Topics in Bioinformatics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

Lecture 6: DNA sequence analysis

1 Introduction

1.a Generating DNA sequences

1.b Historical notes

1.c Application fields

2 Investigating frequencies of occurrencies 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

Markov Chains

3.b Words of length 3

3.c Restriction sites

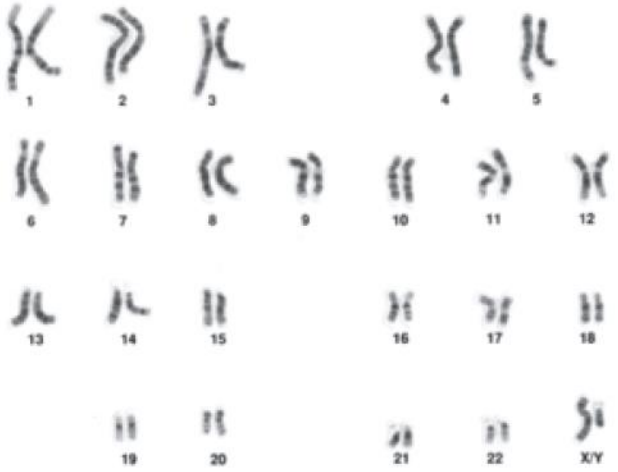
4 Comparing multiple sequences

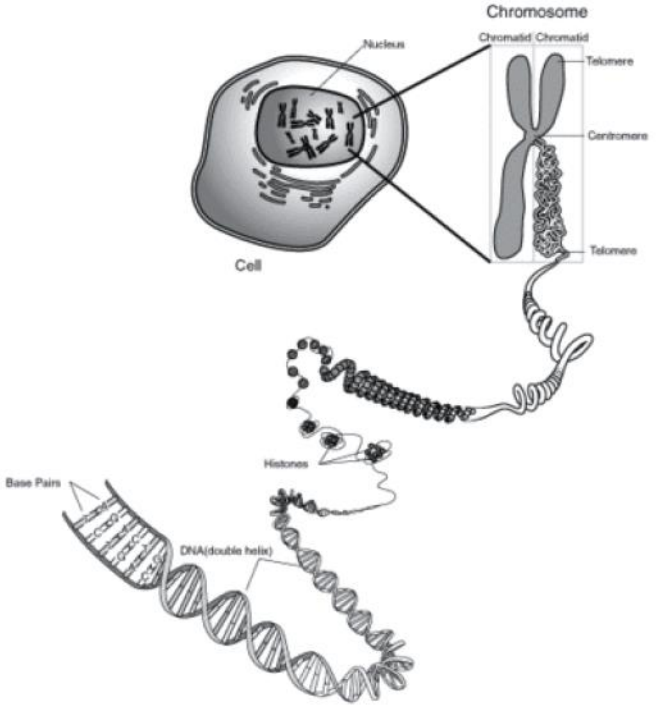
(see practical session)

1 Introduction

1.a Generating DNA sequences

Human genome



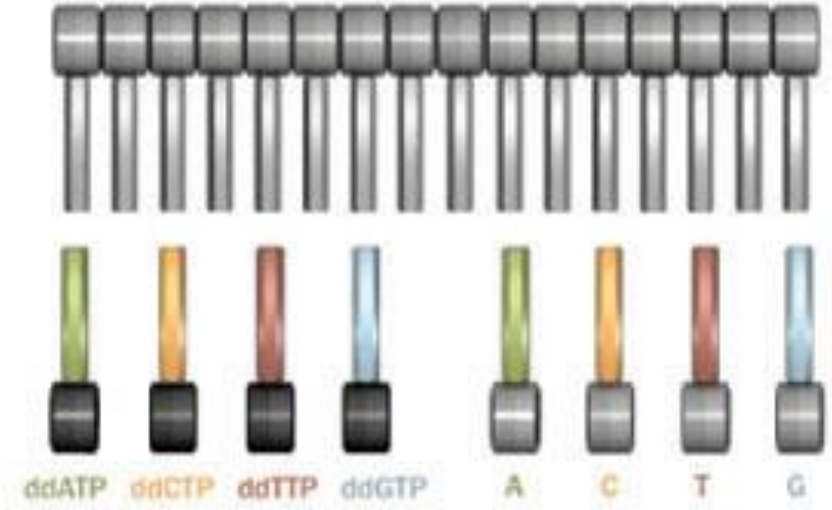


- 3 x 10^9 bases/nucleotides
- < 1 % coding
- 20.000 genes

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, lead by Maxam and Gilbert, used a “chemical cleavage protocol”, while
  - the English, lead 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).

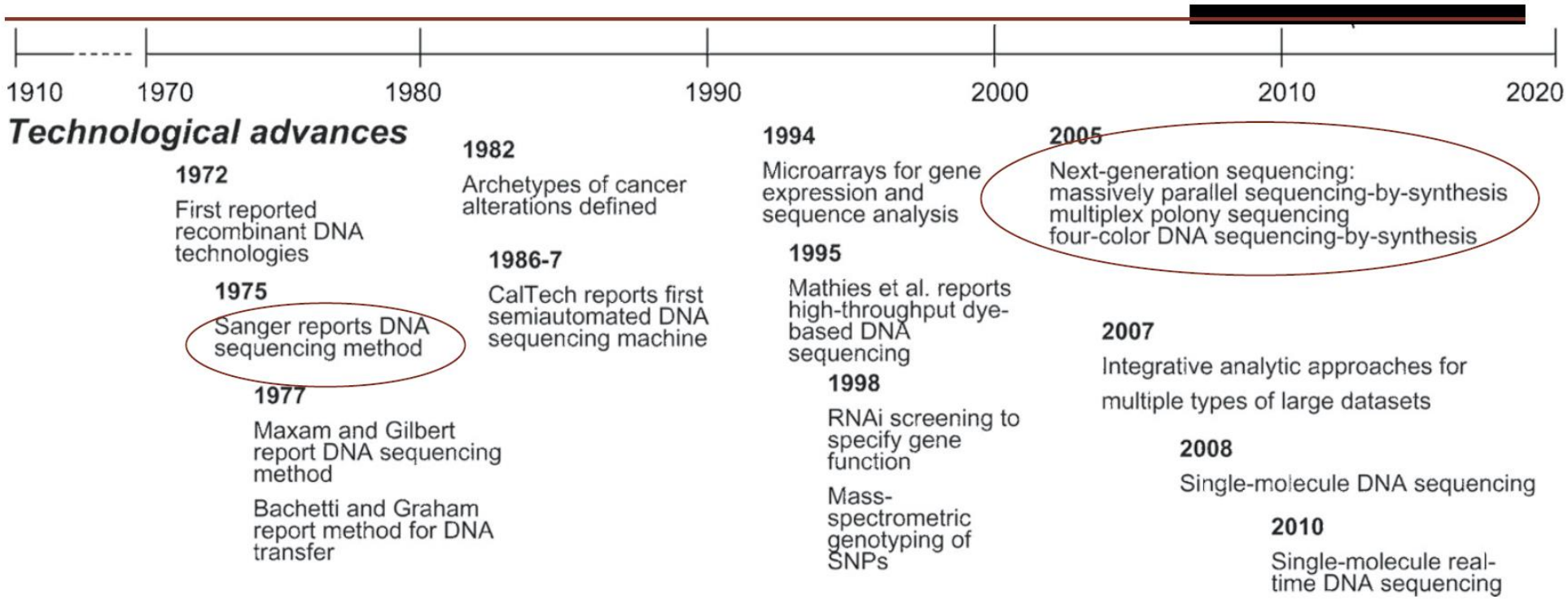
- Sanger's method, which is also referred to as dideoxy sequencing or chain termination, is based on the use of dideoxynucleotides (ddNTP's) in addition to the normal nucleotides (NTP's) found in DNA.



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)

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

Sanger Sequencing
ABI 3730
Electrophoresis
1000 base reads

Roche GS-FLX, LifeTechnologies SOLID5500, Illumina HiSeq2500
Clonal amplification

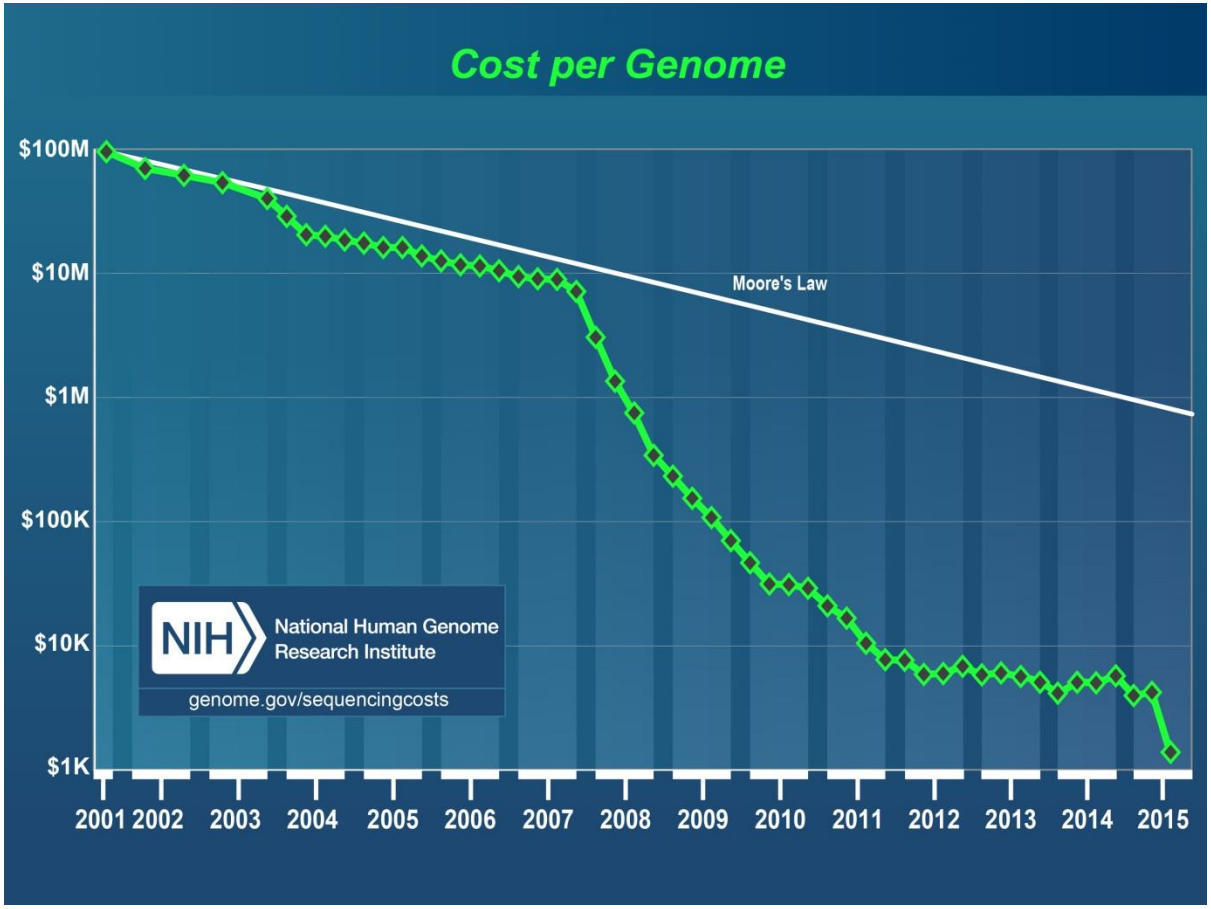
400 base, 75 base, 100 base reads

1st Generation
1977 -

2nd Generation
2005 -

(Ivo Gut, CSCDA2012)

New(er) generations of sequencing methods



Moore's Law is a computing term which originated around 1970; the simplified version of this law states that processor speeds, or overall processing power for computers will double every two years.

