal components for modelling admixture?

- Use principal components where
  - F_{ST} distance between subpopulations is small (< 0.01)
  - genome-wide markers have been typed but markers informative for ancestry cannot be pre-selected
- Use structured association where
  - F_{ST} distance is large
  - markers informative for stratification can be preselected
  - you want to model admixture

Practical implications of availability of statistical methods to control for population stratification

- Family-based designs are unnecessary
  - impractical for late-onset disease
- Strict population-based sampling of cases and controls is unnecessary
  - multiple case collections can use a shared control group
  - selection bias with respect to demographic background is eliminated by controlling for genetic background as if it were a confounder
  - in general this approach to controlling selection bias is valid where selection is not on factors that lie in the causal path between exposure and disease

Applications of statistical modelling of population admixture

- Admixture mapping
  - localizes genes in which risk alleles are distributed differentially between ethnic groups
- Investigating relation of disease risk to individual admixture proportions
  - to distinguish genetic and environmental explanations of ethnic variation in risk
- Controlling for population stratification in genetic association studies
  - eliminates confounding except by alleles at linked loci
- Fine mapping of genetic associations in admixed populations
  - to eliminate long-range signals generated by admixture

Distinguishing between genetic and environmental explanations for ethnic differences in disease risk

- Migrant studies:
  - consistency of high or low risk in varying environments
  - trend of risk ratio with number of generations since migration
  - failure of environmental factors to account for ethnic difference
- Relation of risk to proportionate admixture
  - may be confounded by environmental factors

Ethnic differences in disease risk that (on the basis of migrant studies) are unlikely to have a genetic basis

- Japanese-European: breast cancer, colon cancer, coronary heart disease
  - after 1-2 generations risk in Japanese migrants equals risk in US Whites
- African-European: multiple sclerosis
  - low risk in Europeans who migrated to South Africa before age 12

Type 2 diabetes: prevalence in South Asian migrants and their descendants

		Age	Prevalence
First-generation migrants			
1991	England	40-64	19%
> 5 generations since migration from India			
1977	Trinidad	35-69	21%
1983	Fiji	35-64	25%
1985	South Africa	30-	22%
1990	Singapore	40-69	25%
1990	Mauritius	35-64	20%

Type 2 diabetes: effect of gene flow from European males into a high-risk population (Nauruan islanders)

	% with European HLA types	
Age	Diabetic	Non-diabetic
20-44	6%	12%
45-59	9%	13%
60 +	5%	55%

Odds ratio for diabetes in those with European admixture = 0.31 (95% CI 0.11 - 0.81)

Serjeantson SW. *Diabetologia* 1983;**25**:13

Relation of risk of systemic lupus erythematosus to individual admixture in Trinidad (Molokhia 2003)

- 44 cases and 80 controls resident in northern Trinidad (excluding those with Indian or Chinese ancestry)
- Admixture proportions of each individual estimated from genotypes at 31 marker loci

	Risk ratio (95% CI) for unit change in African admixture	
Unadjusted	32.5	2.0 - 518
Adjusted for socioeconomic status	28.4	1.7 - 485

Exploiting admixture to map genes

- Admixture mapping: infer ancestry at marker locus (0, 1 or 2 copies from the high-risk population) then test for association of ancestry with the trait or disease
  - analogous to linkage analysis of an experimental cross
- Testing for *allelic* association (Chakraborty & Weiss 1988, Stephens et al. 1994 "MALD") does not fully exploit the information about linkage that is generated by admixture
  - efficiency of MALD is limited by information content for ancestry of individual markers (< 40%)
  - cannot use affected-only design

Statistical power of admixture mapping

- Required sample size is determined by the ancestry risk ratio (r)
  - ~800 cases required to detect a locus with $r = 2$
  - ~3000 cases required to detect a locus with $r = 1.5$
  - assuming that:
    - a dense panel of ancestry informative markers is available
    - admixture proportions from the high-risk population are between 20% and 70%
- Affected-only test of N individuals has same statistical power as case-control test of $2N$ cases and $2N$ controls

Advantages of admixture mapping in comparison with other approaches to finding disease susceptibility genes

