

# BIOINFORMATICS

**Kristel Van Steen, PhD<sup>2</sup>**

**Montefiore Institute - Systems and Modeling**

**GIGA - Bioinformatics**

**ULg**

**[kristel.vansteen@ulg.ac.be](mailto:kristel.vansteen@ulg.ac.be)**

## CHAPTER 5: DNA SEQUENCE ANALYSIS

### 1 Introduction

**1.a Historical notes**

**1.b Work flow**

**1.c Applications**

### 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 Restriction sites

## 5 R code

# 1 Introduction

## 1.a 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

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

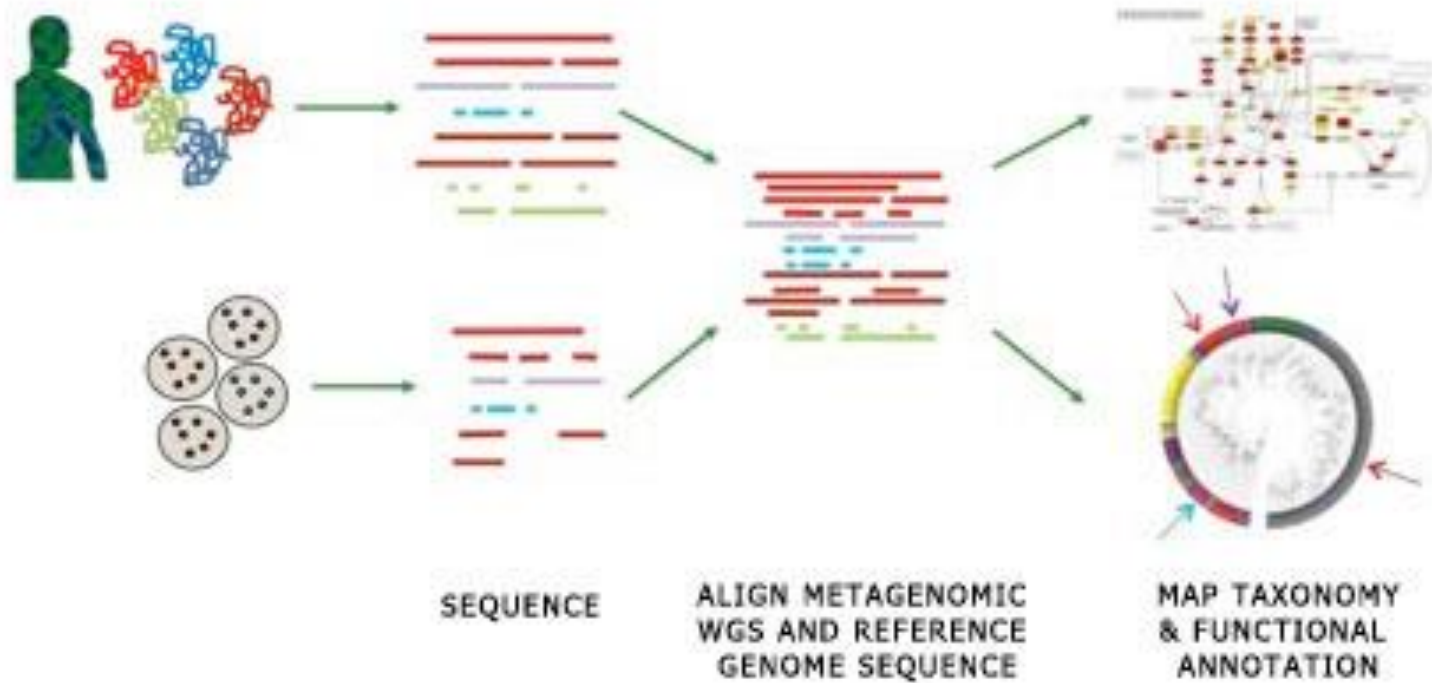
- In human genetics, metagenome sequencing is becoming increasingly important, which lead to the **Human Microbiome Project** (<http://www.hmpdacc.org/>)
  - Metagenome sequencing is defined as an approach for the study of microbial populations in a sample representing a community by analysing the nucleotide 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 to provide a comprehensive pool of genome sequences to aid in the analysis of human metagenomic data.

## 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?

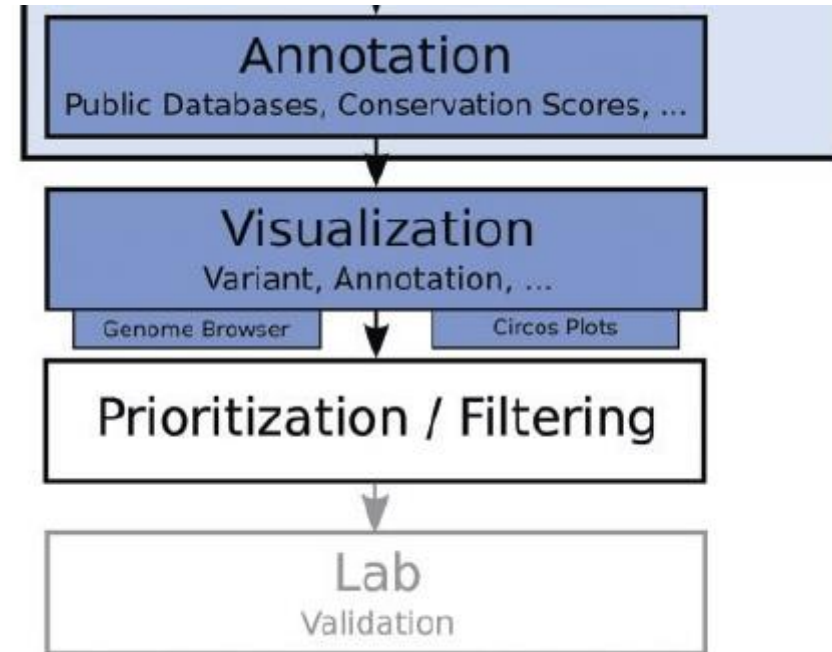
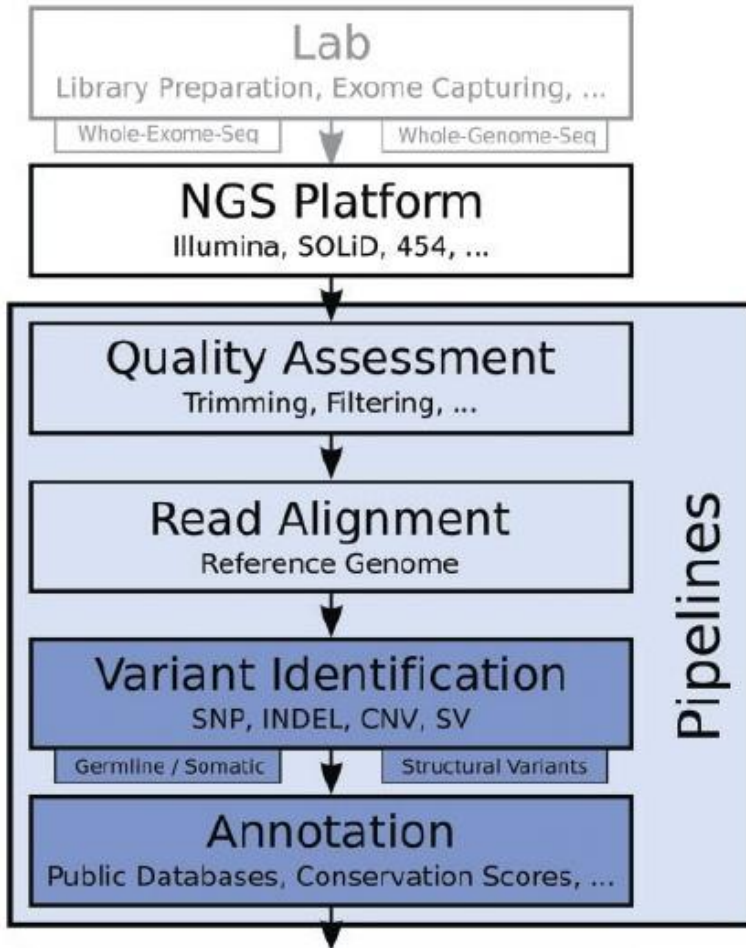


(<http://www.hmpdacc.org/>)

