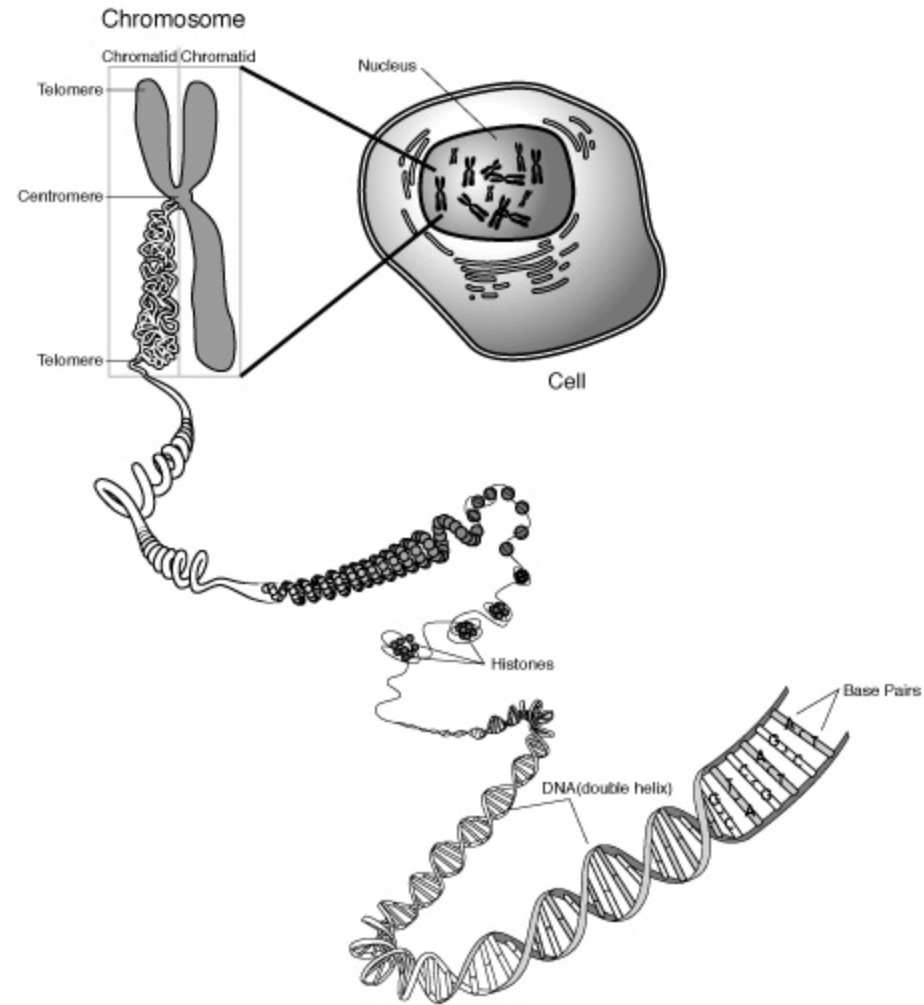
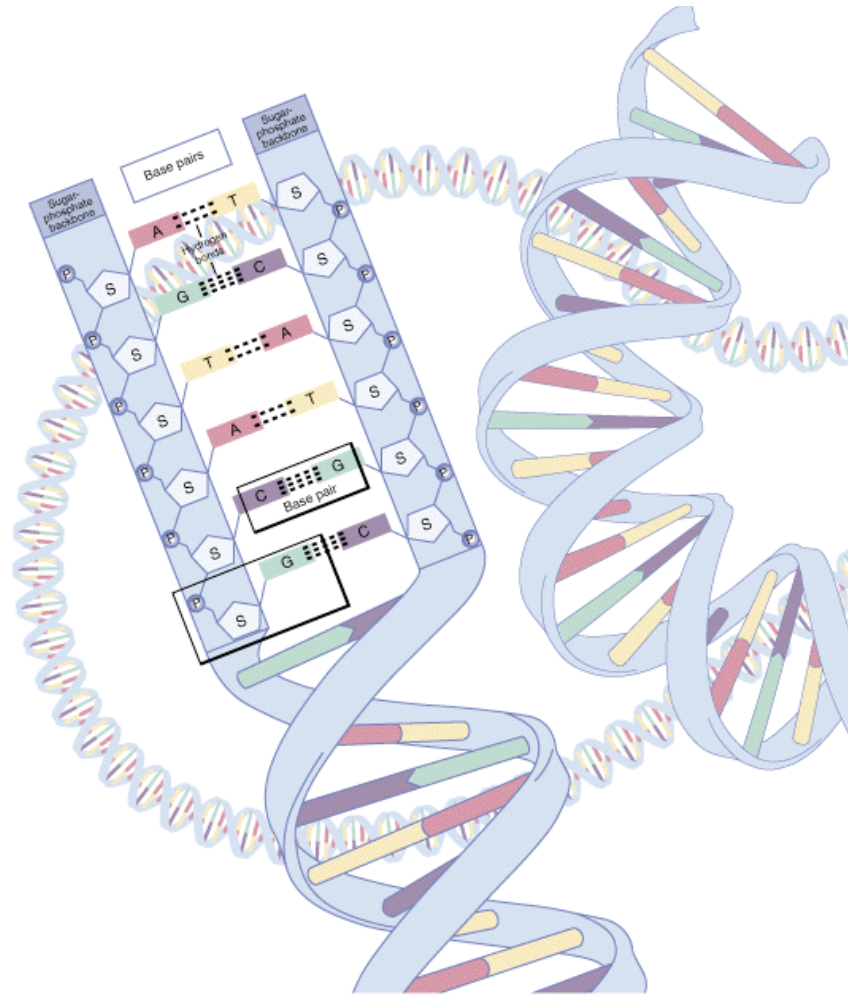


