Genetics and Bioinformatics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

7 When variants become rare – sparse data

7.a Customizing GWAs for rare variants association analyses

7.b From GWAs to Sequence Analyses: recognizing and comparing words

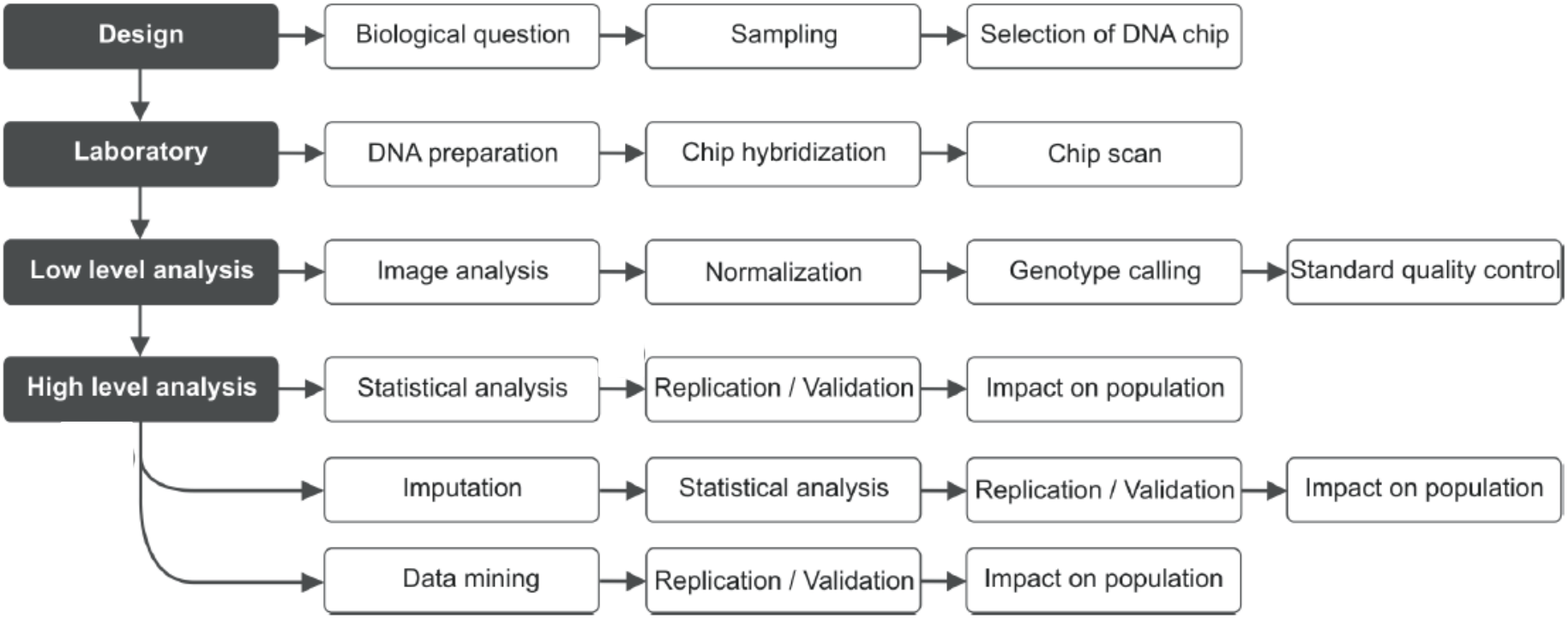
8 When effects become non-independent

Biological vs statistical epistasis (future class)

7 When variants become rare – sparse data – expanding GWAs

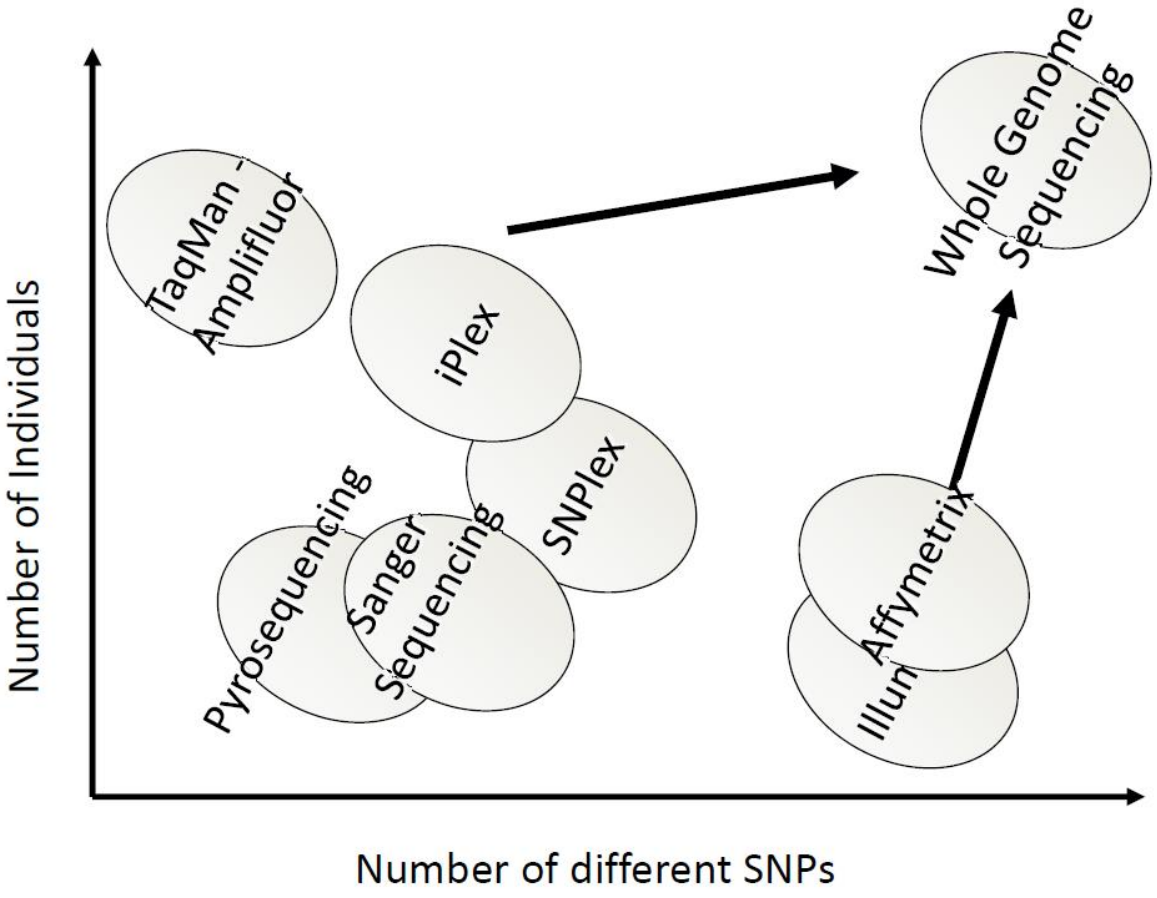
(slide Doug Brutlag 2010)

Detailed flow of a genome-wide association study

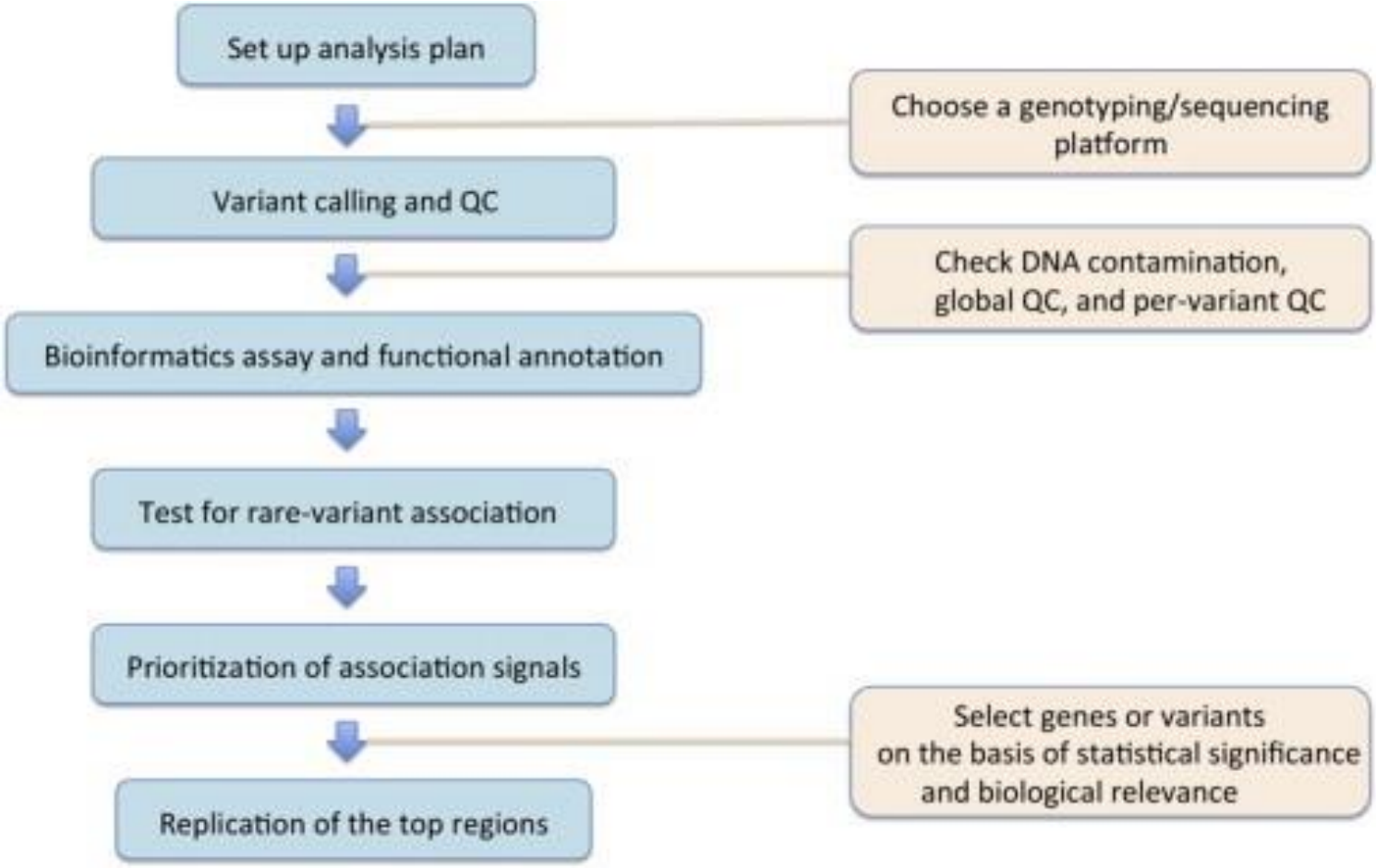


(Ziegler 2009)

