Genetics and Bioinformatics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

Lecture 5: I have my DNA sequences, now what?

1 Introduction

1.a Generating DNA sequences

1.b Historical notes

1.c Application fields

2 Investigating frequencies of occurrences of words

2.a Motivation

2.b Probability distributions

2.c Simulating from a probability distribution

3 Study examples

3.a Words of length 2

3.b Words of length 3

4 Rare variants in humans

4.a Motivation

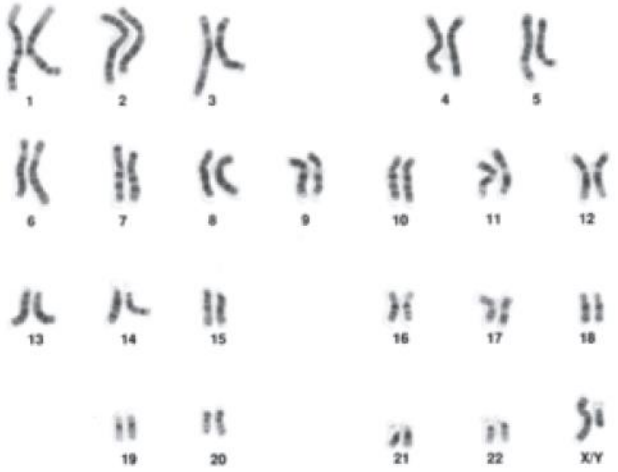
4.b Rare variant association analysis

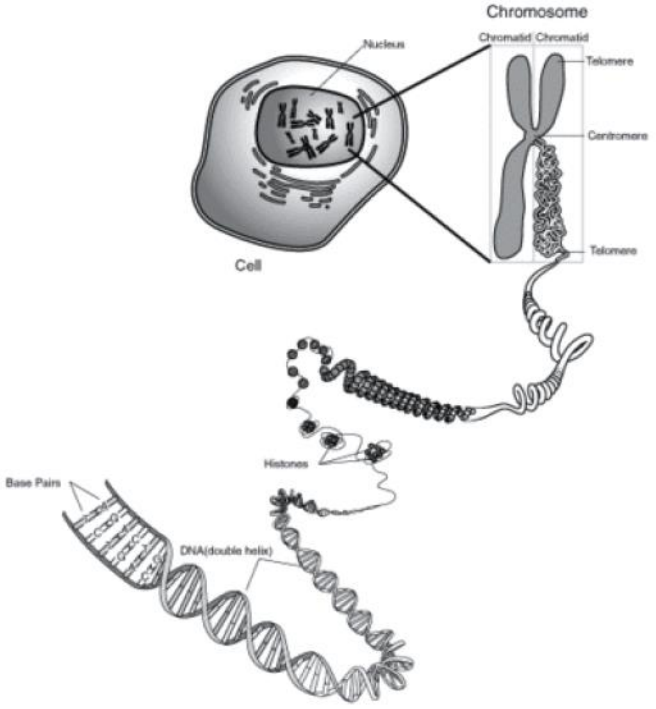
4.c Rare variants to identify population substructure

1 Introduction

1.a Generating DNA sequences

Human genome



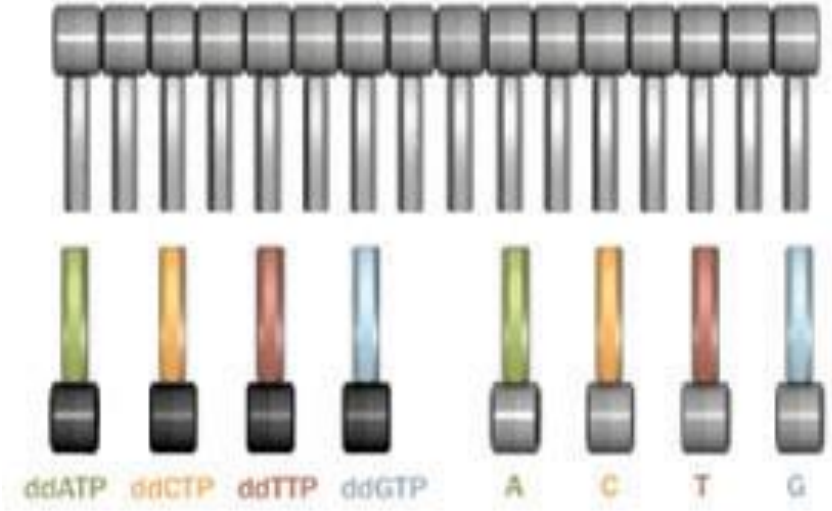


- 3×10^9 bases/nucleotides
- < 1 % coding
- 20.000 genes

Sanger sequencing

- DNA sequencing enables us to perform a thorough analysis of DNA because it provides us with the most basic information of all: the sequence of nucleotides.
- Scientists recognized that this could potentially be a very powerful tool, and so there was competition to create a method that would sequence DNA.
- Then in 1974, two methods were independently developed by an American team and an English team to do exactly this.
  - The Americans, led by Maxam and Gilbert, used a “chemical cleavage protocol”, while
  - the English, led by Sanger, designed a procedure similar to the natural process of DNA replication.
- Even though both teams shared the 1980 Nobel Prize, Sanger’s method became the standard because of its practicality (Speed, 1992).

- Sanger's method, which is also referred to as dideoxy sequencing or chain termination, is based on the use of **dideoxynucleotides** (ddNTP's) in addition to the normal nucleotides (NTP's) found in DNA.



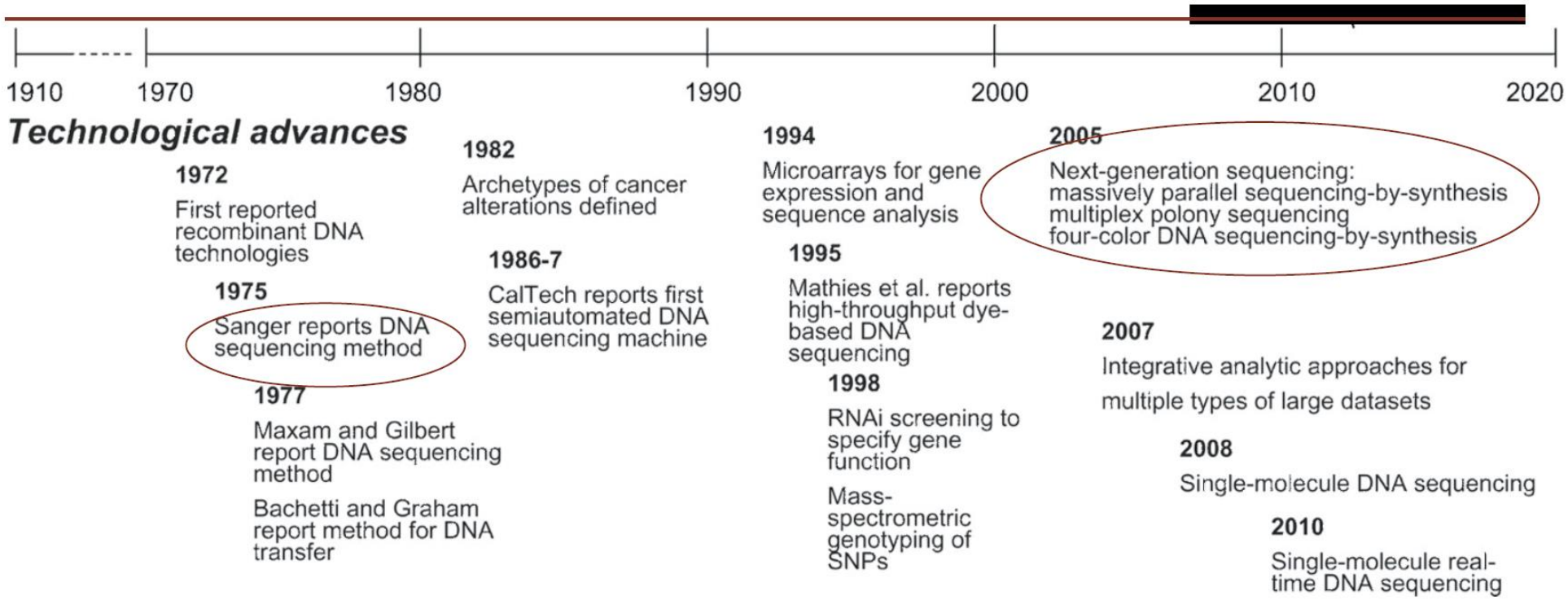
Dideoxynucleotides are essentially the same as nucleotides except they contain a hydrogen group on the 3' carbon instead of a hydroxyl group (OH). These modified nucleotides, when integrated into a sequence, prevent the addition of further nucleotides (Speed, 1992).

(http://www.nature.com/scitable/topicpage/the-order-of-nucleotides-in-a-gene-6525806)

- When Sanger sequencing was first introduced, four separate reagents were used, one for each type of ddNTP.
  - The 4 reaction products were then separated by gel electrophoresis, that organizes DNA fragments in order of size, enabling researchers to assess the lengths of the truncated strands in each sample.
  - The end of each truncated strand was used to determine the position at which a ddNTP was added to the strand, thereby halting DNA elongation.
- More recently, automation of the Sanger technique has made this process more efficient by combining all four ddNTP reactions in a single test tube.
  - Each of the four ddNTPs in the tube is labeled with a different fluorescent color.
  - Rather than being run on a gel and read manually, the reaction products are passed through a small tube containing a gel-like matrix.

(http://www.nature.com/scitable/topicpage/the-order-of-nucleotides-in-a-gene-6525806)

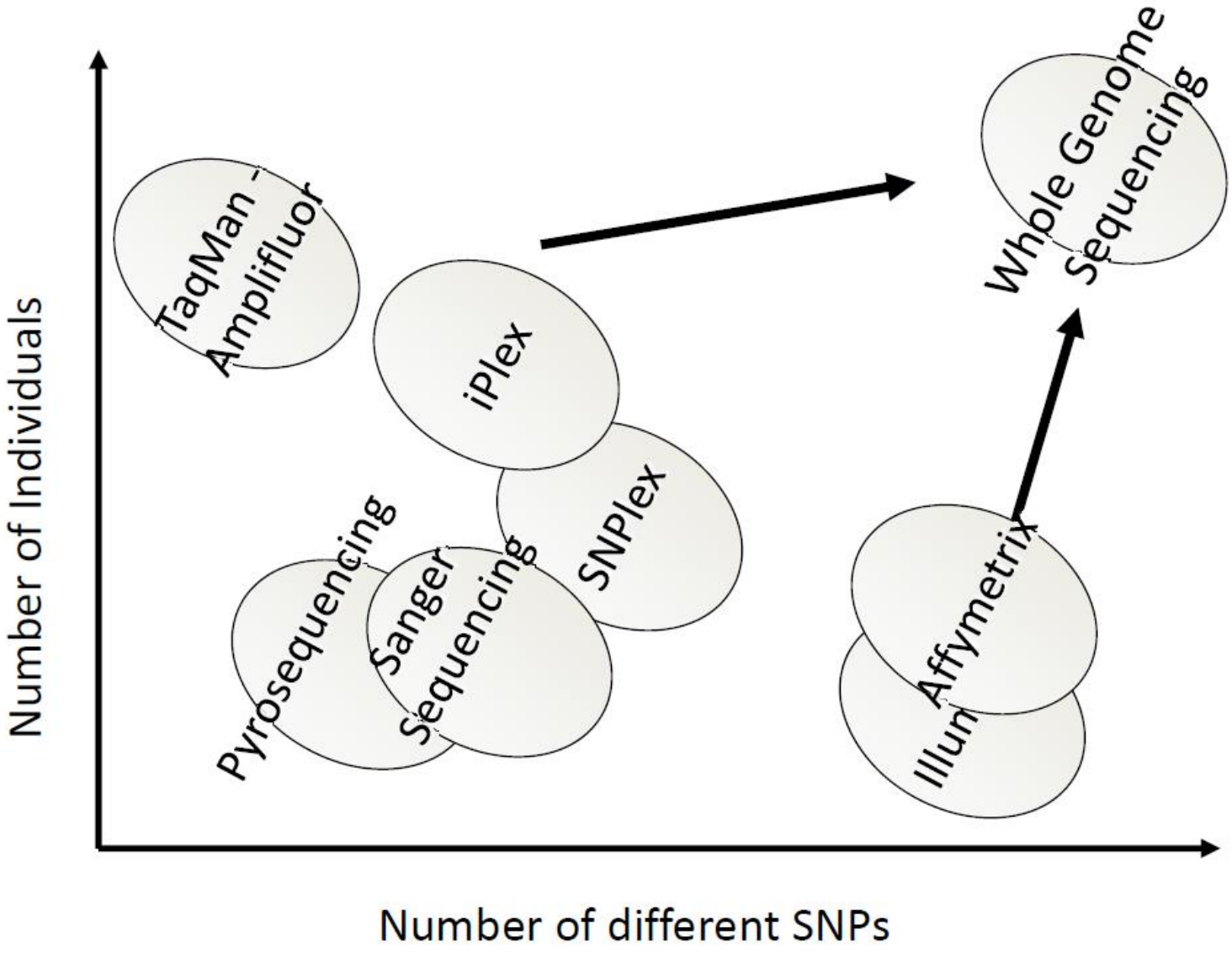
New(er) generations of sequencing methods

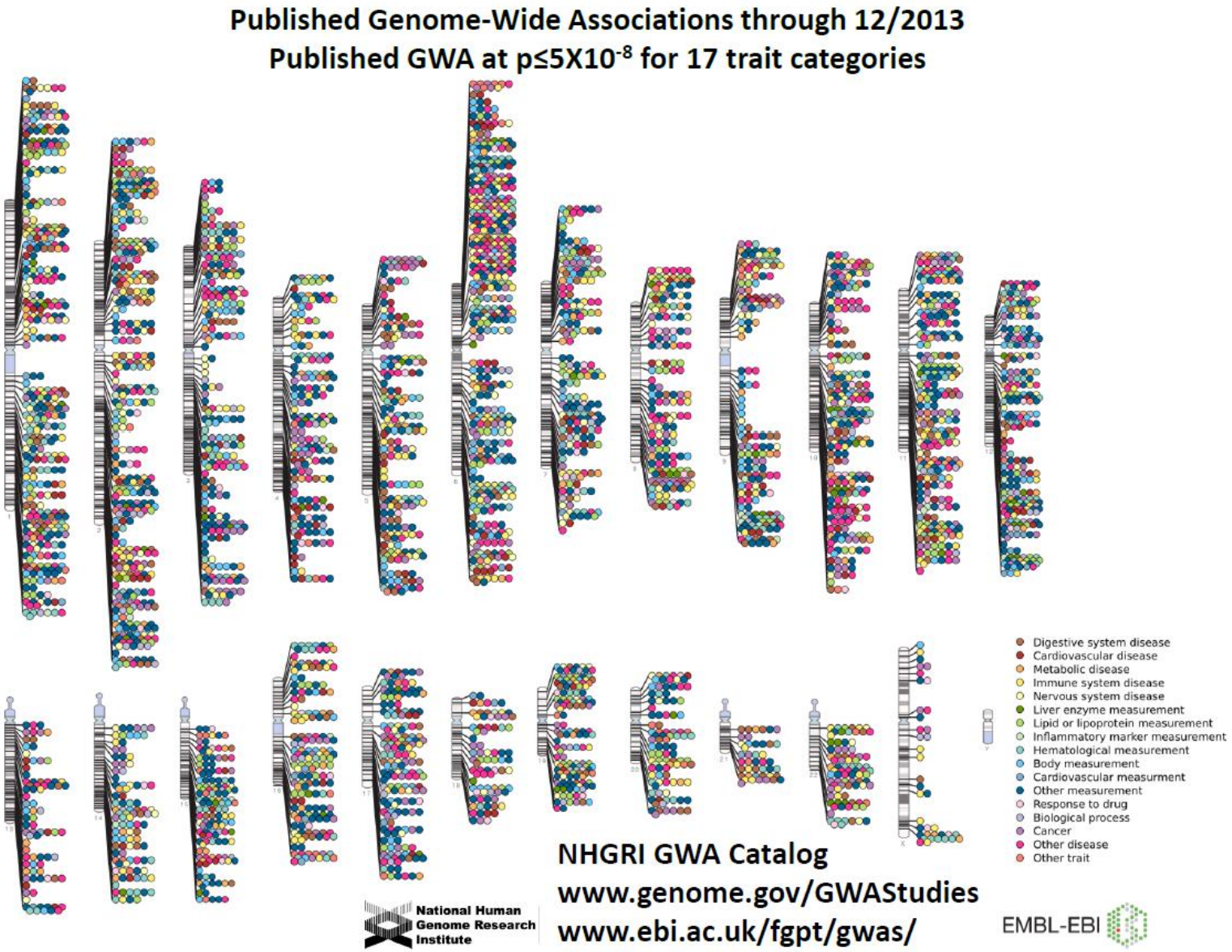


(http://uni-leipzig.de/~strimmer/lab/courses/ss12/current-topics/slides/1-NGS.pdf)