(http://www.genome.gov/sequencingcosts/)

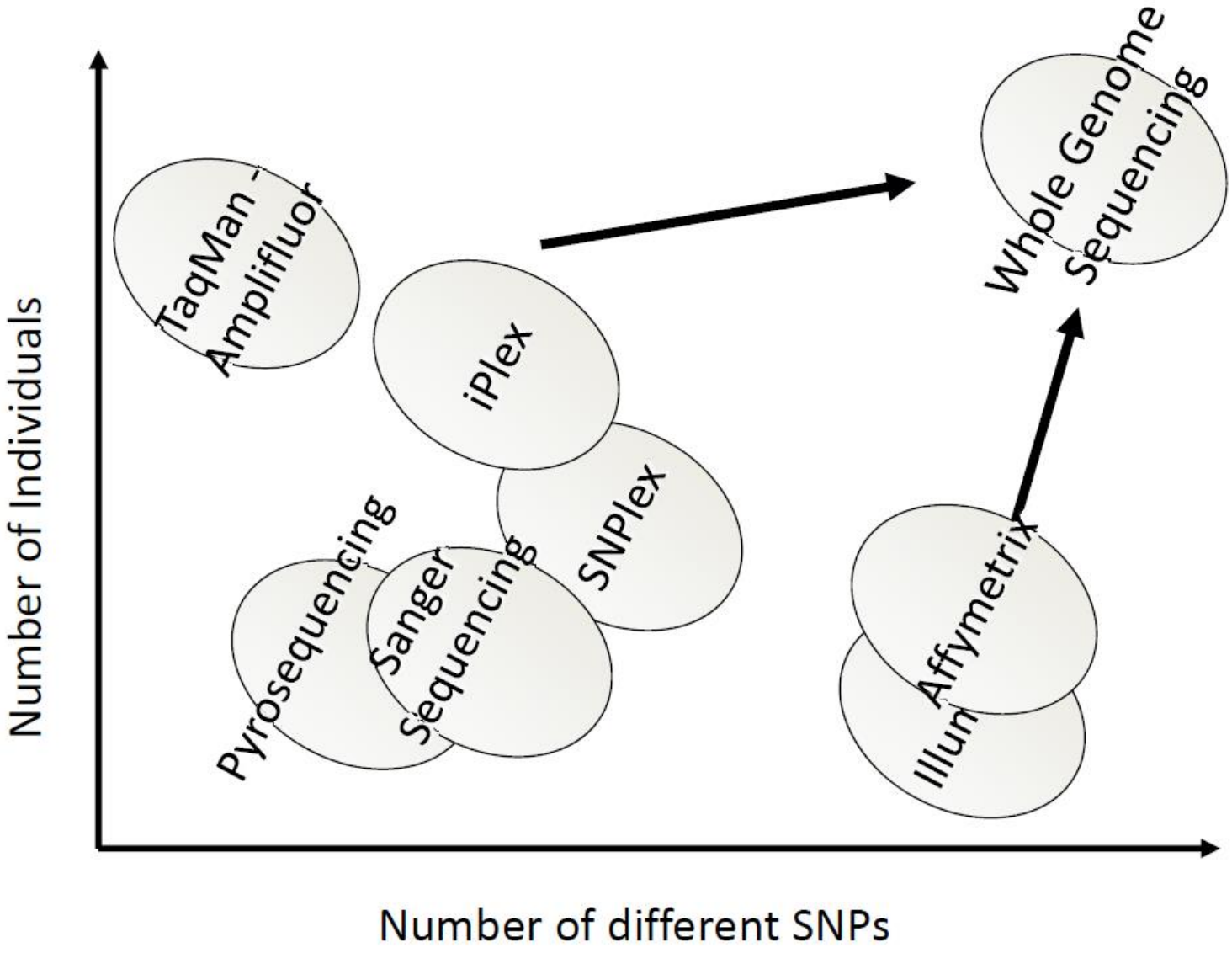
"Cost per Genome" - the cost of sequencing a human-sized genome

- Technology improvements that 'keep up' with Moore's Law are widely regarded to be doing exceedingly well, making it useful for comparison.
- Logarithmic scale on the Y axis
- Sudden and profound out-pacing of Moore's Law beginning in January 2008.
  - The latter represents the time when the sequencing centers transitioned from Sanger-based (dideoxy chain termination sequencing) to 'second generation' (or 'next-generation') DNA sequencing technologies.

(http://www.genome.gov/sequencingcosts/)

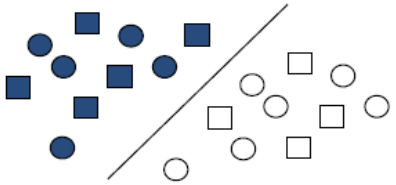
New(er) generations of sequencing methods

- 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
Illun
Whole Genome Sequencing

High throughput has been beneficial for GWA studies

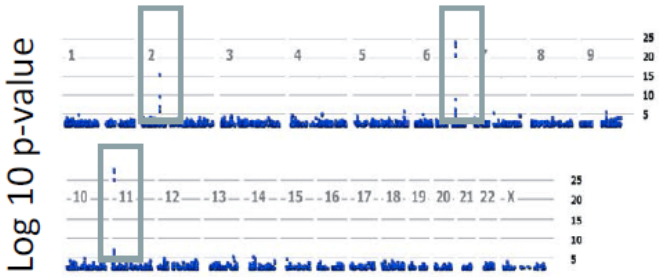
1.



case
control

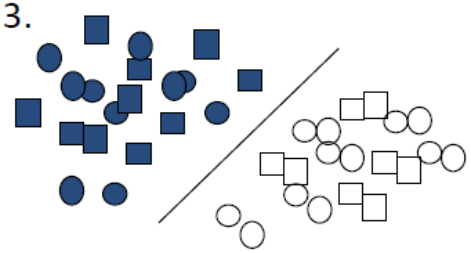
Scan genome in DNA collection
with > 600,000 common variants

2.



Identify regions of marked
differences in frequencies
of variants in cases/controls

3.



case
control

Confirm in additional larger
collections

(Ivo Gut, CSCDA2012)

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 vector itself is sequenced

    A **vector** is a small piece of DNA, taken from a virus, a plasmid, or the cell of a higher organism, that can be stably maintained in an organism, and into which a target DNA sequence can be inserted
  - 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.

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 ~20,000 human genes (at that time, one believed there were about 25,000 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.

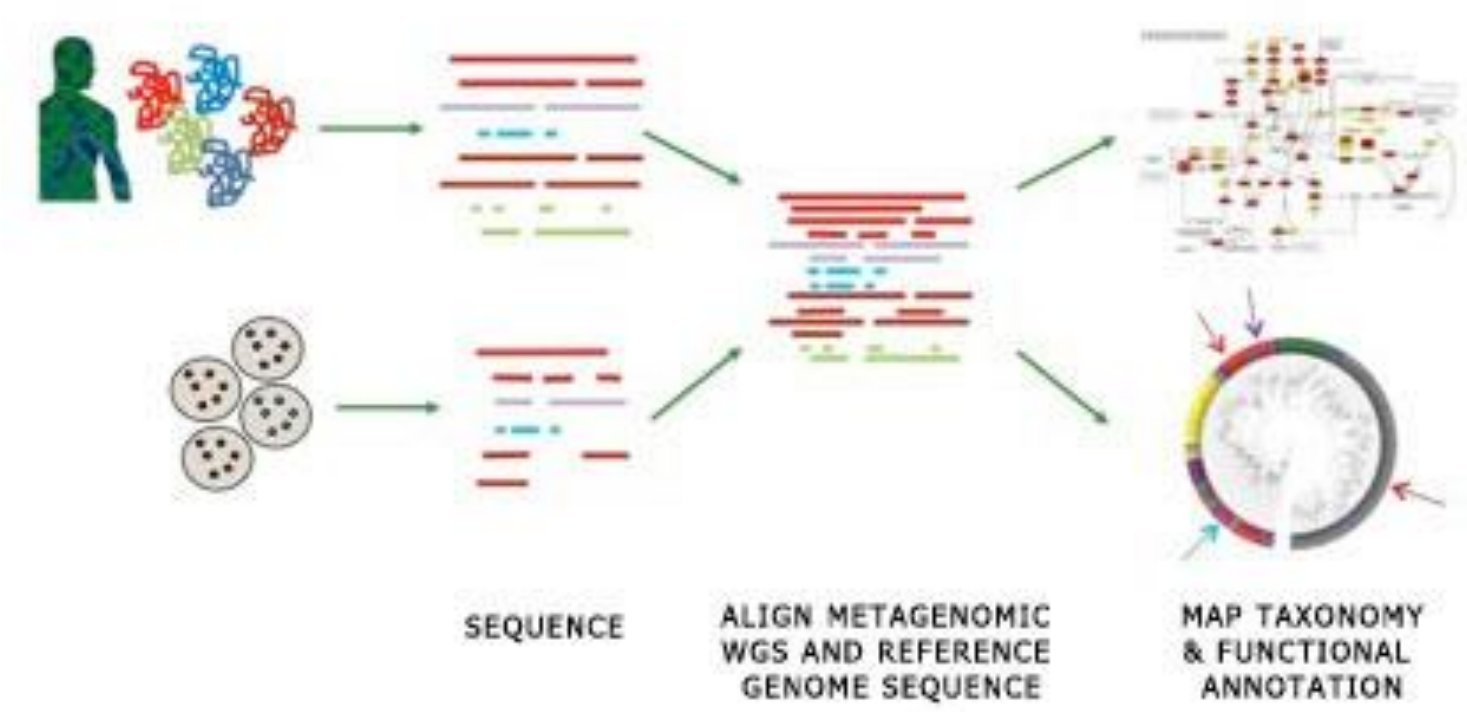
Vertebrates have a series of nerves along the back which need support and protection. That need brings us to the backbones and notochords. Vertebrates have a series of nerves along the back which need support and protection. That need brings us to the backbones and notochords. Notochords were the first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

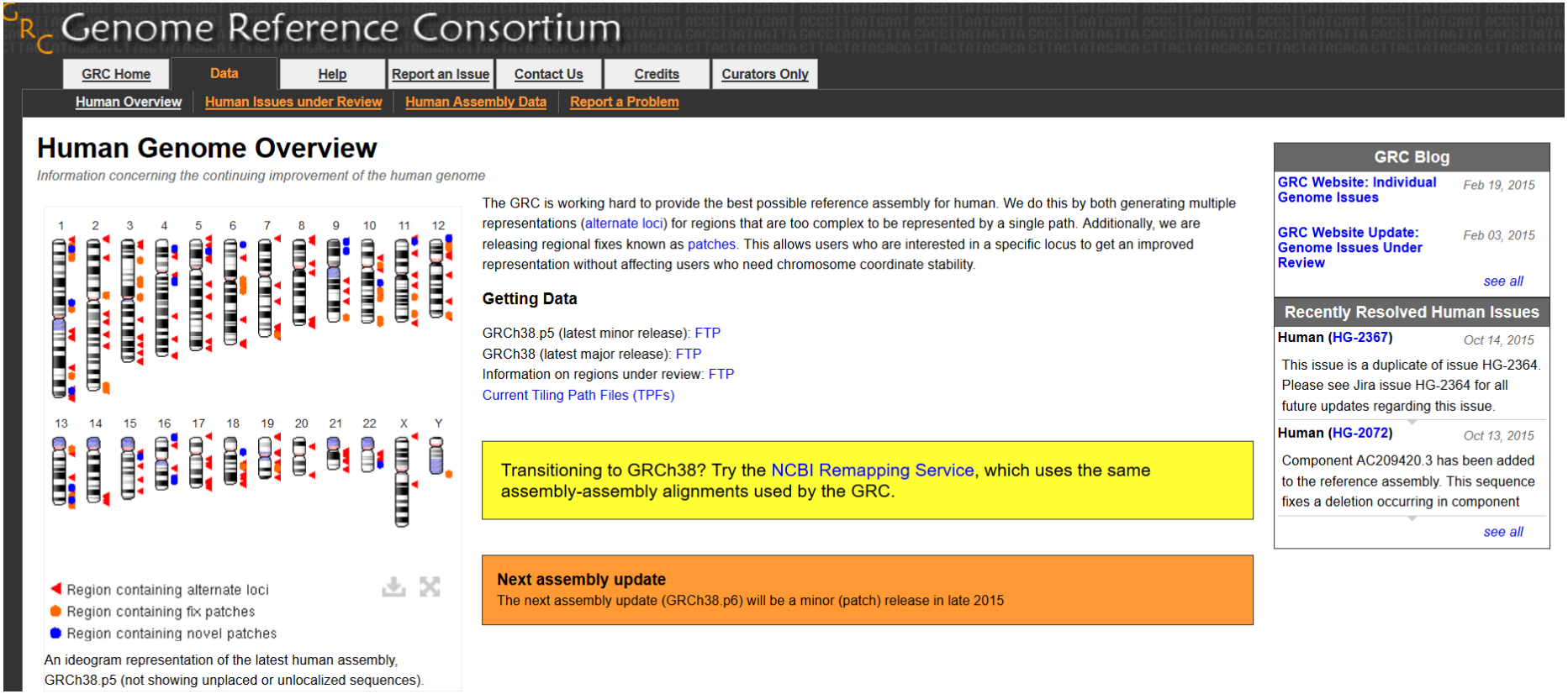
- Within the human body, it is estimated that there are 10x as many microbial cells as human cells.
- Our microbial partners carry out a number of metabolic reactions that are not encoded in the human genome and are necessary for human health (→ human genome = human genes + microbial genes).
- The majority 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?