From arrays to sequence data

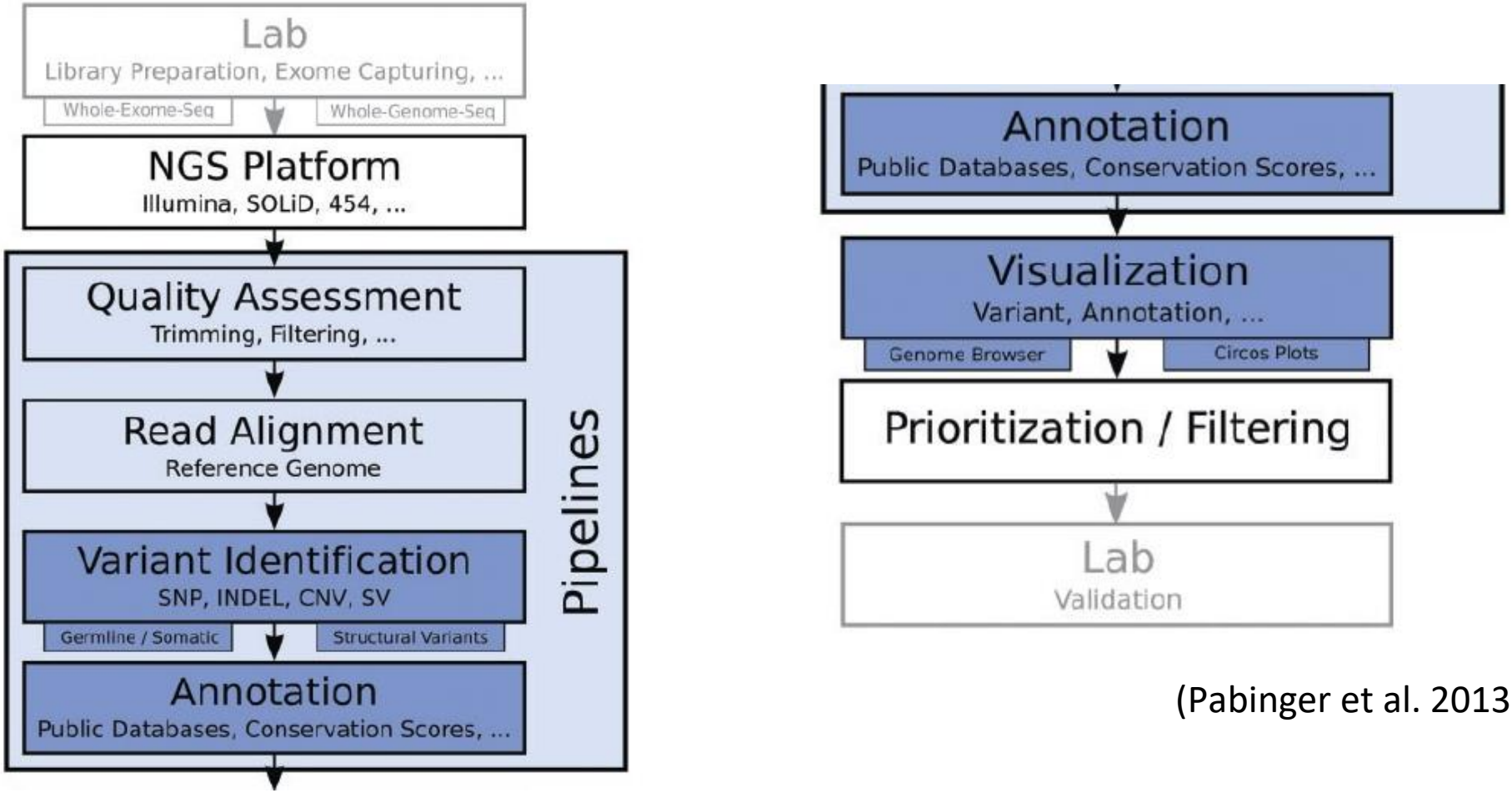


Work flow genome-wide association study with sequence data



(Lee et al. 2014)

Common workflow for whole-exome and whole genome sequencing



(Pabinger et al. 2013)

A primer on rare variant association testing

Fan Li

BIOSTAT 790

January 28, 2016

Outline

Overall goal: to understand the necessity of identifying rare variants and the proposed statistical methods for rare-variant association testing.

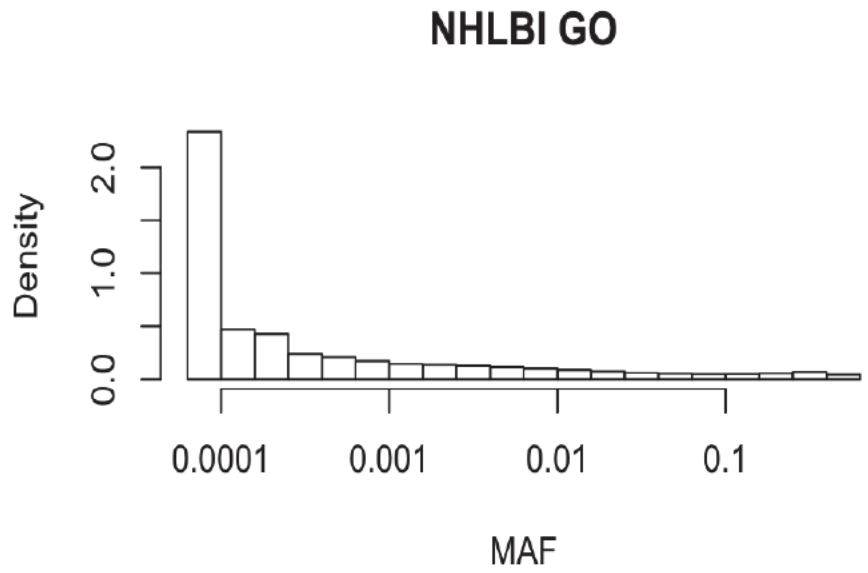
- Rationale for studying rare variants (complex trait)
- Sequencing and study designs
- Rare-variant association tests
- Summary

Common vs rare variants

- MAF: frequency at which the least common allele occurs in population
- Common variants: MAF $\geq 5\%$
- Low frequency variants: $0.5\% \leq$MAF$< 5\%$
- Rare variants: MAF $\leq 0.5\%$

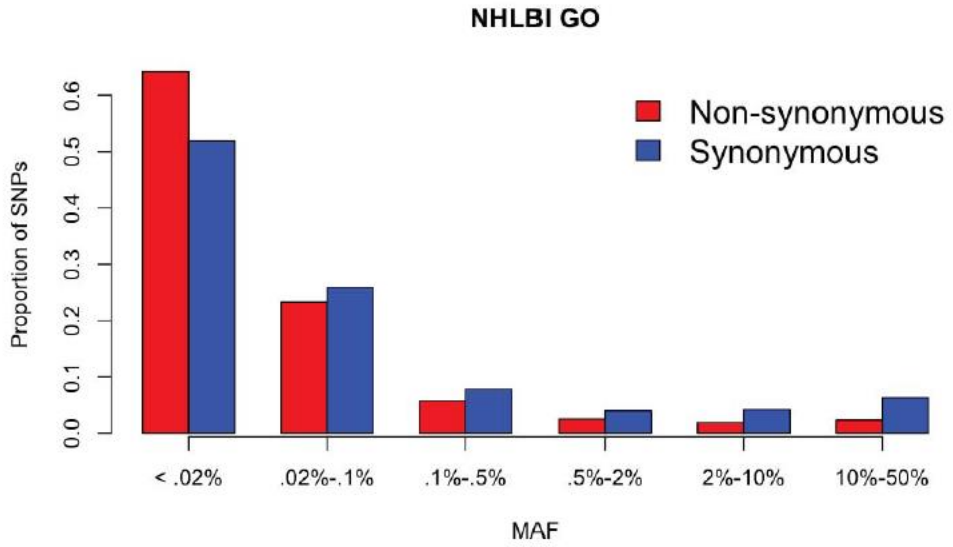
Why rare variants?

- Most of human variants are rare



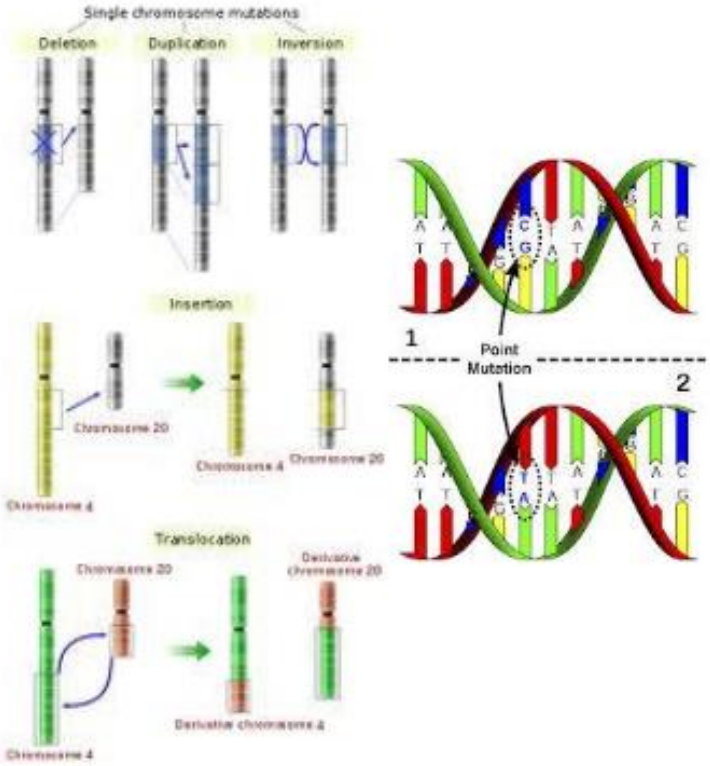
Why rare variants?