- *Statistical power*
  - admixture mapping relies on direct (fixed-effects) comparison
  - family linkage studies rely on indirect (random-effects) comparison
- *Number of markers required for a genome search*
  - ~ 2000 ancestry-informative markers for a genome search, compared with > 300 000 markers for whole-genome association studies
- *Effect of allelic heterogeneity*
  - does not matter whether there are many rare risk alleles or only a few common risk alleles at the disease locus

Recent admixture between low-risk and high-risk populations

	Founding populations	Generations since admixture
Caribbean, USA	W African/European	2 – 15
Australia	Native Aus./European	6 - 8
Americas	Native Am./European	2 - 15
Pacific islands	indigenous/European	
Alaska,Canada, Greenland	Inuit/European	?10
East Africa	Arab/E African	~ 15-20?

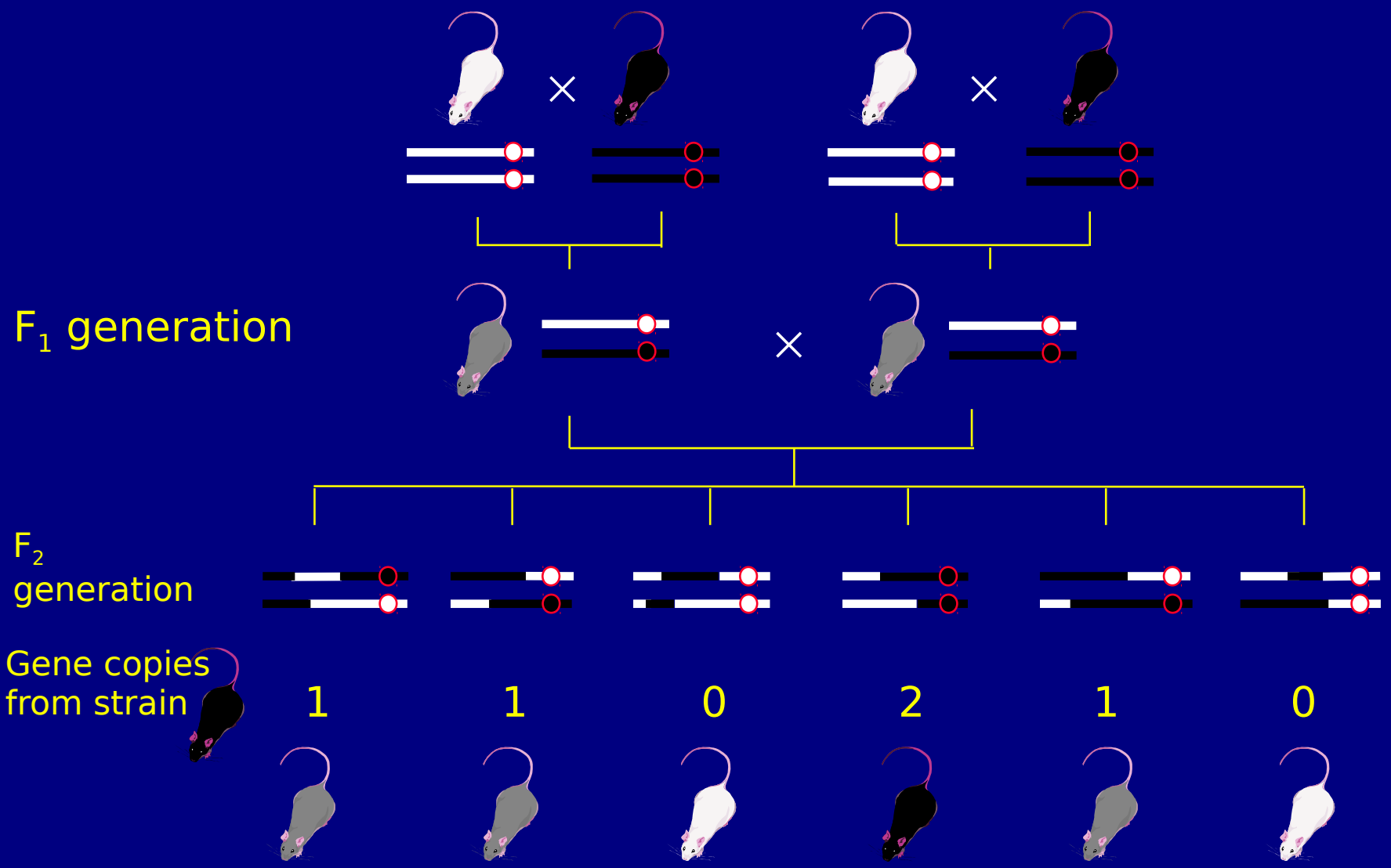
Diseases amenable to admixture mapping in populations of west African/European descent