New(er) generations of sequencing methods

- 1st generation DNA sequencing 1977-2005 – Fred Sanger
  - ~1 million bases/instrument * day
  - Only method in 2005
- 2nd generation DNA sequencing 2005 –
  - 2005 ~10 million bases / instrument * day
  - 2007 ~100 million bases / instrument * day
  - 2010 ~6 billion bases / instrument * day
- 3rd generation DNA sequencing 2011 –
  - Complete human genome
  - < 1k EURO
  - < 1 day


Number of Individuals
Number of different SNPs
TaqMan - Amplifluor
iPlex
SNPlex
Pyrosequencing
Sanger Sequencing
Affymetrix
Illumina
Whole Genome Sequencing


Published Genome-Wide Associations through 12/2013
Published GWA at p≤5X10-8 for 17 trait categories
Digestive system disease
Cardiovascular disease
Metabolic disease
Immune system disease
Nervous system disease
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurment
Other measurement
Response to drug
Biological process
Cancer
Other disease
Other trait
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/
National Human Genome Research Institute
EMBL-EBI

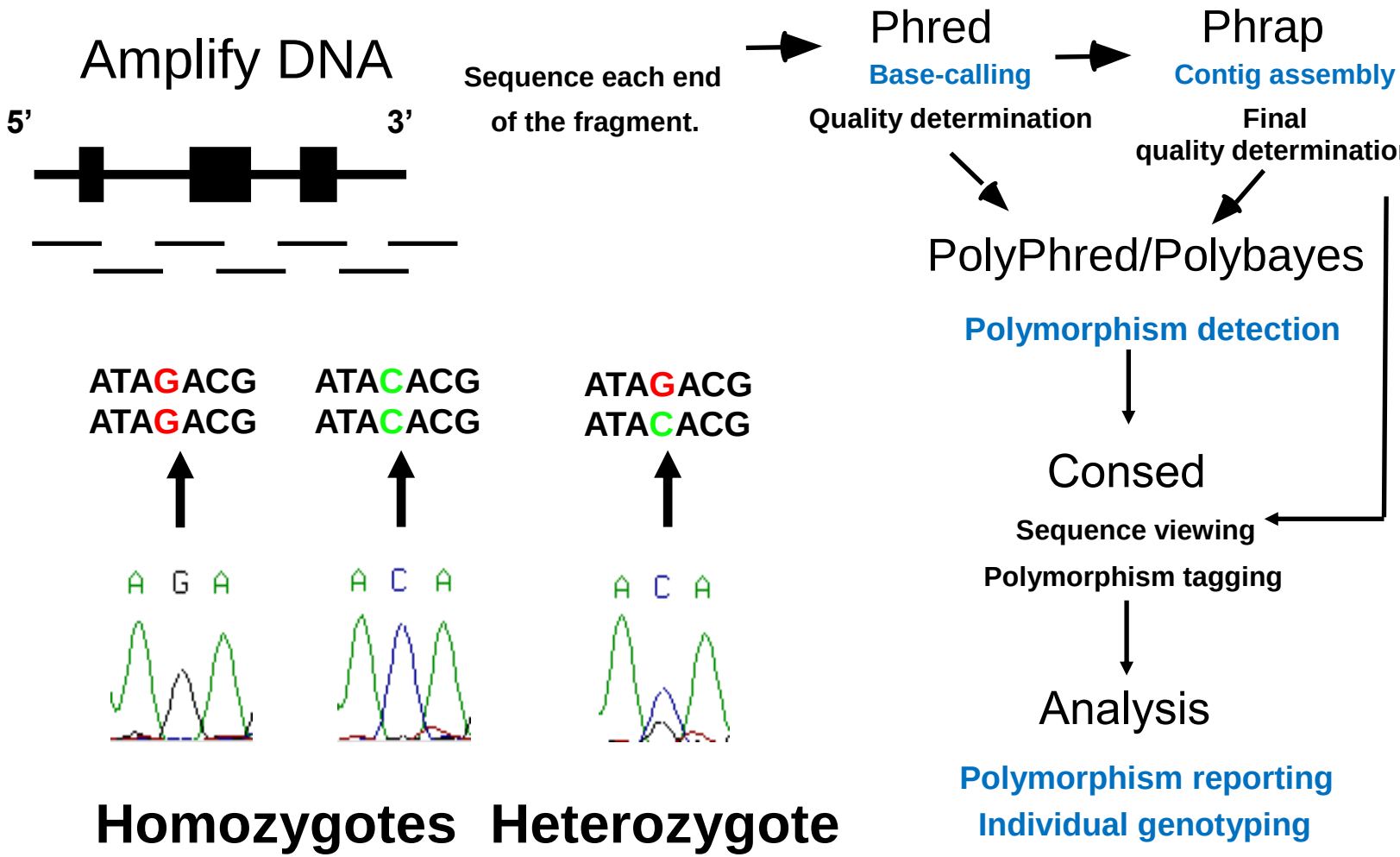
Sequence assembly problems

- In principle, assembling a sequence is just a matter of finding overlaps and combining them.
- In practice:
  - most genomes contain multiple copies of many sequences,
  - there are random mutations (either naturally occurring cell-to-cell variation or generated by PCR or cloning),
  - there are sequencing errors and misreadings,
  - sometimes the cloning vector itself is sequenced
  - sometimes miscellaneous junk DNA gets sequenced
- Getting rid of vector sequences is easy once you recognize the problem: just check for them.

- Repeat sequence DNA is very common in eukaryotes, and sequencing highly repeated regions (such as centromeres) remains difficult even now. High quality sequencing helps a lot: small variants can be reliably identified.
- Sequencing errors, bad data, random mutations, etc. were originally dealt with by hand alignment and human judgment. However, this became impractical when dealing with the Human Genome Project.
- This led to the development of automated methods. The most useful was the phred/phrap programs developed by Phil Green and collaborators at Washington University in St. Louis.

Phred, Phrap, Consed

(slide: J Wegrzyn)



Combining sequences with Phrap

- Phrap first examines all reads for matching "words": short sections of identical sequence. The matching words need to be in the same order and spacing.
- Sequences in both orientations are examined, using the reverse-complement sequence if necessary.
- The entire sequences of pairs with matching words are then aligned using the Smith-Waterman algorithm (a standard technique we will look at later). Phrap then looks for discrepancies in the combined sequences, using phred scores to decide between alternatives. Phrap generates quality scores from the combined phred data.
  Sequencing errors are not necessarily random: homopolymeric regions (several of the same base in a row) are notoriously tricky to sequence

accurately. Using the opposite strand often helps resolve these regions. Also using a different sequencing technology or chemistry.

- Sequences are combined with a greedy algorithm: all pairs of fragments are scored for the length and quality of their overlap region, and then the largest and best-matched pair is merged. This process is repeated until some minimum score is reached.
- The result is a set of contigs: reads assembled into a continuous DNA sequence.
- The ideal result is the entire chromosome assembled into a single contig.

A **greedy algorithm** is a mathematical process that looks for simple, easy-to-implement solutions to complex, multi-step problems by deciding which next step will provide the most obvious benefit. Such algorithms are called greedy because while the optimal solution to each smaller instance will provide an immediate output, the algorithm doesn't consider the larger problem as a whole. Once a decision has been made, it is never reconsidered.

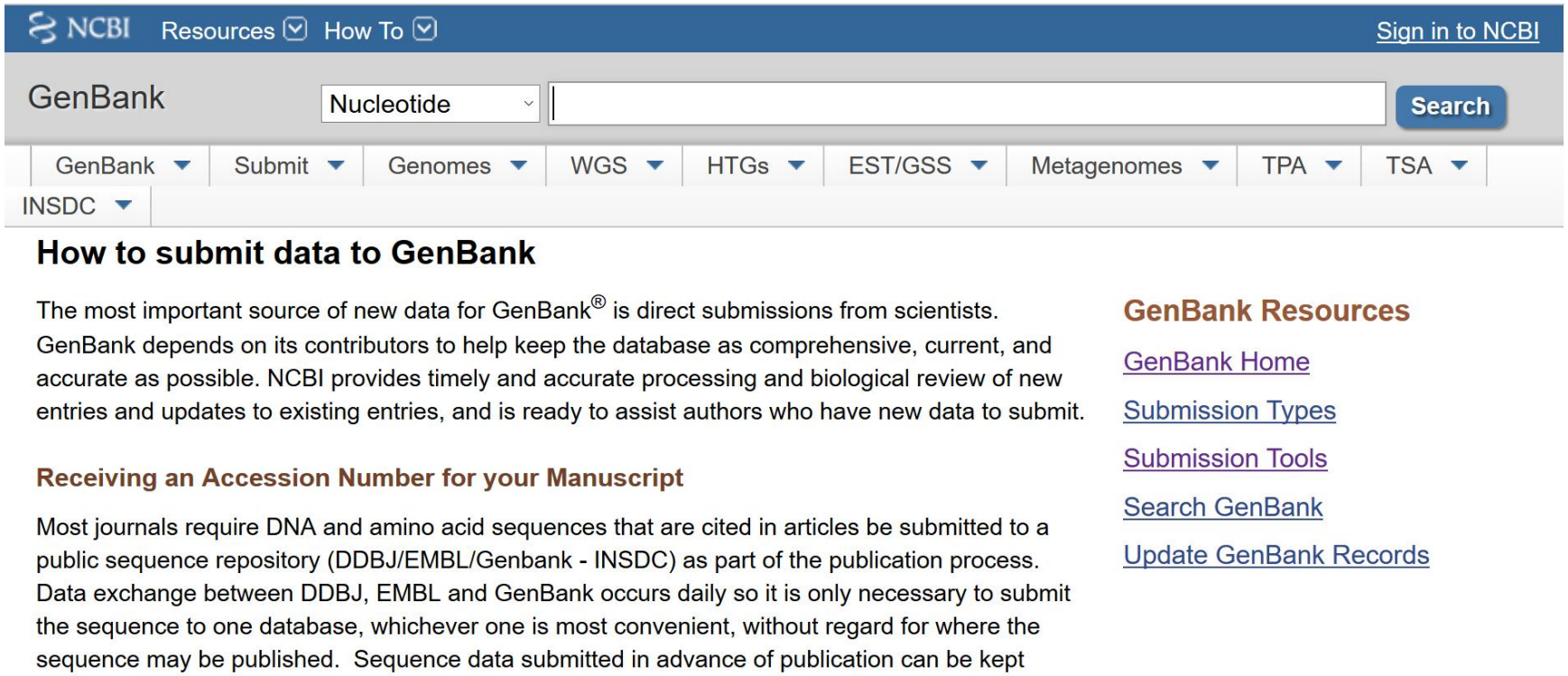
Finishing the sequence

- Shotgun sequencing of random DNA fragments necessarily misses some regions altogether. Also, for sequencing methods that involve cloning (Sanger), certain regions are impossible to clone: they kill the host bacteria.

> **Shotgun sequencing** involves randomly breaking up DNA sequences into lots of small pieces and then reassembling the sequence by looking for regions of overlap.

- Thus it is necessary to close gaps between contigs, and to re-sequence areas with low quality scores. This process is called **finishing**. It can take up to 1/2 of all the effort involved in a genome sequencing project.
- Mostly hand work: identify the bad areas and sequence them by primer walking. Sometimes using alternative sequencing chemistries (enzymes, dyes, terminators, dNTPs) can resolve a problematic region.

- Once a sequence is completed, it is usually analyzed by finding the genes and other features on it: annotation. We will discuss this later.
- Submission of the annotated sequence to Genbank allows everyone access to it: the final step in the scientific method.



(http://www.ncbi.nlm.nih.gov/genbank)

1.b Historical notes

Sequencing projects

- Based on the first Sanger sequencing technique, the **Human Genome Project** (1990–2003), allowed the release of the first human reference genome by determining the sequence of ~3 billion base pairs and identifying the approximately ~25,000 human genes (now we know there are less genes)
- That stood as a great breakthrough in the field of comparative genomics and genetics as one could in theory directly compare any healthy or non-healthy sample against a **golden standard** reference and detect genetic polymorphisms or variants that occur in a genome.

Sequencing projects

- Few years later, as sequencing techniques became more advanced, more accurate, and less expensive, the **1000 Human Genome Project** was launched (January 2008).

> The main scope of this consortium is to sequence, ~1000 anonymous participants of different nationalities and concurrently compare these sequences to each other in order to better understand human genetic variation.

- The **International HapMap Project** (short for "haplotype map") aims to identify common genetic variations among people, making use of data from six different countries.
- Shortly after the 1000 Human Genome Project, the **1000 Plant Genome Project** (http://www.onekp.com) was launched, aiming to sequence and define the transcriptome of ~1000 plant species from different populations around the world.

Notably, out of the 370,000 green plants that are known today, only ~125,000 species have recorded gene entries in GenBank and many others still remain unclassified.

- While the 1000 Plant Genome Project was focused on comparing different plant species around the world, within the **1001 Genomes Project**, 1000 whole genomes of A. Thaliana plants across different places of the planet were sequenced.
- Similar to other consortiums, the **10,000 Genome Project** aims to create a collection of tissue and DNA specimens for 10,000 vertebrate species specifically designated for whole-genome sequencing.

Vertebrates have a series of nerves along the back which need support and protection. That need brings us to the backbones and notochords. Notochords were the first "backbones" serving as support structures.

- The goal of the **1000 Fungal Genome Project** (http://1000.fungalgenomes.org) is to explore all areas of fungal biology.

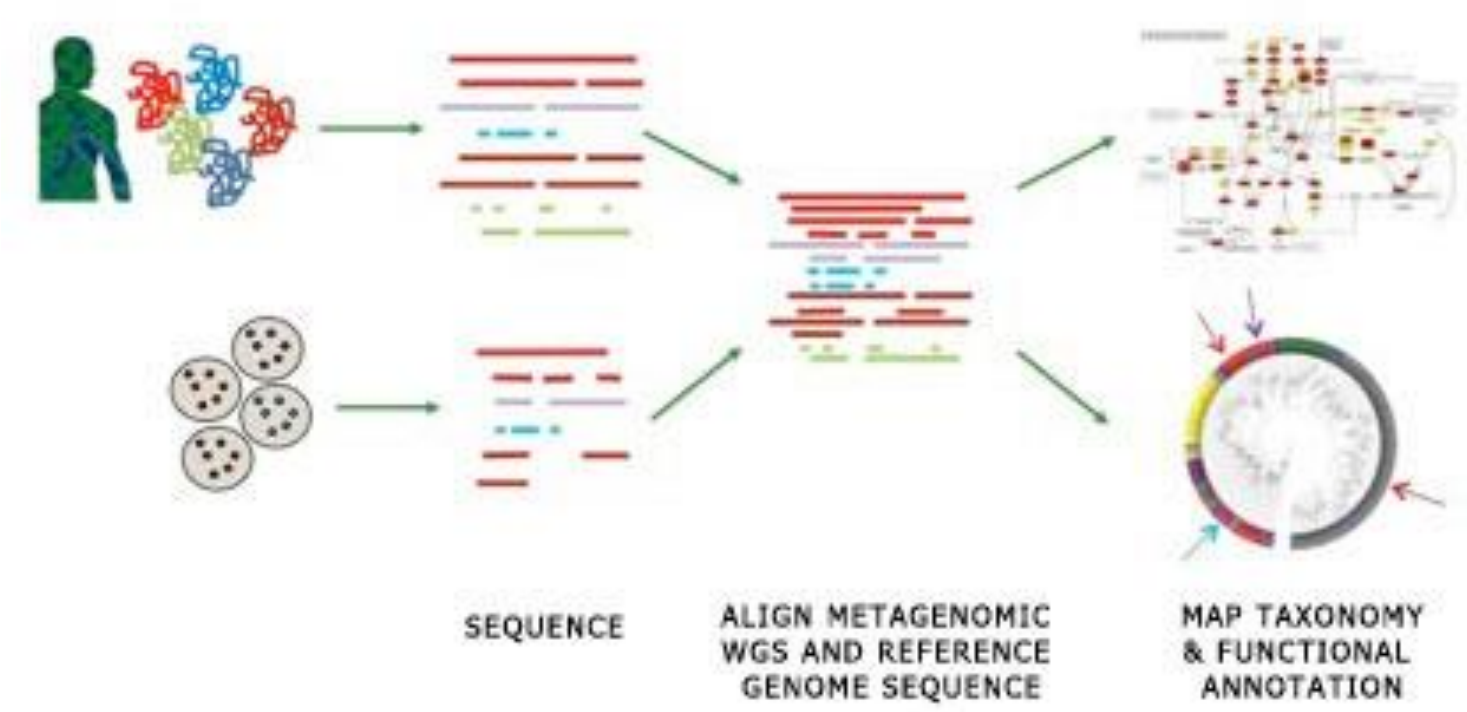
- In human genetics, metagenome sequencing is becoming increasingly important, which lead to the **Human Microbiome Project** (http://www.hmpdacc.org/)

  - Metagenome sequencing is defined as an approach for the study of microbial populations in a sample representing a community by analysing the nucleotide sequence content.
  - The HMP plans to sequence 3000 genomes from both cultured and uncultured bacteria, plus several viral and small eukaryotic microbes isolated from human body sites.
  - This, in conjunction with reference genomes sequenced by HMP Demonstration Projects and other members of the International Human Microbiome Consortium (IHMC), will supplement the available selection of non-HMP funded human-associated reference genomes.