- Functional variants tend to be rare



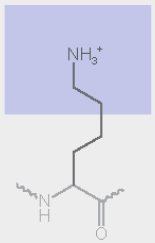
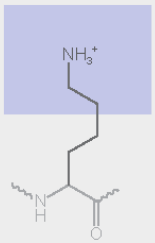
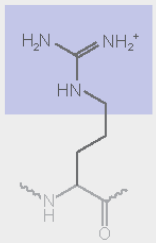
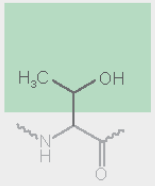
Recall

- Mutations are changes to the genetic information of the cell. There are 2 different types of mutations
  - large scale – Chromosome sections
    - deletions, inversions, translocations, & polyploidy
  - point mutations – single nucleotide
    - Each mutation carries with it the ability to alter the phenotype of the individual and, if in the cells that create the gametes, all the offspring of the individual.



Recall

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**

basic
polar

Why rare variants?

- Further, since common variants explained limited variation in the trait ...
- Some argued rare variants could explain additional trait variability
- Advancement of sequencing technology (NGS), reduction in cost

Challenges

- Require cost-effective study designs to genotype many individuals

  It can be shown that at least $\log(1-\theta)/[2\log(1-MAF)]$ individuals are needed to observe a variant with no less than θ chance. For $\theta = 99.9\%$, we have

MAF(%)	10	1	0.1	0.01
Minimum sample size	33	344	3453	34537

- Classical single-variant tests, developed for common variants detection, are underpowered
- Multiple testing

Challenges

- A variant – genetic association test implies filling in the table below and performing a chi-squared test for independence between rows and columns

	AA	**Aa**	**aa**
Cases			
Controls			

Sum of entries = cases+controls

- How many observations do you expect to have two copies of a rare allele? Example: MAF for a = 0.001 → expected aa frequency is 0.001 x 0.001 or 1 out of 1 million

- **In a chi-squared test of independence setting** (comparing two variables in a contingency table to see if they are related):
  When MAF <<< 0.05 then some cells above will be
  sparse and large-sample statistics (classic chi-squared tests of independence) will no longer be valid. This is the case when there are less than 5 observations in a cell

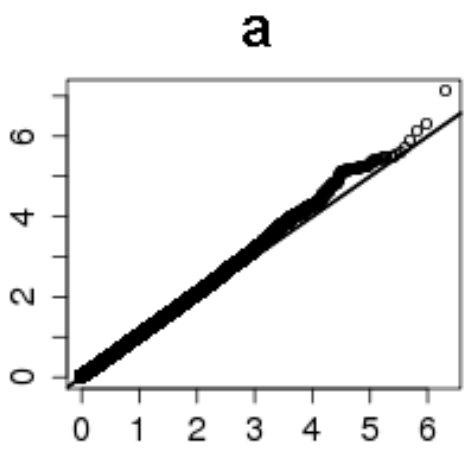
$X^2 = \sum_{all\ cells\ i} \frac{(O_i - E_i)^2}{E_!}$ (contrasting Observed minus Expected)

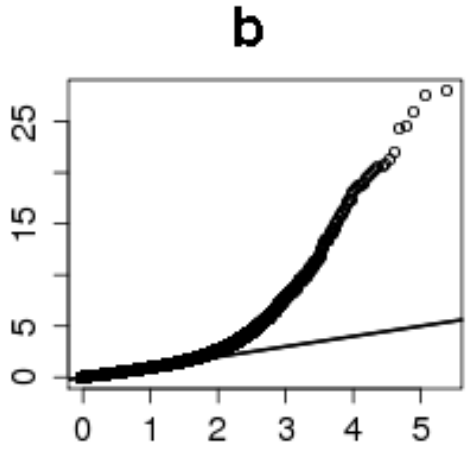
- **In a regression framework:**
  The minimum number of observations per independent variable should be 10, using a guideline provided by Hosmer and Lemeshow (Applied Logistic Regression, one of the main resources for Logistic Regression)

Increased false positive rates

Q-Q plots from GWAS data, unpublished

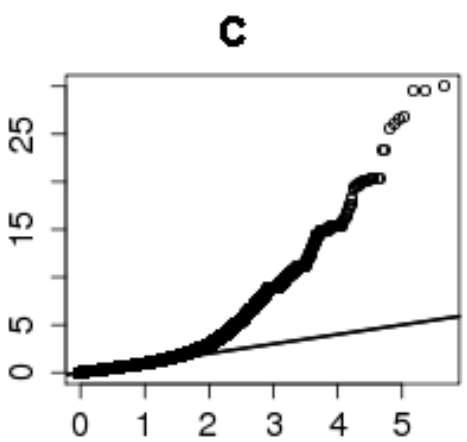
N=~2500
MAF>0.03

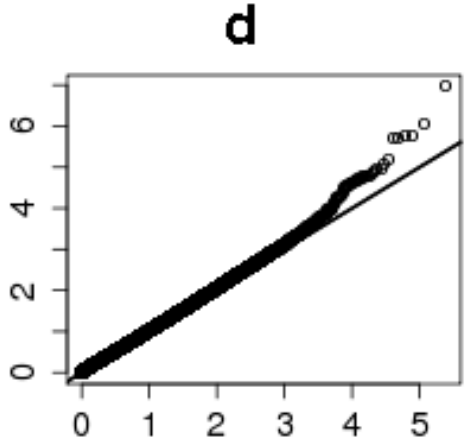




N=~2500
MAF<0.03

N=~2500
MAF<0.03
Permuted





N=50000
MAF<0.03
Bootstrapped

Remediation: do not look at a single variant at a time, but collapse

- Rationale for aggregation tests
  - Alpha level of 0.05, corrected by number of bp in the genome= $1.6*10^{-11}$
  - One needs VERY LARGE samples sizes in order to be able to reach that level, even if you find “the variant”.
- Remedy = aggregate / pool variants
  - Requires specification of a so-called “region of interest” (ROI)
  - A ROI can be anything really:
    - Gene
    - Locus
    - Intra-genic area
    - Functional set

Remediation: design alternatives to deep sequencing

- Low-depth whole-genome sequencing: sequencing depth refers to the average number of reads that cover each base; limited accuracy
- Exome sequencing: limited to exome
- High-priority region sequencing: limited to the target region
- ...

In summary, either the sequenced range or accuracy is compromised

How do aggregation tests for (rare) multiple variants work?

- Region-based: gene, regulatory region
- Identify multiple genetic variants within a region
- Evaluate the joint effects of these variants while adjusting for covariates
- Caution: These tests rely on assumptions for genetic model (e.g.: mode of inheritance), and the power depends on the true disease model $h(\mu(Y))$

Corresponding regression models

- n subjects ($i = 1, \ldots, n$)
- m variants in a region
- Allele counts in a region $\mathbf{G}_i = (G_{i1}, \ldots, G_{im})'$, ($G_{ij} = 0, 1, 2$)
- q-dimensional covariates $\mathbf{X}_i$ (age, gender, PC scores etc.)
- The disease model is given by a GLM

$$h(\mu(Y_i)) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i \qquad (1)$$

- Now the interest is in the null of no genetic-region effect:

$$H_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)' = \mathbf{0}_{m\times 1}$$

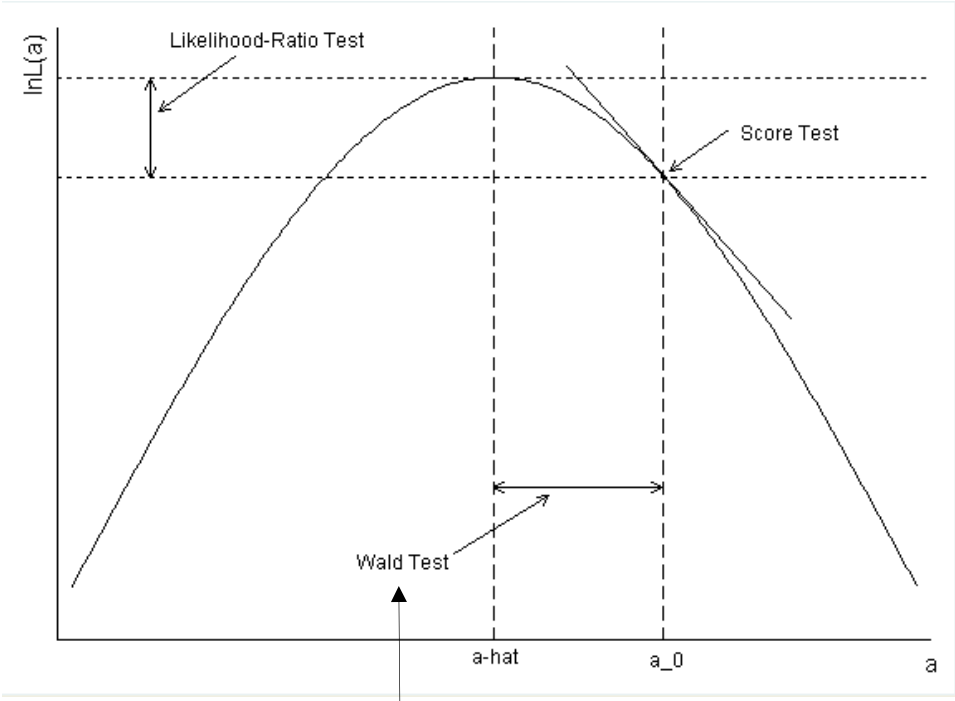
The score statistic

- All the tests are in some sense a modification of the score test for the previous H_0
- Under H_0, the score statistic for a single variant j (marginally)

$$S_j = \sum_{i=1}^{n} G_{ij}(Y_i - \hat{\mu}_i),$$

where $\hat{\mu}_i$ is estimated under the null model with $\boldsymbol{\beta}$ set to zero vector