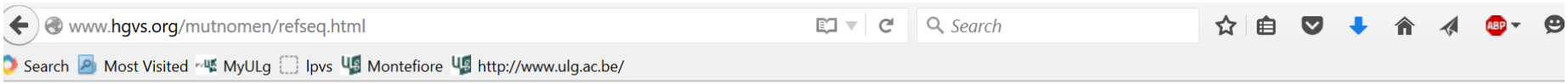
(http://www.hmpdacc.org/)

Which reference sequence?



(http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/)

Which reference sequence?



HGVS
HUMAN GENOME VARIATION SOCIETY

A reference sequence - discussions and FAQs

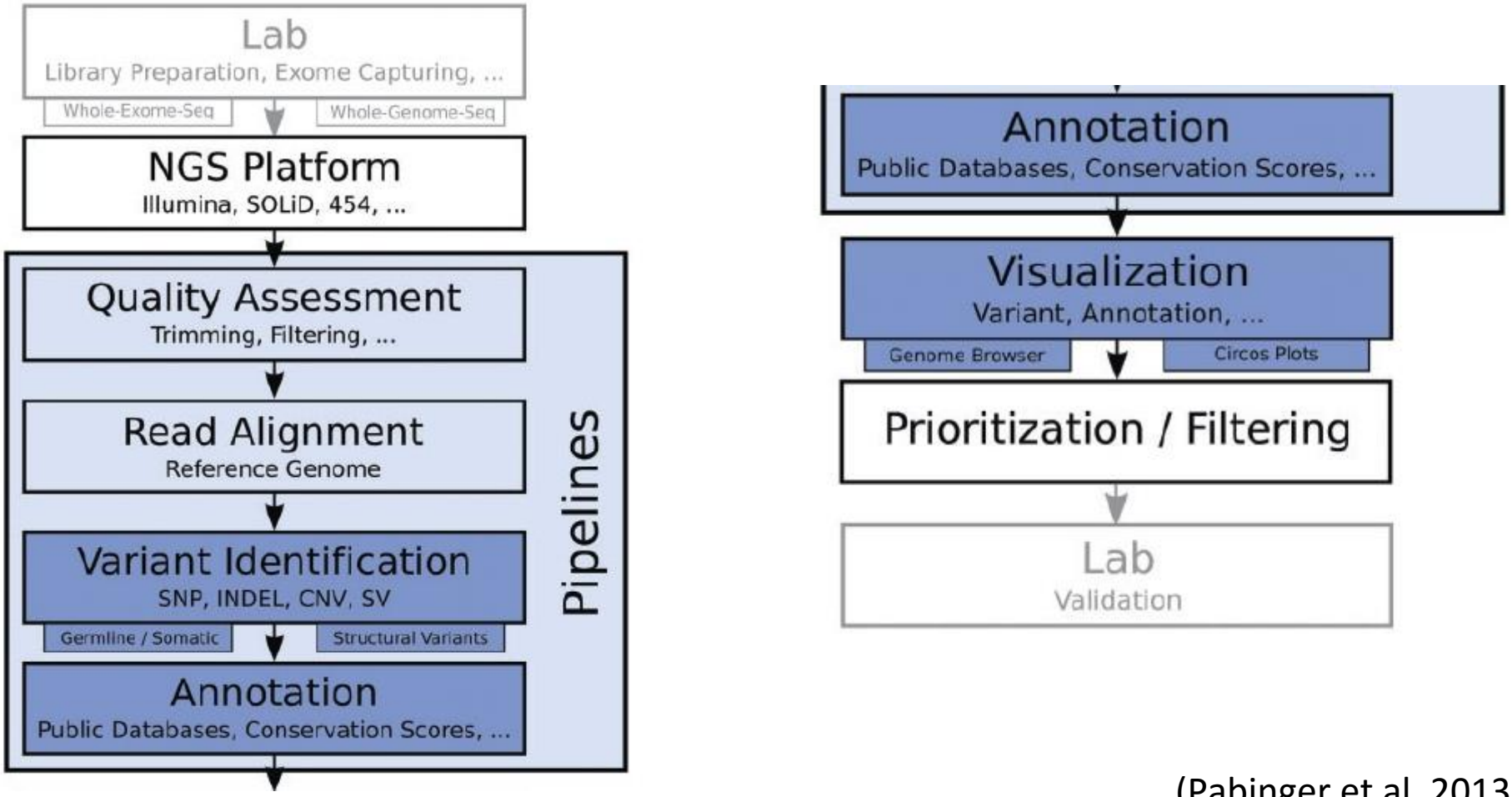
Last modified September 11, 2015

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)

Common workflow for whole-exome and whole genome sequencing



(Pabinger et al. 2013)

Basic workflow for whole-exome and whole-genome sequencing

- After library preparation, samples are sequenced on a certain platform.
- The next steps are quality assessment and read alignment against a reference genome,
- followed by variant identification.
- Detected mutations are then annotated to infer the biological relevance and
- results can be displayed using dedicated tools.
- The found mutations can further be prioritized and filtered, followed by validation of the generated results in the lab.

(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.Open Reading Frames

- In molecular genetics, an **open reading frame** (ORF) is the part of a reading frame that has the potential to code for a protein or peptide. An ORF is a continuous stretch of codons beginning with a start codon (usually AUG) and ending with a stop codon (usually TAA, TAG or TGA)

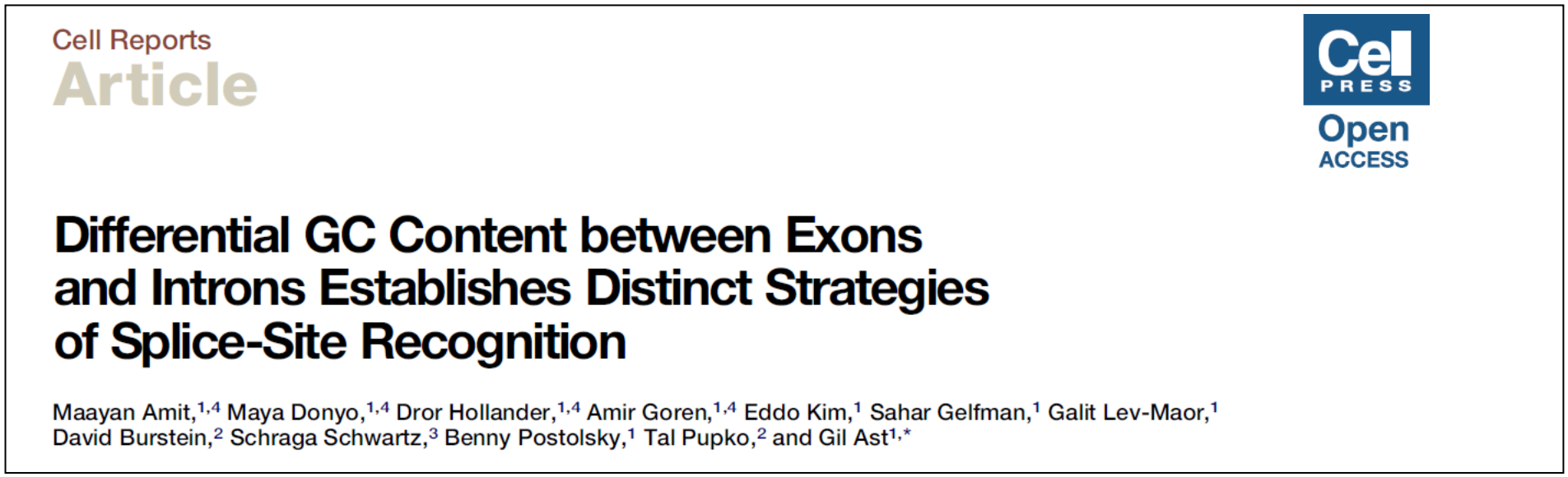
- Sample sequence showing three different possible reading frames. Start codons are highlighted in purple, and stop codons are highlighted in red.

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

(https://en.wikipedia.org/wiki/Open_reading_frame)

B. Counting letters or words

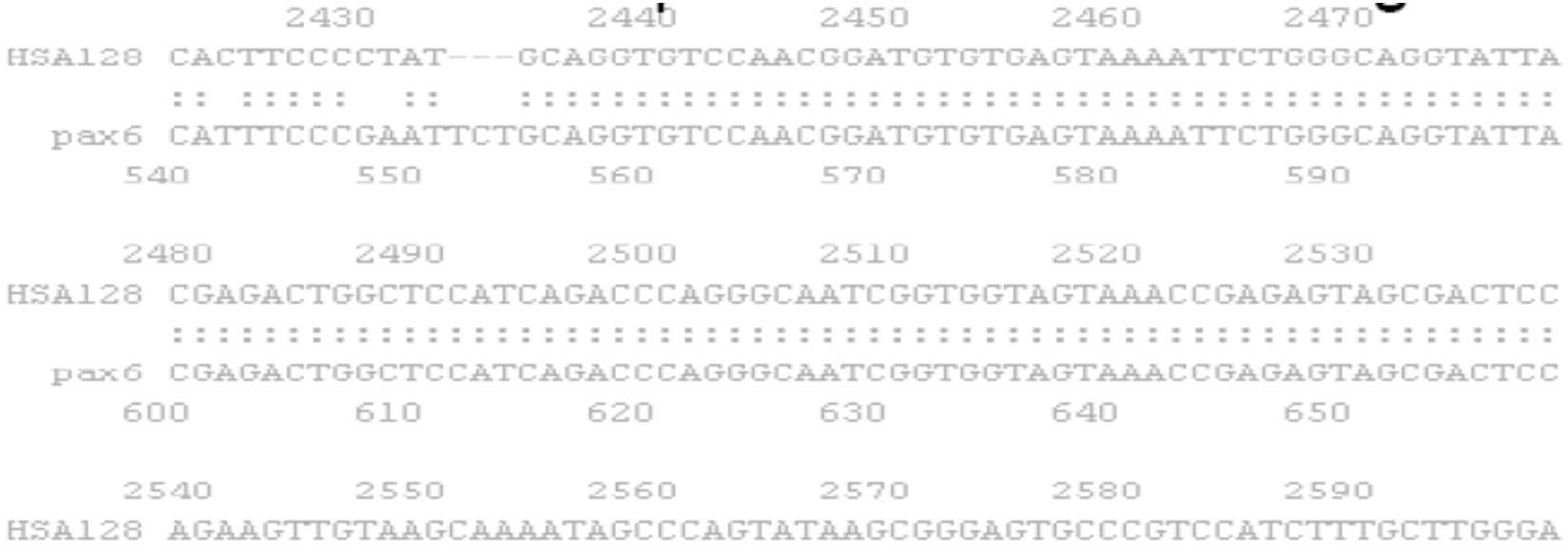
- One of the most fundamental properties of a genome sequence is its GC content, the fraction of the sequence that consists of Gs and Cs, ie. the %(G+C).
- The GC content can be calculated as the percentage of the bases in the genome that are Gs or Cs. That is, GC content = (number of Gs + number of 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)

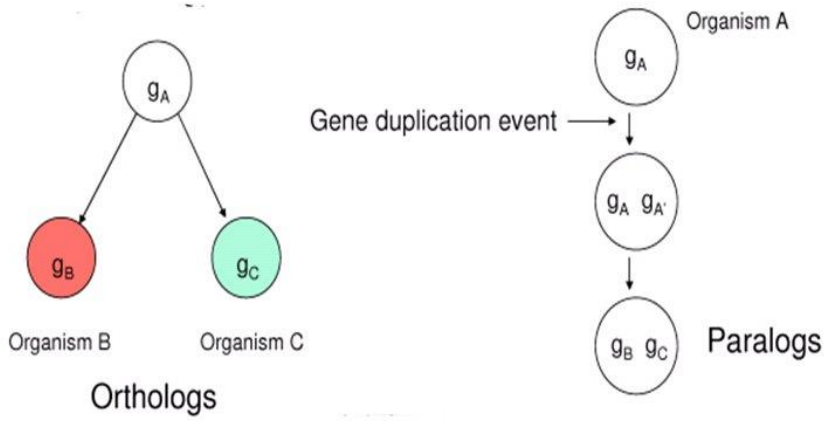
C. Comparing multiple sequences

- 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?