Disease/trait	*Risk difference*
Hypertension	Commoner in west Africans
Systemic lupus erythematosus	
Prostate cancer	
Keloid scarring	
Sarcoidosis	
Focal segmental glomerulosclerosis	
Alzheimer disease	Lower risk in west Africans
Coronary disease / dyslipidaemia	
Osteoporotic fractures	

Diseases amenable to admixture mapping in other populations

Disease/trait	*Type of admixture*
Type 2 diabetes	Native American/European, Pacific islander/European, Native Australian/European Peninsular Arab/east African
Rheumatoid arthritis	Native American/European
Generalized obesity	Pacific islander/European, Native American/European
Central adiposity	South Asian/west African
Dyslipidaemia/coronary disease	South Asian/west African

An experimental cross between inbred strains



Methodological problems of extending linkage analysis of a cross to admixed human populations

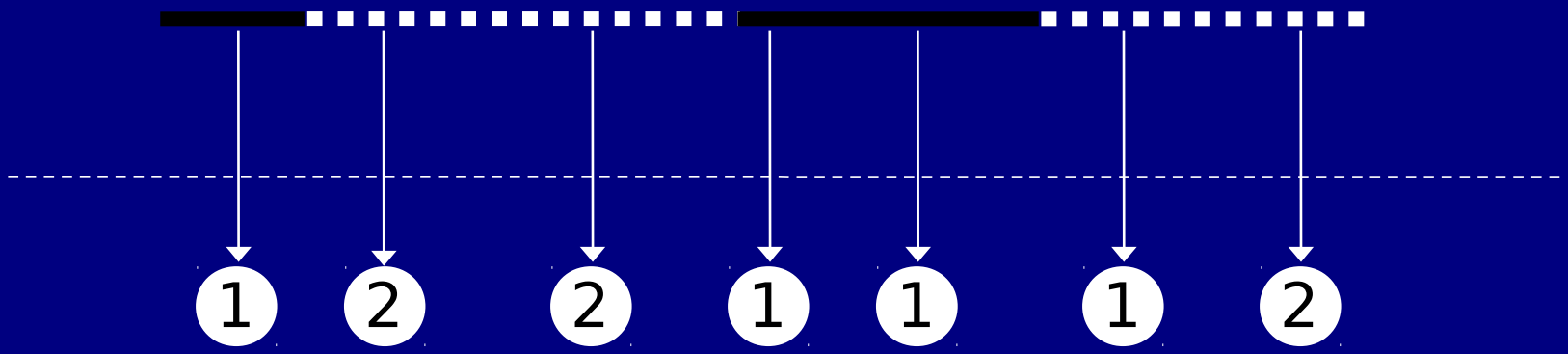
- History of admixture is not under experimental control or even known
  - population structure generates associations with ancestry at loci unlinked to the trait
- Ancestral populations are not available for study
  - cannot sample exact mix of west African populations that contributed to the African-American gene pool
- Human ethnic groups are not inbred strains: $F_{ST} \sim 0.15$
  - markers with 100% frequency differentials are rare
  - cannot unequivocally infer ancestry at locus from marker genotype

Statistical methods that allow linkage analysis of a cross to be extended to admixed humans

Problem	*How to overcome it*
History of admixture is not under experimental control	Condition on parental admixture proportions to eliminate associations with loci unlinked to the trait
Human ethnic groups are not inbred strains	Combine data from all markers in a multipoint analysis to extract information about ancestry at each locus
Ancestral populations are not available for study	Re-estimate ancestry-specific allele frequencies within the admixed population, with priors based on sampling unadmixed modern descendants

Model for stochastic variation of ancestry on chromosomes inherited from an admixed parent