Why do we need reference sequences?

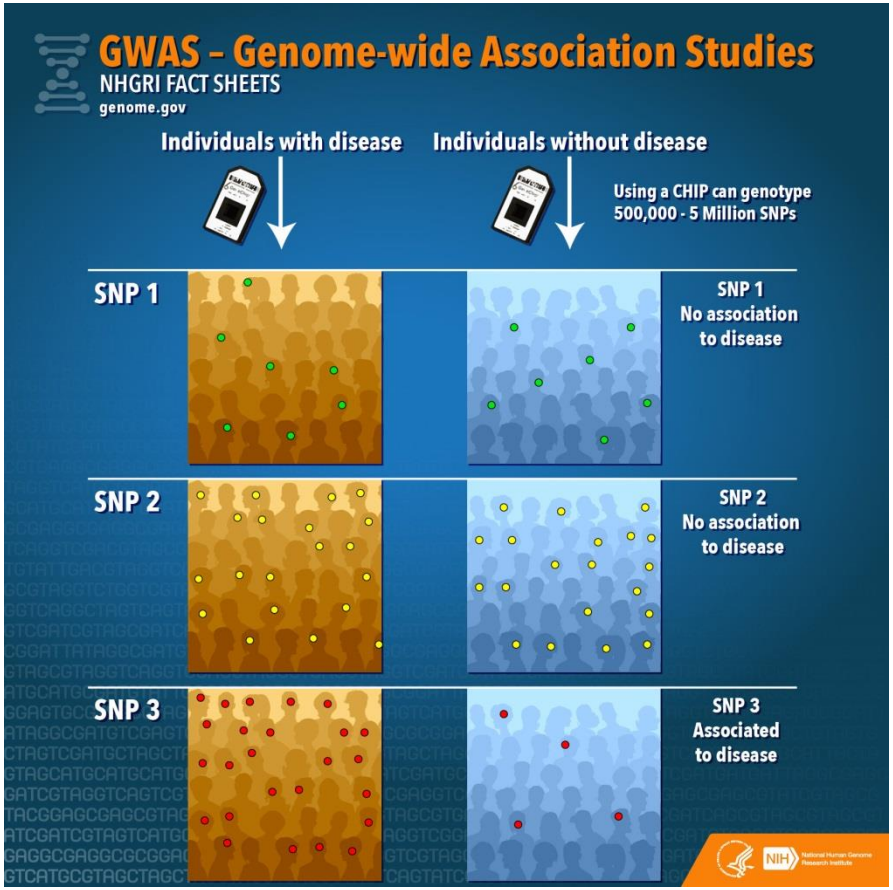
- Within the human body, it is estimated that there are 10x as many microbial cells as human cells.
- Our microbial partners carry out a number of metabolic reactions that are not encoded in the human genome and are necessary for human health (→ human genome = human genes + microbial genes).
- The majority of microbial species present in the human body have never been isolated, cultured or sequenced, typically due to the inability to reproduce necessary growth conditions in the lab (→ study microbial communities – metagenomics)
- In order to assign metagenomic sequence to taxonomic and functional groupings, and to differentiate the novel from the previously described, it is necessary to have a large pool of described genomes from the same environment (reference genomes).

Why Reference Sequences?



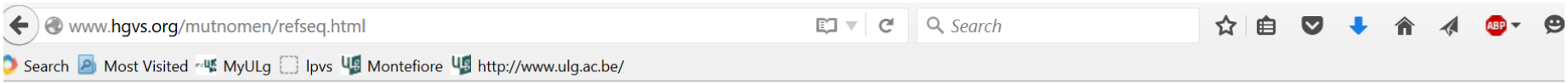
(http://www.hmpdacc.org/)

Why reference sequences?



(https://www.genome.gov/images/content/gwas_infographic.jpg)

Which reference sequence?



HGVS
HUMAN GENOME VARIATION SOCIETY

A reference sequence - discussions and FAQs

Last modified September 11, 2015

Since references to WWW-sites are not yet acknowledged as citations, please mention den Dunnen JT and Antonarakis SE (2000). Hum.Mutat. 15:7-12 when referring to these pages.

Contents

(http://www.hgvs.org/mutnomen/refseq.html)

Which reference sequence?

Practical problems genomic reference sequence

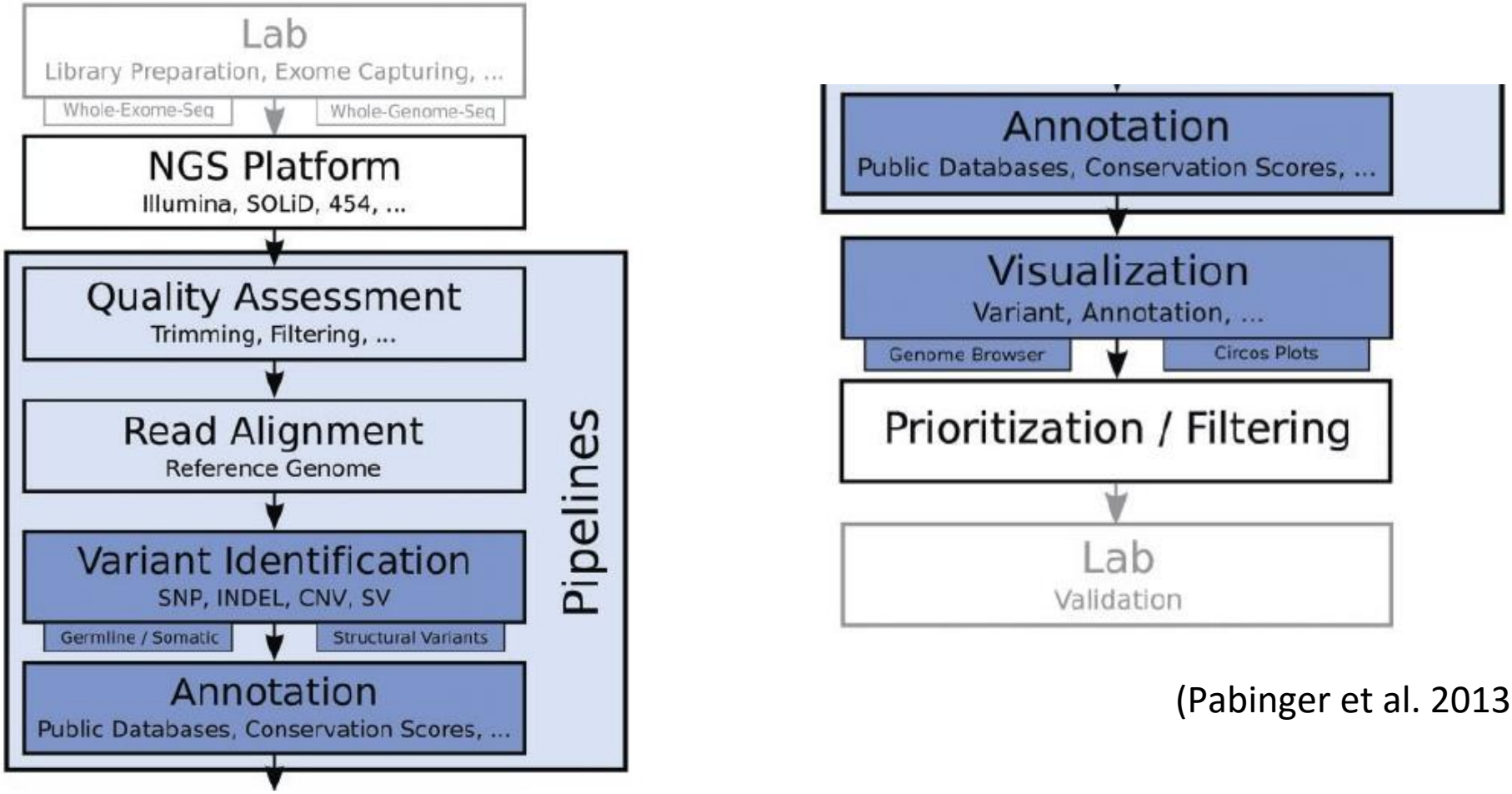
- ***for a human, a genomic reference sequence does not contain any useful information (a coding DNA reference sequence does)***
- a *gene can be very large* (over 2.0 Mb) - this makes nucleotide numbering based on a genomic reference sequence rather impractical (e.g. g.1567234_1567235insTG). Furthermore, genomic reference sequences based on GenBank NT_ files become increasingly long (e.g. the CFTR gene in NT_007933.15, >77 Mb) and consequently loose their informativity. Downloading such large files is, even with good internet connections, time consuming and working with these files is rather difficult.
- when a genomic reference sequence is taken from a complete genome sequence, e.g. a bacterium or the human X-chromosome, the transcriptional orientation of the gene of interest may be on the *minus (-) strand*. This makes the description of sequence variants rather complicated, especially when the consequences on RNA and/or protein level need to be described; nucleotides on DNA and RNA level are complementary and numbering goes in different directions - a confusing situation that should be prevented.
- when different genes (partly) overlap, using the same or the minus (-) DNA strand, which reference sequence should one use to describe the variant and to which gene should the change be assigned ? (see *Recommendations*).
- when the *gene sequence is incomplete* (especially when large introns are present) - a genomic sequence can not be used.
- genes may contain very large introns with many intronic *(length) variants* present in the population - it is thus very difficult to give ***THE*** genomic reference sequence (see *Genomic sequence changes regularly*).

Practical problems coding DNA reference sequence

- the exact *transcriptional start site* (cap-site) of a gene has often not been determined and/or its assignment is debated - the first nucleotide can thus not be assigned with certainty. The same might be true for the translation initiation site (ATG-codon).
- a gene may have *several transcripts*, using different promoters / 5'-first exons, alternatively spliced internal exons, different 3'-terminal exons and polyA-addition sites - ***one*** complete coding DNA reference sequence can thus not be generated (see *Alternatively spliced exons - nucleotide numbering*),
- the different transcripts may *encode different proteins* (isoforms) with, when different promoters are used, different N-terminal sequences and even using different reading frames in one or more exons. ***One*** complete protein reference sequence can thus not be assigned.
- when different genes (partly) overlap, using the same or the minus (-) DNA strand, which reference sequence should one use to describe the variant and to which gene should the change be assigned ? (see *Recommendations*).

(http://www.hgvs.org/mutnomen/refseq.html#standard)

Common workflow for whole-exome and whole genome sequencing



(Pabinger et al. 2013)

1.c Application fields

How is DNA sequencing used by scientists?

A. In recent years, DNA sequencing technology has advanced many areas of science. For example, the field of **functional genomics** is concerned with

- figuring out what certain DNA sequences do, as well as
- which pieces of DNA code for proteins and
- which have important regulatory functions.

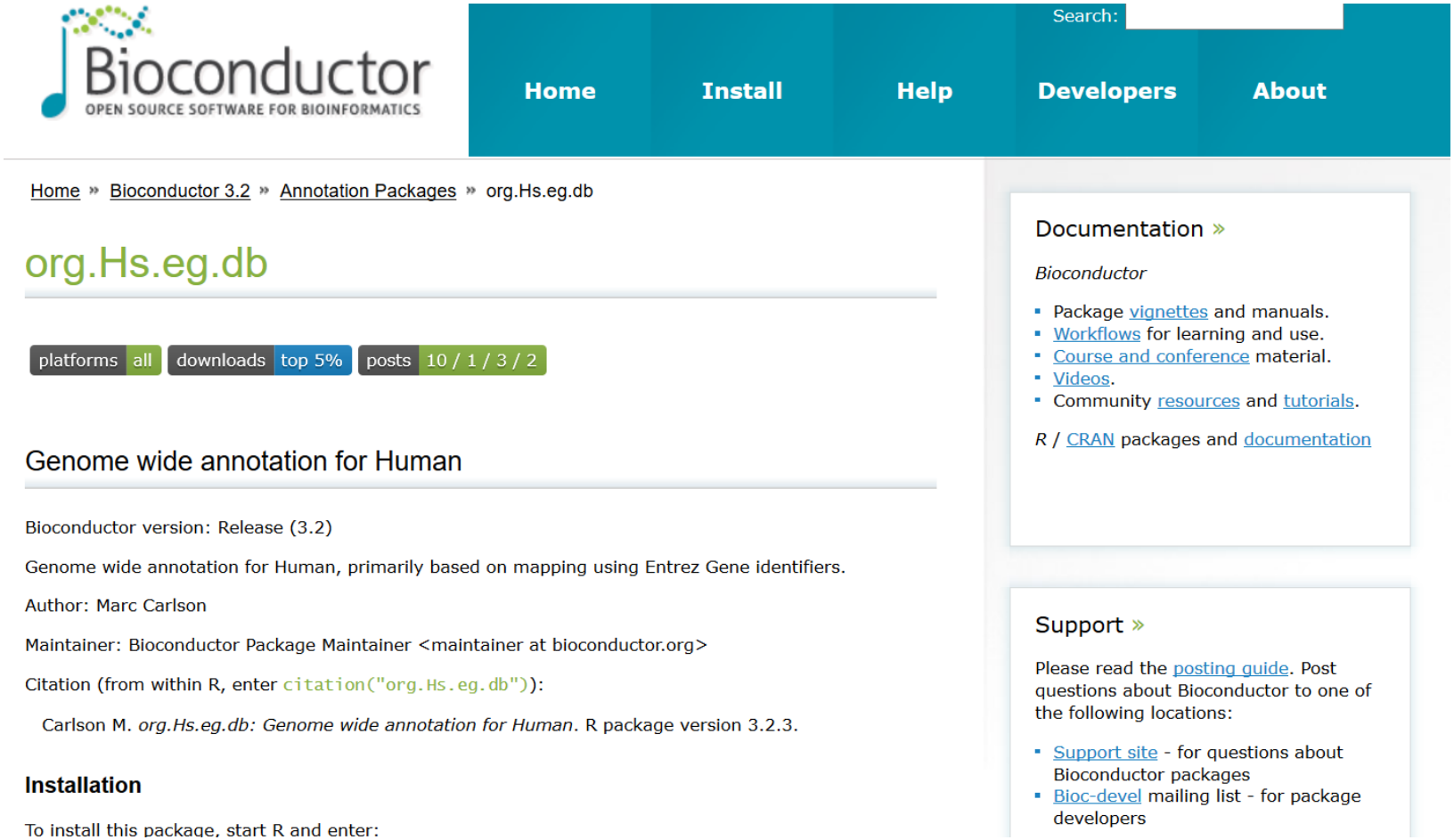
B. An invaluable first step in making these determinations is **learning the nucleotide sequences** of the DNA segments under study.

C. Another area of science that relies heavily on DNA sequencing is comparative genomics, in which researchers compare the genetic material of different organisms in order to learn about their evolutionary history and degree of relatedness.