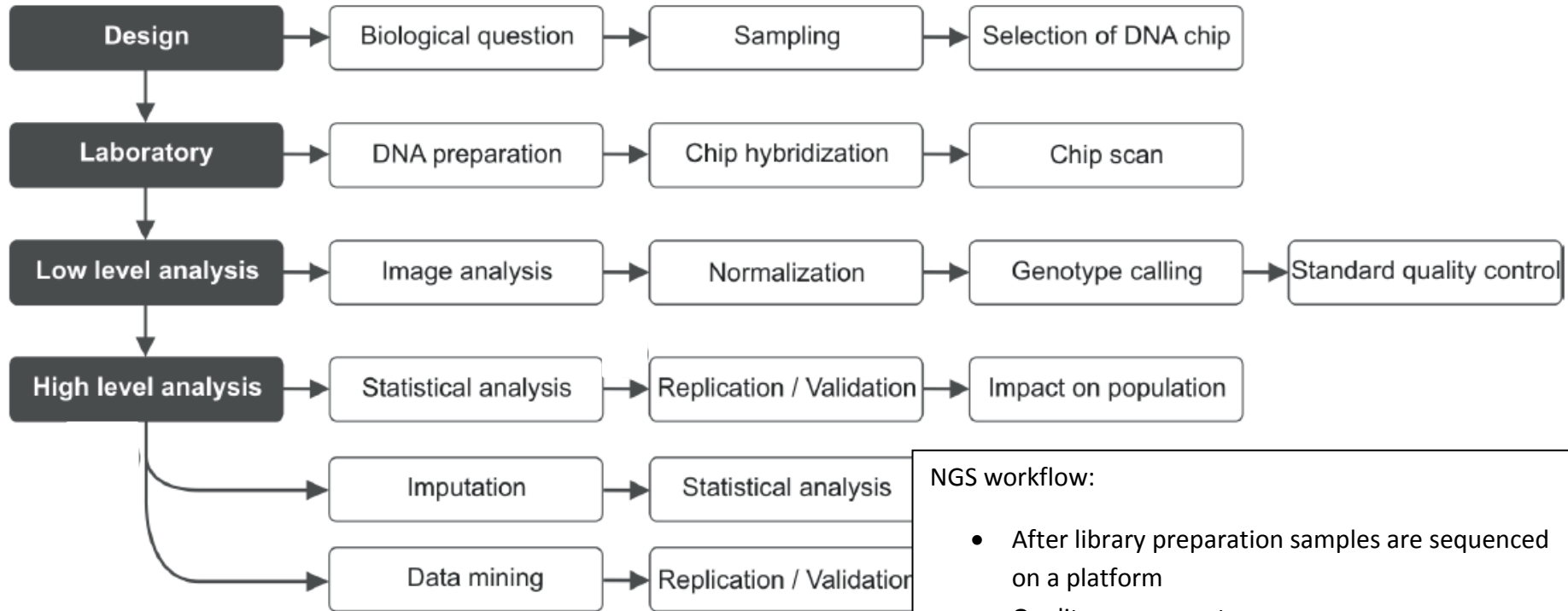
## 1.b Work flow

- 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 (Pabinger et al. 2013).
- As sequencing techniques improve and develop overtime, the amount of data produced increases exponentially and therefore the implementation of efficient platforms to analyze and visualize such large amounts of data in fast and efficient ways has become a necessity (Pavlopoulos et al 2013).
- 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 (Pabinger et al. 2013).

# Basic workflow for whole-exome and whole genome sequencing



# Recall: Detailed flow of a genome-wide association study



**NGS workflow:**

- After library preparation samples are sequenced on a platform
- Quality assessment
- Read alignment against reference genome
- Identify variants
- Detected mutations are then annotated to infer biological relevance and further prioritized or filtered

(Pabinger et al 2013)

(Ziegler 2009)

## 1.c Applications

### 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
- In order to have an idea about potentially interesting questions, we need to survey potential application fields

## Three common scenarios for human geneticists using NGS data

- 1) Identification of causative genes in Mendelian disorders (germline mutations)
- 2) Identification of candidate genes in complex diseases for further functional studies
- 3) Identification of constitutional mutations as well as driver and passenger genes in cancer (somatic mutations)

(Pabinger et al 2013)

A **germline mutation** is one that 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.

## Other scenarios ?


<b>Bioinformatics</b>	<b>Computational biology</b>
Research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data	Development and application of data-analytical, theoretical methods, mathematical modeling and computational simulation to the study of biological, behavioral, and social systems.

(BISTIC Definition Committee, NIH, 2000)

## Other scenarios ?

- 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](http://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?





SEQanswers > Bioinformatics > Bioinformatics

**Software packages for next gen sequence analysis**

User Name  User Name  Remember Me?  
 Password

Register    FAQ    Community ▼    Calendar    Today's Posts    Search

You are currently viewing the SEQanswers forums as a guest, which limits your access. [Click here to register now](#), and join the discussion

**Similar Threads**

Thread	Thread Starter	Forum	Replies	Last Post
<a href="#">ERANGE and other packages for RNAseq analysis</a>	warrenemmett	RNA Sequencing	9	07-02-2013 12:58 PM
<a href="#">Software packages capable of aligning roughly 9000 bp</a>	josecolquitt	Bioinformatics	4	05-18-2010 04:17 AM
<a href="#">DNAnexus free account: next-gen sequence analysis in the cloud</a>	DNAnexus	Vendor Forum	0	04-27-2010 10:46 PM
<a href="#">Sequence Analysis Software Developer</a>	Cofactor Genomics	Industry Jobs!	0	01-27-2010 09:02 AM
<a href="#">Companies offering next gen sequence analysis services</a>	gavin.oliver	Bioinformatics	8	01-12-2010 04:27 AM

Closed Page 1 of 12 [1](#) [2](#) [3](#) [11](#) > Last »

01-23-2008, 10:19 PM #1

**sci\_guy**  
Member  
Location: Sydney, Australia  
Join Date: Jan 2008  
Posts: 81

**Software packages for next gen sequence analysis**

28 Dec 2009: This thread has been closed. Please see our [wiki software portal](#) for information about each of these packages.

**A reasonably thorough table of next-gen-seq software available in the commercial and public domain**

**Integrated solutions**

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

## Web-based programs (1)

### Resource

---

# Galaxy: A platform for interactive large-scale genome analysis

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

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

Accessing and analyzing the exponentially expanding genomic sequence data is a challenge for biomedical researchers. Here we describe an interactive system, Galaxy, that integrates genome annotation databases with a simple Web portal to enable users to perform arbitrary independent queries, and visualize the results. The heart of Galaxy is a workflow engine that runs from each user; performs operations such as intersections, unions, and joins, and integrates existing tools. Galaxy can be accessed at <http://g2.bx.psu.edu>.

[Supplemental material is available online at [www.genome.org](http://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)

**The Genomic HyperBrowser v1.6** (powered by Galaxy)

Analyze Data | Shared Data | Help | User | Using 0 bytes

**Tools** Options

search tools

**HYPERBROWSER ANALYSIS**

**Statistical analysis of tracks**

- Analyze genomic tracks

**Visual analysis of tracks**

**Specialized analysis of tracks**

**Text-based analysis interface**

**HYPERBROWSER TRACK PROCESSING**

**HyperBrowser track repository**

**Customize tracks**

**Generate tracks**

**Format and convert tracks**

**GTrack tools**

**ARTICLE/DOMAIN-SPECIFIC TOOLS**

**The differential disease regulome**

**MCFDR**

**Monte Carlo null models**

**Transcription factor analysis**

**Gene tools**

**microRNA tools**

**HYPERBROWSER INTERNAL TOOLS**

**Admin of genomes and tracks**

**Development tools**

Assorted tools

**History** Options

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

**The Genomic HyperBrowser**

**If you have a genomic track, this is the place to analyze it!**

To analyze a track, simply:

1. Click [Statistical analysis of tracks: Analyze genomic tracks](#) in the left-hand menu.
2. Select tracks from your Galaxy history or browse our collection. (To load a track to your history, click [Get data: Upload file](#))
3. Select the analysis you are interested in:
  - any property of a single track
  - any relation between a pair of tracks

For help using the system:

1. Click [The Genomic Hyperbrowser: Help](#) in the left-hand menu.
2. Or, look through the following screencasts: (further screencasts are available from the help menu)

(<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. “

## Four common scenarios for bioinformaticians using NGS data

- 1) Sequence assembly (see practical sessions /homework assignment)
- 2) Annotation (see practical sessions /homework assignment)
- 3) Comparative genomics (see next class: BLAST-ing, FASTA-ing)
- 4) Pattern recognition

### DNA Sequence Pattern Recognition Methods in GRAIL

Edward C. Uberbacher, Ying Xu, Manesh Shah, Sherri Matis  
Xiaojun Guan and Richard J. Mural<sup>†</sup>  
Informatics Group  
Computer Sciences and Mathematics and <sup>†</sup>Biology Divisions  
Oak Ridge National Laboratory  
Oak Ridge, TN 37831-6364

#### Introduction

The goal of the GRAIL project<sup>1,2,3,4,5</sup> has been to create a comprehensive analysis environment where a host of questions about genes and genome structure can be answered as quickly and accurately as possible. Constructing this system has entailed solving a number of significant technical challenges including: (a) making coding recognition in sequence more sensitive and accurate, (b) compensating for isochore base compositional effects in coding prediction, (c) developing methods to determine which parts of each strand of a long genomic DNA are the coding strand, (d) improving the accuracy of splice site prediction and recognizing non-consensus sites, and (e) recognizing variable regulatory structures such as polymerase II promoters. An additional challenge has been to construct algorithms which compensate for the deleterious effects of insertion or deletion (indel) errors in the coding region recognition process. This paper addresses progress on these technical issues and the current state of sequence feature recognition methods.