Hidden states: states of ancestry at marker loci on chromosome of mixed descent

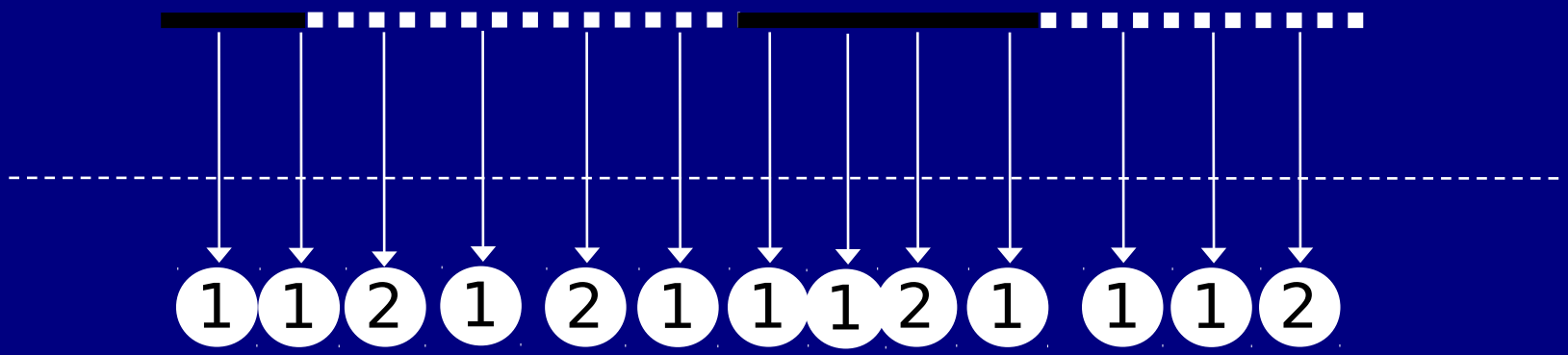


Observed data: marker alleles at each locus

Stochastic variation between K states modelled as sum of K independent Poisson arrival processes

Total arrival rate (sum of intensities) can be interpreted as the effective number of generations back to unadmixed ancestors

Multipoint inference of ancestry at marker loci from genotypes



- Hidden Markov model (HMM) message-passing algorithm yields posterior marginal distribution of ancestry states at each locus, given genotypes at all loci on the chromosome
- Information about locus ancestry depends on marker allele frequencies and marker density

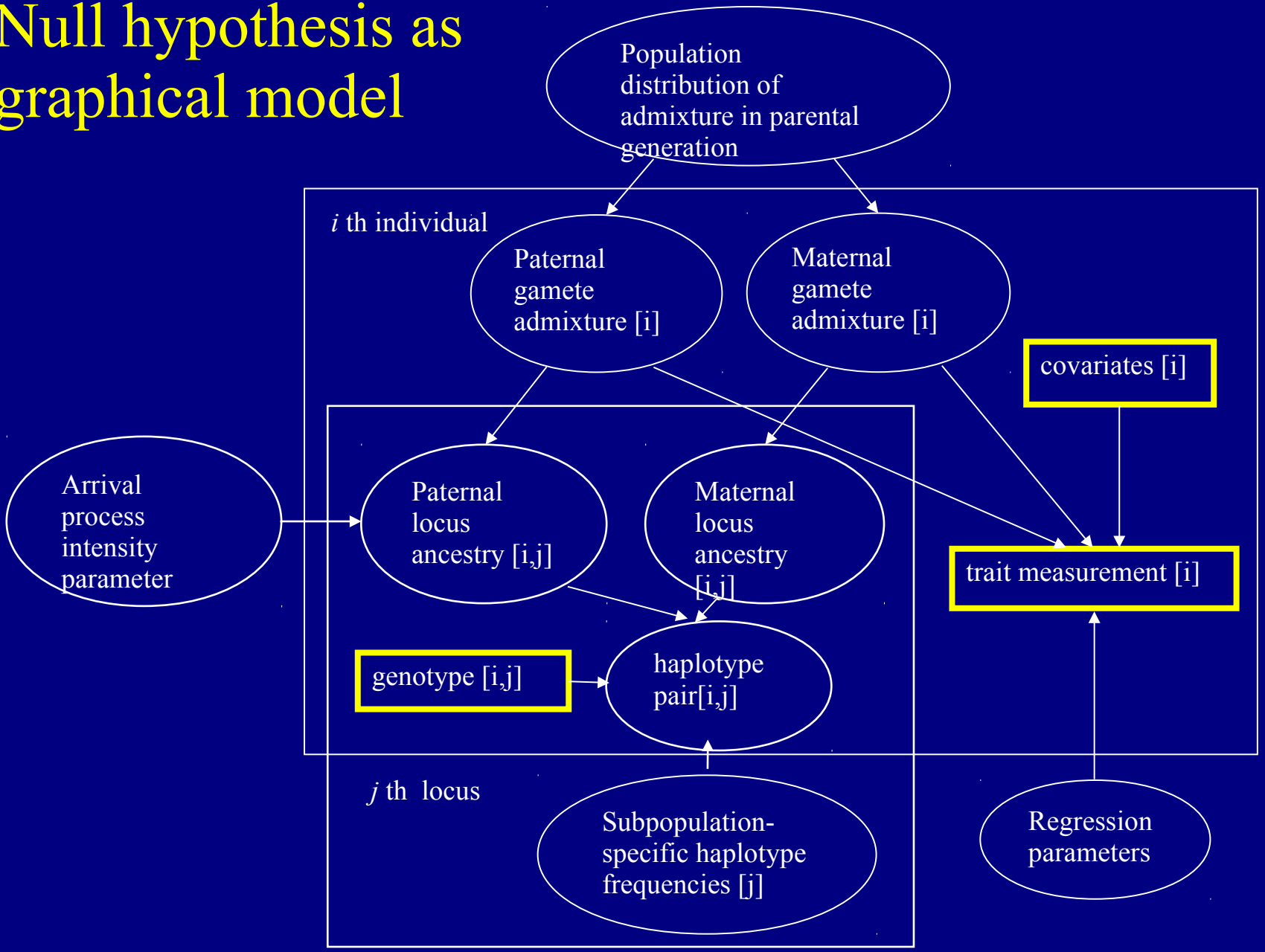
Hidden Markov models based on independent Poisson arrival processes: applications in genetics