D. **Complex disease analysis**

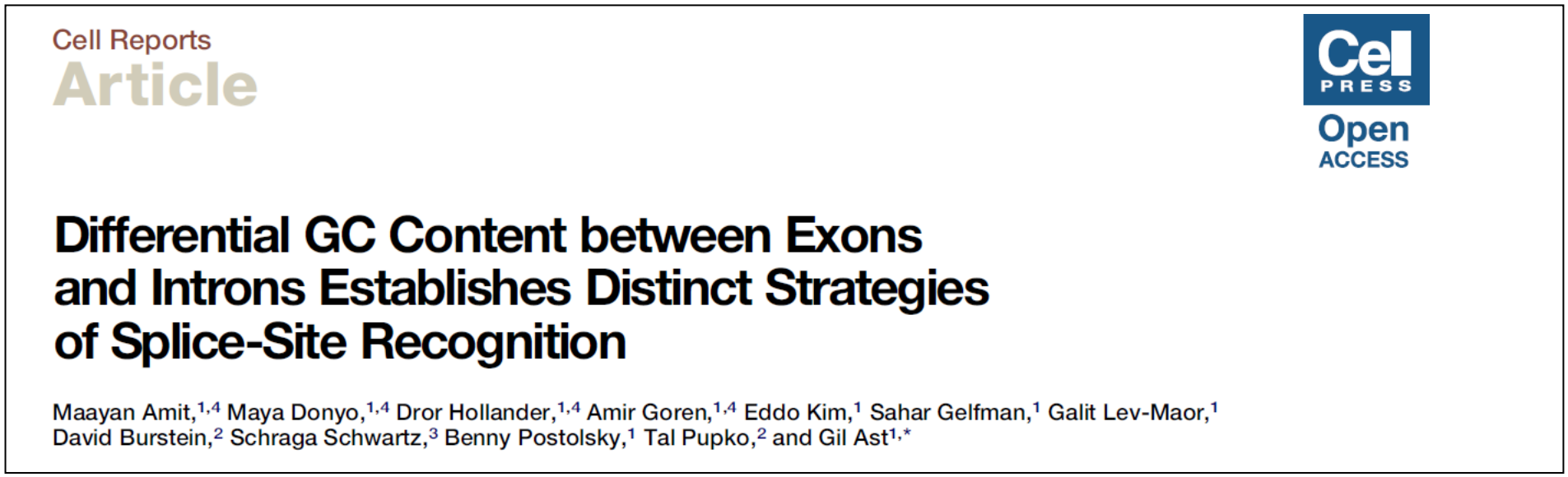
A. Sequence annotation

(see practicals)



B. Counting letters or words

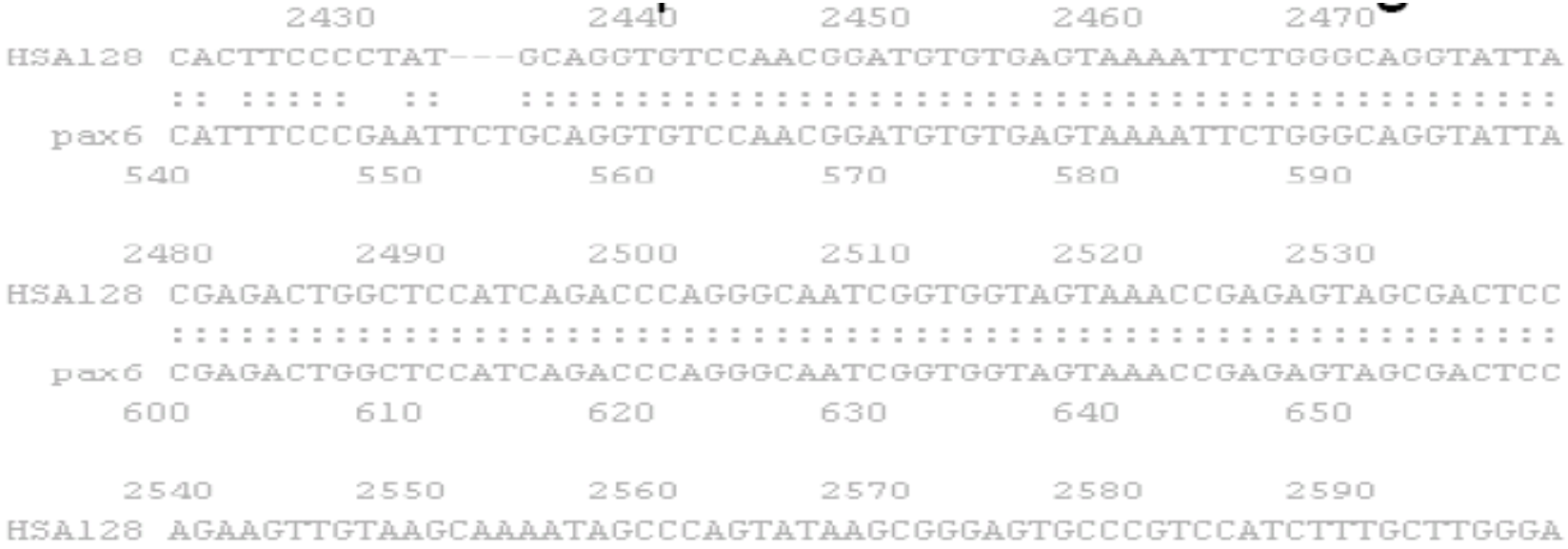
- One of the most fundamental properties of a genome sequence is its GC content, the fraction of the sequence that consists of Gs and Cs, ie. the %(G+C).
- The GC content can be calculated as the percentage of the bases in the genome that are Gs or Cs. That is, GC content = (number of Gs + number of Cs)*100/(genome length). For example, if the genome is 100 bp, and 20 bases are Gs and 21 bases are Cs, then the GC content is (20 + 21)*100/100 = 41%.



- The **CpG sites** or **CG sites** are regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length. "CpG" is shorthand for "—C—phosphate—G—", that is, cytosine and guanine separated by only one phosphate. The "CpG" notation is used to distinguish this linear sequence from the CG base-pairing of cytosine and guanine. (https://en.wikipedia.org/wiki/CpG_site)

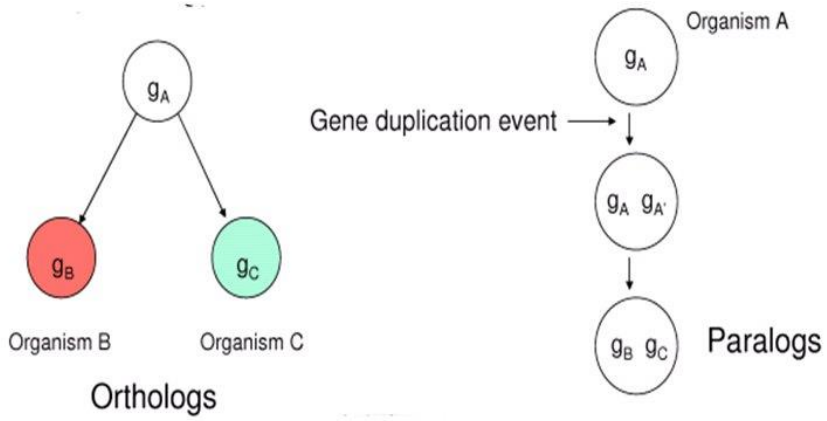
C. Comparing multiple sequences

- 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?

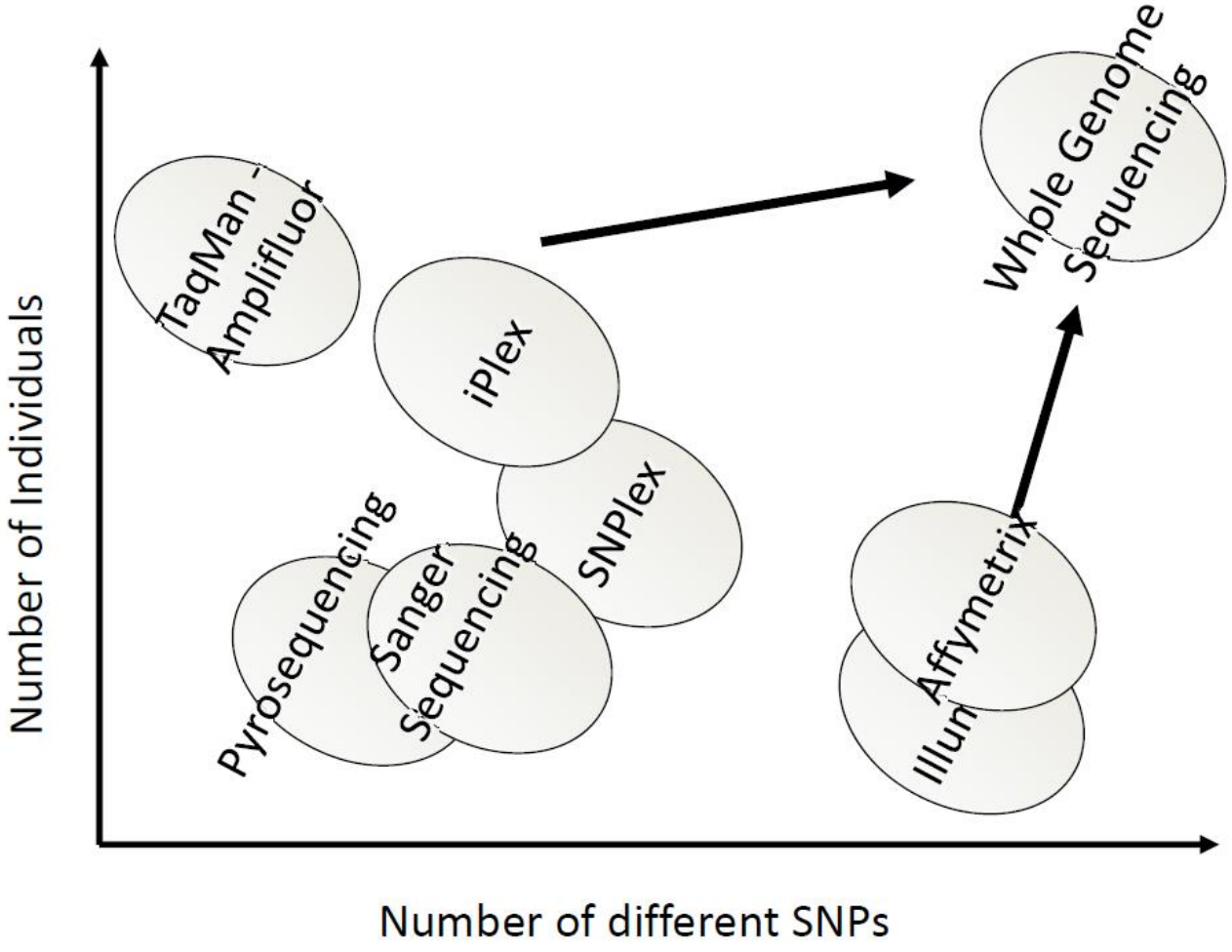


- Are sequences alike?
  - **Heterologs. {*Heterologs differ in both origin and activity.*}**
  - **Homologs. {*Homologs have common origins but may or may not have common activity.*}**
    - Genes that share an arbitrary threshold level of similarity determined by alignment of matching bases are termed **homologous**.
    - **Homology** is a qualitative term that describes a relationship between genes and is based upon the quantitative similarity.
    - **Similarity** is a quantitative term that defines the degree of sequence match between two compared sequences.
    - Homology implies that the compared sequences diverged in evolution from a common origin.

- **Analogs. {*Analogs have common activity but not common origin.*}**
  - Genes or proteins that display the same activity but lack sufficient similarity to imply common origin are said to have **analogous** activity.
  - The implication is that analogous proteins followed evolutionary pathways from different origins to converge upon the same activity.
  - Analogs have homologous activity but heterologous origins.
- **Paralogs. {*Paralogs are homologs produced by gene duplication*.}**



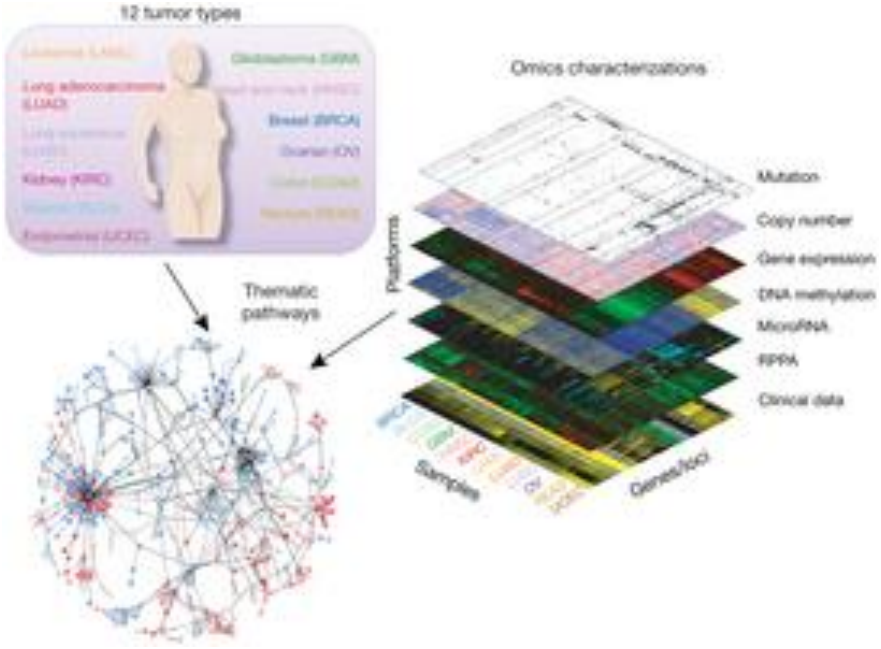
D. Genomic variation for complex diseases



Genomic variation for complex diseases

- High throughput in nr of individuals and variants matters
  - Only identical twins have the same DNA sequences
  - 2 x 10^7 bases in the human genome are variable
  - Average differences between two humans: 0.1% of their genome shared
  - Difference between human and chimpanzee is about 1% ….
- We are targeting very small percentages of the genome in which we can see “differences” or “variation” between ….
- This has been made possible by Next Generation Sequencing achievements…

- Sequencing (DNA, RNA, …) has already aided complex disease research by allowing scientists to catalogue certain genetic variations between individuals that may influence their susceptibility to different conditions or by identifying similar “patterns” between subgroups of patients (i.e., **molecular reclassification of patients**).



The TCGA Pan-Cancer project assembled data from thousands of patients with primary tumors occurring in different sites of the body, covering 12 tumor types. The idea of the TCGA PanCancer project was to integrate data set for comparing and contrasting multiple tumor types … (Weinstein et al. 2013)

Genomic variation for complex diseases

- In general, there are 3 common scenarios for human geneticists using NGS data
  - Identification of causitive genes in Mendelian disorders (germline mutations)
  - Identification of candidate genes in complex diseases for further functional studies (**complex diseases** are governed by multiple genes that are possibly interacting with each other and/or with environment)
  - Identification of constitutional mutations as well as driver and passenger genes in cancer (somatic mutations) (Pabinger et al 2013)

A **germline mutation** is one the was passed on to offspring because the egg or sperm cell was mutated.
A **somatic mutation** is a mutation of the somatic cells (all cells except sex cells) that cannot be passed on to offspring.

The application determines the statistical analysis tool

- The starting point of any sequencing project is the development of an appropriate study design, which starts (should start?) with a biological / research question
- Hence, the work flow for NGS presented earlier is only part of the story …

The application determines the statistical analysis tool

Suppose: You have been given a 5 KB piece of DNA sequence …

What to do next? …

- GeneScan: find any exons in the DNA sequence and generate a predicted protein sequence
- ScanProsite: scan the protein sequence for domains/motifs/patterns found in the prosite database [**Motifs** are structural characteristics and **domains** are functional regions]
- BLASTP: run a BLASTP search against the Swissprot database find some of the best matches (hits) and copy each protein sequence into a word doc for the alignment
- MultAlin: conduct protein sequence alignments from the BLASTP search

The application determines the statistical analysis tool

- The rule of thumb in the genomics community is that every dollar spent on sequencing hardware must be matched by a comparable investment in informatics (www.the-scientist.com/2011/3/1/60/1)
- There is a constant stream of new software
  - What is its quality?
  - How to install it?
  - How to get it working?