The score test vs the Wald Test



$$\frac{(\widehat{\theta}-\theta_0)^2}{\mathrm{var}(\hat{\theta})}$$

Burden tests

- Collapse information on multiple genetic variants into a single genetic score
- Essentially an association test between the score and trait
- Define a weight for each variant $\mathbf{w} = (w_1, \ldots, w_m)'$, the score is developed as

$$\begin{aligned} h(\mu(Y_i)) &= \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i \\ &= \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \tilde{\beta} \underbrace{\mathbf{w}'\mathbf{G}_i}_{scalar\ C_i} \end{aligned} \qquad (2)$$

- Under $H_0 : \tilde{\beta} = 0$, the score statistic is $Q_{burden} = (\sum_{j=1}^{m} w_j S_j)^2$

Burden tests

- If $\mathbf{w} = \mathbf{1}$, we can collapse rare variants the following way

Y	G_1	G_2	G_3	G_4		C
1	1	0	0	0		1
1	0	1	0	0		1
1	0	0	1	1	→	2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
0	0	0	0	0		0
0	0	0	0	0		0
0	0	0	0	0		0

Burden tests

- Choice of **w** accommodates different assumptions about disease mechanism
- e.g., the cohort allelic sums test (CAST)

$$C_i = \begin{cases} 1 & \text{when } \mathbf{1}'\mathbf{G}_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

- Limitation: strong assumption about the same direction/magitude of effect (post to weight adjustment); loss of power

Key features of burden tests

- Several flavors exist:
  - In general they all combine rare variants into a single genetic (risk) score
    Example: Combine minor allele counts into a single risk score (dominant genetic model)
  - Weighted or unweighted versions (f.i., to prioritize certain variant types, based on predictions about damaging effect)

Some problems with burden tests

- Problem 1: When high linkage disequilibrium (LD) [allelic non-independence] exists in the “region”, combined counts may be artificially elevated
- Problem 2: Assumes that all rare variants in a set are causal and associated with a trait in the same direction
  - Counter-examples exist for different directionality (e.g. autoimmune GWAs)
  - Violations of this assumption leads to power loss

Adaptive burden tests

- To obtain tests that are robust to null variants and allow for different effect directions
- Let the data speak!
- e.g. the data-adaptive sum test (aSum)
  - Estimate direction of each variant in marginal models
  - Use the burden test framework with $w_j = 1$ if the coef is likely to be positive and $w_j = -1$ otherwise
  - Require permutation (How?) to obtain the null distribution
  - Further modification based on model-selection allowing for zero weight
- Limitations: although more robust, marginal models are unstable; permutation requires extensive calculation

Variance components tests

- Is there another way to pool/group the rare variants in a region?
- Yes, resort to random-effects models
- To evaluate the distribution of genetic effects for a group of variants
- Suppose $\beta_j \sim N(0, w_j^2\tau)$, $corr(\beta_j, \beta_k) = \rho$
  - e.g., the widely-used sequence kernel association test (SKAT, $\rho = 0$) tests $H_0 : \tau = 0$
  - $Q_{SKAT} = \sum_{j=1}^{m} w_j^2 S_j^2$, a weighted sum of squares of single variant scores, approx follows a mixture of Chi-squared dist
  - Robust to different directions of effects, but ...
  - Can lead to inflated test size in small effective sample size

Omnibus tests

- To achieve robust power
- Often referred to as "combined tests"
- How to combine different tests?
- Fisher's combination method

$$Fisher = -2\log(p_{SKAT}) - 2\log(p_{burden})$$

- Combining test statistic

$$Q_\rho = (1-\rho)Q_{SKAT} + \rho Q_{burden}, \quad \rho \in [0,1]$$

Omnibus tests

- Limitation: it might have lower power than SKAT or burden tests if the assumption underlying one of these tests are largely true
- For unknown genetic architecture, this is an attractive choice

General comments on aggregation tests

- The tests are designed to boost power assuming the rare variants can be grouped together
- This point is shown by simulation work by Li and Neal, 2008
- Power loss occurs (relative to single-variant tests) when only a very few of the variants are associated with the trait and when many variants have no effects
- e.g., Liu et al. studied the association between blood lipids and BCAM and CD300LG, but found weaker signal using gene-level test than single-variant test

Meta-analysis

- Combine data from multiple studies
  - Rare variants association detection requires large sample
- Popular frameworks combine score statistic from different studies instead of combing p-values
  - only requires summary statistics
  - allows study-specific covariates
- Methods should account for heterogeneity of genetic effects (how? see Lee et al 2013 AJHG) across studies to increase power (diff in ancestries)

Variant selection: which variants to use?

- Can use all the variants
- Obtain a refined subset on the basis of MAF, impact of amino acid sequence
- A subset based on the predicted functional role of variants (with bioinformatics tool)

REVIEW

Rare-Variant Association Analysis: Study Designs and Statistical Tests

Seunggeung Lee,[1] Gonçalo R. Abecasis,[1] Michael Boehnke,[1] and Xihong Lin[2,*]

Despite the extensive discovery of trait- and disease-associated common variants, much of the genetic contribution to complex traits remains unexplained. Rare variants can explain additional disease risk or trait variability. An increasing number of studies are underway to identify trait- and disease-associated rare variants. In this review, we provide an overview of statistical issues in rare-variant association studies with a focus on study designs and statistical tests. We present the design and analysis pipeline of rare-variant studies and review cost-effective sequencing designs and genotyping platforms. We compare various gene- or region-based association tests, including burden tests, variance-component tests, and combined omnibus tests, in terms of their assumptions and performance. Also discussed are the related topics of meta-analysis, population-stratification adjustment, genotype imputation, follow-up studies, and heritability due to rare variants. We provide guidelines for analysis and discuss some of the challenges inherent in these studies and future research directions.

(Lee et al. 2014)

Other tests

	Description	Methods	Advantage	Disadvantage	Software Packages[a]
Burden tests	collapse rare variants into genetic scores	ARIEL test,[50] CAST,[51] CMC method,[52] MZ test,[53] WSS[54]	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum,[55] Step-up,[56] EREC test,[57] VT,[58] KBAC method,[59] RBT[60]	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT,[61] SSU test,[62] C-alpha test[63]	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT

(Lee et al. 2014)

Other tests

Combined tests	combine burden and variance-component tests	SKAT-O,[64] Fisher method,[65] MiST[66]	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT
EC test	exponentially combines score statistics	EC test[67]	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal	no software is available yet

Abbreviations are as follows: ARIEL, accumulation of rare variants integrated and extended locus-specific; aSum, data-adaptive sum test; CAST, cohort allelic sums test; CMC, combined multivariate and collapsing; EC, exponential combination; EPACTS, efficient and parallelizable association container toolbox; EREC, estimated regression coefficient; GRANVIL, gene- or region-based analysis of variants of intermediate and low frequency; KBAC, kernel-based adaptive cluster; MiST, mixed-effects score test for continuous outcomes; MZ, Morris and Zeggini; RBT, replication-based test; Rvtests, rare-variant tests; SKAT, sequence kernel association test; SSU, sum of squared score; VAT, variant association tools; VT, variable threshold; and WSS, weighted-sum statistic.
[a]More information is given in Table 3.

(Lee et al. 2014)

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 15 September 2014
doi: 10.3389/fgene.2014.00323

A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required

Carmen Dering[1], Inke R. König[1], Laura B. Ramsey[2], Mary V. Relling[2], Wenjian Yang[2] and Andreas Ziegler[1,3,4]*

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany
[2] Pharmaceutical Department, St. Jude Children's Research Hospital, Memphis, TN, USA
[3] Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck, Germany
[4] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Edited by:
Daniel C. Koboldt, Washington University in St. Louis, USA

Reviewed by:
Michelle Lacey, Tulane University, USA
Jian Li, Tulane University, USA

***Correspondence:**
Andreas Ziegler, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
e-mail: ziegler@imbs.uni-luebeck.de

The advent of next generation sequencing (NGS) technologies enabled the investigation of the rare variant-common disease hypothesis in unrelated individuals, even on the genome-wide level. Analysis of this hypothesis requires tailored statistical methods as single marker tests fail on rare variants. An entire class of statistical methods collapses rare variants from a genomic region of interest (ROI), thereby aggregating rare variants. In an extensive simulation study using data from the Genetic Analysis Workshop 17 we compared the performance of 15 collapsing methods by means of a variety of pre-defined ROIs regarding minor allele frequency thresholds and functionality. Findings of the simulation study were additionally confirmed by a real data set investigating the association between methotrexate clearance and the *SLCO1B1* gene in patients with acute lymphoblastic leukemia. Our analyses showed substantially inflated type I error levels for many of the proposed collapsing methods. Only four approaches yielded valid type I errors in all considered scenarios. None of the statistical tests was able to detect true associations over a substantial proportion of replicates in the simulated data. Detailed annotation of functionality of variants is crucial to detect true associations. These findings were confirmed in the analysis of the real data. Recent theoretical work showed that large power is achieved in gene-based analyses only if large sample sizes are available and a substantial proportion of causing rare variants is present in the gene-based analysis. Many of the investigated statistical approaches use permutation requiring high computational cost. There is a clear need for valid, powerful and fast to calculate test statistics for studies investigating rare variants.