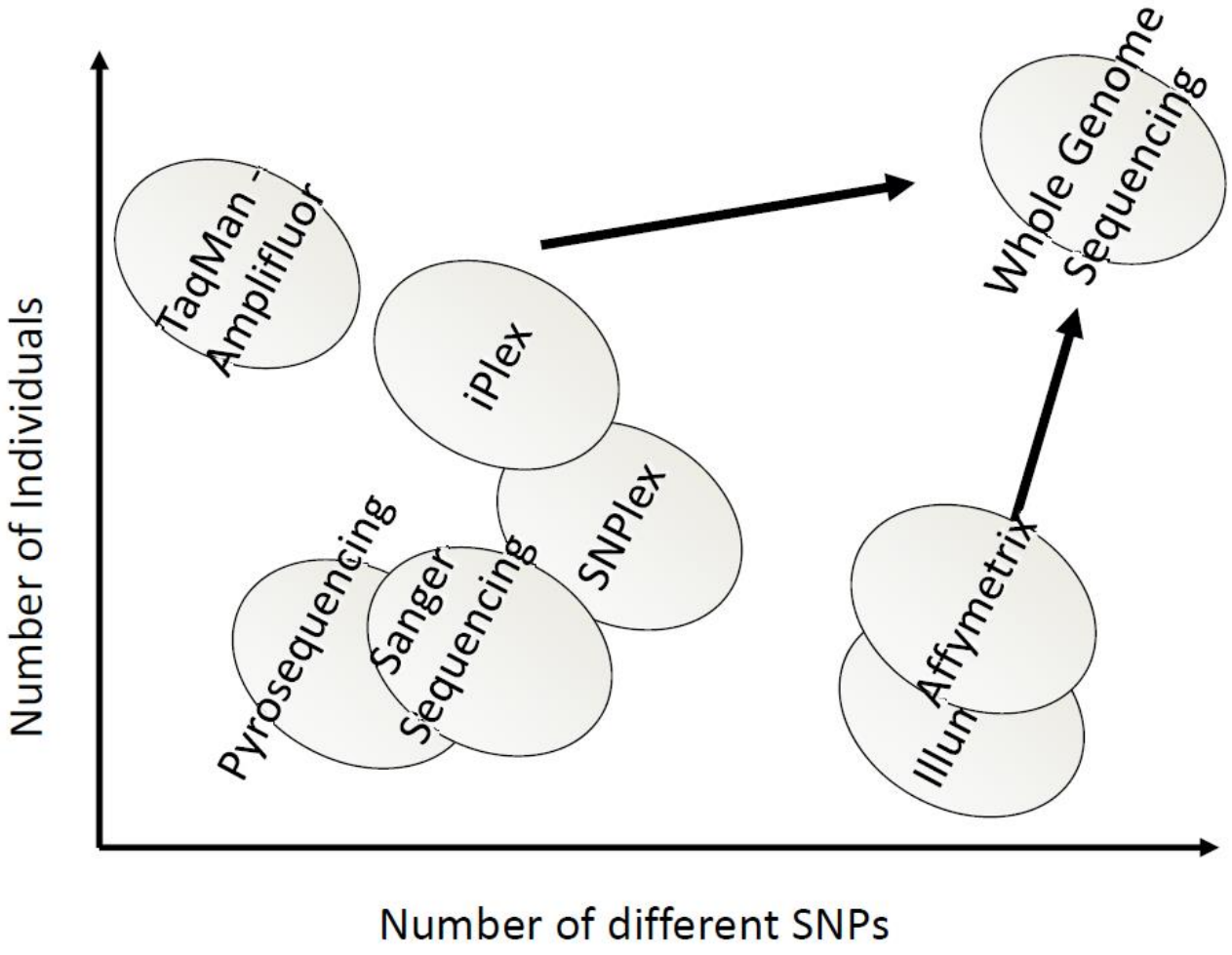


- Are sequences alike?
  - **Heterologs. {*Heterologs differ in both origin and activity.*}**
  - **Homologs. {*Homologs have common origins but may or may not have common activity.*}**
    - Genes that share an arbitrary threshold level of similarity determined by alignment of matching bases are termed **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



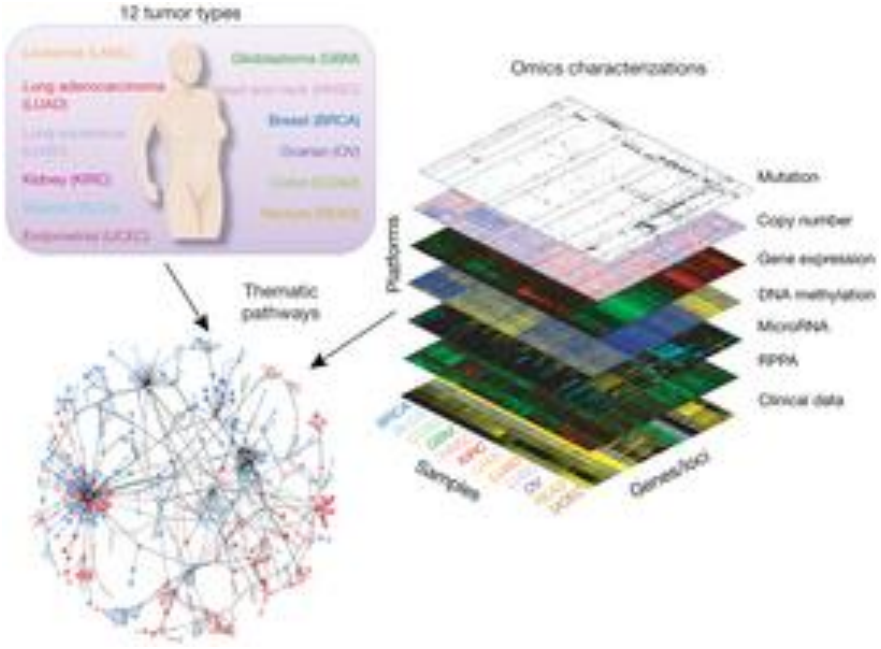
Genomic variation for complex diseases

- 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 indeed 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
  - 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

AATCGGATGCGCGTAGGATCGGTAGGGTAGGCTTTAAGATCATGCTATTTTCGAGA
TTCGATTCTAGCTAGGTTTAGCTTAGCTTAGTGCCAGAAATCGGATGCGCGTAGGAT
CGGTAGGGTAGGCTTTAAGATCATGCTATTTTCGAGATTCGATTCTAGCTAGGTTTT
TAGTGCCAGAAATCGTTAGTGCCAGAAATCGATT

- Can we determine the organism from which this sequence came?
- Is it likely to be a gene?
    - What is the possible expression level?
    - What is the possible protein product?
    - Can we get the protein product?
    - Can we figure out the key residue in the protein product?
- What sort of statistics to be used for describing this sequence?
- Does the description apply to bulk DNA in that organism?

The application determines the statistical analysis tool

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

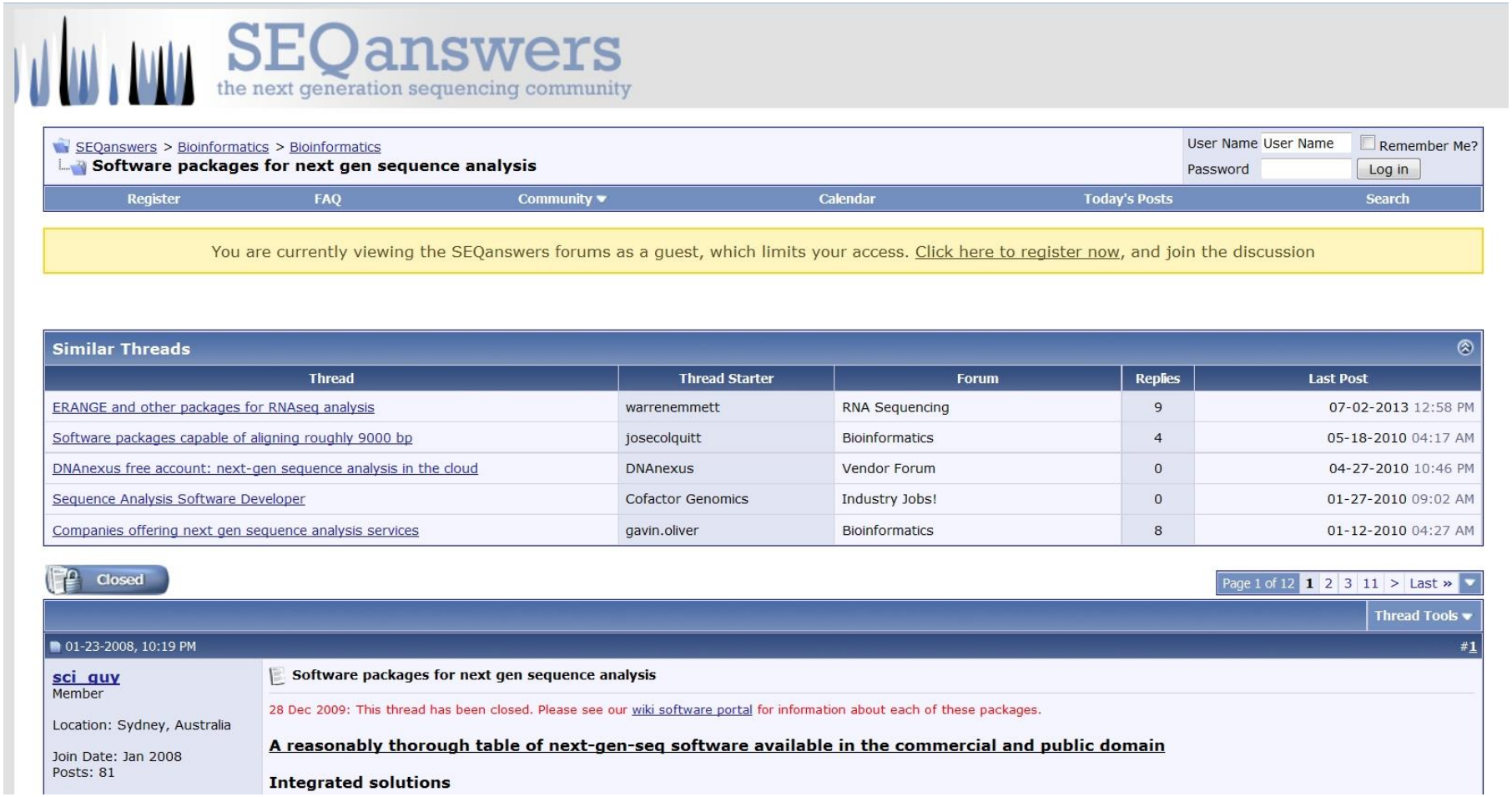
What to do next? … An example:

- 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?

Software packages for next gen sequence analysis



(http://seqanswers.com/forums/showthread.php?t=43)

Web-based programs (1)

Resource

Galaxy: A platform for interactive large-scale genome analysis

Belinda Giardine,[1] Cathy Riemer,[1] Ross C. Hardison,[1] Richard Burhans,[1] Laura Elnitski,[2] Prachi Shah,[1,2] Yi Zhang,[1] Daniel Blankenberg,[1] Istvan Albert,[1] James Taylor,[1] Webb Miller,[1] W. James Kent,[3] and Anton Nekrutenko[1,4]