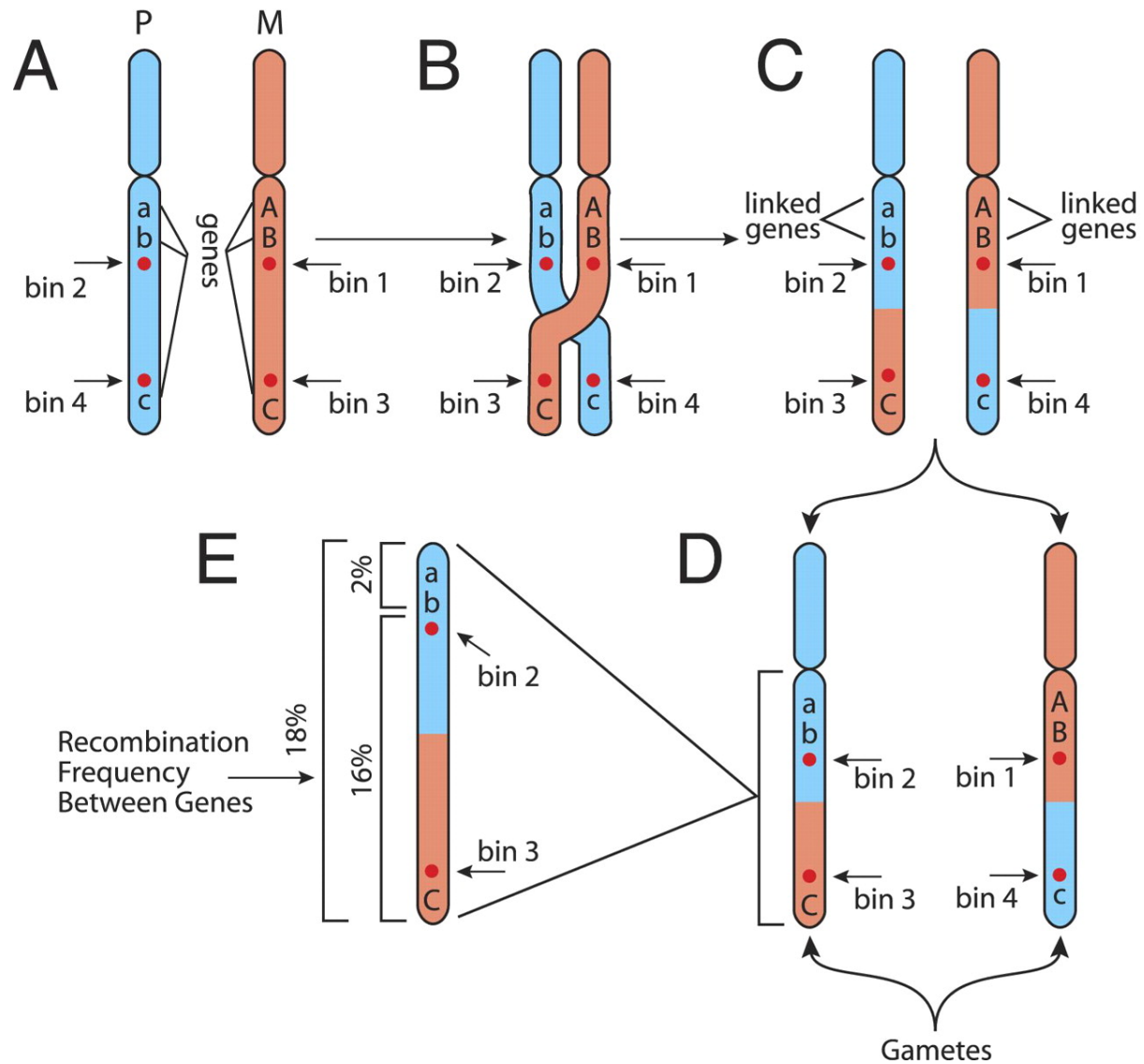
What is a chromosome?