(Dering et al. 2014)

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 15 September 2014
doi: 10.3389/fgene.2014.00323

A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required

Carmen Dering[1], Inke R. König[1], Laura B. Ramsey[2], Mary V. Relling[2], Wenjian Yang[2] and Andreas Ziegler[1,3,4]*

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, Ger
[2] Pharmaceutical Department, St. Jude Children's Research Hospital, Memphis, TN, USA
[3] Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck, Germany
[4] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South A

Edited by:
Daniel C. Koboldt, Washington University in St. Louis, USA

Reviewed by:
Michelle Leavy, Tulane University, USA
Jian Li, Tulane University, USA

***Correspondence:**
Andreas Ziegler, In
Medizinisch
Stat
Univ
Schles
Lübeck,
2356?
e-mail: zie imbs.uni-luebeck.de

The advent of next generatio enabled the investigation of the rare variant-co elated individuals, even on the genome-wide le equires tailored statistical methods as single ma entire class of statistical methods collapses rar of interest (ROI), thereby aggregating rare variants. udy using data from the Genetic Analysis Workshop 17 ormance of 15 collapsing methods by means of a variety of regarding minor allele frequency thresholds and functionality. Findings ulation study were additionally confirmed by a real data set investigating the sociation between methotrexate clearance and the *SLCO1B1* gene in patients with acute lymphoblastic leukemia. Our analyses showed substantially inflated type I error levels for many of the proposed collapsing methods. Only four approaches yielded valid type I errors in all considered scenarios. None of the statistical tests was able to detect true associations over a substantial proportion of replicates in the simulated data. Detailed annotation of functionality of variants is crucial to detect true associations. These findings were confirmed in the analysis of the real data. Recent theoretical work showed that large power is achieved in gene-based analyses only if large sample sizes are available and a substantial proportion of causing rare variants is present in the gene-based analysis. Many of the investigated statistical approaches use permutation requiring high computational cost. There is a clear need for valid, powerful and fast to calculate test statistics for studies investigating rare variants.

Aggregation tests typically do not perform well

(Dering et al. 2014)

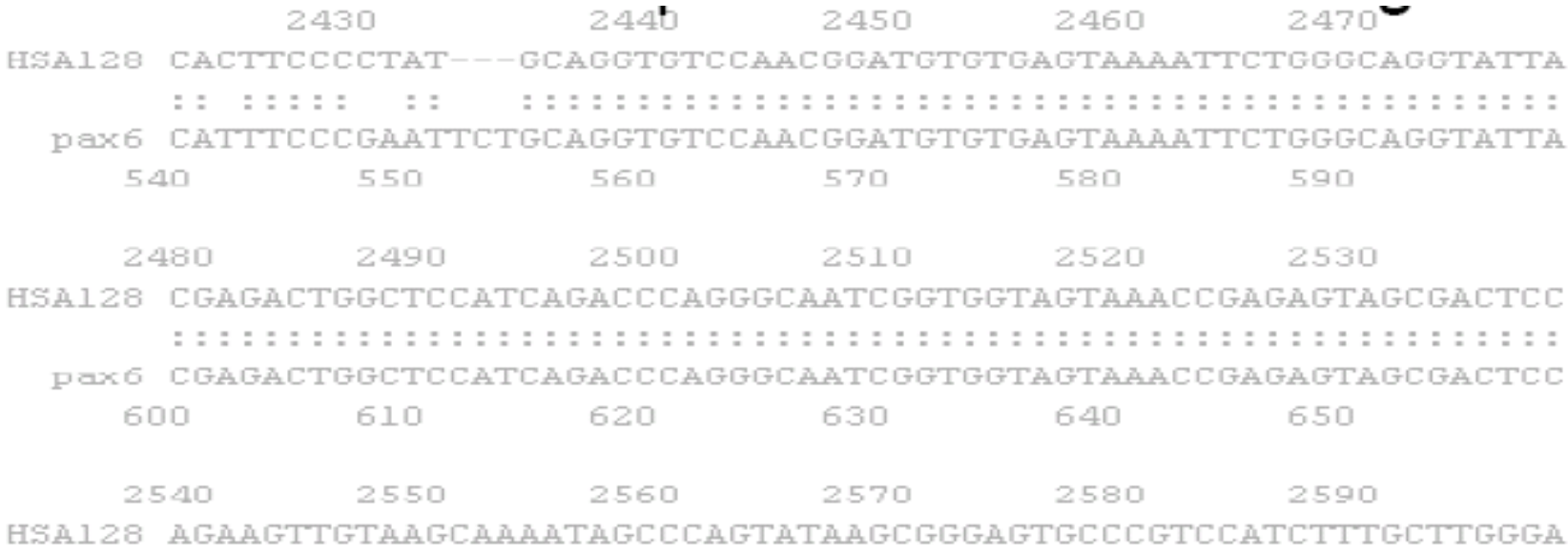
Which tests to use?

- First acknowledge that relative performance depends on the unknown disease architecture
- Use available prior information
  - the region has a large fraction of causal rare variants, majority increase disease risk – burden tests
  - exist both risk-increasing and risk-decreasing variants – variance-component tests
- If no prior information, one can try multiple methods or use the omnibus test

8 DNA Sequence Analyses

Comparing multiple sequences (see practical session)

- After collection of a set of related sequences, how can we compare them as a set?
- How should we line up the sequences so that the most similar portions are together?
- What do we do with sequences of different length?



Sequence alignment

vs

Investigating frequencies of occurrences of words

Introduction

- Words are short strings of letters drawn from an alphabet
- In the case of DNA, the set of letters is A, C, T, G
- A word of length k is called a k-word or k-tuple
- Differences in word frequencies help to differentiate between different DNA sequence sources or regions
- Examples: 1-tuple: individual nucleotide; 2-tuple: dinucleotide; 3-tuple: codon
- The **distributions** of the nucleotides over the DNA sequences have been studied for many years → hidden correlations in the sequences (e.g., CpGs)

Probability distributions

Probability is the science of uncertainty

1. Rules → data: given the rules, describe the likelihoods of various events occurring
2. Probability is about prediction – looking forwards
3. Probability is mathematics

Statistics is the science of data

1. Rules ← data: given only the data, try to guess what the rules were. That is, some probability model controlled what data came out, and the best we can do is guess – or approximate – what that model was. We might guess wrong, we might refine our guess as we obtain / collect more data
2. Statistics is about looking backward. Once we make our best *statistical guess* about what the probability model is (what the rules are), based on looking backward, we can then use that probability model to predict the future
3. Statistics is an art. It uses mathematical methods but it is much more than mathematics alone
4. The purpose of statistics is to make inference about unknown quantities from samples of data.

Statistics is the science of data

- Probability distributions are a fundamental concept in statistics.
- Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set.
- For this chapter, the distribution does not always need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.
- Simulation studies: one way to obtain empirical evidence for a probability model

Expected values and variances

- Mean and variance are two important properties of real-valued random variables and corresponding probability distributions.
- The "mean" of a discrete random variable *X* taking values $x_1, x_2, \ldots$ (denoted EX (or E(X) or E[X]), where E stands for expectation, which is another term for mean) is defined as:

$$\mathrm{E(X)} = \sum_i x_i \, P(X = x_i)$$

  - $\mathrm{E(X_i)} = 1 \times p_A + 0 \times (1 - p_A)$ if $\mathrm{x_i} = \mathrm{A}$ or $\{\text{another letter}\}$
  - If $Y = c\,X$, then E(Y) = c E(X)
  - $\mathrm{E(X_1 + \ldots + X_n)} = \mathrm{E(X_1)} + \ldots + \mathrm{E(X_n)}$
- Because X_i are assumed to be independent and identically distributed (iid):

$$\mathrm{E(X_1 + \ldots + X_n)} = \mathrm{n\ E(X_1)} = \mathrm{n}\ p_A$$

Expected values and variances

- The idea is to use squared deviations of X from its center (expressed by the mean). Expanding the square and using the linearity properties of the mean, the Var(X) can also be written as:

$$Var(X) = E(X^2) - [E(X)]^2]$$

  - If Y=c X then Var (Y) = c^2 Var (X)
  - The variance of a sum of independent random variables is the sum of the individual variances

- For the random variables X_i taking on values A or sth else:
  Var (X_i) = $[1^2 \times p_A + 0^2 \times' (1 - p_A)] - p_A^2 = p_A(1 - p_A)$
  Var (N) = n Var (X_1) = $np_A(1 - p_A)$

Expected values and variances

- The expected value of a random variable X gives a measure of its location. Variance is another property of a probability distribution dealing with the spread or variability of a random variable around its mean.

$$Var(X) = E\left([X - E(X)]^2 \right)$$

  - The positive square root of the variance of X is called its standard deviation sd(X)

Independence

- Discrete random variables X_1, …, X_n are said to **be independent** if for any subset of random variables and actual values, the joint distribution equals the product of the component distributions

Is independence equivalent to correlation 0?

Occurrences of 1-letter words

Assumptions

- Notation for the output of a random string of *n* bases may be: L_1, L_2, …, L_n (L_i = base inserted at position *i* of the sequence)
  - The values l_j for L_j will come from a set χ (with J possibilities)
  - For a DNA sequence, J=4 and $\chi = \{A, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$
- Simple rules specifying a probability model:
  - First base in sequence is either A, C, T or G with prob p_A, p_C, p_T, p_G
  - Suppose the first r bases have been generated, while generating the base at position r+1, no attention is paid to what has been generated before.

- Then we can actually generate A, C, T or G with the probabilities above
- According to our simple model, the L_i are independent and hence

$$P(L_1=l_1, L_2=l_2, \ldots, L_n=l_n)=P(L_1=l_1)\,P(L_2=l_2)\ldots P(L_n=l_n)$$

- If p_j is the prob that the value (realization of the random variable L) l_j occurs, then
  - $p_1, \ldots, p_J \geq O$ and $p_1 + \ldots + p_J = 1$
- The **probability distribution** (probability mass function) of L is given by the collection $p_1, \ldots, p_J$
  - $P(L=l_j) = p_j$, $j=1, \ldots, J$
- The probability that an event S occurs (subset of χ) is $P(L \in S) = \sum_{j:l_j \in S} (p_j)$

Probability distributions of interest

- What is the probability distribution of the number of times a given pattern occurs in a random DNA sequence L_1, …, L_n? Simple pattern = “A”
  - New sequence X_1, …, X_n:

$$X_i=1 \text{ if } L_i=A \text{ and } X_i=0 \text{ else}$$

  - The number of times N that A appears is the sum

$$N=X_1+\ldots+X_n$$

  - The prob distr of each of the X_i:

$$P(X_i=1) = P(L_i=A)=p_A$$

$$P(X_i=0) = P(L_i=C \text{ or } G \text{ or } T) = 1 - p_A$$

- What is a “typical” value of N?
  - Depends on how the individual X_i (for different *i*) are interrelated