R code for DNA seq analysis problems (at home)

- R scripts illustrating relevant R packages for sequence pattern recognition and sequence-based analytics (see also practical session), includes:

  - DNA sequence statistics: http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter1.html
  - Quering sequence data bases: http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter3.html
  - Computational gene finding: http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter7.html

2 Investigating frequencies of occurrences of words

2.a Motivation

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

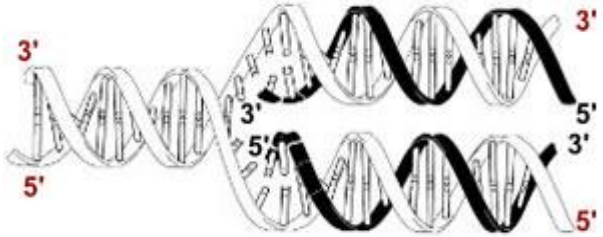
Introduction

R.F. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, Phys. Rev. Lett. 68 (1992) 3805.
W. Li, K. Kaneko, Long-range correlation and partial $1/f$ spectrum in a non-coding DNA sequence, Europhys. Lett. 17 (1992) 655;
W. Li, The study of correlation structures of DNA sequences: a critical review, Comput. Chem. 21 (1997) 257.
C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Long-range correlations in nucleotide sequences, Nature 356 (1992) 168.
S. Karlin, V. Brendel, Patchiness and correlations in DNA sequences, Science 259 (1993) 677.
D. Larhammar, C.A. Chatzidimitriou-Dreissman, Biological origins of long-range correlations and compositional variations in DNA, Nucleic Acids Res. 21 (1993) 5167.
C.L. Berthelsen, J.A. Glazier, M.H. Skolnick, Global fractal dimension of human DNA sequences treated as pseudorandom walks, Phys. Rev. A 45 (1992) 8902.
L. Luo, W. Lee, L. Jia, F. Ji, L. Tsai, Statistical correlation of nucleotides in a DNA sequence, Phys. Rev. 58 (1998) 861.
S. Nee, Uncorrelated DNA walks, Nature 357 (1992) 450.
V.V. Prabhu, J.M. Claverie, Correlations in intronless DNA, Nature 359 (1992) 782.
A.K. Mohanty, A.V.S.S. Narayana Rao, Factorial moments analyses show a characteristic length scale in DNA sequences, Phys. Rev. Lett. 84 (2000) 1832.
R. Román-Roldán, P.B. Galvan, J.L. Oliver, Application of information theory to DNA sequence analysis, Pattern Recogn. 29 (1996) 1187.
A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Characterizing long-range correlations in DNA sequences from wavelet analysis, Phys. Rev. Lett. 74 (1995) 3293.
X. Lu, Z. Sun, H. Chen, Y. Li, Characterizing self-similarity in bacteria DNA sequences, Phys. Rev. E 58 (1998) 3574.
Z. Yu, V.V. Anh, B. Wang, Correlation property of length sequences based on global structure of the complete genome, Phys. Rev. E 63 (2000) 011903-1.

(Som et al. 2003)

Biological words of length 1 – base composition

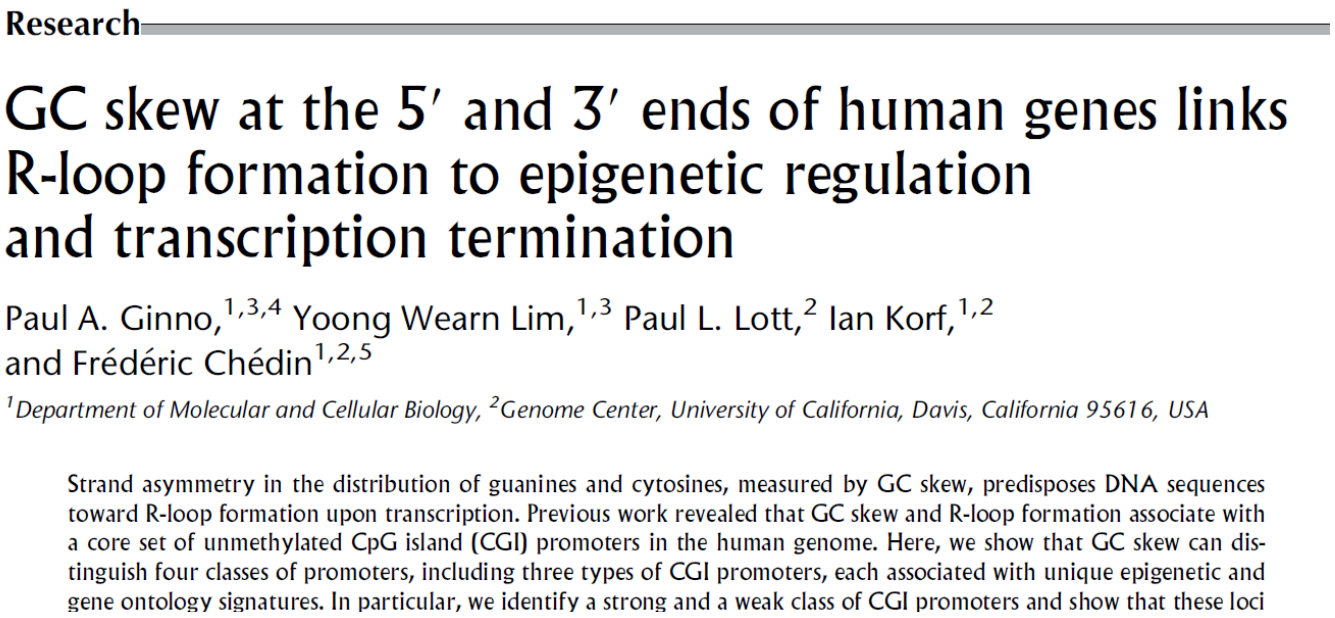
- There are constraints on base composition imposed by the genetic code
- The distribution of individual bases within a DNA molecule is not ordinarily uniform
  - There may be an excess of G over C on the leading strands



  - This can be described by the “GC skew”, characterized by:
    - (#G - #C) / (#G + #C)
    - # = nr of
  - What is the implication for AT skew on the lagging strand?

Biological words of length 1 – base composition

- GC or AT skew sign changes link to where DNA replication starts or finishes.
- Originally this asymmetric nucleotide composition was explained as different mechanism used in DNA replication between leading strand and lagging strand
- But recent research (2013) shows there is much more to it:



Research

GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination

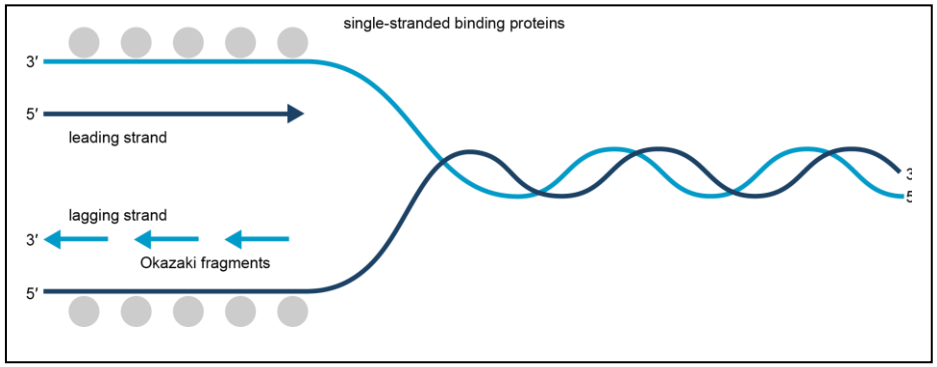
Paul A. Ginno,[1,3,4] Yoong Wearn Lim,[1,3] Paul L. Lott,[2] Ian Korf,[1,2] and Frédéric Chédin[1,2,5]

[1]Department of Molecular and Cellular Biology, [2]Genome Center, University of California, Davis, California 95616, USA

Strand asymmetry in the distribution of guanines and cytosines, measured by GC skew, predisposes DNA sequences toward R-loop formation upon transcription. Previous work revealed that GC skew and R-loop formation associate with a core set of unmethylated CpG island (CGI) promoters in the human genome. Here, we show that GC skew can distinguish four classes of promoters, including three types of CGI promoters, each associated with unique epigenetic and gene ontology signatures. In particular, we identify a strong and a weak class of CGI promoters and show that these loci

Biological words of length 1 - base composition

- DNA biosynthesis proceeds in the 5′- to 3′-direction. This makes it impossible for DNA polymerases to synthesize both strands simultaneously. A portion of the double helix must first unwind, and this is mediated by helicase enzymes.
- The leading strand is synthesized continuously but the opposite strand is copied in short bursts of about 1000 bases, as the lagging strand template becomes available. The resulting short strands are called Okazaki fragments (after their discoverers, Reiji and Tsuneko Okazaki).



- Only one strand is transcribed during transcription; the strand that contains the gene is called the sense strand

2.b 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
3. Statistics is an art. It uses mathematical methods but it is much more than maths alone
4. 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 → 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

Assumptions

- 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
- 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)
- Whatever we would like to do with such strings, we will need to introduce the concept of a random variable

Probability distributions

- Suppose the "machine" we are using produces an output X that takes exactly 1 of the J possible values in a set $\chi = \{l_1, l_2, \dots, l_n\}$
  - In the DNA sequence J=4 and $\chi = \{A, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$
  - L is a discrete random variables (since its values are uncertain)
  - If p_j is the prob that the value (realization of the random variable *L*) l_j occurs, then
    - $p_1, \dots, p_J \geq O$ and $p_1 + \dots + p_J = 1$
- The probability distribution (probability mass function) of L is given by the collection $p_1, \dots, p_J$
  - $P(L=l_j) = p_j$, $j=1, \dots, J$
- 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

- What is the probability distribution of the number of times a given pattern occurs in a random DNA sequence $L_1, \ldots, L_n$?
  - New sequence $X_1, \ldots, X_n$:

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

  - 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

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
- According to our simple model, the L_i are independent and hence

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

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 $Y = c\,X$, then $\mathrm{E}(Y) = c\,\mathrm{E}(X)$
  - $\mathrm{E}(X_1 + \ldots + X_n) = \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n)$
- Because X_i are assumed to be independent and identically distributed (iid):

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

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:
  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)

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 :

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

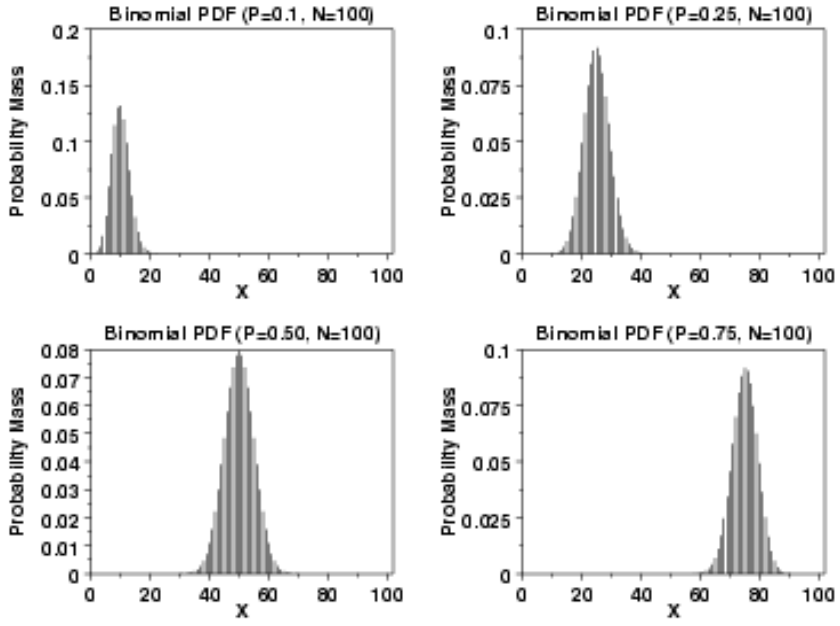
with the binomial coefficient $\binom{n}{j}$ determined by

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

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.



2.c 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_n having the same distribution as N
  - We can use the sample mean to estimate the expected value E(N):

$$\bar{N} = (N_1 + \ldots + N_n)/n$$

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

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

Why do we use (n-1) and not n 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

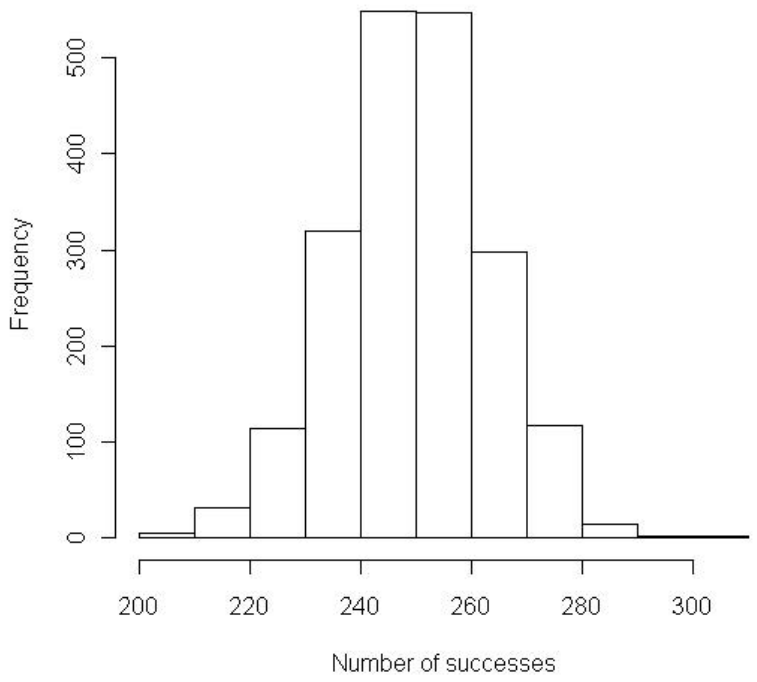
Simulating from probability distributions

- Simulate a sequence of bases $L_1, \ldots, L_n$:
  - Divide the interval (0,1) in 4 intervals with endpoints
    $$p_A, p_A + p_C, p_A + p_C + p_G, 1$$
  - 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 $L_1, \ldots, L_n$
- 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 and analyzing each of them, we can get a feel for whether or not our particular pattern of interest is unusual




```
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)

```
> rbinom(1,1000,0.25)
[1] 250
```

Simulating from probability distributions

- Using R code:

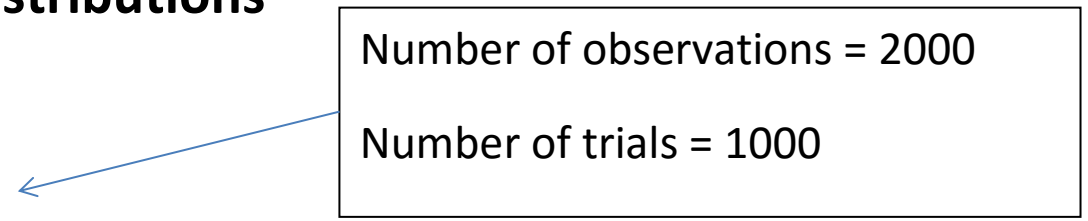
```
x<- rbinom(2000,1000,0.25)
mean(x)
sd(x)^2
hist(x,xlab="Number of successes",main="")
```

What is the number of observations?

- 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

Simulating from probability distributions

- Using R code:




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

What is the number of observations?

- 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

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

$$= 0.00019359032194965841$$

	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	0.0001935903219496584 1
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	0.0001935903219496584 1	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 distribution, $P(N \geq 300)$ can be estimated as 0.000196 via

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

- 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)
```