## Comparative genomics

- **Motivations:** the study of the genomic sequence of organisms that are related to humans could ultimately help to identify targets for drug development; conserved regions must be important to life ....

### 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

**C**onfucius 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)

## Pattern recognition

**Motivation:** Human DNA sequences carry relevant biophysical information in the form of a one-dimensional chain of 4 nucleotide bases: Adenine, Guanine, Cytosine, Thymine. One way of retrieving this information is by looking at patterns / distributions of (collections or sequences of) these letters.

- These chains are broadly divided into 3 parts:
  - Regions called **genes**
  - These are connected by **intergenic regions** (sometimes called the flanks)
  - Inside the genes these sequences divide into **exons and introns**. The exons code for proteins, the introns come in between exons. The ***coding DNA sequence (CDS)*** is obtained by taking the genes and splicing the non-coding intron regions out of it.

## Pattern recognition

- In general, given the DNA sequence

```
AATCGGATGCGCGTAGGATCGGTAGGGTAGGCTTTAAGATCATGCTATTTTCGAGATTTCGATTCT
AGCTAGGTTTAGCTTAGCTTAGTGCCAGAAATCGGATGCGCGTAGGATCGGTAGGGTAGGCTTTA
AGATCATGCTATTTTCGAGATTTCGATTCTAGCTAGGTTTTTAGTGCCAGAAATCGTTAGTGCCAGA
AATCGATT
```

several questions can be asked:

- Is it a gene? (What is the possible expression level? What is the possible protein product? How can we obtain the protein product?)
- Can we determine the organism from which this sequence came?
- What sort of statistics to be used for *describing* this sequence?
- Do parameters describing the sequence differ from those describing bulk DNA in that organism?
- Can we spot motifs?

## DNA sequence motifs

atgaccgggatactgatAAAAAAGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg  
accctatTTTTTgagcagatttagtgacctggaaaaaaatttgagtacaaaactTTTccgaataAAAAAAGGGGGGga  
tgagtatccctgggatgacttAAAAAAGGGGGGtgctctcccgattTTTgaatatgtaggatcattcgccaggggccga  
gctgagaattggatgAAAAAAGGGGGGtccacgcaatcgcgaaaccaacgcggacccaaaggcaagaccgataaaggaga  
tccTTTTgCGGtaatgtgcccgggaggctggttacgtagggaagccctaacggacttaataAAAAAAGGGGGGcttatag  
gtcaatcatgttcttgtgaatggatttAAAAAAGGGGGGgaccgcttggcgcacccaaattcagtgtgggCGagCGcaa  
cggTTTTgCCcttgttagaggccccgTAAAAAAGGGGGGcaattatgagagagctaattctatcgCGtgcgtgttcat  
aacttgagttAAAAAAGGGGGGctggggcacatacaagaggagtcttcttatcagttaatgctgtatgacactatgta  
ttggccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAGGGGGGaccgaaaggaag  
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAGGGGGGga



## DNA sequence motifs

atgaccgggatactgatAAAAAAGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgcccg  
accctatTTTTTgagcagatttagtgacctggaaaaaaatttgagtacaaaactTTTccgaataAAAAAAGGGGGGga  
tgagtatccctgggatgacttAAAAAAGGGGGGtgctctcccgattTTTgaatatgtaggatcattcgccaggggtccga  
gctgagaattggatgAAAAAAGGGGGGtccacgcaatcgcgaaaccaacgcggacccaaaggcaagaccgataaaggaga  
tccTTTTgCGGtaatgtgcccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAGGGGGGcttatag  
gtcaatcatgttcttgtgaatggatttAAAAAAGGGGGGgaccgcttggcgcacccaaattcagtgtgggCGagcgcaa  
cggTTTTgCCcttgtttagaggccccgtAAAAAAGGGGGGcaattatgagagagctaattctatcgCGtgCGtgttcat  
aacttgagttAAAAAAGGGGGGctggggcacatacaagaggagtcttcttatcagttaatgctgtatgacactatgta  
ttggccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataAAAAAAGGGGGGaccgaaaggaag  
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAGGGGGGga

## DNA sequence motifs

atgaccgggatactgat**AAAAAAAGGGGGG**ggcgtacacattagataaacgtatgaagtacgtagactcggcgccgccg  
 acccctatTTTTTgagcagatttagtgacctggaaaaaaatttgagtacaaaactTTTccgaata**AAAAAAAGGGGGGA**  
 tgagtatccctgggatgactt**AAAAAAAGGGGGG**TgctctcccgatTTTTgaatatgtaggatcattcgccagggtccga  
 gctgagaattggatg**AAAAAAAGGGGGG**tccacgcaatcgcaaccaacgcggacccaaaggcaagaccgataaaggaga  
 tccTTTTgCGgtaatgtgCCgggaggctggttacgtagggaagccctaacggacttaat**AAAAAAAGGGGGG**cttatag  
 gtcaatcatgttcttTgtgaatggattt**AAAAAAAGGGGGG**gaccgcttggcgcacccaaattcagtgTgggcgagcgcaa  
 cgTTTTggcccttTtagaggccccgt**AAAAAAAGGGGGG**caattatgagagagctaattctatcgCGtgCGTgttcat  
 aacttgagtt**AAAAAAAGGGGGG**ctggggcacatacaagaggagtcttcttatcagttaatgctgtatgacactatgta  
 ttggccattggctaaaagcccaacttgacaaatggaagatagaatccttgcat**AAAAAAAGGGGGG**accgaaaggggaag  
 ctggtgagcaacgacagattcttacgtgcattagctcgcttccgggatctaatagcacgaagctt**AAAAAAAGGGGGGA**

atgaccgggatactgat**AgAAgAAAGGttGGG**ggcgtacacattagataaacgtatgaagtacgtagactcggcgccgccg  
 acccctatTTTTTgagcagatttagtgacctggaaaaaaatttgagtacaaaactTTTccgaata**CAAtAAACGGCGGGA**  
 tgagtatccctgggatgactt**AAAAtAAtGGaGtGG**TgctctcccgatTTTTgaatatgtaggatcattcgccagggtccga  
 gctgagaattggatg**CAAAAAAGGGattG**tccacgcaatcgcaaccaacgcggacccaaaggcaagaccgataaaggaga  
 tccTTTTgCGgtaatgtgCCgggaggctggttacgtagggaagccctaacggacttaat**AtAAAtAAAGGaaGGG**cttatag  
 gtcaatcatgttcttTgtgaatggattt**AACAAtAAGGGctGGG**gaccgcttggcgcacccaaattcagtgTgggcgagcgcaa  
 cgTTTTggcccttTtagaggccccgt**AtAAACAAGGaGGGc**caattatgagagagctaattctatcgCGtgCGTgttcat  
 aacttgagtt**AAAAAtAGGGaGcc**ctggggcacatacaagaggagtcttcttatcagttaatgctgtatgacactatgta  
 ttggccattggctaaaagcccaacttgacaaatggaagatagaatccttgcat**ActAAAAAGGaGCGG**accgaaaggggaag  
 ctggtgagcaacgacagattcttacgtgcattagctcgcttccgggatctaatagcacgaagctt**ActAAAAAGGaGCGGA**