The binomial distribution

- The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. These outcomes are appropriately labeled "success" and "failure". The binomial distribution is used to obtain the probability of observing x successes in a fixed number of trials, with the probability of success on a single trial denoted by p. The binomial distribution assumes that p is fixed for all trials.
- The formula for the binomial probability mass function is :

$$P(N = j) = \binom{n}{j} p^j (1-p)^{n-j}, \text{ j = 0,1, ...,n}$$

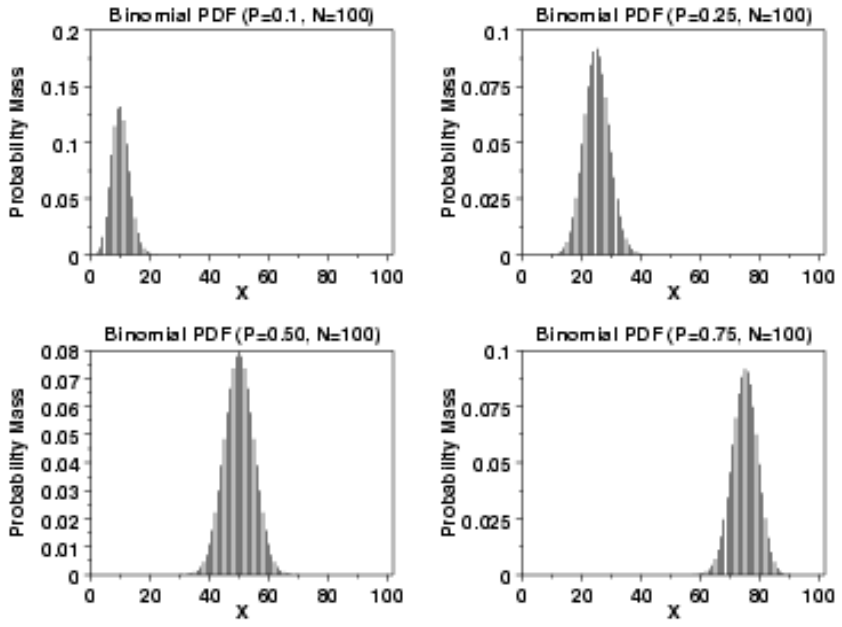
with the binomial coefficient $\binom{n}{j}$ determined by

$$\binom{n}{j} = \frac{n!}{j!\,(n-j)!},$$

and *j!=j(j-1)(j-2)…3.2.1, 0!=1*

The binomial distribution

- The mean is n*p* and the variance is n*p(1-p)*
- The following is the plot of the binomial probability density function for four values of *p* and n = 100.



Simulating from probability distributions

- The idea is that we can study the properties of the distribution of N when we can get our computer to output numbers N_1, …, N_k having the same distribution as N
  - We can use the sample mean to estimate the expected value E(N):

$$\bar{N} = (N_1 + \ldots + N_k)/k$$

  - Similarly, we can use the sample variance to estimate the true variance of N:

$$s^2 = \frac{1}{k-1}\sum_{i=1}^{k}(N_i - \bar{N})^2$$

Why do we use (k-1) and not k in the denominator?

Simulating from probability distributions

- What is needed to produce such a string of observations?
  - Access to **pseudo-random numbers**: random variables that are uniformly distributed on (0,1): any number between 0 and 1 is a possible outcome and each is equally likely
- In practice, simulating an observation with the distribution of X_1:
  - Take a uniform random number u
  - Set X_1=1 if $U \leq p \equiv p_A$ and 0 otherwise.
  - Why does this work? … $P(X_1 = 1) = P(U \leq p_A) = p_A$
  - Repeating this procedure n times results in a sequence X_1, …, X_n from which N can be computed by adding the X's

Simulating from probability distributions

- FYI: Simulate a general DNA sequence of bases A, C, T, G:
  - Divide the interval (0,1) in 4 intervals with endpoints

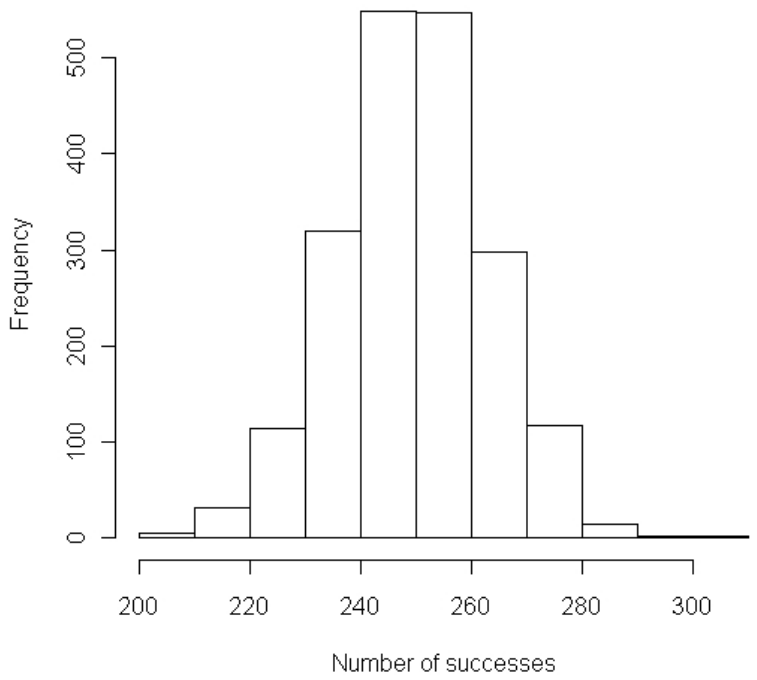
$$0, p_A, p_A + p_C, p_A + p_C + p_G, 1$$

  - If the simulated u lies in the leftmost interval, L_1=A
  - If u lies in the second interval, L_1=C; if in the third, L_1=G and otherwise L_1=T
  - Repeating this procedure n times with different values for U results in a sequence L_1, …, L_n
- Use the "sample" function in R:

```
pi <- c(0.25,0.75)
x<-c(1,0)
set.seed(2009)
sample(x,10,replace=TRUE,pi)
```

Simulating from probability distributions

- By looking through a given simulated sequence, we can count the number of times a particular pattern arises (for instance, the base A)
- By repeatedly generating sequences (k times) and analyzing each of them, we can get a feel for whether or not our particular pattern of interest is unusual



Simulating from a known probability distribution

- Using R code:

```
x<- rbinom(2000,1000,0.25)
mean(x)
sd(x)^2
hist(x,xlab="Number of successes",main="")
```

R documentation

Binomial {stats} R Documentation

The Binomial Distribution

Description

Density, distribution function, quantile function and random generation for the binomial distribution with parameters `size` and `prob`.

This is conventionally interpreted as the number of 'successes' in `size` trials.

Usage

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

Arguments

`x, q`
: vector of quantiles.

`p`
: vector of probabilities.

`n`
: number of observations. If `length(n) > 1`, the length is taken to be the number required.

`size`
: number of trials (zero or more).

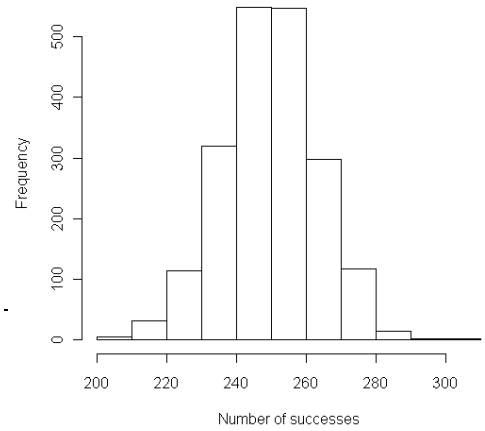
(https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Binomial.html)

Simulating from a known probability distribution

- Using R code:

```
x<- rbinom(2000,1000,0.25)
mean(x)
sd(x)^2
hist(x,xlab="Number of successes",main="")
```

How many entries are taken to compute the mean(x)?



Number of sequences = 2000 = k

Number of trials = 1000 = n

teen K

Back to our original question

- Suppose we have a sequence of 1000bp and assume that every base occurs with equal probability. How likely are we to observe at least 300 A's in such a sequence?
  - Exact computation using a closed form of the relevant distribution
  - Approximate via simulation
  - Approximate using the Central Limit Theory

Exact computation via closed form of relevant distribution

- The formula for the binomial probability mass function is :

$$P(N = j) = \binom{n}{j} p^j (1-p)^{n-j}, \text{ j = 0,1, ...,n}$$

and therefore

$$P(N \geq 300) = \sum_{j=300}^{1000} \binom{1000}{j} (1/4)^j (1-1/4)^{1000-j}$$

$$= 0.00019359032194965841$$

- Note that the probability $P(N \geq 300)$ is estimated to be 0.0001479292 via

```
1-pbinom(300,size=1000,prob=0.25)
pbinom(300,size=1000,prob=0.25,lower.tail=FALSE)
```

	P: exactly 300 out of 1000
Method 1. exact binomial calculation	0.00004566114740576488
Method 2. approximation via normal	0.000038
Method 3. approximation via Poisson	------
	P: 300 or fewer out of 1000
Method 1. exact binomial calculation	0.9998520708293378
Method 2. approximation via normal	0.999885
Method 3. approximation via Poisson	------
	P: 300 or more out of 1000
Method 1. exact binomial calculation	0.00019359032194965841
Method 2. approximation via normal	0.000153
Method 3. approximation via Poisson	------

For hypothesis testing P: 300 or more out of 1000

	One-Tail	Two-Tail
Method 1. exact binomial calculation	0.00019359032194965841	0.0003025705168772097
Method 2. approximation via normal	0.000153	0.000306
Method 3. approximation via Poisson	------	------

(http://faculty.vassar.edu/lowry/binomialX.html)

Approximate via simulation

- Using R code and simulations from the theoretical (“known”) distribution, $P(N \geq 300)$ can be estimated as 0.000196 via

```
x<- rbinom(1000000,1000,0.25)
sum(x>=300)/1000000
```

Approximate via Central Limit Theory

- The central limit theorem offers a 3rd way to compute probabilities of a distribution
- It applies to sums or averages of iid random variables
- Assuming that X_1, …, X_n are iid random variables with mean μ and variance σ^2, then we know that for the sample average

$$\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n),$$

$$\mathrm{E}(\bar{X}_n) = \mu \text{ and Var } (\overline{X_n}) = \frac{\sigma^2}{n}$$