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,

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

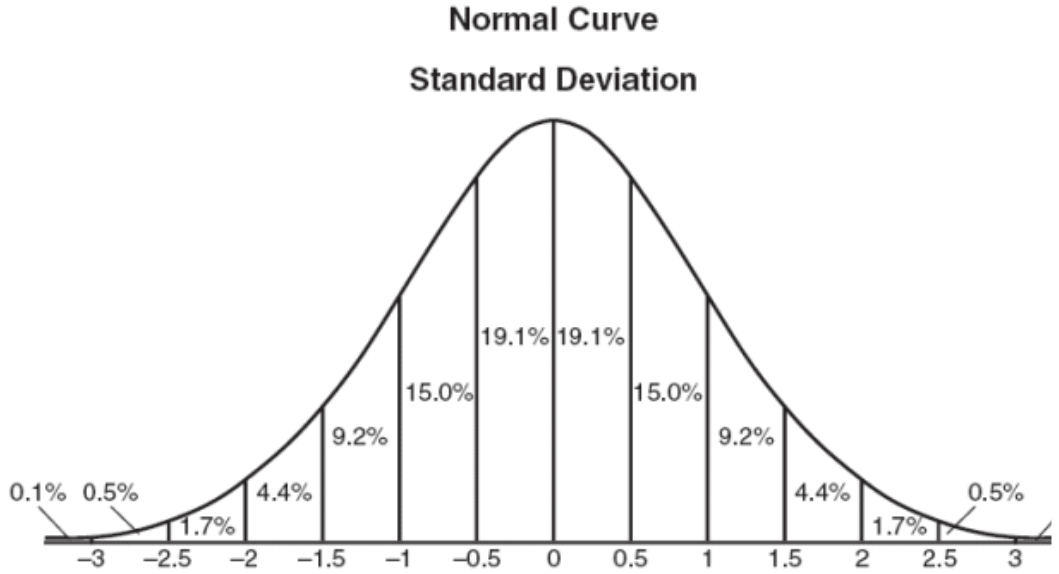
Approximate via Central Limit Theory

- 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

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

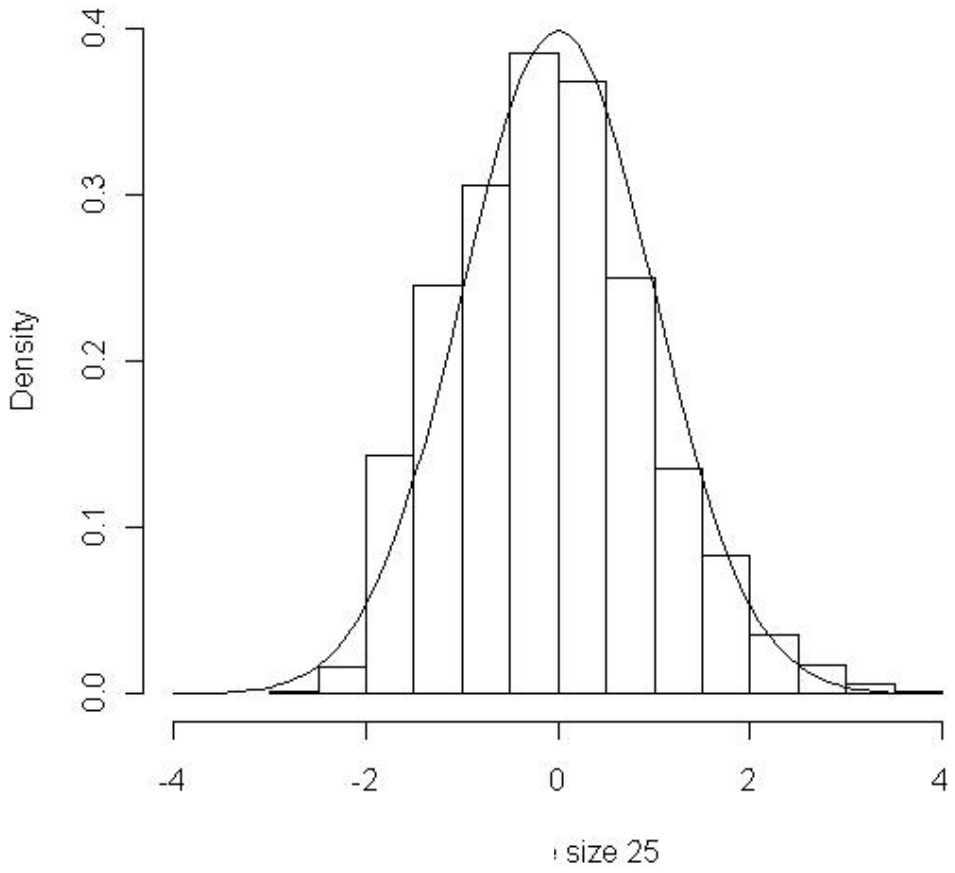


Approximate via Central Limit Theory

- 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



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?

3 Study examples

3.a Studying words of length 2

Introduction

- Dinucleotides are important because physical parameters associated with them can describe the trajectory of the DNA helix through space (such as DNA bending), which may affect gene expression.
    - For example: CC dinucleotides contribute to the bending of DNA in chromatin (Bolshoy 1995)
- Also occurrences of CGs are of interest … (see before)
- Recall: the CpG sites or CG sites are regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length.

CpG sites

Left: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG consitutes the start codon.
Right: CpG sites present at every 1/100 nucleotides, consituting a more normal example of the genome, or a region of the genome that is commonly methylated.

CpG sites

CpG Dinucleotide Distribution and DNA Methylation

Tom Shimizu[1,2] tom@sfc.keio.ac.jp
Kouichi Takahashi[1,3] t94249kt@sfc.keio.ac.jp
Masaru Tomita[1,3] mt@sfc.keio.ac.jp

[1] Laboratory for Bioinformatics, [2] Graduate School of Media and Governance,
[3] Department of Environmental Information,
Keio University
5322 Endo, Fujisawa, Kanagawa 252 Japan

It is known that the dinucleotide CpG is significantly underrepresented in genomic sequences of organisms which extensively methylate their DNA[1]. In these species, most cytosine bases of CpG dinucleotides are found to be methylated and this extensive CpG methylation is thought to have caused the depletion of the dinucleotide over the course of evolution[2]. Thus, the extent of CpG depletion in the genomic sequence can serve as an index of the extent of CpG methylation in an organism.

CpG islands are small regions of these CpG-depleted genomes which have remained relatively CpG-rich, and are usually unmethylated[3]. They are associated with most housekeeping genes and many tissue-specific genes and are most often found in the 5′ flanking region[4]. It is also known that the methylation state of CpG islands is sometimes associated with gene suppression.

Occurrences of 2-words

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

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

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

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

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

3.b Studying words of length 3

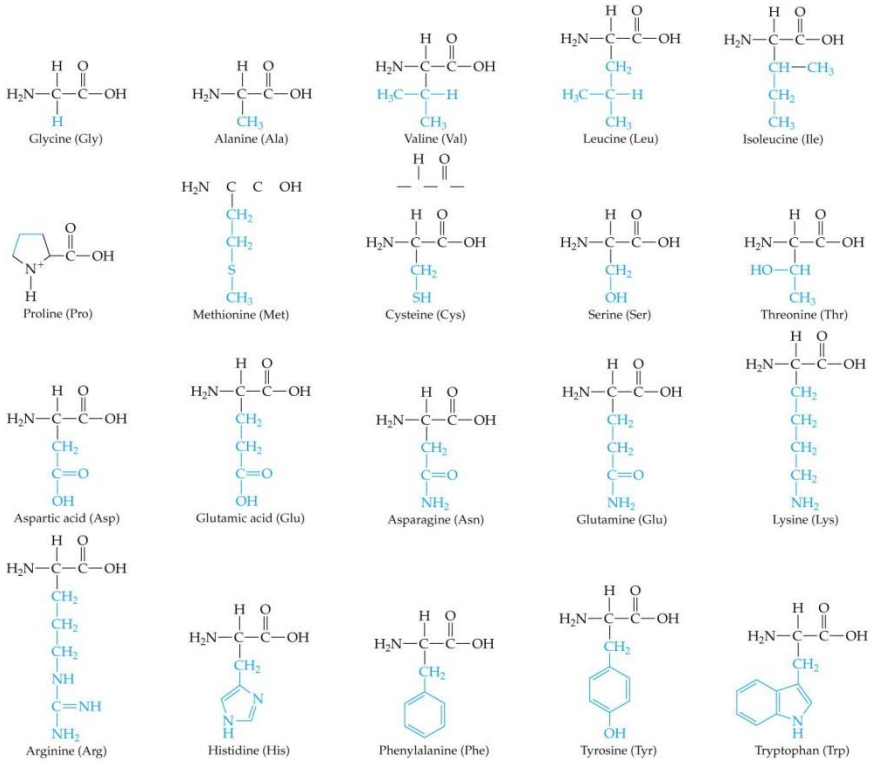
Amino acids

- 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?)

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 e 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 class II genes
- 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

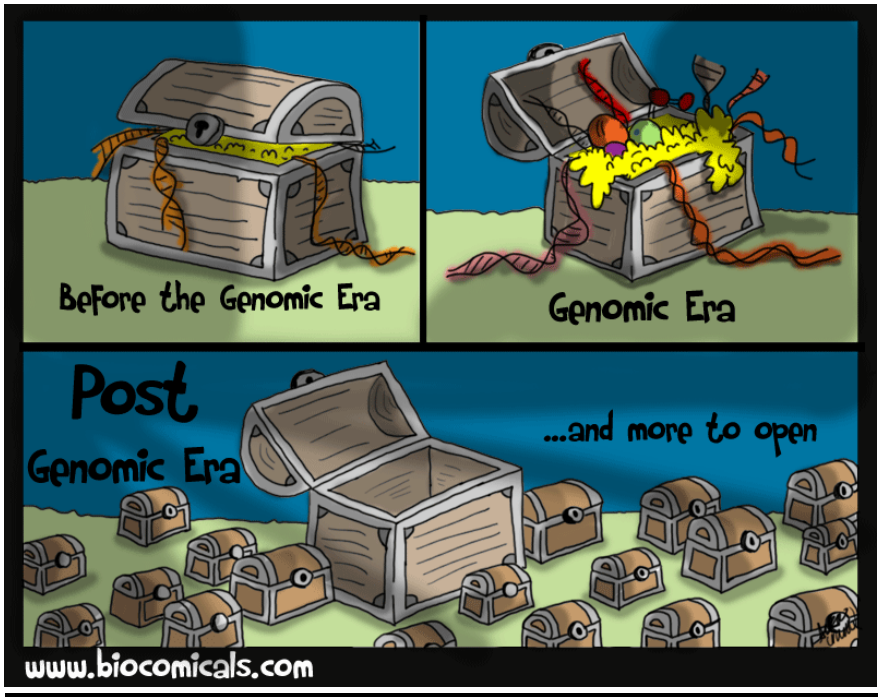
Class I : Moderately expressed genes

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

4 Rare variants in humans

4.1 Motivation

- There has been a lot of discussion on rare variants (loosely defined as <1% population frequency) and its effect on complex traits in the genetics literature lately.



Studies < 2011 pursuing rare variant association analyses

Phenotype	Method	Sample	Genes	Variants	Associated	Comments	Ref
HTG levels	CAST	438/327	4	187	154	Associated variants across 4 genes	133
Type 1 Diabetes	CS/FET	480/480	10	212	4	Four rare variants in one gene	22
Plasma HDL levels	FET	3551	4	93	NP	Rare NS cSNPs more frequent in low TG subjects	134
Plasma HDL levels	Observe	154/102	1	NP	3	5 carriers of rare variants with low HDL	135
Folate response	FET	564	1	14	5	Functional evaluation of NS mutations	136
Blood pressure	FET	3125	3	138	30	Rare mutations affect blood pressure	137
Plasma HDL levels	FET	95/95	1	51	3	Variants in ABCA1 influence HDL-C	138
Colorectal cancer	FET	691/969	1	61	NP	Rare NS variants in patients	139
Pancreatitis	CS	216/350	1	20	18	Rare variants common in patients	140
Tuberculosis	FET	1312	5	179	NP	Rare NS variants in tuberculosis cases	141
BMI	CS	379/378	58	1074	NP	Rare NS variants in obese vs. lean	142
HTG levels	CS	110/472	3	NP	10	Single common variant combined with rare variants = HTG	143
Heart Disease	CS	3363	1	2	2	Rare variants associated with lower plasma LDL	144
Plasma LDL levels	FET	3543	4	17	1	PCSK9 variants associated with low LDL	145
Plasma LDL levels	NP	512	1	26	NP	Variants in NPC1L1 associated with low cholesterol	146
Plasma LDL levels	NP	128	1	2	2	2 missense mutations associated with low LDL	147
Plasma AGT levels	FET	29/28	1	93	11	Rare haplotypes associated with high AGT levels	45
Plasma HDL levels	FET	519	3	NP	NP	Used collapsing of rare variants	148
Colorectal Adenoma	NP	124/483	4	NP	NP	25% vs. 12% rare variants in cases vs controls	149
Complex I	Observe	Pooled	103	898	151	More likely deleterious variants in Complex I Deficiency	150

Reference previous slide: Table 1 from Bansal et al. 2010

Key:

ABCA1: ATP-binding cassette transporter 1; HTG: Hypertryglycerides; HDL: High density lipoproteins; LDL: Low density lipoproteins; BMI: Body mass index; AGT: angiotensinogen;

CAST: Cohort allelic sums test[30]; CS: Contingency table Chi-square test; FET: Fisher's Exact Test; NP: Not Provided in the text in an obvious way;

NS, non-synonymous; cSNP, SNPs that occur in cDNAs; TG, triglycerides; PCSK9: Proprotein convertase subtilisin/kexin type 9; NPC1L1: Niemann-Pick C1 Like 1;

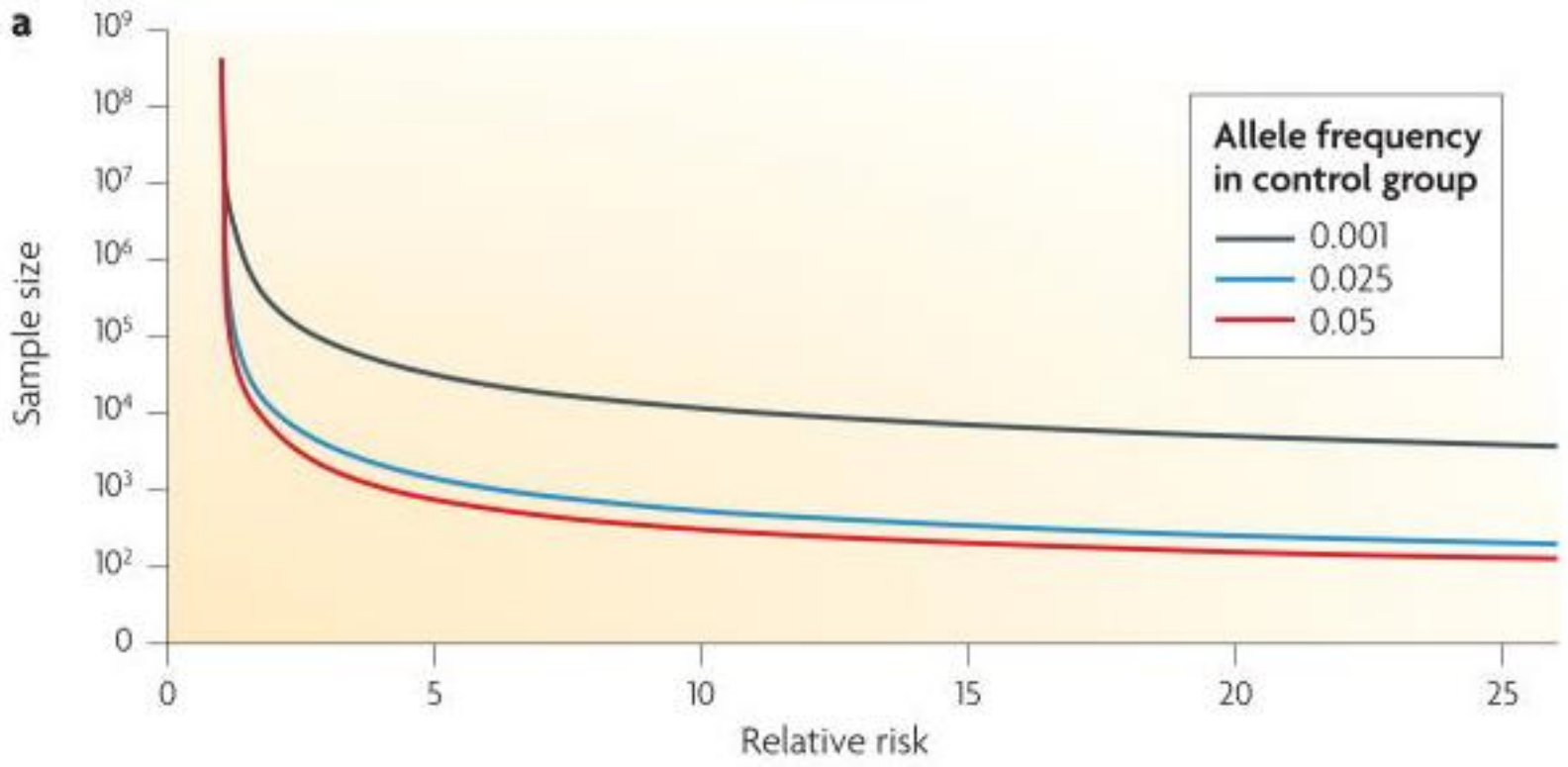
Genes = number of genes/genomic regions sequenced;

Variants = total number of variants found;

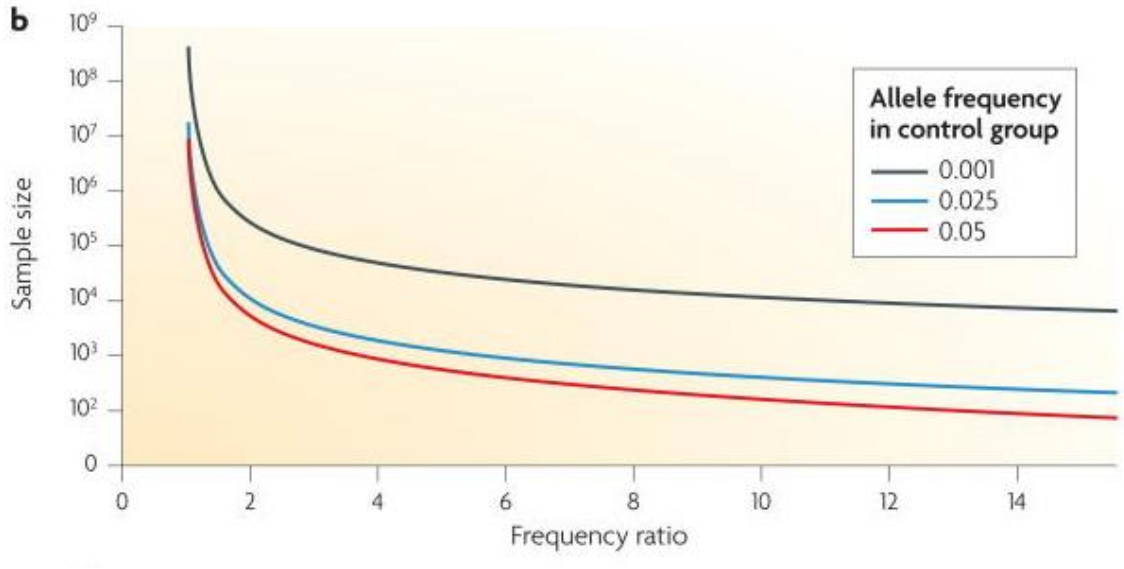
Associated = number of variants associated with the phenotype.