- Linkage analysis with Haldane mapping function: 2 independent Poisson arrival processes each with intensity 1 per morgan (by definition)
- Admixture: K independent Poisson arrival processes with total intensity (effective number of generations since admixture) inferred from data
- Phasing and imputation: H independent Poisson arrival processes with total intensity varying across genome
  - IMPUTE, MACH: H=120 known haplotypes in HapMap)
  - fastPHASE: H ~ 8 modal haplotypes (not fixed)

Hidden Markov model algorithms for this class of models

- Matrix of transition probabilities has a regular structure that allows fast computation
  - Poisson arrival process implies exponential waiting times
- Standard algorithms can compute
  - Likelihood of model parameters given observed data
  - Posterior distribution of hidden states (segregation indicators, locus ancestry or haplotype) at each locus

Null hypothesis as graphical model



Statistical approach to model fitting

- Bayesian model of null hypothesis: all observed and missing data are random variables
  - *Observed data*: genotypes, trait values, covariates
  - *Missing data*:-
    - model parameters (admixture proportions, arrival rate)
    - locus ancestry states
- Posterior distribution of model parameters is generated by Markov chain Monte Carlo (MCMC) simulation
- For each realization of the model parameters, marginal distribution of locus ancestry is calculated by an HMM algorithm
- Three programs based on this approach are currently available: ADMIXMAP, ANCESTRYMAP, STRUCTURE

Limitations of the standard statistical model

- Assumption of no LD between markers within ancestral subpopulations limits marker density to 1 per cM.
- Alternative approaches:
- SABER: standard model extended to allow for 1st-order LD between loci
- HAPMIX: Model admixed gametes as mosaic of haplotypes in HAPMAP source populations
- LAMP: sliding window of SNPs to reconstruct most likely ancestry state of individuals within each window.

Alternative approach to modelling locus ancestry: HAPMIX

- Extension of methods for imputing genotypes at untyped HAPMAP loci from tag SNP genotyping arrays (MACH, IMPUTE)
- Requires phased haplotype data sampled from the two ancestral populations
- Target gametes (admixed) are modelled as mosaics of the source haplotypes (phased data from HAPMAP or similar).
  - copying from source to target is allowed to be noisy

Another approach to detecting admixture with genome-wide SNP data: LAMP (Sankaraman 2008)

- Sliding window of SNPs
  - window must be short enough that most individuals do not have an ancestry breakpoint within the window
- Within each window, cluster individuals according to most likely ancestry state (3 states with 2-way admixture model
- Can use ancestry-specific allele frequencies if known, or learn weights from the data (principal components analysis would work).