- Hence,

$$E\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = 0, Var\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = 1$$

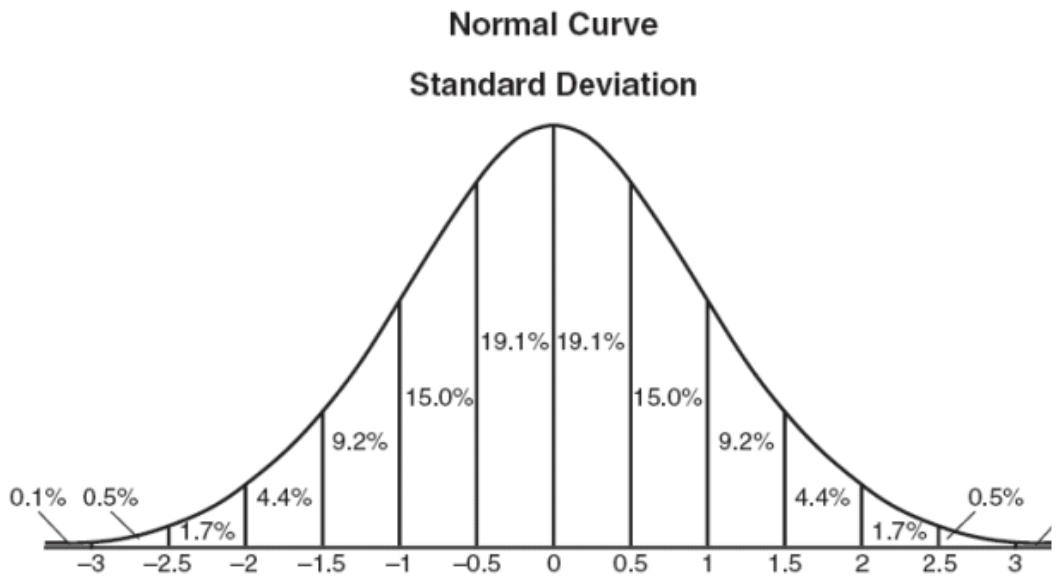
Approximate via Central Limit Theory

- The central limit theorem states that if the sample size n is large enough,

$$P\left(a \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \approx \phi(b) - \phi(a),$$

with $\phi(.)$ the standard normal distribution defined as

$$\phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(x)dx$$


Normal Curve
Standard Deviation
0.1%
0.5%
1.7%
4.4%
9.2%
15.0%
19.1%
19.1%
15.0%
9.2%
4.4%
1.7%
0.5%
–3
–2.5
–2
–1.5
–1
–0.5
0
0.5
1
1.5
2
2.5
3

Approximate via Central Limit Theory

- Estimating the quantity $P(N \geq 300)$ when N has a binomial distribution with parameters n=1000 and p=0.25,

$$E(N) = n\mu = 1000 \times 0.25 = 250,$$

$$sd(N) = \sqrt{n}\,\sigma = \sqrt{1000 \times \frac{1}{4} \times \frac{3}{4}} \approx 13.693$$

$$P(N \geq 300) = P\left(\frac{N - 250}{13.693} > \frac{300 - 250}{13.693}\right)$$

$$\approx P(Z > 3.651501) = 0.0001303560$$

- R code:

```
pnorm(3.651501,lower.tail=FALSE)
```

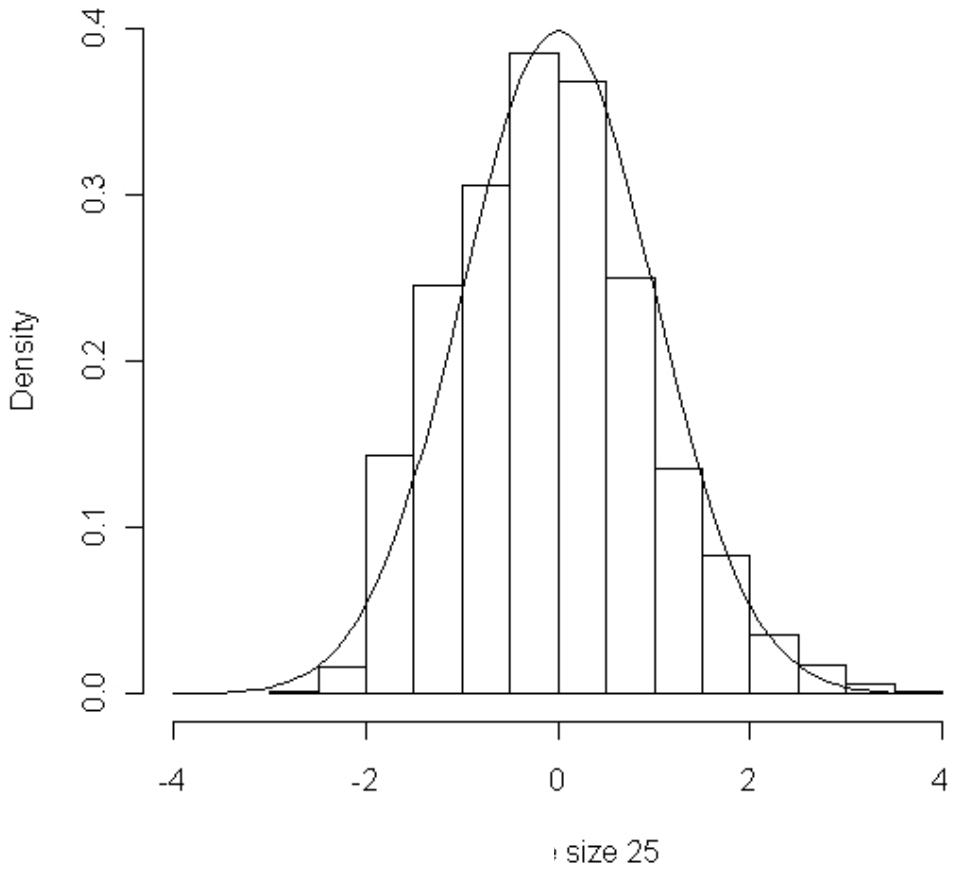
How do the estimates of $P(N \geq 300)$ compare?

Approximate via Central Limit Theory

- The central limit theorem in action using R code:

```
bin25<-rbinom(1000,25,0.25)
av.bin25 <- 25*0.25
stdev.bin25 <- sqrt(25*0.25*0.75)
bin25<-(bin25-av.bin25)/stdev.bin25
hist(bin25,xlim=c(-4,4),ylim=c(0.0,0.4),prob=TRUE,xlab="Sample size
25",main="")
x<-seq(-4,4,0.1)
lines(x,dnorm(x))
```

Approximate via Central Limit Theory



Occurrences of 2-letter words

- Concentrating on abundances, and assuming the iid model for L_1, …, L_n:

$$P(L_i = l_i = C, L_{i+1} = l_{i+1} = G) = p_{l_i}\, p_{l_{i+1}}$$

- Has a given sequence an unusual dinucleotide frequency compared to the iid model?
  - Compare observed O with expected E dinucleotide numbers

$$\chi^2 = \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}},$$

with $E = (n-1)p_{l_i}p_{l_{i+1}}$.

Where have we seen this statistic before? How many df?

Why (n-1) as factor in E above? How many df?

Comparing to the reference

- How to determine which values of χ^2 are unlikely or extreme?
  - If the observed nr is close to the expected number, then the statistic will be small. Otherwise, the model will be doing a poor job of predicting the dinucleotide frequencies and the statistic will tend to be large…
  - Recipe:
    - Compute the number c given by

$$c = \begin{cases} 1 + 2p_{l_i} - 3p_{l_i}^2, & \text{if } l_i = l_{i+1} \\ 1 - 3p_{l_i}p_{l_{i+1}}, & \text{if } l_i \neq l_{i+1} \end{cases}$$

    - Calculate the ratio $\frac{\chi^2}{c}$, where χ^2 is given as before
    - If this ratio is larger than 3.84 then conclude that the iid model is not a good fit. Note that qchisq(0.95,1) = 3.84

Occurrences of 3-letter words

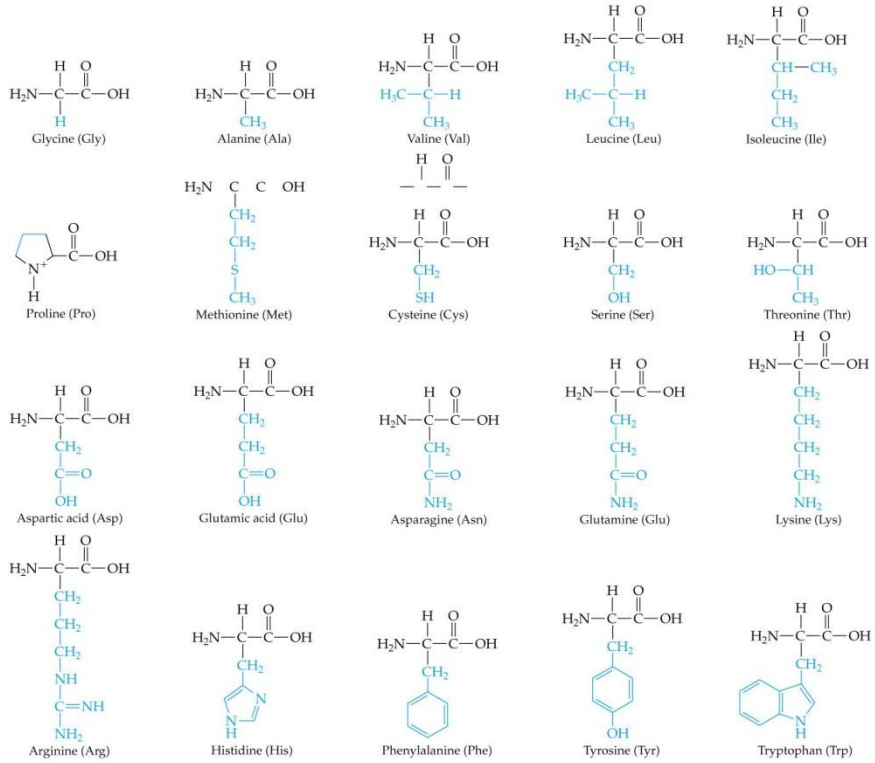
Amino acids

- There are 61 codons that specify amino acids and three stop codons → 64 meaningful 3-words.
- Since there are 20 common amino acids, this means that most amino acids are specified by more than one codon.