# Why finding (15,4) motif is difficult

atgaccgggatactgat**AgAAgAAAGGttGGG**ggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg  
 acccctatTTTTTgagcagatttagtgacctggaaaaaaatttgagtacaaaactTTTccgaata**CAAtAAACGGCGGG**a  
 tgagtatccctgggatgactt**AAAAtAAtGGaGtGG**tgctctcccgattTTTgaatatgtaggatcattcggcagggtccga  
 gctgagaattggatg**CAAAAAAGGGattG**tccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga  
 tccTTTTgCGgtaatgtgccgggaggctggttacgtagggaagccctaacggacttaat**AtAAtAAAGGaaGGG**cttatag  
 gtcaatcatgTtcttgatggatt**AACAAtAAGGGctGG**gaccgcttggcgcacccaaattcagtgTgggCGagcgcaa  
 cggtTTTggcccttgtagaggccccgt**AtAAACAAGGaGGG**ccaattatgagagagctaattctatcgcgTgcgtgttcat  
 aacttgagtt**AAAAAtAGGGaGcc**ctggggcacatacaagaggagtcttcttatcagttaatgctgtatgacactatgta  
 ttggccattggctaaaagcccaacttgacaaatggaagatagaatccttgc**ActAAAAGGaGcGG**accgaaaggggaag  
 ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagctt**ActAAAAGGaGcGGa**

**AgAAgAAAGGttGGG**  
 ..|..|||.||..|||  
**CAAtAAACGGCGGG**

## DNA sequence motifs

- Definition: ***Sequence motifs*** are short, recurring patterns in DNA that are presumed to have a biological function.
  - Often they indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF).
  - Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing (e.g., splicing, editing) and transcription termination.
- Potential research question:  
Find a motif in a sample of
  - 20 “random” sequences (e.g., of length 600 nucleotides) where
  - each sequence contains an implanted pattern of length 15, and
  - each pattern appears with 4 mismatches as (15,4)-motif.

## 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

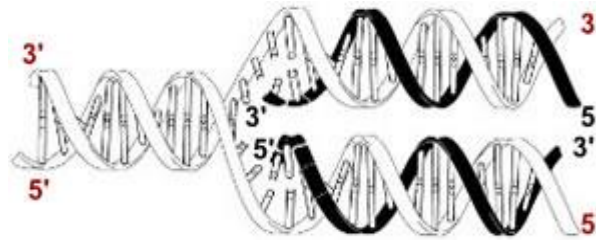
## Introduction

- 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,<sup>1,3,4</sup> Yoong Wearn Lim,<sup>1,3</sup> Paul L. Lott,<sup>2</sup> Ian Korf,<sup>1,2</sup>  
and Frédéric Chédin<sup>1,2,5</sup>

<sup>1</sup>Department of Molecular and Cellular Biology, <sup>2</sup>Genome Center, University of California, Davis, California 95616, USA

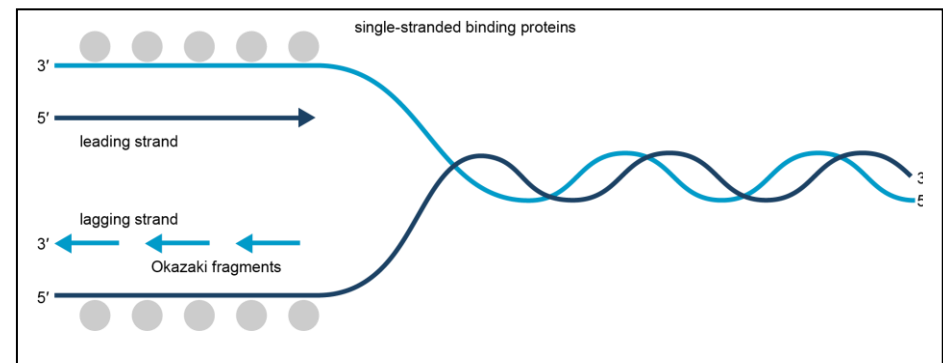
Strand asymmetry in the distribution of guanines and cytosines, measured by GC skew, predisposes DNA sequences toward R-loop formation upon transcription. Previous work revealed that GC skew and R-loop formation associate with a core set of unmethylated CpG island (CGI) promoters in the human genome. 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
- Only one strand is transcribed during transcription; the strand that contains the gene is called the sense strand

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).



## 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  $\leftarrow$  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  $\rightarrow$  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.
- A, C, T or G is generated with the probabilities above
- Notation for the output of a random string of  $n$  bases may be:  $L_1, L_2, \dots, L_n$  ( $L_i$  = 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  $\chi = \{x_1, x_2, \dots, x_n\}$ 
  - In the DNA sequence  $J=4$  and  $\chi = \{A, C, T, G\}$
  - $X$  is a discrete random variables (since its values are uncertain)
  - If  $p_j$  is the prob that the value (realization of the random variable  $X$ )  $x_j$  occurs, then
    - $p_1, \dots, p_J \geq 0$  and  $p_1 + \dots + p_J = 1$
- The probability distribution (probability mass function) of  $X$  is given by the collection  $p_1, \dots, p_J$ 
  - $P(X=x_j) = p_j, j=1, \dots, J$
- The probability that an event  $S$  occurs (subset of  $\chi$ ) is  $P(X \in S) = \sum_{j:x_j \in S} (p_j)$

## Probability distributions

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

- New sequence  $X_1, \dots, X_n$ :

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

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

$$N=X_1+\dots+X_n$$

- 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  $X_1, \dots, X_n$  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

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



## 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  $x_1, x_2, \dots$  (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  $E(Y) = c E(X)$
  - $E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n)$
- Because  $X_i$  are assumed to be independent and identically distributed (iid):

$$E(X_1 + \dots + X_n) = n E(X_1) = n p_A$$

## 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  $\text{Var}(X)$  can also be written as:

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

- If  $Y=c X$  then  $\text{Var}(Y) = c^2 \text{Var}(X)$
  - The variance of a sum of independent random variables is the sum of the individual variances
- 
- For the random variables  $X_i$ :  
 $\text{Var}(X_i) = [1^2 \times p_A + 0^2 \times (1 - p_A)] - p_A^2 = p_A(1 - p_A)$   
 $\text{Var}(N) = n \text{Var}(X_1) = np_A(1 - p_A)$

## 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.

$$\text{Var}(X) = E ( [X - E(X)]^2 )$$

- The positive square root of the variance of  $X$  is called its standard deviation  $\text{sd}(X)$

## 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 :

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

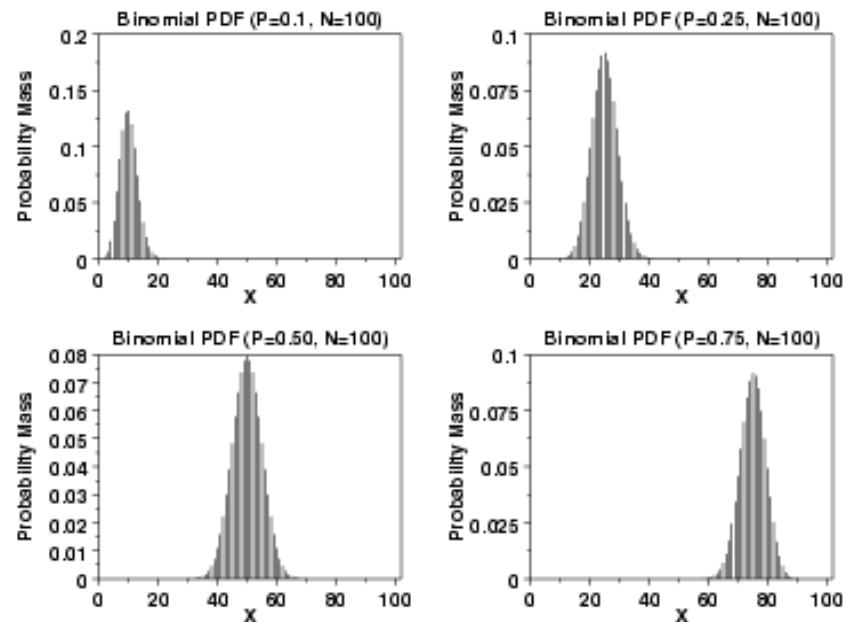
with the binomial coefficient  $\binom{n}{j}$  determined by

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

and  $j! = j(j-1)(j-2)\dots 3.2.1$ ,  $0! = 1$

## The binomial distribution

- The mean is  $np$  and the variance is  $np(1-p)$
- 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  $N_1, \dots, N_n$  having the same distribution as  $N$

