# BTRY 4830/6830: Quantitative Genomics and Genetics
# Spring 2012

Lists of Notation and Abbreviations

Version 2: November 16

This document will be added to throughout the semester as we cover more concepts (just be sure that you have the latest version!)

Note that as quantitative genomics borrows concepts from multiple fields (statistics, population genetics, etc.) we will often use the sample notation to refer to different terms. While this has the downside of adding some confusion, this will allow us to be consistent as possible with the notation in the fields where these concepts originated (which will make it easier to read the literature!). We will use the note '(multiple usage)' at the end of the definition for symbols where the symbol is used for more than one purpose.

## 1 List of Notation

### 1.1 Set Related

$\{\mathcal{A}_i, \mathcal{A}_j, ...\} \equiv$ a set (within the brackets are elements of the set)

$\mathcal{A}_i \equiv$ element $i$ of a set (we will also make use of $\mathcal{B}$)

$\subset \equiv$ a subset, i.e. $A \subset B = A$ is a subset of $B$

$\cup \equiv$ union operator ($\cup_i^\infty$ union of elements from $i$ to $\infty$)

$\cap \equiv$ intersection operator ($\cap_i^\infty$ is intersection of elements from $i$ to $\infty$)

$\emptyset \equiv$ the empty set (also represented as $\{\}$)

$\infty \equiv$ infinite ($-\infty$ is negative infinite)

$\mathbb{R} \equiv$ real numbers

## 1.2 Random Variables and Related

$S \equiv$ Sample space (also represented as $\Omega$)

$P_X(x) = Pr(X = x) \equiv$ probability mass function

$f_X(x) \equiv$ probability density function

$F_X(x) \equiv$ cumulative mass or cumulative density function

$X \equiv$ a random variable, where this can be indexed as $X_i$ for random variable $i$, often used to represent a genotype (multiple usage)

$x \equiv$ realization of a random variable $X$, used to represent an observation in a sample, often used to represent a genotype (this can be indexed as $x_i$ for observation $i$)

$\mathbf{X} \equiv$ vector or matrix of random variables

$\mathbf{x} \equiv$ vector or matrix of observations of a vector or matrix of random variables $\mathbf{X}$

$Y \equiv$ random variable often used to represent phenotype, this can be indexed as $Y_i$ for random variable $i$, (multiple usage).

$y \equiv$ realization of a random variable $Y$, often used to represent a phenotype observation (this can be indexed as $y_i$ for observation $i$).

$\mathbf{Y} \equiv$ vector or matrix of $Y_i$ random variables.

$\mathbf{y} \equiv$ vector or matrix of $y_i$ observations or random variable $Y_i$.

## 1.3 Functions and Operators

$Y = f(X) \equiv$ a function that takes $X$ to $Y$ ($f : X \to Y$ maps from $X$ to $Y$)

$f^{-1}(Y) = X \equiv$ inverse of a function, i.e. a function that takes $Y$ to $X$ ($f^{-1} : Y \to X$)

$X = f(X) \equiv$ the identity function, i.e. a function that takes $X$ and returns the same

value of $X$

$\mathbf{Y} = f(\mathbf{X}) \equiv$ a vector valued function, ($f : \mathbf{X} \rightarrow \mathbf{Y}$ maps from vector space $X$ to vector space $Y$)

$df(\mathbf{Y})/d\mathbf{X} \equiv$ derivative of a function with respect to the entire set of input variables $\mathbf{X}$, also written $f'(\mathbf{Y})$

$d^2f(\mathbf{Y})/d\mathbf{X}^2 \equiv$ second derivative of a function with respect to the entire set of input variables $\mathbf{X}$, also written $f''(\mathbf{Y})$

$\partial f(\mathbf{Y})/\partial X \equiv$ partial derivative of a function with respect to one or more (but not all) of the input variables X

$\partial f^2(\mathbf{Y})/\partial X^2 \equiv$ partial second derivative of a function with respect to one or more (but not all) of the input variables X

$\partial f^2(\mathbf{Y})/\partial X_i \partial X_j \equiv$ partial derivative of a function with respect to variables $X_i$ and $X_j$

$Pr() \equiv$ probability function

$| \equiv$ conditional operator, e.g. $Pr(\mathcal{A}_i|\mathcal{A}_j)$ the probability of even $\mathcal{A}_i$ given $\mathcal{A}_j$ has occurred

E() $\equiv$ expectation function (the range of which are two sets: the random variable and the Pr function of the random variable)

Var() or V() $\equiv$ variance function (the range of which is two sets: the random variable and the Pr function of the random variable)

Cov() $\equiv$ covariance function (the range of which are two sets: two random variable and the joint Pr function of the random variable)

corr() $\equiv$ correlation function (the range of which are two sets: two random variable and the joint Pr function of the random variable)