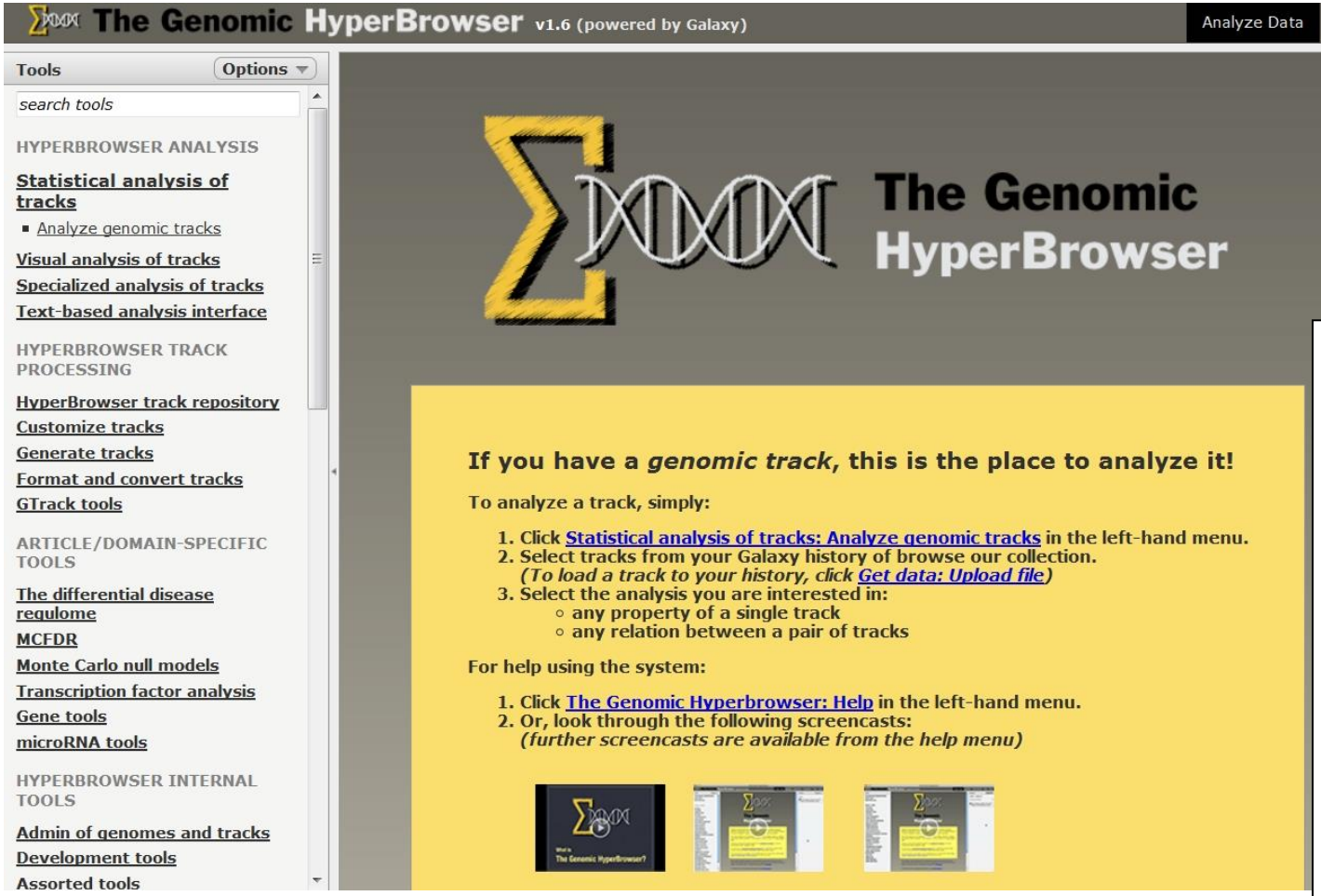
[1]Center for Comparative Genomics and Bioinformatics, Huck Institutes for Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA; [2]National Human Genome Research Institute, Bethesda, Maryland 20892, USA; [3]Department of Computer Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA

Accessing and analyzing the exponentially expanding genomic seque
biomedical researchers. Here we describe an interactive system, Galax
annotation databases with a simple Web portal to enable users to
independent queries, and visualize the results. The heart of Galaxy is
from each user; performs operations such as intersections, unions, and
tools. Galaxy can be accessed at http://g2.bx.psu.edu.

[Supplemental material is available online at www.genome.org.]

" **Galaxy** is a scientific workflow, data integration, and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience. It serves as a general bioinformatics workflow management system. "

Web-based programs (2)



(https://hyperbrowser.uio.no/hb/)

" **Genomic HyperBrowser** 's focus is on statistical inference on relations between genomic tracks. An example of analysis is to investigate the relationship between histone modifications and gene expression, using ChIP-based tracks of histone modifications versus tracks of genes marked with expression values from a microarray experiment. "

What will WE learn from sequence data?

- Recognizing **motifs, sites, signals, domains**
  - Sequence motifs are extremely convenient descriptors of conserved, functionally important short portions of proteins
  - A conserved motif can be represented by a **consensus sequence**
  - Often the above words (in bold) are used interchangeably to describe recurring elements of interest (patterns – **pattern recognition**)

- Why **(probability) models** for biomolecular motifs?
  - To characterize them
  - To help identify them
  - For incorporation into larger models (e.g., an entire gene)

Towards a theoretical understanding of false positives in DNA motif finding

Amin Zia and Alan M Moses*

Abstract

Background: Detection of false-positive motifs is one of the main causes of low performance in *de novo* DNA motif-finding methods. Despite the substantial algorithm development effort in this area, recent comprehensive benchmark studies revealed that the performance of DNA motif-finders leaves room for improvement in realistic scenarios.

Results: Using large-deviations theory, we derive a remarkably simple relationship that describes the dependence of false positives on dataset size for the one-occurrence per sequence motif-finding problem. As expected, we predict that false-positives can be reduced by decreasing the sequence length or by adding more sequences to the dataset. Interestingly, we find that the false-positive strength depends more strongly on the number of sequences in the dataset than it does on the sequence length, but that the dependence on the number of sequences diminishes, after which adding more sequences does not reduce the false-positive rate significantly. We compare our theoretical predictions by applying four popular motif-finding algorithms that solve the one-occurrence-per -sequence problem (MEME, the Gibbs Sampler, Weeder, and GIMSAN) to simulated data that contain no motifs. We find that the dependence of false positives detected by these softwares on the motif-finding parameters is similar to that predicted by our formula.

(Zia and Moses 2012)

What will WE learn from sequence data?

- **Comparative genomics**:
  the study of the genomic sequence of organisms that are related to humans – could ultimately help to identify targets for drug development

LEADING EDGE

SHAKING THE TREE OF LIFE

Comparative genomics – the study of the genomic sequence of organisms that are related to humans – could ultimately help to identify targets for drug development. **BY JACK MCCAIN,** Contributing Editor

Confucius said that the measure of man is man, but curious creatures may be useful yardsticks in determining the workings of the human body. Careful comparisons of the tree (Figures 1–4). Note that the tree's true shape is unknown in many instances and is subject to substantial ongoing revision.

The National Human Genome Research Institute (NHGRI), part of the National Institutes of Health, is Genome Research, Cambridge, Mass.; The Institute for Genomic Research [TIGR], Rockville, Md.; Washington University Medical Center, St. Louis). Organisms selected for sequencing include many with a long history of use as models

(McCain 2004)

What will WE learn from sequence data?

- Identifying disease relevant mutations and variants

ARTICLE SERIES: **Applications of next-generation sequencing**

Sequencing studies in human genetics: design and interpretation

David B. Goldstein, Andrew Allen, Jonathan Keebler, Elliott H. Margulies, Steven Petrou, Slavé Petrovski & Shamil Sunyaev

Nature Reviews Genetics **14**, 460–470 (2013) doi:10.1038/nrg3455
Published online 11 June 2013

Abstract

Next-generation sequencing is becoming the primary discovery tool in human genetics. There have been many clear successes in identifying genes that are responsible for Mendelian diseases, and sequencing approaches are now poised to identify the mutations that cause undiagnosed childhood genetic diseases and those that predispose individuals to more common complex diseases. There are, however, growing concerns that the complexity and magnitude of complete sequence data could lead to an explosion of weakly justified claims of association between genetic variants and disease. Here, we provide an overview of the basic workflow in next-generation sequencing studies and emphasize, where possible, measures and considerations that facilitate accurate inferences from human sequencing studies.

(Goldstein et al. 2013)

A survey of tools for variant analysis of next-generation genome sequencing data

Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski

Submitted: 20th August 2012; Received (in revised form): 4th December 2012

Abstract

Recent advances in genome sequencing technologies provide unprecedented opportunities to characterize individual genomic landscapes and identify mutations relevant for diagnosis and therapy. Specifically, whole-exome sequencing using next-generation sequencing (NGS) technologies is gaining popularity in the human genetics community due to the moderate costs, manageable data amounts and straightforward interpretation of analysis results. While whole-exome and, in the near future, whole-genome sequencing are becoming commodities, data analysis still poses significant challenges and led to the development of a plethora of tools supporting specific parts of the analysis workflow or providing a complete solution. Here, we surveyed 205 tools for whole-genome/whole-exome sequencing data analysis supporting five distinct analytical steps: quality assessment, alignment, variant identification, variant annotation and visualization. We report an overview of the functionality, features and specific requirements of the individual tools. We then selected 32 programs for variant identification, variant annotation and

(Pabinger et al 2013)

Public resources that help in mapping complex diseases in humans

http://www.ncbi.nlm.nih.gov/ - The National Center for Biotechnology Information (**NCBI**) advances science and health by providing access to biomedical and genomic information.
http://www.ensembl.org/index.html – The **Ensembl** project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.
http://hapmap.ncbi.nlm.nih.gov/ – **HapMap** – multi-country effort to identify and catalog genetic similarities and differences in human beings.
http://lynx.ci.uchicago.edu/ - **LYNX** – Gene Annotations, Enrichment Analysis and Genes Prioritization.
http://www.genemania.org/ – **GeneMANIA** - Indexing 1,421 association networks containing 266,984,699 interactions mapped to 155,238 genes from 7 organisms.

We will focus on human data and Bioconductor / R Environment

- R scripts illustrating relevant R packages for sequence pattern recognition and sequence comparison:

  - 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
  - Pairwise sequence alignment: http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter4.html
  - Multiple alignments and phylogenetic analysis: http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter5.html
  - Computational gene finding: http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter7.html
  - Comparative genomics: http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter9.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

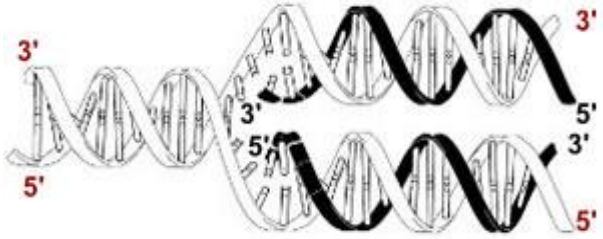
Introduction

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

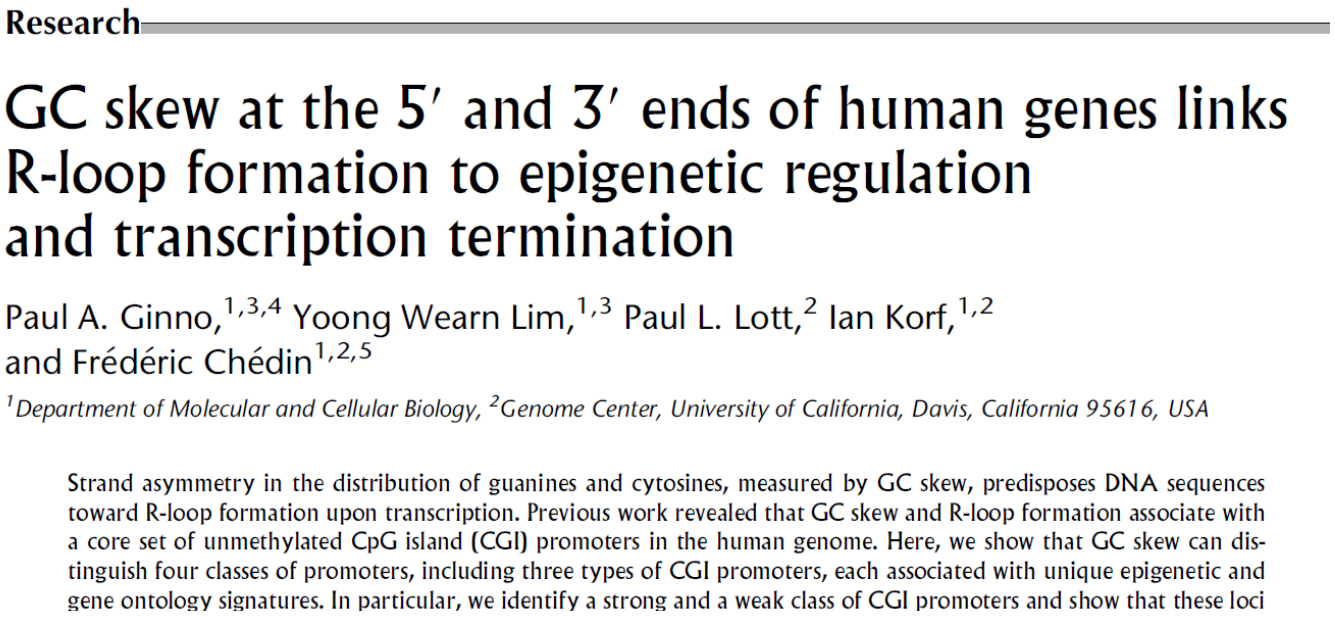
- There are constraints on base composition imposed by the genetic code
- The distribution of individual bases within a DNA molecule is not ordinarily uniform
  - There may be an excess of G over C on the leading strands



  - 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

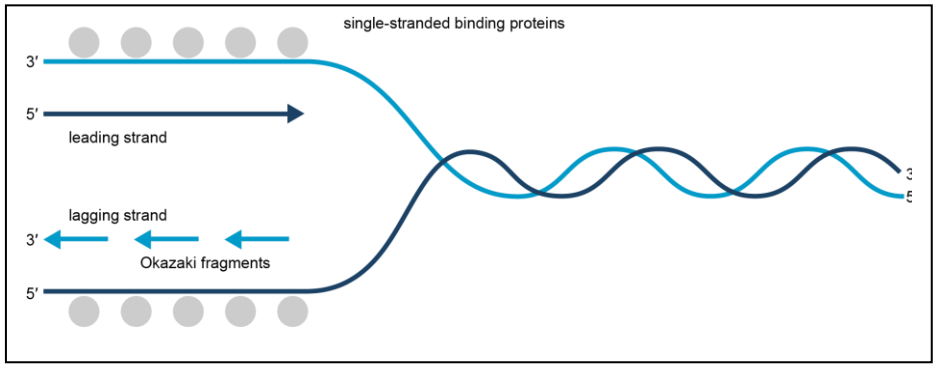
Paul A. Ginno,[1,3,4] Yoong Wearn Lim,[1,3] Paul L. Lott,[2] Ian Korf,[1,2] and Frédéric Chédin[1,2,5]