DNA



Recombination



QTL

- Genetic variation arises from
 - Differences in DNA sequence
 - Mutation
 - Recombination
- For any given sequence polymorphism (in a population of individuals), we want to know whether individuals carrying different alleles are also different for a trait
- Locus = location in the genome
- Allele = specific variant of sequence at a location
- QTL = quantitative trait locus

Linkage Disequilibrium

- $P_{AB} = P_A * P_B$
- $D = P_{AB} - P_A * P_B$
- $r^2 = D^2 / [P_A(1-P_A)P_B(1-P_B)]$
- In a population, after one generation of random mating LD between unlinked loci = 0
- LD between linked loci decays over a number of generations
- rate of decay is proportional to the distance between the loci

Linkage vs Association

- Linkage mapping uses regular family structure to establish a direct relationship between genetic distance (cM) and r^2 between two loci
- In association mapping high r^2 implies physical linkage but there is no simple relationship between cM and r^2

QTL Mapping Methods

- Regression (fixed effects, random effects)
- Maximum Likelihood based on normal mixture models
- Single Marker, Multiple Marker, Interval
- Bayesian
- Decision Trees

Where to start?

- Regression
- Fixed effects
- Single Markers, Marker pairs
- Stepwise regression
- Bootstrap forward regression
- Permutation tests

Fixed Effects Linear Model

- dependent variable, $y = (y_1, y_2, \dots, y_n)'$
- independent variables, incidence matrix
 - for a discrete variable or factor equals columns of 1's and 0's corresponding to presence or absence of a factor level in each observation
 - for a continuous variable a single column containing the values of that variable
- p unknown parameters, β

Fixed Effects Linear Model

- The model : $y = X\beta$, solve for β
- Since $n > p$, the problem is overdetermined and there is no exact solution, so look for the least squares solution
- If X is singular the problem is also under-determined and there is no unique solution
- Standard methods exist for dealing with these issues

Computational Issues

- Because the number of observations may be in the thousands and the number of fixed effects often relatively small, it is often much faster to solve the least squares equation
- $X'X\beta = X'y$
- It is often faster to form $X'X$ from data rather than using matrix multiplication
- For testing a large number of candidate SNPs, it may be faster to test each SNP by updating the base linear model rather than solving the entire model (Partitioned Linear Model)

Hypothesis testing

- Test the hypothesis, $H: K'\beta = m$
- $F = [(K'\beta - m)'(K'GK)^{-1}(K'\beta - m)] / r_K \sigma^2$
- $G = (X'X)^{-}$, $r_K = \text{rank of the } K \text{ matrix}$
- $\sigma^2 = \text{SSE} / (n - r_X)$
- $\text{SSE} = \text{SST} - \text{SSM}$
- $\text{SST} = y'y$
- $\text{SSM} = \beta'X'y$
- Usually easy to calculate, K' has a single row, $m = 0$
- Calculate $p = P(F > F_{0.05, r_K, \text{error df}})$
- Alternatively F can be calculated from the reduction in the model SS resulting from adding a term to the model

Workflow, Basic Regression

Inputs:

Phenotype (y)
vector of doubles
length = n

Fixed Effects
matrix of doubles or integers
dimension $n \times p$ ($p \ll n$)

Genotypes
matrix of doubles or integers
dimension $n \times m$

For each genotype (one or more columns of Genotypes):

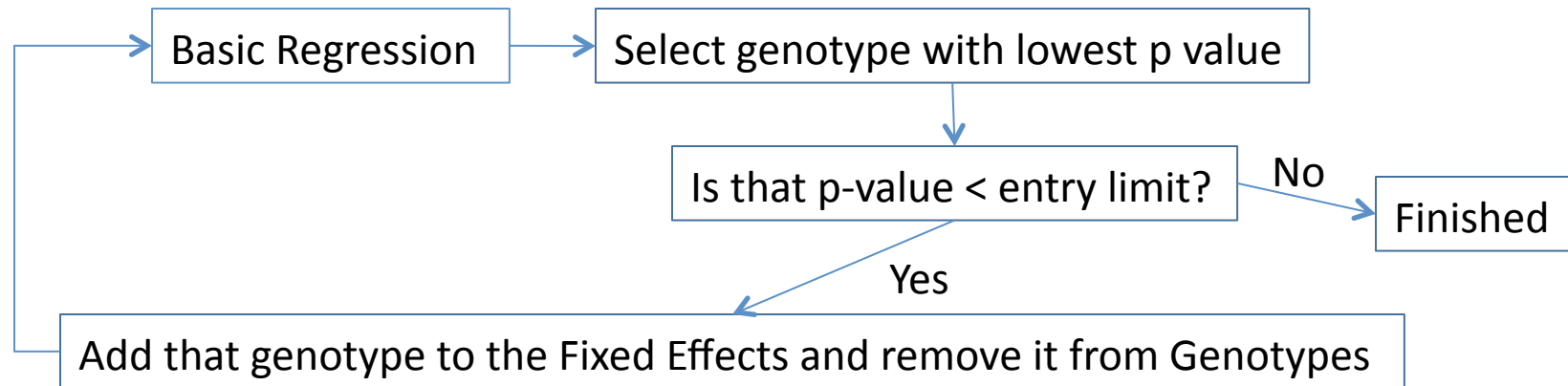
Form the incidence matrix X and $X'X$ from the Genotypes column and Fixed Effects