$T() \equiv$ a statistic (a function on a sample), where we often write $T(X) = t$, i.e. $t$ is the output of the function

$\bar{x} \equiv$ mean of the sample $\mathbf{x}$, i.e. a statistic ('bar' notation is the same for any random sample vector)

$argmax_{\theta \in \Theta}() \equiv$ argument (value) of $\theta$ that maximizes ()

$\int_{-\infty}^{\infty} \equiv$ integral operator, i.e. integral of a function from $-\infty$ to $\infty$

$\sum_{i}^{n} \equiv$ sum operator, i.e. the sum of variables (numbers) from $i$ to $n$

$\Pi_{i}^{n} \equiv$ product operator, i.e. the product of variables (numbers) from $i$ to $n$

$log() \equiv$ log function

$ln() \equiv$ natural log function, i.e. log base e

$L(\theta|y) \equiv$ Likelihood function (also written $Pr(y|\theta)$ in a Bayesian setting)

$l(\theta|y) \equiv$ log-likelihood function, i.e. $ln(L(\theta|y))$

$exp() \equiv e^{()}$

$|\mathbf{M}| \equiv$ determinant function of a matrix $\mathbf{M}$.

$tr(\mathbf{M}) \equiv$ trace function of a matrix $\mathbf{M}$.

$Pr(\theta|y) \equiv$ posterior probability.

$Pr(\theta) \equiv$ prior probability.

### 1.4 Distributions

$Binom(p) \equiv$ binomial distribution with parameter $p$

$Bern(p) \equiv$bernoulli distribution with parameter $p$

$N(\mu, \sigma^2) \equiv$ a normal distribution with parameters $\mu$ and $\sigma^2$

$\chi^2_{df} \equiv$ chi-square distribution with degrees of freedom df

$unif[a, b] \equiv$ uniform distribution with parameters $a$ and $b$

$t\text{-}dist(df) \equiv$ t-distribution with degrees of freedom $df$.

$F_{[df1,df2]} \equiv$ F-distribution with degrees of freedom $df1$ and $df2$.

$multiN(\mu, \mathbf{M}) \equiv$ a multivariate normal distribution with 'center' parameter vector $\mu$ and 'covariance' parameter matrix $\mathbf{M}$.

$multi{-}t(\mu, \mathbf{M}, df) \equiv$ a multivariate normal distribution with 'center' parameter vector $\mu$, 'covariance' parameter matrix $\mathbf{M}$, and degrees of freedom $df$.

## 1.5 Estimation and Hypothesis Testing

$\theta \equiv$ a single parameter or vector of parameters (multiple usage)

$\Theta \equiv$ the range of values a parameter $\theta$ may take

$\hat{\theta} \equiv$ estimate of a parameter (the 'hat' notation above any parameter indicates an estimate)

$\Theta_0 \equiv$ the range of values a parameter $\theta_0$ may take, restricted to the null hypothesis

$\hat{\theta_0} \equiv$ estimate of a parameter under the null hypothesis

$\Theta_1 \equiv$ the range of values a parameter $\theta_1$ may take, restricted to either the alternative hypothesis or to the union of the null and alternative hypotheses (also represented as $\Theta_A$)

$\hat{\theta_1} \equiv$ estimate of a parameter under the alternative hypothesis or under the union of the null and alternative hypotheses (also represented as $\hat{\theta}_A$)

$H_0 \equiv$ null hypothesis

$H_A \equiv$ alternative hypothesis

$\beta \equiv$ power of a statistical test (multiple usage)

$\alpha \equiv$ Type I error of a hypothesis test (multiple usage)

$c_\alpha \equiv$ The critical value of a statistic for a hypothesis test

$\Lambda \equiv$ Likelihood ratio test statistic

## 1.6 Regression Models and Genetics

$A_i \equiv$ allele $i$ at locus $A$ (also using B, C, etc.)

$A_i \rightarrow A_j \equiv$ substitution of allele $A_1$ for allele $A_2$.

$\Delta\bar{Y} \equiv$ change in the mean (average) value of a phenotype phenotype.

$S_g \equiv$ genotype sample space.

$S_P \equiv$ phenotype sample space.

$g_i \equiv$ genotype of an individual $i$, e.g. for two genotype with two alleles $g_i = A_iA_j$.

$G \equiv$ random variable representing genotypic value (for a specific genotype, a subscript is used $G_{A_iA_jB_kB_l}$.

$n \equiv$ sample size ($n \rightarrow \infty$ describes $n$ approaching infinite).

$N \equiv$ number of genotypes.

$\beta \equiv$ a parameter in a regression model, also used as a vector defined as $\beta = [\beta_\mu, \beta_a, \beta_d]$ in a multiple regression of phenotype on genotype.

$\beta_\mu \equiv$ the intercept parameter for a multiple linear regression of phenotype on genotype.