4.2 Rare variant association analysis

- A new paradigm that a lot of labs are pursuing at the moment with their favorite traits is to use NGS to identify novel variants and then take these variants on a genotyping ride with larger cohorts.
- How many samples does it take? (see next slides and Bansal et al. 2010)
  - Links to “power”
  - Links to “test statistic”
  - Links to “hypotheses”



Sample sizes necessary to detect an association between an allele with a specific effect size and a binary trait. The plots assume a standard z-test for the difference in the frequency of the allele between the two phenotypic categories. A genome-wide type I error rate of 10^{-9} was assumed, under the assumption that one may perform 2 orders of magnitude more tests in a complete sequence-based GWAS than a standard GWAS.



X axis gives the ratio of the frequency of the allele in the case vs. control groups. The curves give insight into the power gains associated with the collapsing strategy. Consider the black line in Figure B and testing a single rare variant with a frequency of 0.01 in the controls and 0.02 in the cases. This difference would require approximately 250,000 cases and controls to detect with 80% power at a super genome-wide level of significance. However, if one were to test 5 such variants with the same frequencies after collapsing them (assuming they are independent and no individual has more than one such variant), then one would effectively be testing a 0.05 frequency among the controls and a 0.10 frequency among the cases. From the red line in Figure B this difference would require only 3000 cases and controls.

Issues impacting the interpretation of a rare variants association method

- Appropriately sophisticated methods for identifying variants, assigning genotypes, and sampling individuals are crucial for rare variant analyses

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)

- ***Sequencing and Genotyping errors***
  - It has been shown that differential genotyping error rate can have substantial impact on common-variant based GWA studies.
  - Given that current sequencing protocols have inherent error rates, more research is needed to understand how false positive variant calls and nucleotide mis-assignments in sequence-based association studies of rare variants will impact inferences.
- ***Phasing***
  - Leveraging phase information in an association study of rare variants may be crucial, but obtaining phase from sequence data alone is not trivial.

(Bansal et al. 2010)

- ***Stratification***
  - The potential for false positive associations due to population stratification is large in studies involving rare variants since specific rare variants are more likely to be unique to a particular geoethnic group.
  - Thus, even if focus in a rare variant study is on a particular gene or genomic region, it is important to genotype the individuals in the study on enough additional markers to assess and control for stratification using standard strategies.
- ***The Use of In Silico Controls***
  - The practice of identifying and quantifying allele frequencies in a group of individuals and comparing them with historical or publicly available control sets in studies involving rare variants is highly problematic due to the potential for stratification and sampling variation effects.
  - In order to avoid this, either sophisticated genetic background matching strategies or **de novo** sequencing of a case and control group are recommended, but more work in this area is needed.

- ***Genomic Units of Analysis***
  - Different strategies for testing a genomic region for association involving rare variants exist.
  - For example, one could test all the variants in a region (depending on its size) for collective frequency differences between, e.g., cases and controls, define particular regions of interest, such as exons or transcription factor binding sites, or pursue a moving window analysis in which variants in contiguous, possibly overlapping, subregions are tested.
  - Each of these strategies impacts the number and nature of multiple testing problems.

(Bansal et al. 2010)

Integrative Web-Servers for Variant Annotation

Server Name	URL	Types of variant annotated
FASTSNP	http://fastsnp.ibms.sinica.edu.tw/	Precalculated SNPs
F-SNP	http://compbio.cs.queensu.ca/F-SNP/	Precalculated SNPs
Human Splicing Finder	http://www.umd.be/HSF/	Any Sequence / Splicing Only
MutDB	http://mutdb.org/	Precalculated SNPs
PharmGKB	http://www.pharmgkb.org/index.jsp	Pharmacogenetic SNPs
PolyDoms	http://polydoms.cchmc.org/polydoms/	Precalculated SNPs
PupaSuite	http://pupasuite.bioinfo.cipf.es/	Precalculated SNPs
SeattleSeq	http://gvs.gs.washington.edu/SeattleSeqAnnotation/	Any Sequence
Sequence Variant Analyzer	http://www.svaproject.org/	Any Sequence
SNP@Domain	http://bioportal.net/.	Precalculated SNPs
SNPeffect	http://snpeffect.vib.be/	Precalculated SNPs
SNP Functional Portal	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx	Precalculated SNPs
Trait-o-matic	http://snp.med.harvard.edu/	SNPs Associated with Traits

(Bansal et al. 2010)

A note about multiple testing

- Recall we are testing ~1 Million markers, more or less
- Several strategies to adjust the p-values for doing so many tests
  - Bonferroni
    - 0.05/{# tests, i.e., # markers, M}
    - most widely used in practice
    - Pr(Reject any test | null hypothesis true) = 0.05
  - False Discovery Rate (FDR)
  - Permutation

A note about multiple testing – FDR

- False Discovery Rate (FDR) limits the expected number of false positives
- Less stringent control than Bonferroni, e.g.
- *"Another way to look at the difference is that a p-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives. The latter is clearly a far smaller quantity."*

  (http://www.nonlinear.com/support/progenesis/samespots/faq/pq-values.aspx)

Number of errors committed when testing M null hypotheses

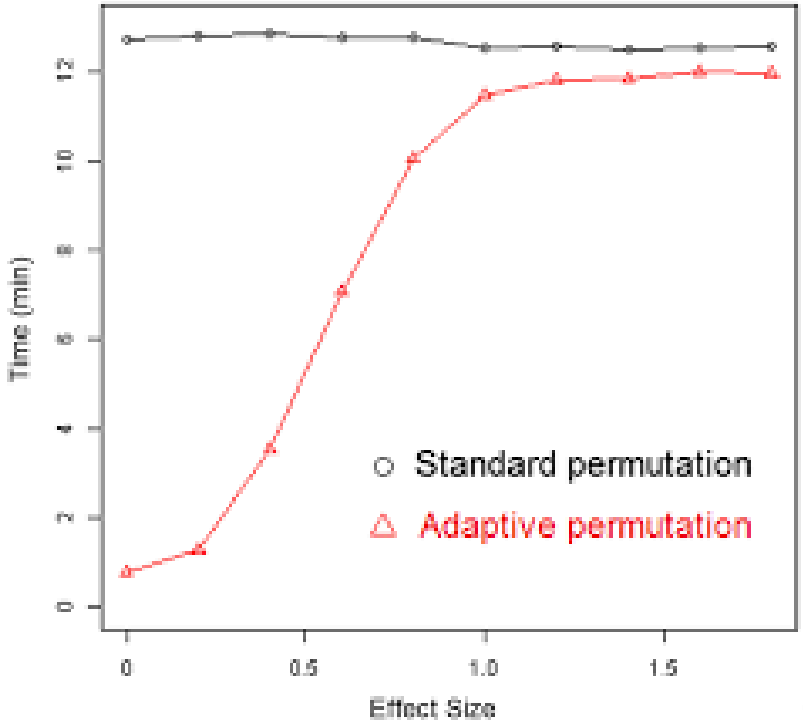
	Declared non-significant	Declared significant	Total Total
true null hypotheses	U	V	M_0
non-true null hypotheses	T	S	$M - M_0$
	$M - R$	R	M

Then the false discovery rate is given by

$$E\left(\frac{V}{V+S}\right).$$

A note about multiple testing – permutations

- Many of the tested genotype markers are correlated with each other (in LD), and so the tests are correlated
- Bonferroni adjusts as if they were completely independent
- Permutation will be more powerful, but… ?



(Che et al. 2014)

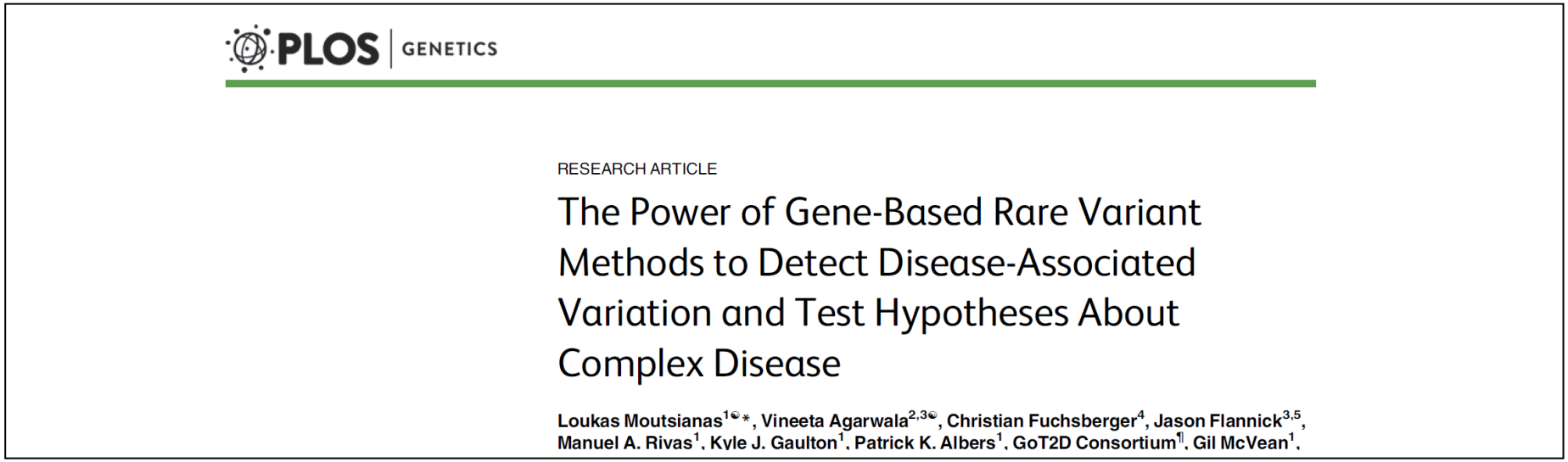
A note about multiple testing – in conclusion

- Nan Laird comments:

> "Given the many false positive findings
> in the history of genetic association studies,
> one rather errs
> on being too conservative."

- Initial GWAS had a lot of false positives (replication, replication, replication...)

Increasing power … (Moutsianas 2014)


PLOS | GENETICS

RESEARCH ARTICLE

The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease

Loukas Moutsianas[1☯*], Vineeta Agarwala[2,3☯], Christian Fuchsberger[4], Jason Flannick[3,5], Manuel A. Rivas[1], Kyle J. Gaulton[1], Patrick K. Albers[1], GoT2D Consortium[¶], Gil McVean[1],

"For loci explaining ~1% of phenotypic variance underlying a common dichotomous trait, we find that all methods have **low absolute power** to achieve exome-wide significance (~5-20% power at α=2.5×10-6) in 3K individuals; even in 10K samples, power is modest (~60%). The **combined application of multiple methods** increases sensitivity, but does so at the expense of a higher false positive rate. MiST, SKAT-O, and KBAC have the highest individual mean power across simulated datasets, but we observe wide **architecture-dependent variability** in the individual loci detected by each test …"

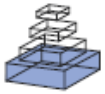
Types of association tests

- Because we have **less statistical power** with rare variants there are a lot of different approaches proposed to **aggregate variants** for association analysis (as opposed to the GWAS-type analysis).

- For a comprehensive review, see

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]**

Types of association tests

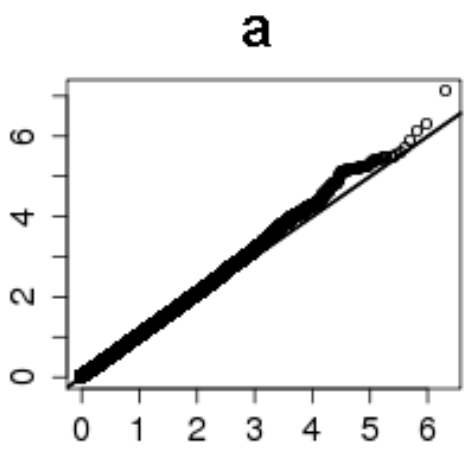
- There are three levels of rare variant data:
  - **Level 1**: Individual-level
  - **Level 2**: Summarized over subjects
  - **Level 3**: Summarized over both subjects and variants

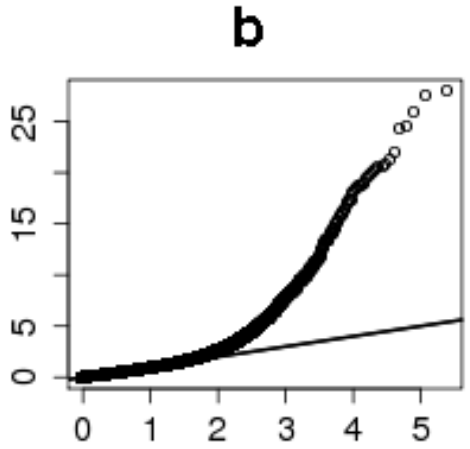
- When not cautious enough, an increased false positive number of findings is to be expected as well (lessons learned from GWAs).