[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. But 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 $\chi = \{l_1, l_2, \ldots, l_n\}$
  - 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
    - $p_1, \ldots, p_J \geq O$ 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

- What is the probability distribution of the number of times a given pattern occurs in a random DNA sequence L_1, …, L_n?
  - 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

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:

$$E(X) = \sum_i x_i \, P(X = x_i)$$

  - $E(X_i) = 1 \times p_A + 0 \times (1 - p_A)$
  - If $Y = c\,X$, then E(Y) = c E(X)
  - $E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n)$
- Because X_i are assumed to be independent and identically distributed (iid):

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

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 :

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

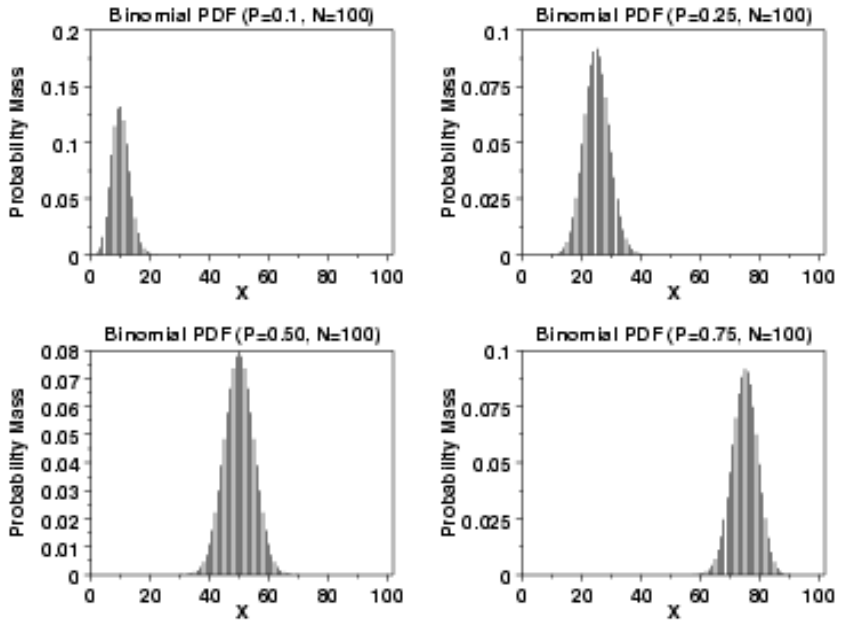
with the binomial coefficient $\binom{n}{j}$ determined by

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

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):

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

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

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

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

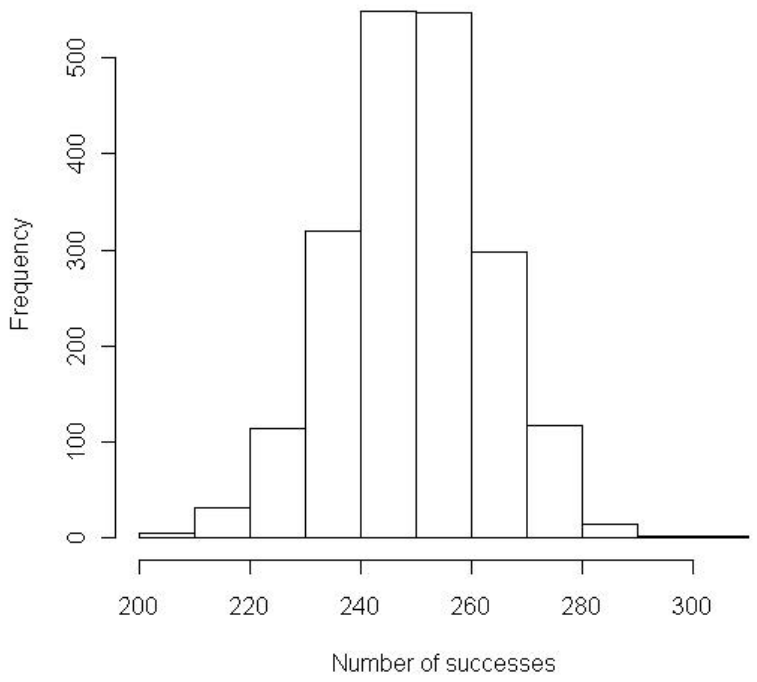
Simulating from probability distributions

- 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:

Number of observations = 2000

Number of trials = 1000

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

- 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,

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

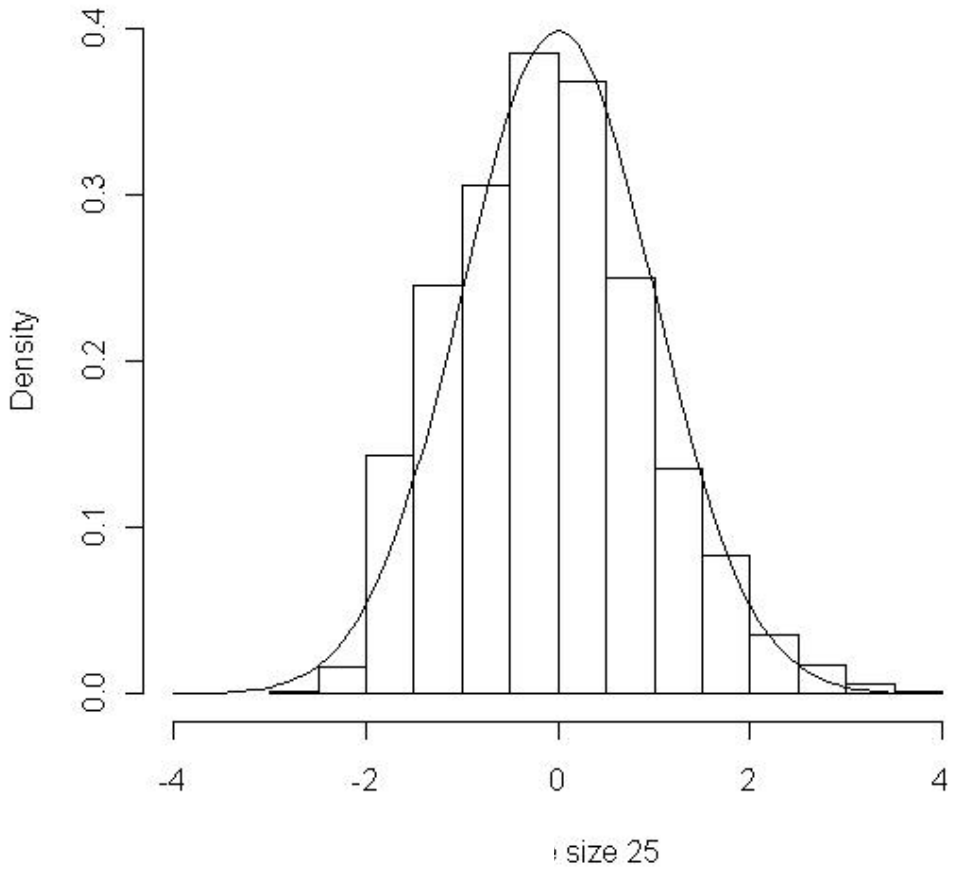
with $\phi(.)$ the standard normal distribution defined as

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

- 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.
    - CC dinucleotides contribute to the bending of DNA in chromatin (Bolshoy 1995)
- Also occurrences of CGs are of interest …

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.

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?
  - 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: qchisq(0.95,1) = 3.84
  - How can you verify that this recipe is correct?
    (see handbook Deonier et al. p 63)

Markov chains

- When moving from bacteria (such as **E**. **coli**, a common type of bacteria that can get into food, like beef and vegetables) to real genomes, a more complicated probabilistic model is required then the iid model before to capture the dinucleotide properties
- One approach is to use Markov chains.
- Markov chains are a direct generalization of independent trials, where the character at a position may depend on the characters of preceding positions, hence may be conditioned on preceding positions

Conditional probabilities

- If Ω refers to the set of all possible outcomes of a single experiment, A to a particular event, and A^c to the complement Ω-A of A, then

$$P(A) + P(A^c) = 1,$$

- The conditional probability of A given B is $P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$
- As a consequence: $P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$, also known as Bayes' Theorem
- For B_1, …, B_k forming a partition of Ω, this is the B_i are disjoint and the Bi are exhaustive (their union Is Ω), the law of total probability holds:

$$P(A) = \sum_{i=1}^{k} P(A \cap Bi) = \sum_{i=1}^{k} P(A|Bi)\,P(Bi)$$

The Markov property

- The property will be explained via studying a sequence of random variables X_t, t=0,1,… taking values in the state space {A,C,T,G}
- The sequence $\{X_t, t \geq 0\}$ is called a **first-order Markov chain** if only the previous neighbor influences the probability distribution of the character at any position, and hence satisfies the Markov property:

$$P(X_{t+1} = j \,|X_t = i, X_{t-1} = i_{t-1}, \dots , \quad X_0 = i_0) = P(X_{t+1} = j | X_t = i)$$

for $t \geq 0$ and for all $i, j, i_{t-1}, \dots, i_0$ in the state space

- We consider Markov chains that are **homogeneous**:
  $P(X_{t+1} = j | X_t = i) = p_{ij}$ (i.e. independent of the position t)

The Markov property

- The p_{ij} are the elements of a matrix P called the one-step transition matrix of the chain.
- Stepping from one position to the next is one issue, how to start is another issue
  - An initial probability distribution is needed as well
  - It is determined by a vector of probabilities corresponding to every possible initial state value i: $\pi_i^{(0)} = \pi_i \ = P(X_0 = i)$
- The probability distribution for the states at position 1 can be obtained as follows:

$$P(X_1 = j) = \sum_{i \in \chi} P(X_0 = i, X_1 = j)$$

$$= \sum_{i \in \chi} P(X_0 = i)\, P(X_1 = j | X_0 = i) \ = \ \sum_{i \in \chi} \pi_i p_{ij}$$

The Markov property

- To compute the probability distribution for the states at position 2, we first show that $P(X_2 = j|X_0 = i)$ is the ij-th element of PP=P^2

$$P(X_2 = j|X_0 = i) = \sum_{k \in \chi} P(X_2 = j, X_1 = k|\, X_0 = i)$$
$$= \sum_{k \in \chi} P(X_2 = j|X_1 = k, X_0 = i) P(X_1 = k|\, X_0 = i)$$
$$= \sum_{k \in \chi} P(X_2 = j|X_1 = k) P(X_1 = k|\, X_0 = i)$$
$$= \sum_{k \in \chi} p_{ik} p_{kj} = (PP)_{ij}$$

- Therefore

$$\pi_j^{(2)} = P(X_2 = j) = \sum_{i \in \chi} \pi_i \, P_{ij}^2$$

The Markov property

- In a similar way it can be shown that

$$\pi_j^{(t)} = P(X_t = j) = \sum_{i \in \chi} \pi_i \, P_{ij}^t$$

- In principle, it can happen that the distribution $\pi^{(t)}$ is independent of t. This event is then referred to as a **stationary distribution** of the chain.
  - It occurs when $\sum_{i \in \chi} \pi_i p_{ij} = \pi_{j,}$ for all j, or stated differently when $\pi = \pi P$

Creating our own Markov chain simulation in practice

- Assume the observed dinucleotide relative frequencies (each row specifies a base and each column specifies the following base):

	A	C	G	T
A	0.146	0.052	0.058	0.089
C	0.063	0.029	0.010	0.056
G	0.050	0.030	0.028	0.051
T	0.087	0.047	0.063	0.140

- How to compute the individual base frequencies?
- How to compute the transition matrix?
- How to propose initial state parameters to build a Markov chain?

How to compute the individual base frequencies?

	A	C	G	T
A	0.146	0.052	0.058	0.089
C	0.063	0.029	0.010	0.056
G	0.050	0.030	0.028	0.051
T	0.087	0.047	0.063	0.140

$P_A = 0.345$

How to compute the transition matrix?

$$P(X_1 = A | X_0 = A) = 0.146/0.345$$

giving the transition probability matrix

P =

	A	C	G	T
A	0.423	0.151	0.168	0.258
C	0.399	0.184	0.063	0.354
G	0.314	0.189	0.176	0.321
T	0.258	0.138	0.187	0.415

How to propose initial state parameters to build a Markov chain?