$\beta_a \equiv$ the additive parameter for a multiple linear regression of phenotype on genotype.

$\beta_d \equiv$ the dominance parameter for a multiple linear regression of phenotype on genotype.

$\beta_{aa} \equiv$ the additive by additive epistasis parameter for a multiple linear regression of phenotype on genotype.

$\beta_{ad} \equiv$ the additive by dominance epistasis parameter for a multiple linear regression of phenotype on genotype.

$\beta_{da} \equiv$ the dominance by additive epistasis parameter for a multiple linear regression of phenotype on genotype.

$\beta_{dd} \equiv$ the dominance by dominance epistasis parameter for a multiple linear regression

of phenotype on genotype.

$\sigma^2_\epsilon \equiv$ the variance parameter for the $\epsilon$ term of a multiple linear regression of phenotype on genotype.

$X_{i,a} \equiv$ the additive dummy random variable for a genotype for individual $i$ (similarly for the dominance dummy variable $X_{i,d}$).

$x_{i,a} \equiv$ the observed additive random variable for individual $i$ (similarly for the dominance dummy variable $X_{i,d}$).

$X_{a,i} \equiv$ the additive random variable for a genotype for a marker $i$ (similarly for the dominance dummy variable $X_{d,i}$).

$\mathbf{x} \equiv$ a matrix of genotypic vectors $[\mathbf{1}, \mathbf{x}_a, \mathbf{x}_b]$.

$\hat{y} \equiv$ predicted value of $y$ for a regression model.

$\epsilon \equiv$ the error term for a regression (often used as a vector including error terms $\epsilon_i$ for individuals $i$).

$X' \equiv$ genotype coding for a tag polymorphism (often a tag SNP), i.e. a genotype that is correlated with a causal genotype.

$X_A$ (or $X_B$) $\equiv$ genotype coding for locus $A$ (or $B$).

$A_i - B_j \equiv$ linkage between alleles $A_i$ and $B_j$, i.e. allele $A_i$ and $B_j$ are on the same chromosome.

$h_i \equiv$ a haplotype allele.

$fr(A_i) \equiv$ frequency of allele $A_i$ (or for a haplotype $fr(h_i)$.

$p_i \equiv$ frequency of the allele $i$, e.g. if $A_1$ is the minor allele, then $p_1 = f(A_1)$ (multiple usage).

$\gamma \equiv$ link function of a glm.

$\gamma^{-1} \equiv$ inverse of a link function of a glm.

$t \equiv$ time step of an algorithm.

$\theta^{[t]} \equiv$ the values of the parameters $\theta$ at time step $t$ in an algorithm, e.g. for the vector of glm parameters, this is $\beta^{[t]}$.

$\alpha \equiv$ equivalent to the parameter $\beta_a$ in a glm where the dominance term is not included.

$h^2_m \equiv$ marginal heritability.

$\mathbf{I} \equiv$ the identity matrix (a diagonal matrix with ones on the diagonal.

$\mathbf{Z} \equiv$ the 'incidence matrix' associated with the random effect in a mixed model.

$\mathbf{a} \equiv$ the random effect vector in a mixed model.

$\mathbf{A} \equiv$ the matrix defining the covariance for random effects in a mixed model.

$\sigma^2_a \equiv$ the matrix defining the covariance for random effects in a mixed model.

$n_{ij} \equiv$ cell counts in a table test (e.g Chi-square, Fisher's).

## 2 List of Abbreviations

r.v. $\equiv$ Random variable (or random vector)

pmf $\equiv$ Probability Mass Function

cmf $\equiv$ Cumulative Mass Function

pdf $\equiv$ Probability Density Function or Probability Distribution Function (where the latter includes both pmf's and pdf's)

cdf $\equiv$ Cumulative Density Function or Cumulative Distribution Function (where the latter includes both pmf's and pdf's)

Var $\equiv$ Variance

Cov $\equiv$ Covariance

Corr $\equiv$ Correlation

MLE ≡ Maximum Likelihood Estimate

p-val ≡ p value

LRT ≡ Likelihood Ratio Test

df ≡ Degrees of Freedom (also abbreviated d.f.)

CI ≡ Confidence Interval

ci ≡ Credible Interval

FDR ≡ False Discovery Rate.

SNP ≡ Single Nucleotide Polymorphism.

LD ≡ Linkage Disequilibrium.

GLM or glm ≡ generalized linear model.

IRLS ≡ Iterative Re-weighted Least Squares algorithm.

D ≡ deviance criterion.

QQ ≡ Quantile-Quantile plot (qq is also used).

RR ≡ Relative Risk.

OR ≡ Odds Ratio.

AR ≡ Attributable Risk.

EM ≡ Expectation-Maximization algorithm.

MCMC ≡ Markov chain Monte Carlo algorithm.