High-density (HAPMIX, LAMP) versus standard approach to admixture modelling (ADMIXMAP, ANCESTRYMAP)

- Advantages of HAPMIX, LAMP
  - Can use all SNPs on a genotyping array
    - no need to select panel of ancestry-informative markers
  - higher proportion of information extracted
- Disadvantages of HAPMIX, LAMP
  - Posterior probs of locus ancestry are not correctly calibrated: statistical tests that depend on averaging over posteriorprobs will behave strangely
  - Limited to 2-way admixture and to unrelated individuals

Combining high-density methods (HAPMIX, LAMP) and low-density (ADMIXMAP, ANCESTRYMAP) methods

- use HAPMIX/LAMP with all SNPs to infer locus ancestry
- thin these inferred locus ancestry states to spacing of 1 per cM, and code them as pseudo-genotypes
- run ADMIXMAP with these pseudo-genotypes, and allow program to learn pseudo-allele freqs

Statistical approaches to hypothesis testing

- Null hypothesis: $\theta = 0$ (where θ is the log ancestry risk ratio generated by the locus under study)
- By averaging over the posterior distribution of missing data under the null, we can evaluate two types of test:-
- *Likelihood ratio test* (implemented in ANCESTRYMAP):
  - evaluates $L(\theta) / L(0)$
  - averaging over prior on θ yields Bayes factor (ratio of integrated likelihoods) for an effect at the locus under study compared with the null
  - averaging over all positions on genome yields Bayes factor for an effect somewhere on the genome compared with the null
- *Score test* (implemented in ADMIXMAP):
  - evaluates gradient and second derivative of $\log L(\theta)$ at $\theta = 0$, to obtain a classical *p*-value

Evaluation of score test by averaging over posterior distribution of missing data

- For each realization of complete data, evaluate:
  - score (gradient of log-likelihood) at $\theta = 0$
  - information (curvature of log-likelihood) at $\theta = 0$
- Score U = posterior mean of realized score
  - Complete info = posterior mean of realized info
  - Missing info = posterior variance of realized score
- Observed info V = complete info – missing info
- Test statistic = $UV^{-½}$

Advantages of the score test algorithm (compared with likelihood ratio)

- All calculations are at $\theta = 0$
  - computationally efficient, no ascertainment problems
- Meta-analyses are straightforward: just add the score and information across studies
- Ratio of observed to complete information provides a useful measure of the efficiency of the study design
- Can be used to calculate model diagnostics:
  - test for departure from Hardy-Weinberg equilibrium
  - test for residual LD between pairs of adjacent marker loci

Information about ancestry conveyed by a diallelic marker

Marker allele 1 has ancestry-specific frequencies p_X, p_Y given ancestry from populations X, Y respectively

In an equally-admixed population, the proportion of Fisher information about ancestry of an allele (X or Y by descent) extracted by typing the allele is

$$f = \frac{\left(p_X - p_Y\right)^2}{4\,\bar{p}\left(1-\bar{p}\right)} \qquad \text{where} \qquad \bar{p} = \frac{1}{2}\left(p_X + p_Y\right)$$

40% ancestry information content (f = 0.4) is equivalent to allele frequency differentials of about 0.6