Take the initial base frequencies

R code to generate a string having characteristics of Mycoplasma DNA

Mycoplasma is a genus of bacteria that lack a cell wall around their cell membrane. Without a cell wall, they are unaffected by many common antibiotics. In man, *mycoplasmas* may be found in the airway and urinary tract

- x is a vector with elements 1, 2, 3, 4 representing A, C, G, T
- pi is a vector with initial state probabilities
- P is a 4x4 transition matrix
- n is the length of the simulated sequence [we are going to generate a sequence with 50,000 positions]

```
markov1 <- function(x,pi,P,n){
mg <- rep(0,n)
mg[1] <- sample(x,1,replace=TRUE,pi)
for (k in 1:(n-1)){
  mg[k+1] <- sample(x,1,replace=TRUE,P[mg[k],])
  }
return(mg)
}
```

- In concreto, we initialize the parameters as follows:

```
x<-c(1:4)
pi <- c(0.342,0.158, 0.158, 0.342)
P <- matrix(scan(),ncol=4,nrow=4,byrow=T)
0.423 0.151 0.168 0.258
0.399 0.184 0.063 0.354
0.314 0.189 0.176 0.321
0.258 0.138 0.187 0.415
```

- Creating the sequence; executing the R function:

```
tmp <- markov1(x,pi,P,50000)
```

- Checking the simulation output:

```
A<- length(tmp[tmp[]==1])
C<- length(tmp[tmp[]==2])
G<- length(tmp[tmp[]==3])
T<- length(tmp[tmp[]==4])
(C+G)/(A+C+G+T)   # fraction of G+C
```

  and comparing it to $p_C + p_G$ derived from the transition matrix

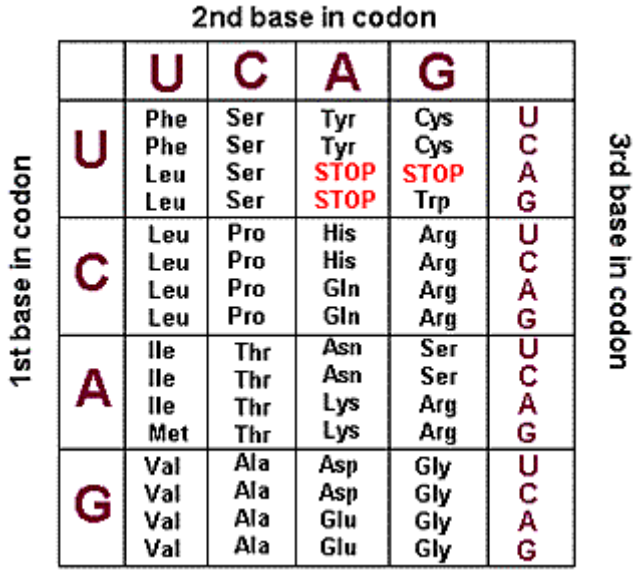
- Does tmp contain an appropriate fraction of GC dinucleotides?

```
count <-0
for (i in 1:49999){
 if (tmp[i]==2 && tmp[i+1]==3)
 count <- count+1
 }
count/49999 # abundance of CG dinucleotide as estimated by the model
```

  and compare (0.0096) with pCG in the transition matrix (= 0.010)

3.b Studying words of length 3

Amino acids



1st base	U	C	A	G	3rd base
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

- 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 – biological problem

- 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 (i.e., 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)

Towards the codon adaptation index

- Consider a sequence of amino acids X = x_1, x_2, ... , x_L representing protein X, with x_k representing the amino acid residue corresponding to codon *k* in the gene.
- Question: How does the actual codon usage compare with a model that states that the codons employed are the most probable codons for highly expressed genes?
- For the codon corresponding to a particular amino acid at position *k* in protein *X*, let p_k be the probability that *this* particular codon is used to code for the amino acid
- Let q_k correspond to the probability for *the most frequently used* codon of the corresponding amino acid in highly expressed genes.

The codon adaptation index (Sharp and Li 1987)

- The CAI is defined as

$$\mathrm{CAI} = \left[\prod_{k=1}^{L} p_k / q_k \right]^{1/L}$$

- An alternative way of writing this is

$$\log(\mathrm{CAI}) = \frac{1}{L} \sum_{k=1}^{L} \log(p_k / q_k).$$

- So the formula for CAI is the geometric mean of the ratios of the probabilities for the codons *actually* used to the probabilities of the codons *most frequently* used in highly expressed genes.

The codon adaptation index

- An example:

  Consider the amino acid sequence from the amino terminal end of the *himA* gene of E. coli (which codes for one of the two subunits of the protein IHF: length L = 99).

```
M   A   L   T   K   A   E   M   S   E   Y   L   F   ...
ATG GCG CTT ACA AAA GCT GAA ATG TCA GAA TAT CTG TTT ...
```

The codon adaptation index

- Top lines: amino acid sequence and corresponding codons.
- Upper table: probabilities for codons in lower table.
- The probability of the most frequently used codon in highly expressed genes is underlined (Fig 2.2 – Deonier et al 2005).

M	A	L	T	K	A	E	M	S	E	Y	L	F	...
ATG	GCG	CTT	ACA	AAA	GCT	GAA	ATG	TCA	GAA	TAT	CTG	TTT	...
1.000	0.469	0.018	0.451	0.798	0.469	0.794	1.000	0.428	0.794	0.193	0.018	0.228	
	0.057	0.018	0.468	0.202	0.057	0.206		0.319	0.206	0.807	0.018	0.772	
	0.275	0.038	0.035		0.275			0.033			0.038		
	0.199	0.033	0.046		0.199			0.007			0.033		
		0.007						0.037			0.007		
		0.888						0.176			0.888		
ATG	GCT	TTA	ACT	AAA	GCT	GAA	ATG	TCT	GAA	TAT	TTA	TTT	
	GCC	TTG	ACC	AAG	GCC	GAG		TCC	GAG	TAC	TTG	TTC	
	GCA	CTT	ACA		GCA			TCA			CTT		
	GCG	CTC	ACG		GCG			TCG			CTC		
		CTA						AGT			CTA		
		CTG						AGC			CTG		

The codon adaptation index

- The CAI for this fragment of coding sequence is given by

$$\mathrm{CAI} = \left[\frac{1.000}{1.000} \times \frac{0.199}{0.469} \times \frac{0.038}{0.888} \times \frac{0.035}{0.468} \cdots \right]^{1/99}.$$

M	A	L	T	K	A	E	M	S	E	Y	L	F	...
ATG	GCG	CTT	ACA	AAA	GCT	GAA	ATG	TCA	GAA	TAT	CTG	TTT	...
1.000	0.469	0.018	0.451	0.798	0.469	0.794	1.000	0.428	0.794	0.193	0.018	0.228	
	0.057	0.018	0.468	0.202	0.057	0.206		0.319	0.206	0.807	0.018	0.772	
	0.275	0.038	0.035		0.275			0.033			0.038		
	0.199	0.033	0.046		0.199			0.007			0.033		
		0.007						0.037			0.007		
		0.888						0.176			0.888		
ATG	GCT	TTA	ACT	AAA	GCT	GAA	ATG	TCT	GAA	TAT	TTA	TTT	
	GCC	TTG	ACC	AAG	GCC	GAG		TCC	GAG	TAC	TTG	TTC	
	GCA	CTT	ACA		GCA			TCA			CTT		
	GCG	CTC	ACG		GCG			TCG			CTC		
		CTA						AGT			CTA		
		CTG						AGC			CTG		

The codon adaptation index

- Interpretation:
  - If every codon in a gene corresponded to the most frequently used codon in highly expressed genes, then the CAI would be 1.0.
  - The CAI for a gene sequence in genomic DNA provides a first approximation of its expression level:
    - if the CAI is relatively large, then we would predict that the expression level is also large.
  - Hence, the CAI can be shown to be correlated with mRNA levels

- In E. coli a sample of 500 protein-coding genes displayed CAI values in the range from 0.2 to 0.85 (Whittam, 1996)

3.c Restriction sites

The biological problem

If we were to digest the DNA with a restriction endonuclease such as EcoR1, then

1) approximately how many fragments would be obtained, and

2) what would be their size distribution?

The number of restriction sites

- Restriction endonuclease recognition sequences have length *t* (4, 5, 6 or 8 typically), where *t* is much smaller than *n.*
- Our model assumes that cleavage can occur between any two successive positions on the DNA:

    - This is wrong in detail because, depending upon where cleavage occurs within the bases of the recognition sequence (which may differ from enzyme to enzyme), there are positions near the ends of the DNA that are excluded from cleavage.
    - However, since *t* is much smaller than *n,* the ends of the molecule do not affect the result too much

The number of restriction sites

- We again use X_i to represent the outcome of a trial occurring at position *i*, but this time X_i does not represent the identity of a base (one of four possible outcomes) but rather whether position i is or is not the beginning of a restriction site.
- In particular,

$$X_i = \begin{cases} 1, \text{ if base } i \text{ is the start of a restriction site,} \\ 0, \text{ if not.} \end{cases}$$

- We denote by *p* the probability that any position i is the beginning of a restriction site:

$$X_i = \begin{cases} 1, \text{ with probability } p, \\ 0, \text{ with probability } 1-p. \end{cases}$$

The number of restriction sites

- Unlike with tossing a fair coin, for the case of restriction sites on DNA, *p* depends upon
  - the base composition of the DNA and
  - the identity of the restriction endonuclease.
- For example:
  - Suppose that the restriction endonuclease is *Eco*RI, with recognition sequence 5'-GAATTC-3'.

  - Suppose furthermore that the DNA has equal proportions of A, C, G, and T.

The number of restriction sites

- The probability that any position is the beginning of a site is the probability that this first position is G, the next one is A, the next one is A, the next one is T, the next one is T, and the last one is C.
- Since, by the iid model, the identity of a letter at any position is independent of the identity of letters at any other position, we see from the multiplication rule that

$$p = \mathbb{P}(\mathtt{GAATTC}) = \mathbb{P}(\mathtt{G})\mathbb{P}(\mathtt{A})\mathbb{P}(\mathtt{A})\mathbb{P}(\mathtt{T})\mathbb{P}(\mathtt{T})\mathbb{P}(\mathtt{C}) = (0.25)^6 \sim 0.00024.$$

- Notice that *p* is small, a fact that becomes important later.

The number of restriction sites

- The appearance of restriction sites along the molecule is represented by the string $X_1, X_2, \ldots, X_n$,
- The number of restriction sites is $N = X_1 + X_2 + \ldots + X_m$, where $m = n - 5$.
    - The sum has m terms in it because a restriction site of length 6 cannot begin in the last five positions of the sequence, as there aren't enough bases to fit it in.
- For simplicity of exposition we take m = n in what follows.
- What really interests us is the number of "successes" (restriction sites) in n trials.

The number of restriction sites

- If $X_1, X_2, \ldots, X_n$ were independent of one another, then the probability distribution of *N* would be a binomial distribution with parameters n and *p;*
  - The expected number of sites would therefore be *np*
  - The variance would be *np(*1 - *p).*
- We remark that despite the X_i are not in fact independent of one another (because of overlaps in the patterns corresponding to X_i and X_{i+1}, for example), the binomial approximation usually works well.
- Computing probabilities of events can be cumbersome when using the probability distribution

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

Poisson approximation to the binomial distribution

- In what follows, we assume that *n* is large and *p* is small, and we set λ= *np.*
- We know that for *j* = 0, 1, ... , *n,*

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

- Writing

$$\mathbb{P}(N = j) = \frac{n(n-1)(n-2)\cdots(n-j+1)}{j!(1-p)^j} p^j (1-p)^n.$$

and given that the number of restriction sites (j) is small compared to the length of the molecule (n), such that

$$n(n-1)(n-2)\ldots(n-j+1) \approx n^j, (1-p)^j \approx 1,$$

Poisson approximation to the binomial distribution

$$\mathbb{P}(N = j) \approx \frac{(np)^j}{j!}(1-p)^n = \frac{\lambda^j}{j!}\left(1 - \frac{\lambda}{n}\right)^n.$$

in which $\lambda = np$.