- We can use the sample mean to estimate the expected value  $EN$ :

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

- 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  $X_1, \dots, X_n$  from which  $N$  can be computed by adding the  $X$ 's

## Simulating from probability distributions

- Simulate a sequence of bases  $L_1, \dots, L_n$ :
  - 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  $L_1, \dots, L_n$

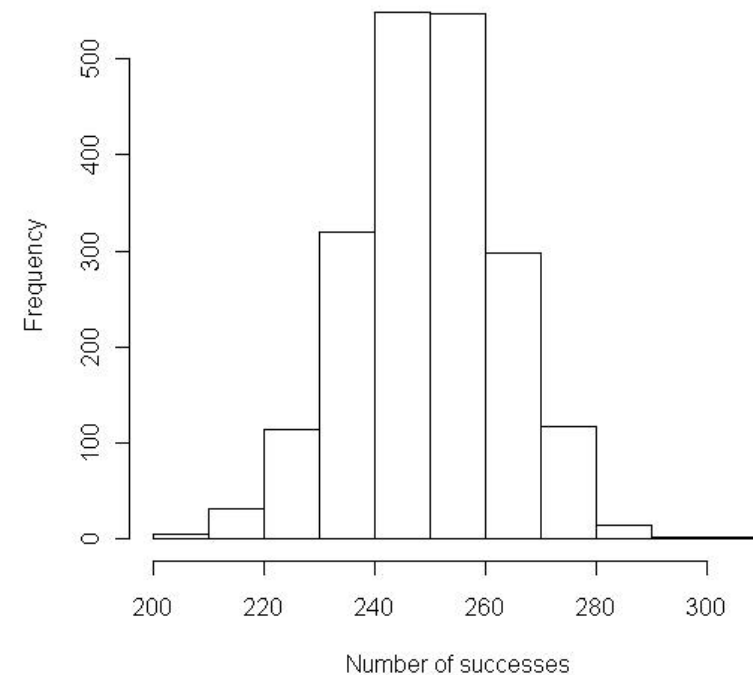
- Use the “sample” function in R:

```
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="")
```

## 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="")
```

Number of observations = 2000

Number of trials = 1000

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 :

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

and therefore

$$\begin{aligned} P(N \geq 300) &= \sum_{j=300}^{1000} \binom{1000}{j} (1/4)^j (1 - 1/4)^{1000-j} \\ &= 0.00019359032194965841 \end{aligned}$$

	P: exactly 300 out of 1000	
Method 1. exact binomial calculation	0.00004566114740576488	
Method 2. approximation via normal	0.000038	
Method 3. approximation via Poisson	-----	
	P: 300 or fewer out of 1000	
Method 1. exact binomial calculation	0.9998520708293378	
Method 2. approximation via normal	0.999885	
Method 3. approximation via Poisson	-----	
	P: 300 or more out of 1000	
Method 1. exact binomial calculation	0.00019359032194965841	
Method 2. approximation via normal	0.000153	
Method 3. approximation via Poisson	-----	
For hypothesis testing	P: 300 or more out of 1000	
	One-Tail	Two-Tail
Method 1. exact binomial calculation	0.00019359032194965841	0.0003025705168772097
Method 2. approximation via normal	0.000153	0.000306
Method 3. approximation via Poisson	-----	-----

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

## Approximate via simulation

- Using R code and simulations from the theoretical 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 3<sup>rd</sup> way to compute probabilities of a distribution
- It applies to sums or averages of iid random variables
- Assuming that  $X_1, \dots, X_n$  are iid random variables with mean  $\mu$  and variance  $\sigma^2$ , then we know that for the sample average

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

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

- Hence,

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

## Approximate via Central Limit Theory

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

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

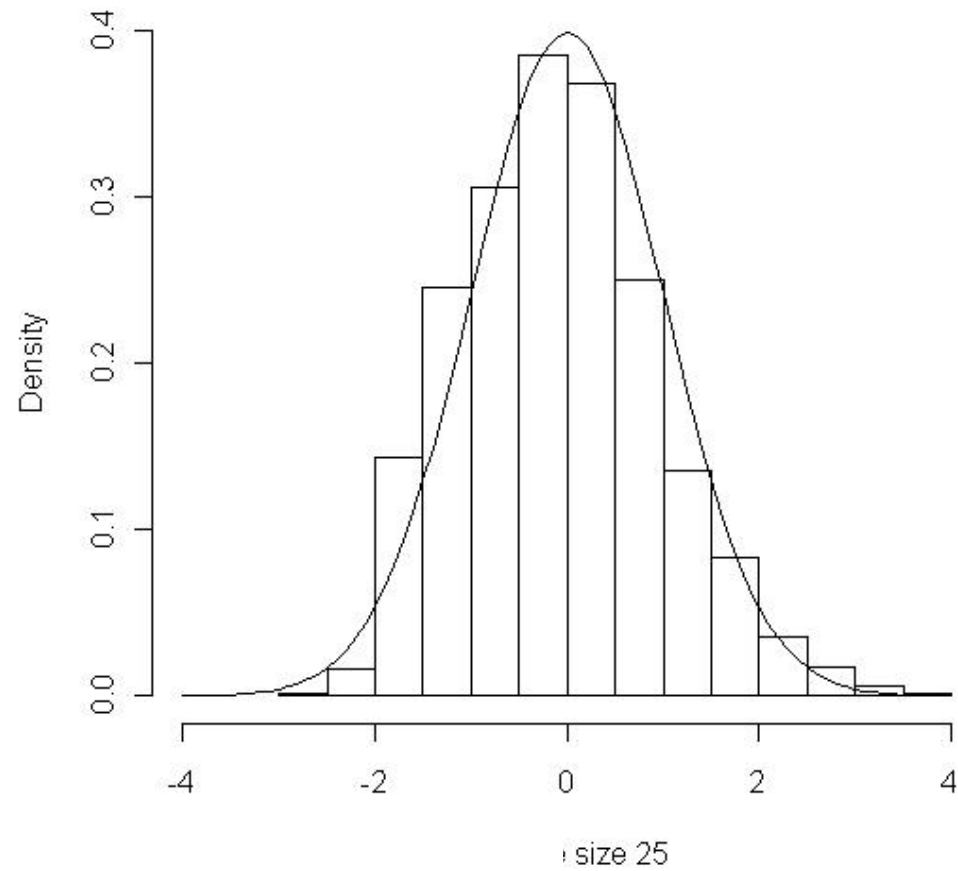
with  $\phi(\cdot)$  the standard normal distribution defined as

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

- 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 ...
- 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 " $\text{—C—phosphate—G—}$ " (cytosine and guanine separated by only one phosphate; phosphate links any two nucleosides together in DNA)

## CpG sites