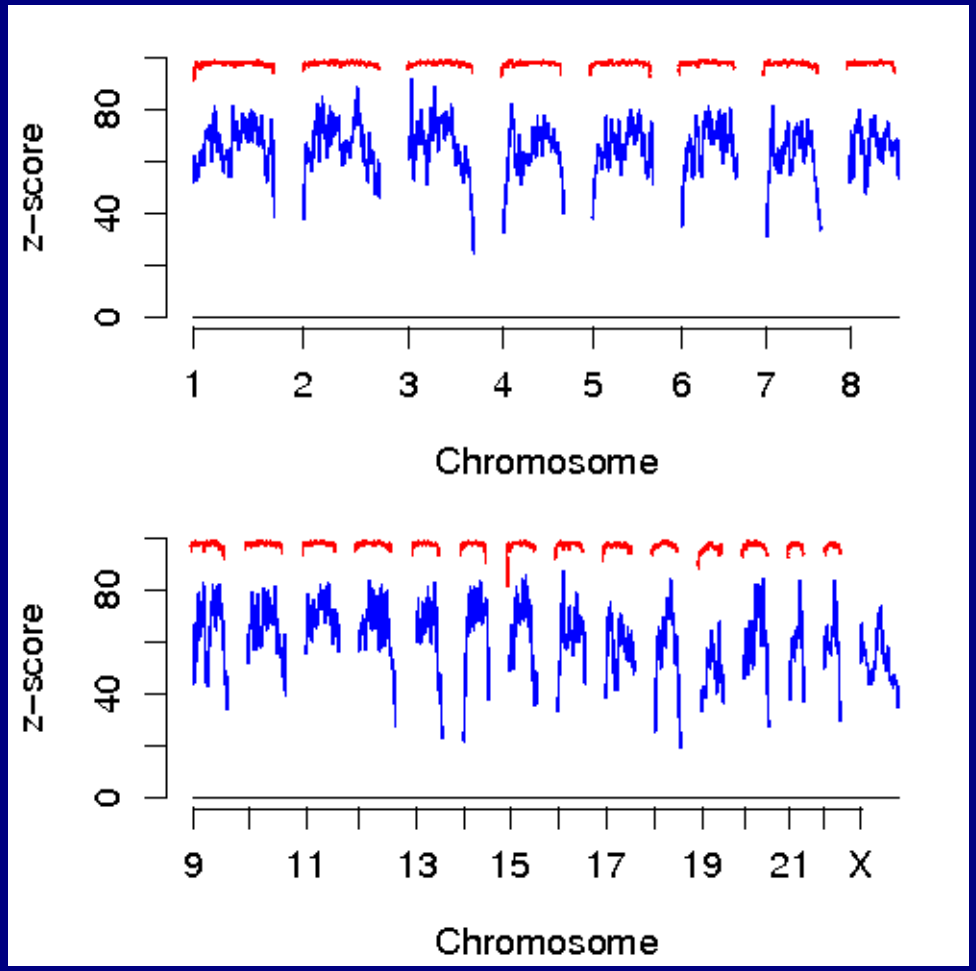
How many markers are required for genome-wide admixture mapping?

- Simulation studies based on typical African-American population: 80%/20% admixture, sum of intensities 6 per 100 cM, markers with 36% information content for ancestry
  - 64% of information about ancestry is extracted with markers spaced at 3 cM
  - 80% of information about ancestry is extracted with markers spaced at 1 cM