- From calculus, for any x,

$$\lim_{n\to\infty}\left(1 - \frac{x}{n}\right)^n = \mathrm{e}^{-x}.$$

- Since *n* is large (often more than 10^4), we replace $(1 - \frac{\lambda}{n})^n$ by $e^{-\lambda}$ to get our final approximation in the form

$$\mathbb{P}(N = j) \approx \frac{\lambda^j}{j!}\,\mathrm{e}^{-\lambda}, \quad j = 0, 1, 2, \ldots$$

- This is the formula for the Poisson distribution with parameter $\lambda = np$

Poisson approximation to the binomial distribution

- Example:
  - To show how this approximation can be used, we estimate the probability that there are no more than two *Eco*RI sites in a DNA molecule of length 10,000, assuming equal base frequencies
  - Earlier we obtained p=0.00024 for this setting:

$$p = \mathbb{P}(\mathtt{GAATTC}) = \mathbb{P}(\mathtt{G})\mathbb{P}(\mathtt{A})\mathbb{P}(\mathtt{A})\mathbb{P}(\mathtt{T})\mathbb{P}(\mathtt{T})\mathbb{P}(\mathtt{C}) = (0.25)^6 \sim 0.00024.$$

  - The problem is to compute $P(N \leq 2)$
    - Therefore $\lambda = np = 2.4$
    - Using the Poisson distribution: $P(N \leq 2) \approx 0.570$
    - Interpretation: More than half the time, molecules of length 10,000 and uniform base frequencies will be cut by *Eco*RI two times or less

Distribution of restriction fragment lengths

- With this generalization, we assume that restriction sites now occur according to a Poisson process with rate λ per bp. Then the probability of *k* sites in an interval of length *l* bp is

$$\mathbb{P}(N = k) = \frac{\mathrm{e}^{-\lambda l}(\lambda l)^k}{k!}, \ k = 0, 1, 2, \ldots \ .$$

- We can also calculate the probability that a restriction fragment length X is larger than x. If there is a site at y, then the length of that fragment is greater than x if there are no events in the interval (y, y + x):

$$\mathbb{P}(X > x) = \mathbb{P}(\text{no events in } (y, y + x)) = \mathrm{e}^{-\lambda x}, \ x > 0.$$

Distribution of restriction fragment lengths

- The previous has some important consequences:

$$\mathbb{P}(X \leq x) = \int_0^x f(y)\mathrm{d}y = 1 - \mathrm{e}^{-\lambda x},$$

so that the density function for X is given by

$$f(x) = \lambda \mathrm{e}^{-\lambda x},\ x > 0.$$

- The distance between restriction sites therefore follows an exponential distribution with parameter λ
  - The mean distance between restriction sites is $1/\lambda$

Simulating restriction fragment lengths (sizes)

- If we *simulated* a sequence using the iid model, we could compute the fragment sizes in this simulated sequence and visualize the result
- R code simulating a DNA sequence having 48500 positions and uniform base probabilities:

```
x<-c(1:4)
propn <- c(0.25,0.25,0.25,0.25)
seq2 <- sample(x,48500,replace=TRUE,prob=propn)
seq2[1:15]
length(seq2[])
```

Simulating restriction fragment lengths

- R code identifying the restriction sites in a sequence string, with bases coded numerically:

```
rsite <- function(inseq, seq){
  # inseq: vector containing input DNA sequence,
  # A=1, C=2, G=3, T=4
  # seq: vector for the restriction site, length m
  # Make/initialize vector to hold site positions found in inseq
  xxx <- rep(0,length(inseq))
  m <-length(seq)
  # To record whether position of inseq matches seq
  truth <- rep(0,m)
```

```
# Check each position to see if a site starts there
  for (i in 1:(length(inseq) - (length(seq) -1))){
    for (j in 1:m){
      if (inseq[i+j-1]==seq[j]){
      truth[j] <- 1 # Record match to jth position
      }
    }
    if (sum(truth[]) ==m){ # Check whether all positions match
    xxx[i] <- i          # Record site if all positions match
    }
    truth <- rep(0,m)    # Reinitialize for next loop cycle
  }
  # Write vector of restriction sites positions stored in xxx
  L <- xxx[xxx>0]
  return(L)
}
```

Simulating restriction fragment lengths

- The restriction sites we look for are for A*lu*I, AGCT.
- R code envoking the appropriate function:

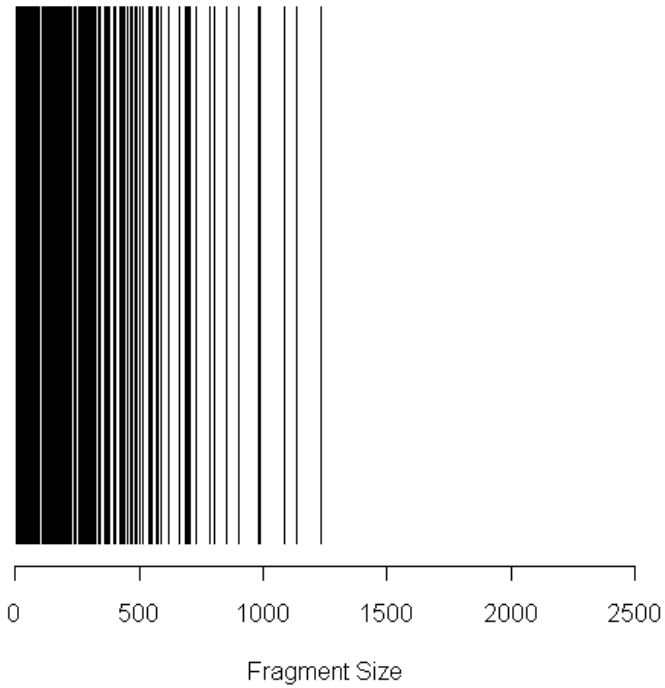
```
alu1 <- c(1,3,2,4)
alu.map <- rsite(seq2,alu1)
length(alu.map)
alu.map[1:10]
```

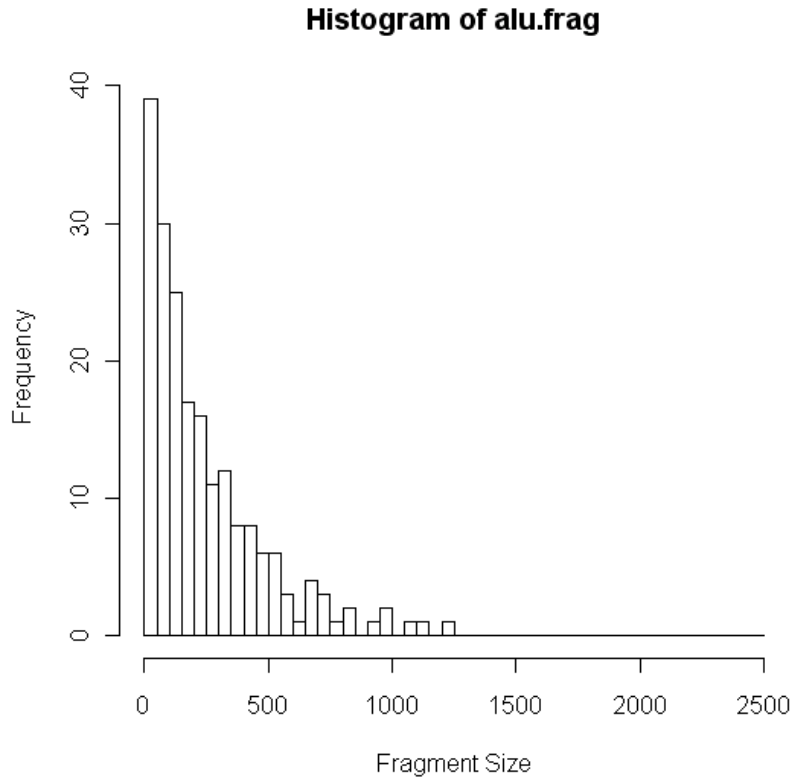
Simulating restriction fragment lengths

- The fragment lengths can be obtained by subtracting positions of successive sites
- R code doing it for you:

```
flengthr <- function(rmap,N){
  # rmap is a vector of restriction sites for a linear molecule
  # N is the length of the molecule
  frags <- rep(0,length(rmap))
  # Vector for substraction results: elements initialized to 0
  rmap <-c(rmap,N)
  # Adds length of molecule for calculation of end piece
  for(i in 1:(length(rmap)-1)){
  frags[i] <- rmap[i+1]-rmap[i]
  }
  frags <- c(rmap[1],frags) # First term is left end piece
  return(frags)
}
```

Simulating restriction fragment lengths





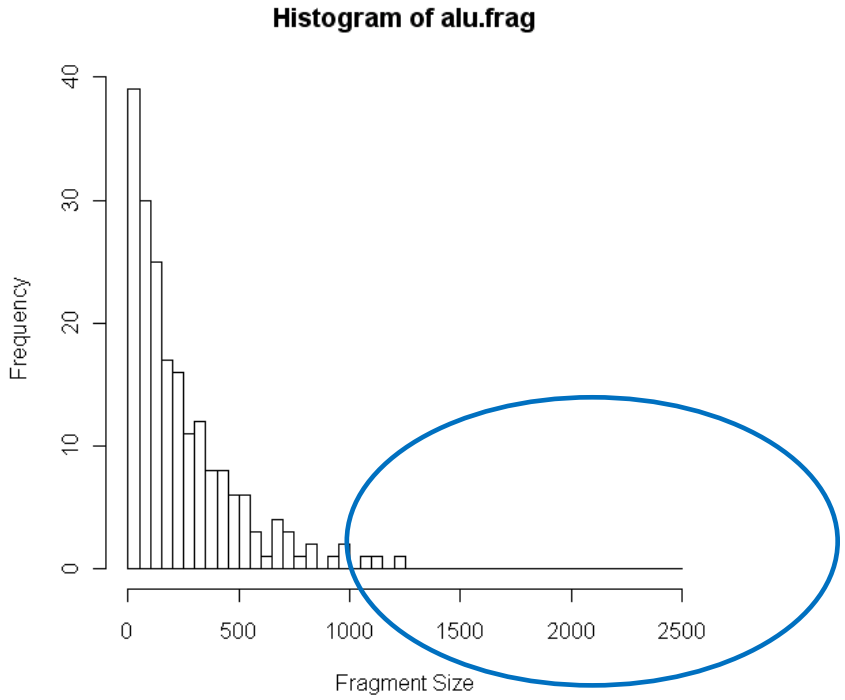
Simulating restriction fragment lengths

- R code corresponding to the figures:

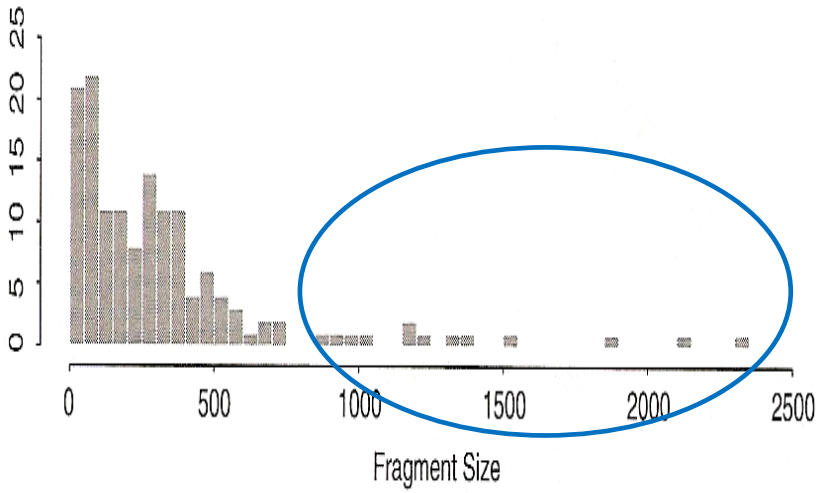
```
plot(c(0,2500),c(3,1),xlab="Fragment Size",ylab="",type="n",axes=F)
axis(1,c(0,500,1000,1500,2000,2500))
for (i in 1:length(alu.frag)){
lines(c(alu.frag[i],alu.frag[i]),c(1,3))
}
hist(alu.frag,breaks=seq(0,2500,50), freq = TRUE,xlab="Fragment Size")
```

Simulating restriction fragment lengths

- Is our theoretical model to simulate restriction fragment lengths valid?



Histogram based on theoretical model



Histogram of fragment sizes (bp) produced by AluI digestion of bacteriophage lambda DNA

Simulating restriction fragment lengths

- To determine whether the actual distribution differs significantly from the mathematical model (exponential distribution), we could break up the length axis into a series of "bins" and calculate the expected number of fragments in each bin by using the model-based (theoretical) density.
- We could then compare the observed with expected number of fragments (using the same bin boundaries) via for instance a χ^2 – test.

4 Comparing sequences

- Sequence comparisons are important for a number of reasons.
  - First, they can be used to establish evolutionary relationships among organisms using methods analogous to those employed for anatomical characters.
  - Second, comparison may allow identification of functionally conserved sequences (e.g., DNA sequences controlling gene expression).
  - Finally, comparisons between humans and other species may identify corresponding genes in model organisms, which can be genetically manipulated to develop models for human diseases.

(see practical session)