```

CATTCCGCCTTCTCTCCGCAGGTGGCGCGTGGGA
GGTGTTTTGCTCGGGTTCTGTAAGAATAGGCCAGG
CAGCTTCCCGCGGGATGCGCTCATCCCCTCTCGG
GGTTCGGCTCCCACCGCGCGCGGTTCCGCCGGT
CCGCCTGCGAGATGTTTTCCAGCGACAATGATTC
CACTCTCGCGCCTCCCATGTTGATCCCAGCTCCT
CTGCGGGCGTCAGGACCCCTGGGCCCGCCCG
CTCCACTCAGTCAATCTTTGTCCCCTATAAGGCG
GATTATCGGGTGGCTGGGGGCGGCTGATTCGA
CGAATGCCCTTGGGGGTCACCGGGAGGGAATC
CGGGCTCGGCTTTGGCCAGCCCGCACCCCTGTT
TGAGCCGGCCCGAGGGCCACCAGGGGGCGCTCG
ATGTTCTGCAGCCCCCGCAGCAGCCCCACTCC
CCGGCTCACCCCTACGATTGGCTGGCCCGCCCGAG
CTCTGTGCTGTGATTGGTCACAGCCCGTGCCGTC
CGGGCGCCCGGGCGGATACGAGGTGACCGCGCA
GAGGCCAGCTCGGGCGGTGTCCCGCGCCGGC
GACTCGGGCGGAGTTTCCCGAGGGCCGAAGCG
GGCAGTGTGACCGCAGCGGTCCTGGGAGGCGC
CCGGCGCGCGTCCGAGCAGCTCCCCTCCTCGCA
GCCTCACCGCGCGCGTCCCGCGCCCTGGCC
TCCCGCACTCGCGCACTCCTGTCCCGCGCCACG
GCCCACCTCCCACCTCGATGCGGTGCCTGGCTGC
TGCGTGATGGGGCTGCGGAGCGGCGCCCTGCGG
CTCGCGCGCGCGCTGCTGCTCGCGCTGAGGTGCGT
CGGTGCCCGGCCCGCGCCCGCGCGCGCGCG
GGCTCCTGTTGACCCTGTCGCGCCCGTCCGTCTGC
AGCGCGGCTGAGGTAAGGCGGCGGGGCTGGCCG
CGGTTGGCGCGCGGTCCCGGGGTTGGGGAGG
GGCCGCTTCCCGCGGGGAGGAGCGGCCTGGCCG
GGTCCGGCGGGTCTGAGGGGA
CTCTTAGTTTTGGGTGCATTTGTCTGGTCTTCCAAA
CTAGATTGAAAGCTCTGAAAAAAAAAACTATCTTGT
GTTTCTATCTGTTGAGCTCATAGTAGGTATCCAGGA
AGTAGTAGGGTTGACTGCATTGATTTGGGACTACAC
TGGGAGTTTTCTTCCCATCTCCCTTAGTTTTCT
TTTTTCTTTCTTTCTTTCTTTTCTTTTCTTTTTTTT
TTGAGATGTCTTCTTGTCTCAGTCCCCCAGGCTGGA
GTGCAGTGGTGCGATCTTGGCTCACTGTAGCCTCC
ACCTCCCAGGTTCAAGCAATCTACTGCCTTAGCCT
CCCGAGTAGCTGGGATTACAAGCACCCCGCCACCAT
TCCTGGCTAATTTTTTTTTTTGTATTTTAGTTGAGA
CAGGGTTTACCATGTTGGTGATGCTGGTCTCAGA
CTCCTGGGGCCTAGCGATCCCCCTGCCTCAGCCT
CCCAGAGTGTTAGGATTACAGGCATGAGCCACTGT
ACCCGCCTCTCTCCAGTTTCCAGTTGGAATCCAA
GGGAAGTAAGTTTAAGATAAAGTTACGATTTTGAAT
TTTTGGATTCAGAAGAATTTGTACCTTTAACACCT
AGAGTTGAACGTTTCATACCTGGAGAGCCTTAACATT
AAGCCCTAGCCAGCCTCCAGCAAGTGGACATTGGT
CAGGTTTGGCAGGATTCTCCCTGAAGTGGACT
GAGAGCCACACCCTGGCCTGTACCATACCCATCC
CCTATCCTTAGTGAAGCAAACTCCTTTGTTCCCTT
CTCCTTCTCCTAGTGACAGGAAATATTGTGATCCTA
AAGAATGAAAATAGCTTGTACCTCGTGGCCTCAG
GCCTCTTGACTTCAGGCGGTTCTGTTAATCAAGT
GACATCTTCCCGAGGCTCCTGAATGTGGCAGATG
AAAGAGACTAGTTCAACCCTGACCTGAGGGGAAAG
CCTTTGTGAAGGGTCAGGAG

```

**Left:** CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG constitutes the start codon.

**Right:** CpG sites present at every 1/100 nucleotides, constituting 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<sup>1,2</sup>  
tom@sfc.keio.ac.jp

Kouichi Takahashi<sup>1,3</sup>  
t94249kt@sfc.keio.ac.jp

Masaru Tomita<sup>1,3</sup>  
mt@sfc.keio.ac.jp

<sup>1</sup> Laboratory for Bioinformatics, <sup>2</sup> Graduate School of Media and Governance,  
<sup>3</sup> 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  $L_1, \dots, L_n$ :

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

- Has a given sequence an unusual dinucleotide frequency compared to the iid model?

- Compare observed  $O$  with expected  $E$  dinucleotide numbers

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

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

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

## Comparing to the reference

- How to determine which values of  $\chi^2$  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  $\chi^2$  is given as before
- If this ratio is larger than 3.84 then conclude that the iid model is not a good fit
- Note:  $qchisq(0.95,1) = 3.84$

### 3.b Studying words of length 3

- 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.
- 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  $L_1, L_2, \dots, L_n$  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

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



## The codon adaptation index

- Comparison of predicted and observed triplet frequencies in coding sequences for a subset of genes and codons from E. coli.
- Figures in parentheses below each gene class show the number of genes in that class.

	Codon	Predicted	Observed	
			Gene Class I (502)	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**

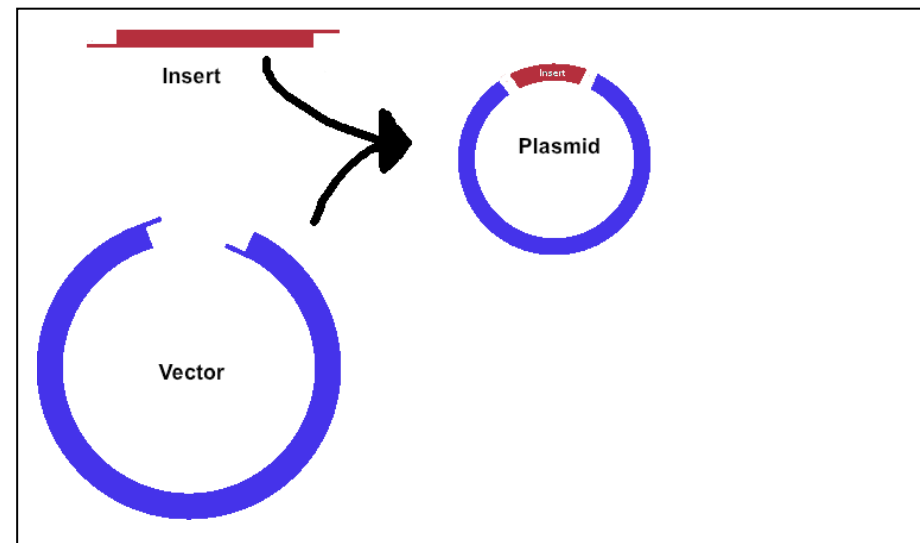
Class I : Moderately expressed genes

(Table 2.3 from Deonier et al 2005)

## 4 Restriction sites

### Introduction

- Because DNA can be long but is very thin, it is easily broken during processing. Note that the DNA in human chromosome 1, at 245,000,000bp, is 8.33cm long and only  $20 \times 10^{-8}$  cm thick
- **Molecular scissors: Restriction endonucleases** provides the means for precisely and reproducibly cutting the DNA into fragments of manageable size (usually in the size range of 100s to 1000s of base pairs)