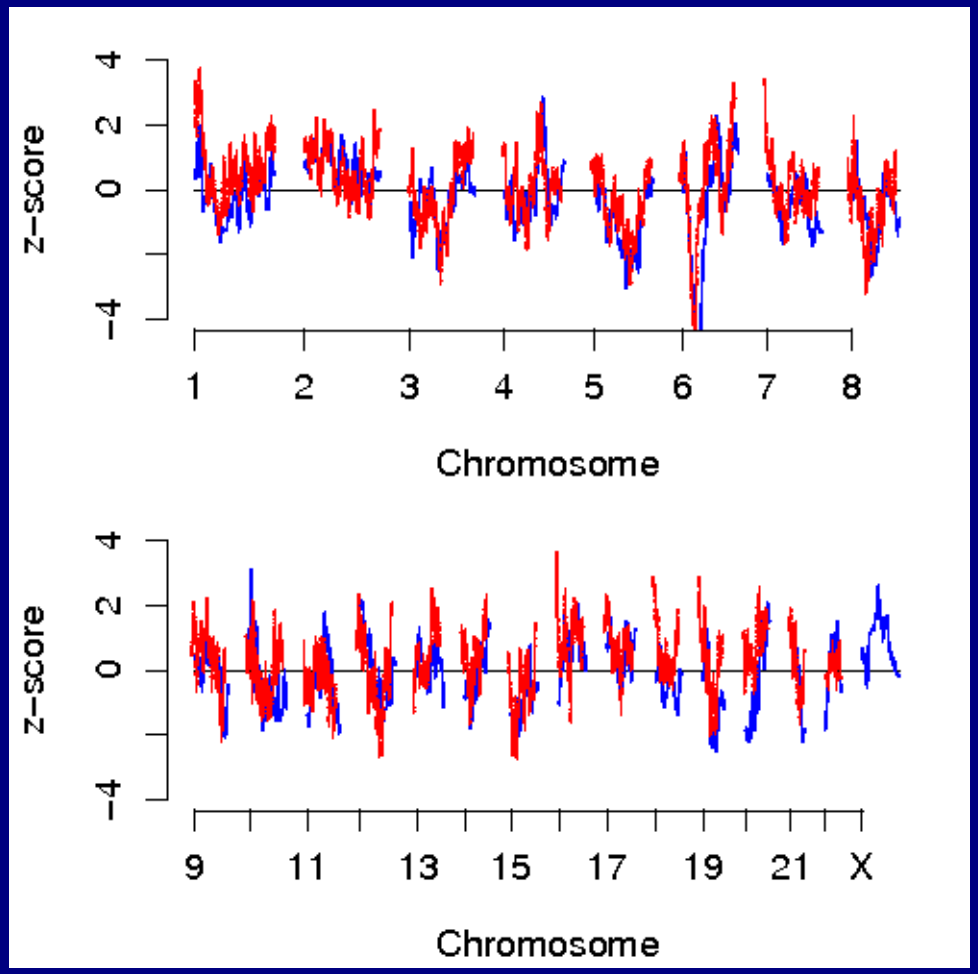
Panels of ancestry-informative markers

- Assembly of a panel of ~ 3000 ancestry-informative markers (AIMs) requires screening several hundred thousand SNPs for which allele frequency data are available
- Marker panels are now available for
  - west African / European admixture (Smith 2004, Tian 2006)
  - Native American / European admixture (Mao 2007, Tian 2007, Price 2007)

Percent information extracted in Mexico diabetes study: ADMIXMAP with ancestry informative markers, HAPMIX with tag SNP array



Mexico City diabetes admixture mapping study: affected-only tests for linkage shows excess Eur ancestry in HLA region



Do admixture mapping studies require a control group?

- Affected-only design is the most efficient if model assumptions hold
- Control group is useful:-
  - as a source of unbiased information on allele frequencies
  - as a sanity check, and specifically to test the assumption of no ancestry state heterogeneity across the genome
  - for subsequent fine mapping
  - Control data from studies of other disease in the same population can be re-used

Successes with admixture mapping

- Detection of disease genes
  - *APOL1* identified as underlying cause of differential susceptibility to focal segmental glomerulosclerois and to (possibly misdiagnosed) hypertensive kidney disease in African-Americans (Kopp 2008)
    - apolipoprotein L1 has trypanolytic activity ? selection
  - Ancestry peak on chr 17q for sarcoidosis in African-Americans: tentatively identified as *XAF1*
- Identification of QTLs
  - Detection of a functional SNP in *SLC24A5* that accounts for ~25% of European/African difference in skin melanin content (Lamason 2005)
  - Detection of a functional SNP in *IL6R* that accounts for 33% of variance in interleukin 6 soluble receptor levels (Reich 2007)

Fine mapping in admixed populations

- For fine mapping, conditioning on locus ancestry so as to eliminate long-range signals generated by admixture
- Standard model of admixture requires minimal spacing of 0.5 cM to ensure no residual association between marker loci
  - For inference of locus ancestry (as in admixture mapping), ~3000 ancestry-informative markers are sufficient
- For fine mapping with ~500 000 tag SNPs, we can model all loci but omit feedback of information about locus ancestry from all but a subset of ~3000 AIMs

Other applications of statistical modelling of admixture / stratification

- Admixture mapping in outbred animal populations
  - livestock, heterogeneous stocks of mice
- Inferring the genetic background of an individual
  - forensic applications, restricting samples by genetic background, classification of domestic animals and livestock

Revision exercises: stratification

- How is genetic stratification defined?
- What size of stratification effect (Fst) attributable to a single component could be detected with 10^5 thinned SNPs (not in short-range LD) typed in 1000 individuals?
- What steps are necessary to eliminate artefacts when using principal components to correct for genetic stratification?