An Example of High False Positive Rate

Q-Q plots from GWAS data, unpublished

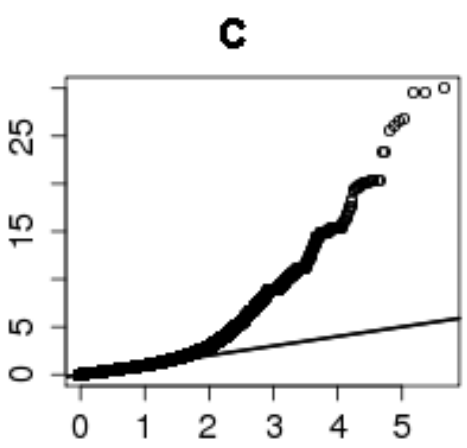
N=~2500
MAF>0.03

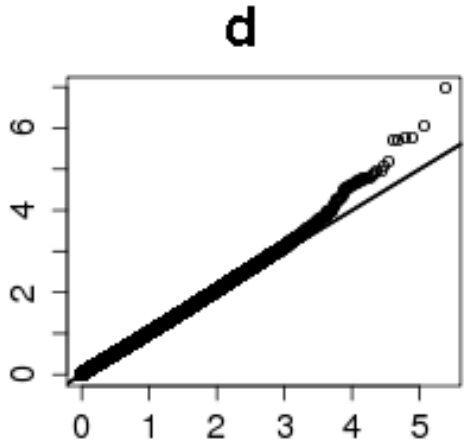




N=~2500
MAF<0.03

N=~2500
MAF<0.03
Permuted





N=50000
MAF<0.03
Bootstrapped

Level 1 data: individual level

Subject	V1	V2	V3	V4	Trait-1	Trait-2
1	1	0	0	0	90.1	1
2	0	1	0	.	99.2	1
3	0	0	0	0	105.9	0
4	0	0	0	0	89.5	0
5	0	.	0	0	97.6	0
6	0	0	0	0	110.5	0
7	0	0	1	0	88.8	0
8	0	0	0	1	95.4	1

Level 1 data analysis: Collapsing C test

(Li and Leal,The American Journal of Human Genetics 2008(83): 311–321)

- **Step 1:**

define an indicator variable X for the j^{th} case individual as

$$X_j = \begin{cases} 1 & \text{rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

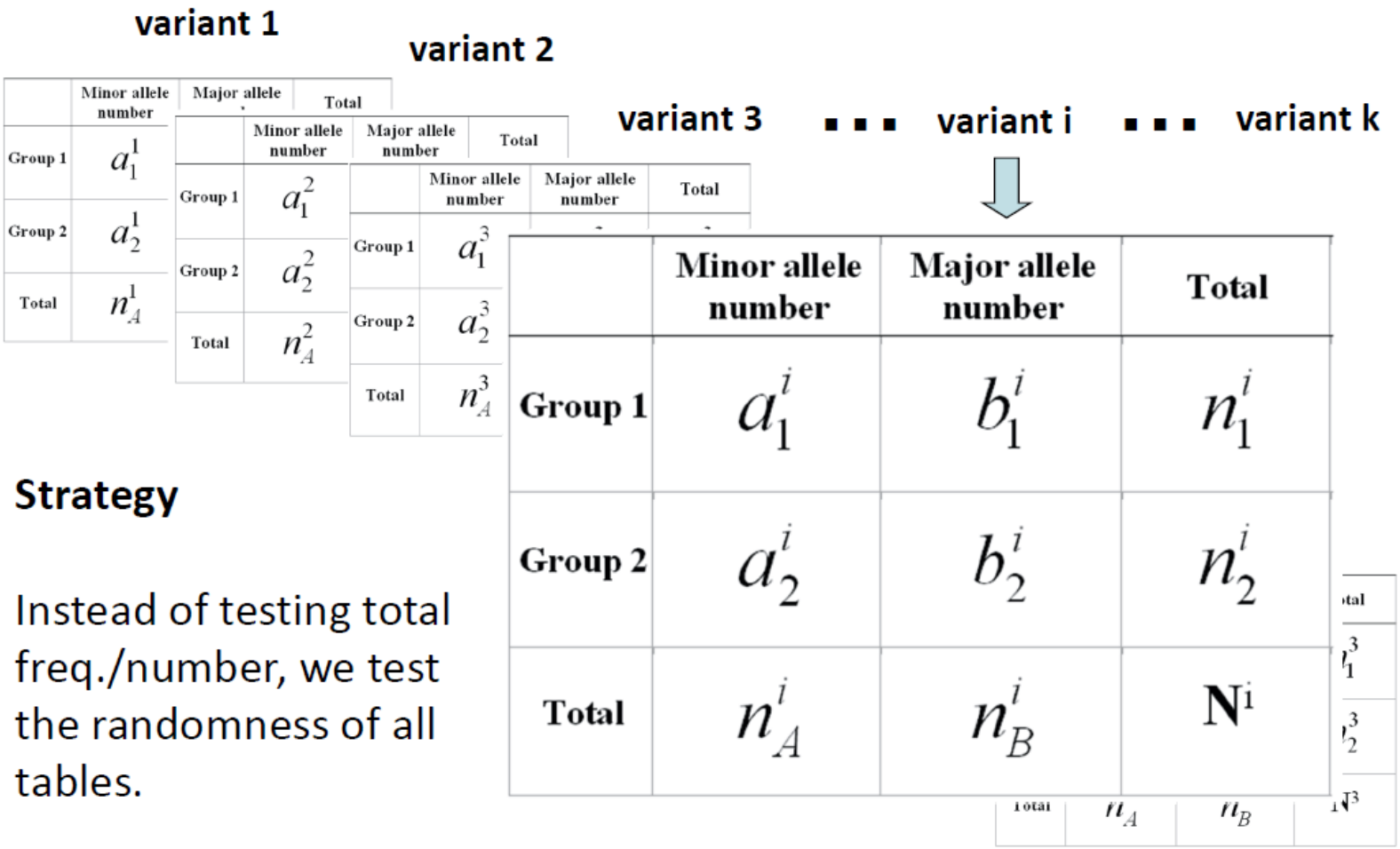
- **Step 2:**

logit(y)=a + b* X + e (logistic regression)

Level 2 data: summarized over subjects (groups)

	Low-HDL group			High-HDL group			
Variants in ABCA1	variant number	n	2n	variant number	n	2n	P-value
c.593C_A	1	128	256	0	128	256	1
c.742G_A	1	128	256	0	128	256	1
c.1201A_C	1	128	256	0	128	256	1
c.1769G_C	1	128	256	0	128	256	1
c.1913G_A	1	128	256	0	128	256	1
c.2320A_T	4	128	256	0	128	256	0.12359
c.2320A_T	1	128	256	0	128	256	1
c.2444A_G	1	128	256	0	128	256	1
c.3542C_T	1	128	256	0	128	256	1
c.4022G_C	1	128	256	0	128	256	1
c.4126A_G	1	128	256	0	128	256	1
c.4844G_A	1	128	256	0	128	256	1
c.5008G_A	1	128	256	0	128	256	1
c.5398A_C	4	128	256	0	128	256	0.12359
c.1486C_T	0	128	256	1	128	256	1
c.5039G_A	0	128	256	1	128	256	1

Level 2 data analysis



Strategy

Instead of testing total freq./number, we test the randomness of all tables.

Level 2 data analysis

- Calculating the probability P_i of each table based on hypergeometric distribution

  Hypergeometric Formula: Suppose a population consists of *N* items, *k* of which are successes. And a random sample drawn from that population consists of *n* items, *x* of which are successes. Then the hypergeometric probability is: C(k,x) C(N-k, n-x) / C(N,n)

- $P_i = C\left(n_1^i, a_1^i\right) \times C\left(n_2^i, a_2^i\right) / C\left(N^i, n_A^i\right)$
- Calculating the logarized joint probability (L) for all *k* tables: $L = \sum_{i=1}^{k} \log(P_i)$
- Enumerating all possible tables and L scores

  $\hat{\mathrm{L}} \ni (L_1, L_2, \ldots, L_{\mathrm{j}}, \ldots \mathrm{L_M})$

- Calculating p-value P= Prob.($L_j \geq L$)

Level 3: Summarized over subjects (by group) and variants (e.g., gene)

	Low-HDL group			High-HDL group			
Variants in ABCA1	variant number	n	2n	variant number	n	2n	P-value
c.593C_A	1	128	256	0	128	256	1
c.742G_A	1	128	256	0	128	256	1
c.1201A_C	1	128	256	0	128	256	1
c.1769G_C	1	128	256	0	128	256	1
c.1913G_A	1	128	256	0	128	256	1
c.2320A_T	4	128	256	0	128	256	0.12359
c.2320A_T	1	128	256	0	128	256	1
c.2444A_G	1	128	256	0	128	256	1
c.3542C_T	1	128	256	0	128	256	1
c.4022G_C	1	128	256	0	128	256	1
c.4126A_G	1	128	256	0	128	256	1
c.4844G_A	1	128	256	0	128	256	1
c.5008G_A	1	128	256	0	128	256	1
c.5398A_C	4	128	256	0	128	256	0.12359
c.1486C_T	0	128	256	1	128	256	1
c.5039G_A	0	128	256	1	128	256	1
total	20	128	256	2	128	256	0.000107

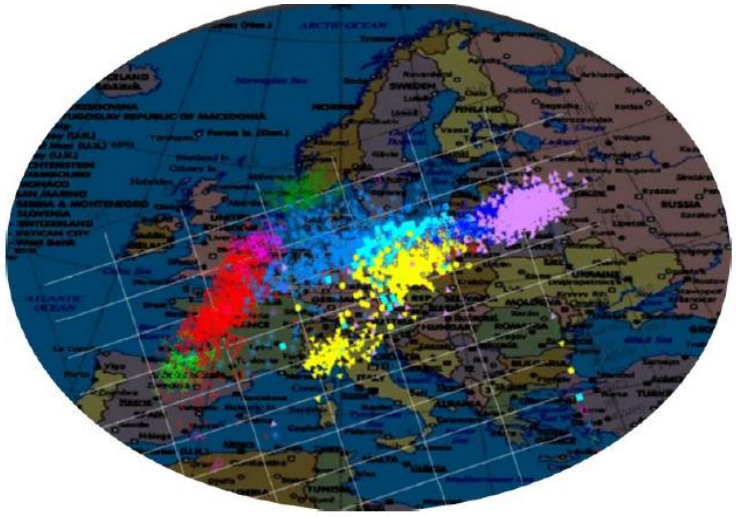
Level 3 data analysis – the total frequency test

	Low-HDL group			High-HDL group			
Variants in ABCA1	variant number	n	2n	variant number	n	2n	P-value
c.593C_A	1	128	256	0	128	256	1
c.742G_A	1	128	256	0	128	256	1
c.1201A_C	1	128	256	0	128	256	1
c.1769G_C	1	128	256	0	128	256	1
c.1913G_A	1						
c.2320A_T	4						
c.2320A_T	1						
c.2444A_G	1						
c.3542C_T	1						
c.4022G_C	1						
c.4126A_G	1						
c.4844G_A	1						
c.5008G_A	1						
c.5398A_C	4						
c.1486C_T	0						
c.5039G_A	0						
total	20						

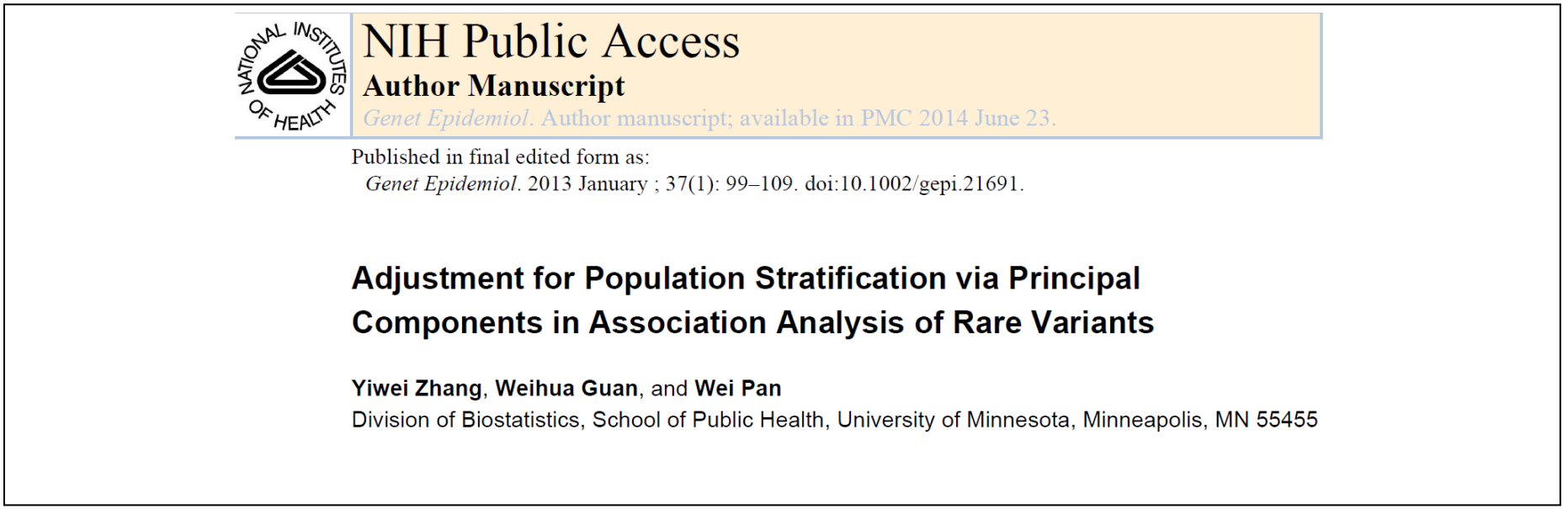
	Variant allele number	Reference allele number	Total
Low-HDL group	20	236	256
High-HDL group	2	254	256
Total	22	490	512

4.3 Rare variants to identify population substructure

- Popular method in GWAs to “correct” for population stratification = add principal components as covariates in a regression model
- Principal components analysis is a procedure for identifying a smaller number of uncorrelated variables, called **principal components**, from a large set of data.
- The **goal of principal components** analysis is to explain the maximum amount of variance with the fewest number of principal components.



- However, it is less clear how it would perform in analysis of low-frequency variants (LFVs, MAF between 1% and 5%), or of rare variants (RVs, MAF < 5%).
- Furthermore, with next-generation sequencing data, it is unknown whether PCs should be constructed based on CVs, LFVs or RVs.
- Although a few top PCs based on LFVs could better separate two continental groups considered in Zhang et al. (2012) than those based on CVs, the use of the former could lead to over-adjustment in the sense of substantial power loss in the absence of population stratification ….



Published in final edited form as:
Genet Epidemiol. 2013 January ; 37(1): 99–109. doi:10.1002/gepi.21691.

Adjustment for Population Stratification via Principal Components in Association Analysis of Rare Variants

Yiwei Zhang, **Weihua Guan**, and **Wei Pan**
Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

- … and to make it more complicated: what is the role of DNA methylation in characterizing population substructure?

Published in final edited form as:
Genet Epidemiol. 2014 April ; 38(3): 231–241. doi:10.1002/gepi.21789.

Accounting for Population Stratification in DNA Methylation Studies

Richard T. Barfield[1], **Lynn M. Almli**[2], **Varun Kilaru**[2], **Alicia K. Smith**[2], **Kristina B. Mercer**[2], **Richard Duncan**[4], **Torsten Klengel**[3], **Divya Mehta**[3], **Elisabeth B. Binder**[2,3], **Michael P. Epstein**[4], **Kerry J. Ressler**[2], and **Karen N. Conneely**[4]

Total number of sites associated with race, before and after correction for population stratification

Correction method used	# markers used to compute PCs	# FDR-significant CpG sites	# Holm-significant CpG sites	λ_{GC}
No correction	-	12827	912	2.09
GC	-	578	90	1
PC_{gwas}	54,610	13	3	1
$PC_{GWAS\ TW}$	54,610	19	4	1
$PC_{unpruned}$	469,142	1	1	1.08
$PC_{r2<0.25}$	225,440	0	0	1.06
$PC_{r2<0.1}$	121,855	0	0	1.11
PC_{0bp}	7,326	0	0	1.16
PC_{1bp}	17,105	1	1	1.18
PC_{2bp}	20,336	1	1	1.18
PC_{5bp}	31,178	1	1	1.12
PC_{10bp}	48,998	1	1	1.10
PC_{50bp}	174,510	1	1	1.02
PC_{100bp}	271,877	1	1	1.05

Type I error rate and power for analysis of a continuous trait, by method of correction for population stratification

	Rate of type I error		Power	
Correction method	**No population stratification**	**Stratification present**	**No population Stratification**	**stratification present**
No correction	0.0364	0.2690	0.964	---
Race included as covariate	0.0344	0.0344	0.963	0.963
GC	0.0116	0	0.908	0.662
PC_{gwas}	0.0348	0.0326	0.879	0.871
PC_{GWAS_TW}	0.0340	0.0322	0.962	0.951
$PC_{unpruned}$	0.0466	0.0478	0.885	0.860
$PC_{r2<0.25}$	0.0464	0.0514	0.888	0.861
$PC_{r2<0.1}$	0.0448	0.0500	0.893	0.857
PC_{0bp}	0.0418	0.0412	0.832	0.828
PC_{1bp}	0.0380	0.0374	0.880	0.858
PC_{2bp}	0.0390	0.0376	0.887	0.852
PC_{5bp}	0.0382	0.0436	0.888	0.856
PC_{10bp}	0.0404	0.0430	0.893	0.860
PC_{50bp}	0.0496	0.0462	0.894	0.869
PC_{100bp}	0.0464	0.0450	0.884	0.860

Reference previous slide: table 1 from Barfield et al. 2014

Key:
the set that were significant according to a Tracey-Widom test [Patterson et al., 2006] (PC_{GWAS_TW}) and the top 10 PCs (PC_{GWAS});
the top 10 PCs based on: the complete unpruned data ($PC_{unpruned}$), data pruned to keep only CpG sites with $r^2 < 0.25$ ($PC_{r^2<0.25}$), or data pruned to keep only CpG sites with $r^2 < 0.1$ ($PC_{r^2<0.1}$).
top 10 PCs based on CpG sites located: directly on a genetic variant (PC_{0bp}), within one (PC_{1bp}), two (PC_{2bp}), five (PC_{5bp}), 10 (PC_{10bp}), 50 (PC_{50bp}), or 100 bps (PC_{100bp}) of a genetic variant

Conclusion: principal components based on SNPs or methylation markers outperform genomic control