Amino acids

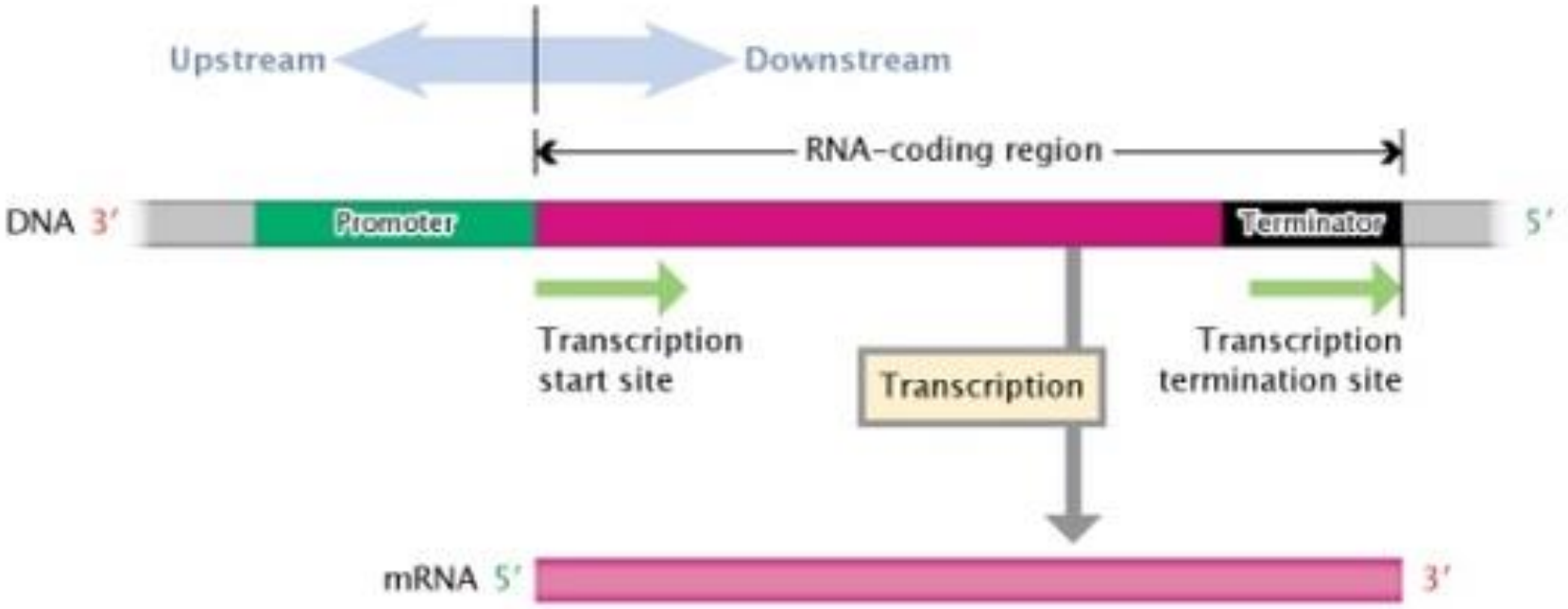
2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G



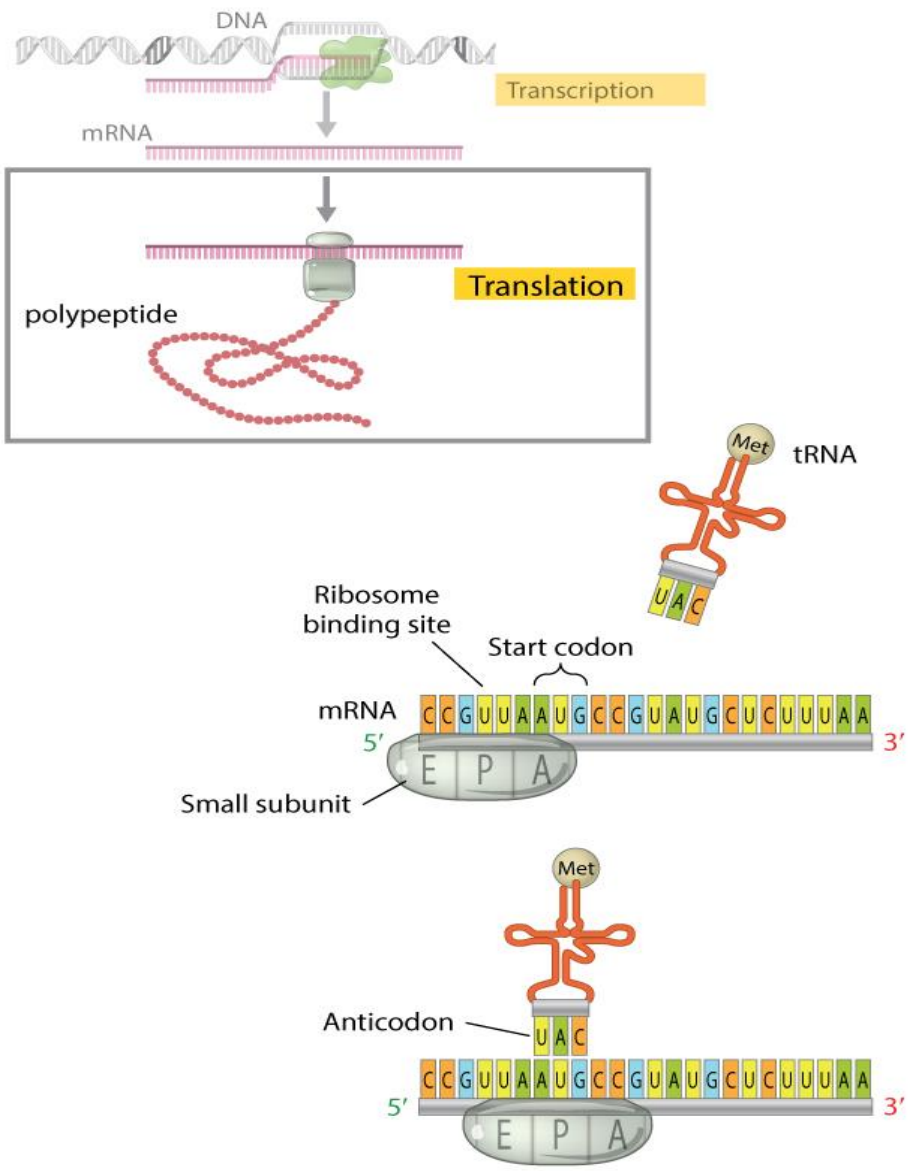
- This has led to the use of a number of statistics to summarize the "bias" in codon usage: An amino acid may be coded in different ways, but **perhaps some codes have a preference**? (higher frequency?)

Transcription



(https://www.nature.com/scitable)

Transcription and Translation



(https://www.nature.com/scitable)

Predicted relative frequencies

- For a sequence of independent bases L_1, L_2, ... , L_n the expected 3-tuple relative frequencies can be found by using the logic employed for dinucleotides we derived before
- The probability of a 3-word can be calculated as follows:

$$\mathbb{P}(L_i = r_1, L_{i+1} = r_2, L_{i+2} = r_3) = \mathbb{P}(L_i = r_1)\mathbb{P}(L_{i+1} = r_2)\mathbb{P}(L_{i+2} = r_3).$$

assuming the iid model

- This provides the expected frequencies of particular codons, using the individual base frequencies. It follows that among those codons making up the amino acid Phe, the expected proportion of TTT is

$$\frac{\mathrm{P(TTT)}}{\mathrm{P(TTT)} + \mathrm{P(TTC)}}$$

The codon adaptation index

- One can then compare predicted and observed triplet frequencies in coding sequences for a subset of genes and codons from E. coli.
- Médigue et al. (1991) clustered different genes based on codon usage patterns, and they observed three classes.
- For instance for Phe, the observed frequency differs considerably from the predicted frequency, when focusing on highly expressed genes (so-called “class II genes” in the work of Médigue et al. (1999) - see also next slide
- Checking the gene annotations for class II genes: highly expressed genes (ribosomal proteins or translation factors)

- Table 2.3 from Deonier et al 2005: figures in parentheses below each gene class show the number of genes in that class.

	Codon	Predicted	Observed Gene Class I (502)	Observed Gene Class II (191)
Phe	TTT	0.493	0.551	0.291
	TTC	0.507	0.449	0.709
Ala	GCT	0.246	0.145	0.275
	GCC	0.254	0.276	0.164
	GCA	0.246	0.196	0.240
	GCG	0.254	0.382	0.323
Asn	AAT	0.493	0.409	0.172
	AAC	0.507	0.591	0.828

Class II : Highly expressed genes

Class I : Moderately expressed genes

Main reference of foregoing material in this chapter: Deonier et al. *Computational Genome Analysis*, 2005, Springer (Ch 6,7)

Supporting doc to this class (complementing course slides)

√

REVIEW

Rare-Variant Association Analysis: Study Designs and Statistical Tests

Seunggeung Lee,[1] Gonçalo R. Abecasis,[1] Michael Boehnke,[1] and Xihong Lin[2,*]

Despite the extensive discovery of trait- and disease-associated common variants, much of the genetic contribution to complex traits remains unexplained. Rare variants can explain additional disease risk or trait variability. An increasing number of studies are underway to identify trait- and disease-associated rare variants. In this review, we provide an overview of statistical issues in rare-variant association studies with a focus on study designs and statistical tests. We present the design and analysis pipeline of rare-variant studies and review cost-effective sequencing designs and genotyping platforms. We compare various gene- or region-based association tests, including burden tests, variance-component tests, and combined omnibus tests, in terms of their assumptions and performance. Also discussed are the related topics of meta-analysis, population-stratification adjustment, genotype imputation, follow-up studies, and heritability due to rare variants. We provide guidelines for analysis and discuss some of the challenges inherent in these studies and future research directions.

AJHG 2014; 95, 5-23

Questions?