Solve the least squares equations: $X'X\beta = X'y$ for β

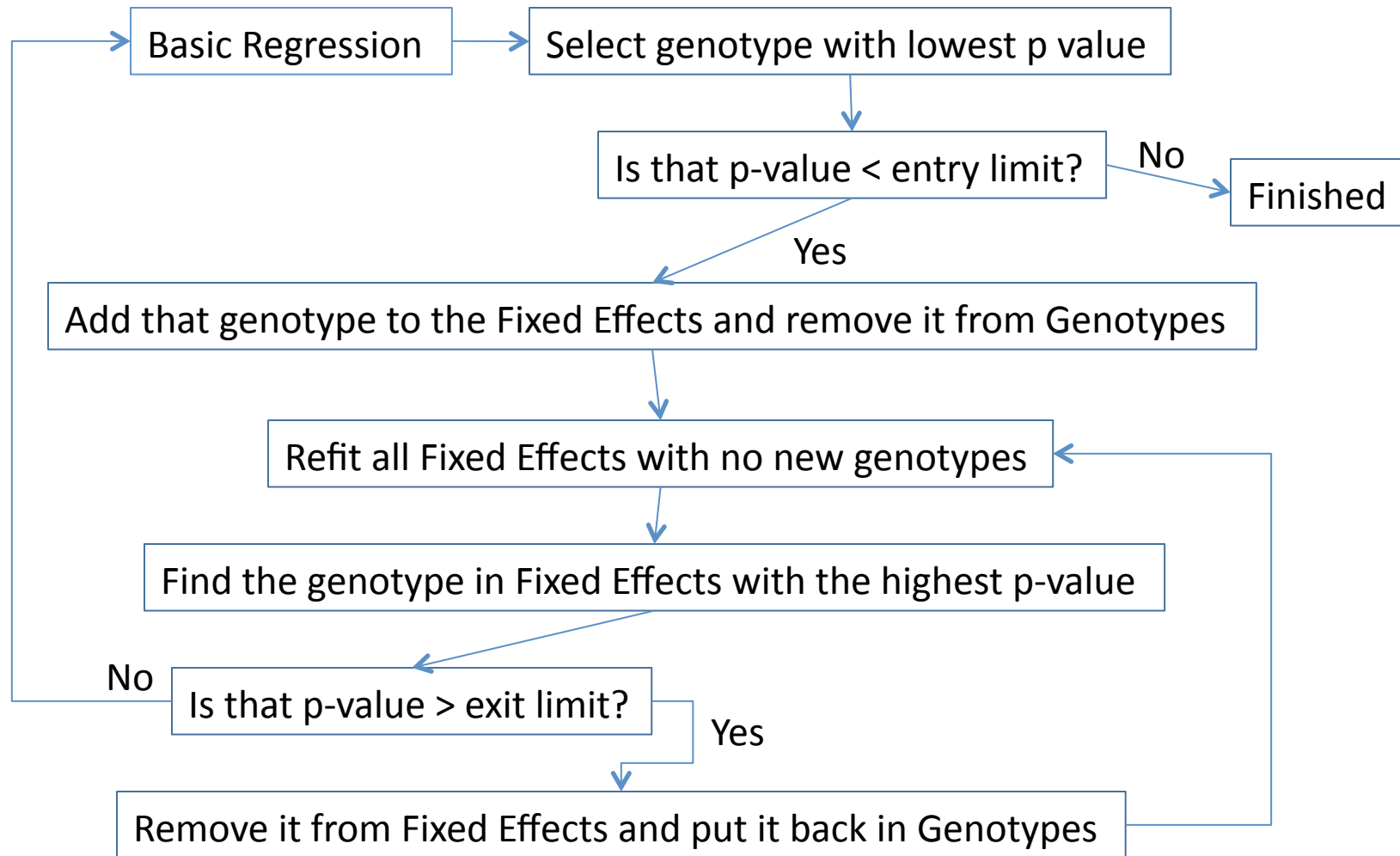
Test hypotheses about β

Output: β and results (F , p) from the hypothesis tests

Workflow Forward Regression



Workflow, Stepwise Regression



Workflow Bootstrap