- Cloning puts DNA of manageable size into vectors that allow the inserted DNA to be amplified

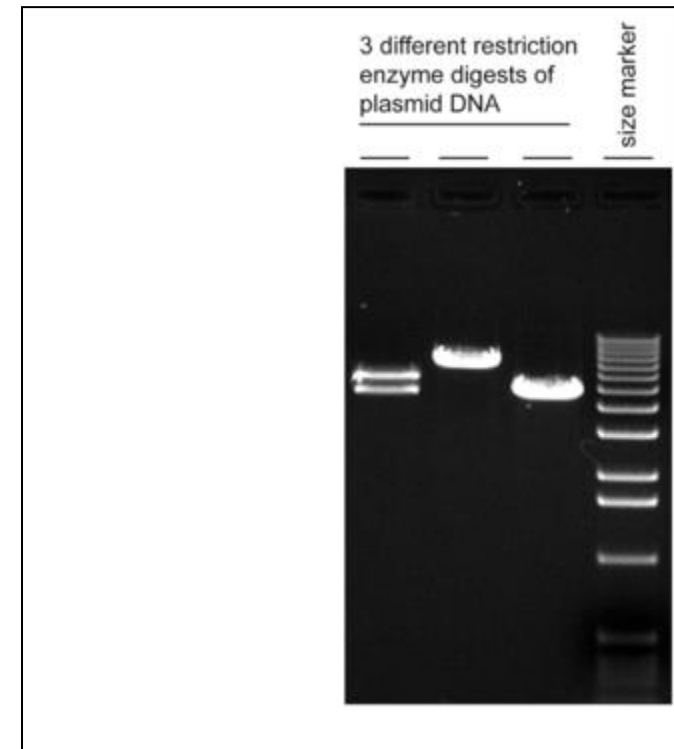
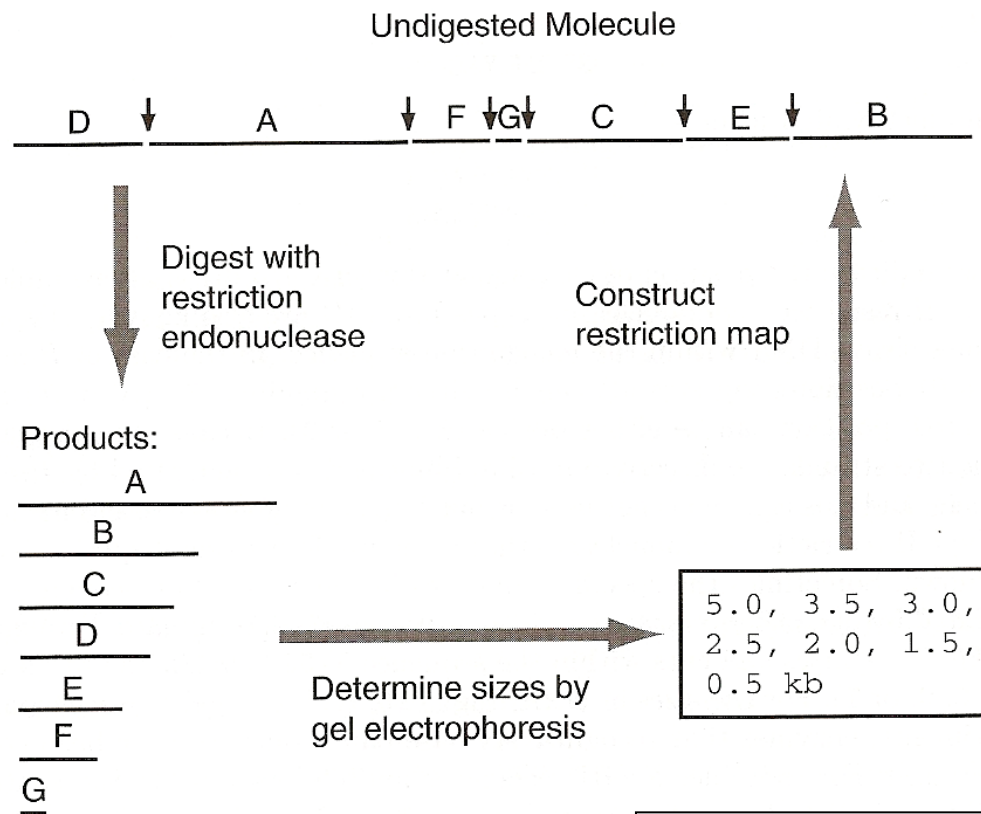


## Introduction

- A **restriction map** is a display of positions on a DNA molecule where cleavage by one or more restriction endonucleases can occur.
- It is created by determining the ordering of the DNA fragments generated after digestion with one or more restriction endonucleases.
- The restriction map is useful not only for dissecting a DNA segment for further analysis but also as a "fingerprint" or bar code that distinguishes that molecule from any other molecule.
- A graphical summary is given in the following figure (Figure 3.1 – Deonier et al 2005)

## Introduction

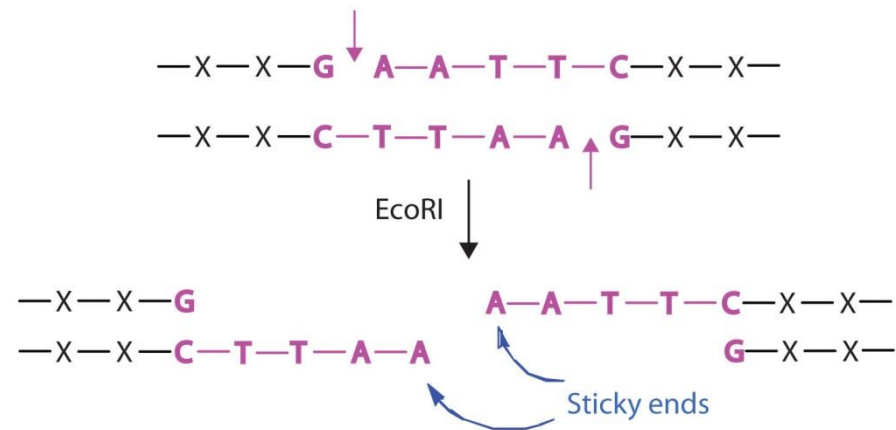
- The order of fragments (D, A, F, G, C, E, B) is originally unknown. A variety of techniques may be employed to determine this order.



Fragments of linear DNA migrate through agarose gels with a mobility that is inversely proportional to the log<sub>10</sub> of their molecular weight.

## Introduction

- An example of a restriction enzyme is EcoRI
- The EcoRI restriction enzyme, the first restriction enzyme isolated from *E. Coli* bacteria, is able to recognize the base sequence 5' GAATTC 3'.
- Each strand of DNA is cut between the G and the A in this sequence. This leaves "sticky ends" or single stranded overhangs of DNA. Each single stranded overhang has the sequence 5' AATT 3'.



## Introduction

- If we were to digest the DNA with a restriction endonuclease such as EcoR1, then we can ask ourselves the following questions:
  - 1) approximately how many fragments would be obtained (how many times was the sequence cut), and
  - 2) what would be their size distribution (which lengths are obtained for the restriction fragments)?

## 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,

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

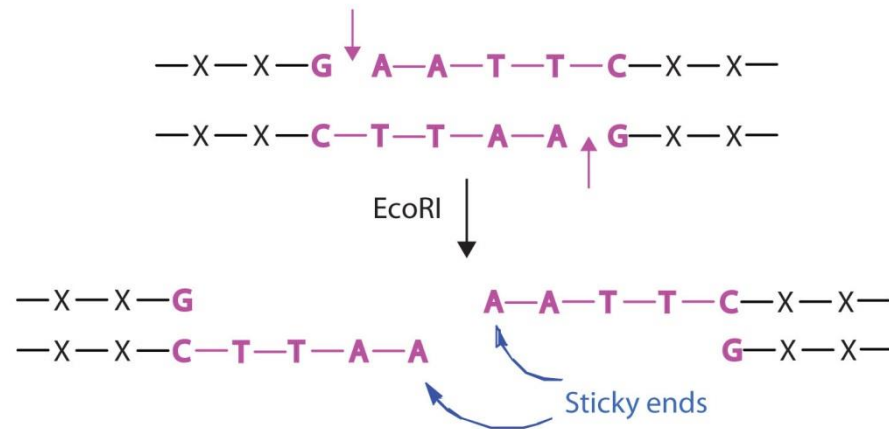
- We denote by  $p$  the probability that any position  $i$  is the beginning of a restriction site:

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



## 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.



## The number of restriction sites

- **Model assumptions**

- The DNA has equal proportions of A, C, G, and T (e.g.  $P(G)=0.25$ ).
- 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

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

- 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  $X_1, X_2, \dots, X_n$ ,
- The number of restriction sites is  $N = X_1 + X_2 + \dots + X_m$ , with  $m = n - 5$ .
  - 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  $X_1, X_2, \dots, X_n$  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  $np(1 - p)$ .
- The binomial approximation usually works well, even though we know that the  $X_i$  are in fact NOT independent of one another (because of overlaps in the patterns corresponding to  $X_i$  and  $X_{i+1}$ , for example).
- We have already seen that computing probabilities of events can be cumbersome when using the probability distribution

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

- Since  $n$  is large and  $p$  is small (see before), we can rely on the Poisson approximation of the binomial distribution

## Poisson approximation to the binomial distribution

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

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

- 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

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

## 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  $x$ ,

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

- Since  $n$  is large (often more than 104), we replace  $\left(1 - \frac{\lambda}{n}\right)^n$  by  $e^{-\lambda}$  to get our final approximation in the form

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

- This is the formula for the Poisson distribution with parameter  $\lambda = np$  (note: this parameter represents both the mean and variance)

## 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 *EcoRI* sites in a DNA molecule of length 10,000, assuming equal base frequencies
- Earlier we obtained  $p=0.00024$  for this setting.
- 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 *EcoRI* two times or less

- R code:

```
ppois(2,2.4)
```

## Distribution of restriction fragment lengths

- There is a more general version of the Poisson distribution: it generalizes  $n$  into “length” and  $p$  into “rate”. We suppose that “events” (restriction sites) occur on a line at rate  $\mu$ .
- Then the probability of  $k$  sites in an interval of length  $l$  bp is

$$\frac{e^{-\mu l} (\mu l)^k}{k!}, \quad k=0,1,2, \dots$$

- 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  $(y, y + x)$ :

$$P(X > x) = e^{-\mu x} = e^{-\lambda \left(\frac{x}{n}\right)}$$



## Distribution of restriction fragment lengths

- The previous has some important consequences:

$$P(x \leq x) = \int_0^x f(y)dy = 1 - e^{-\mu x}$$

so that the density function for  $X$  is given by

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

- The distance between restriction sites therefore follows an exponential distribution with parameter  $\mu$ ; the mean distance between restriction sites is  $1/\mu$

## Simulating restriction fragment lengths

- 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:

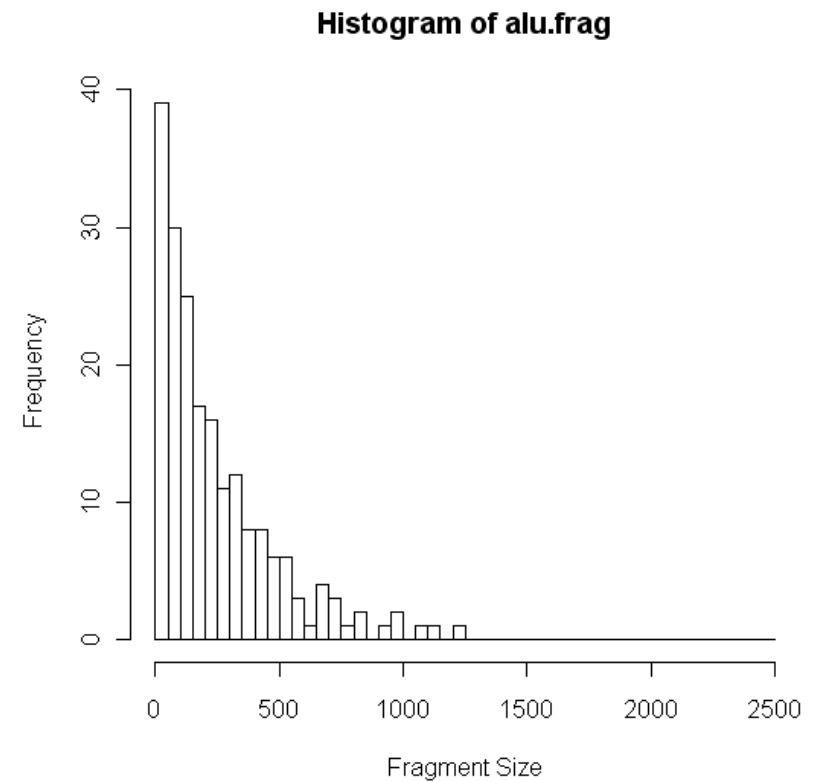
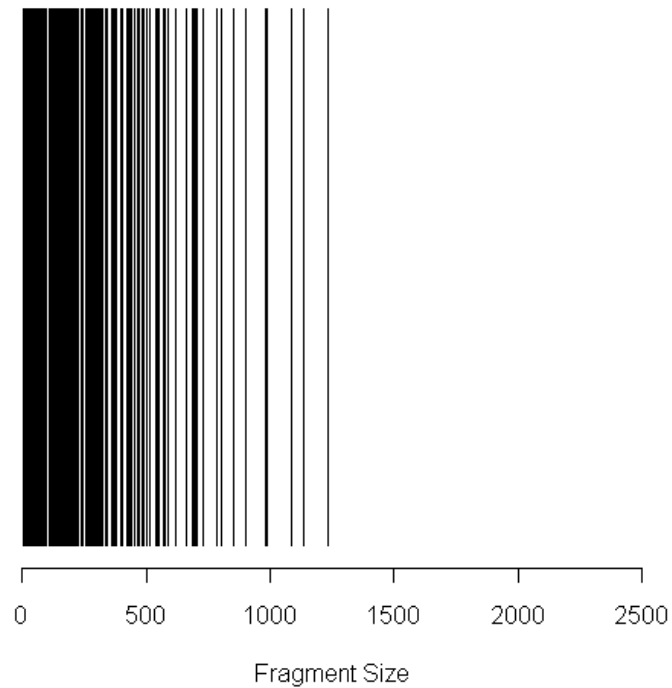
```
x<-c(1:4)
propn <- c(0.25,0.25,0.25,0.25)
seq2 <- sample(x,48500,replace=TRUE,prob=propn)
seq2[1:15]
length(seq2[])
```

## Simulating restriction fragment lengths

- What else is needed?
  - R code identifying the restriction sites in a sequence string, with bases coded numerically: function *rsite*
  - Code of the restriction sites we are looking for: e.g., for AluI it would be AGCT.
  - R code to compute the fragment lengths: subtract positions of successive sites

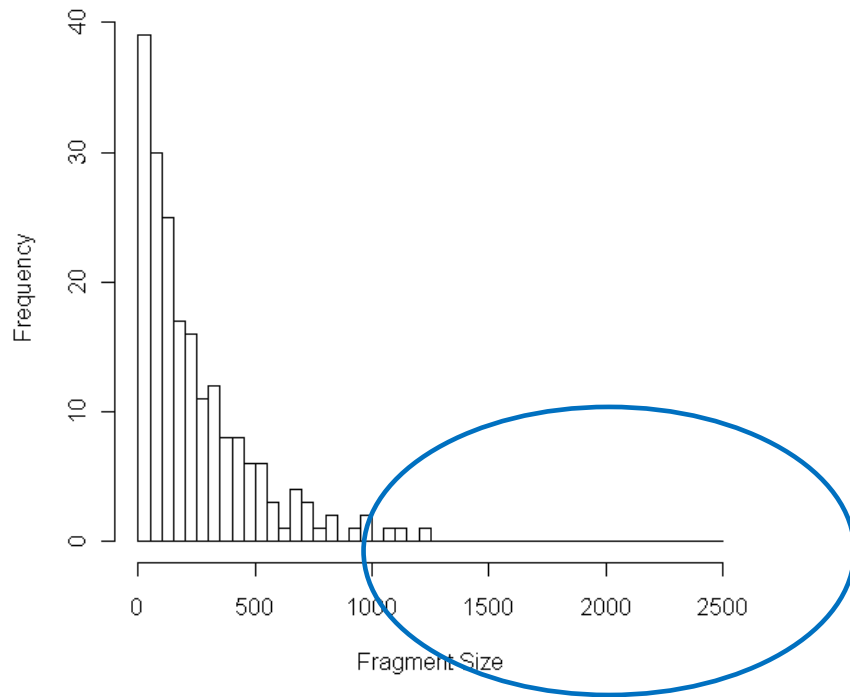
(R code posted online)

## Simulating restriction fragment lengths

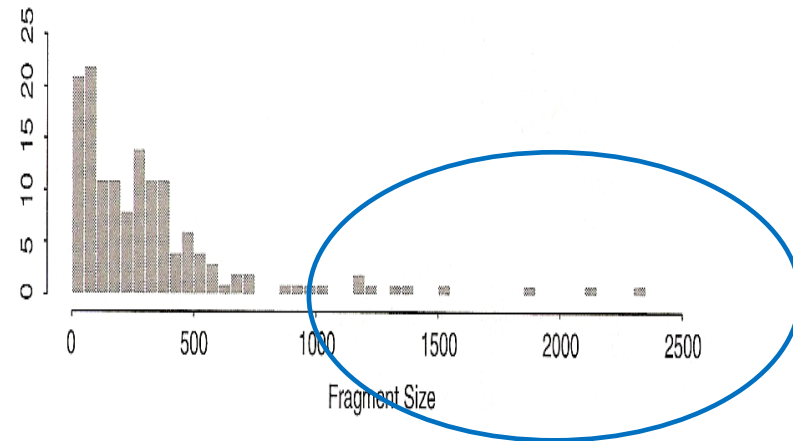


# Is our theoretical model to simulate restriction fragment lengths valid?

Histogram of alu.frag



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  $\chi^2$  – test.

## 5 R code (at home)

- R scripts used throughout this Chapter can be replayed via the code included in the file

[RCode to Chapter5 and BackgroundInfo.7z](#)

- R scripts illustrating relevant R packages for sequence pattern recognition and sequence comparison (see also practical session):
  - DNA sequence statistics: <http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter1.html>
  - Querying 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>

## Main reference

- Deonier et al. *Computational Genome Analysis*, 2005, Springer. (Chapters 6,7)

## Background reading

- Pabinger et al. 2013. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*.
- Pavlopoulos et al. 2013. Unraveling genomic variation from next generation sequencing data. *BioData Mining* 6:13.