Welcome to Bioinformatics

Christopher Lee

Department of Chemistry and Biochemistry, UCLA

Welcome to Bioinformatics!

Bioinformatics is the study of the inherent structure of biological information.

- There is inherent structure: many kinds of patterns.
- An ascending scale of problems from here to eternity...
- Algorithms that solve the *pattern finding* problem.
- Statistical metrics that distinguish *information* from random *noise*.

Who Is This Class For?

- This course is for people who may want to invent new kinds of bioinformatics methods in their future work.
- Therefore it focuses on understanding how to use the "core" concepts that arise in virtually every bioinformatics problem, to solve these problems.
- Very broadly applicable to any data mining problem -- what Google, Amazon, Netflix and every Big Data company uses every day.
- *Not* a "survey course" that gives a *comprehensive* overview of standard methods in the field.
- *Not* an "applications course" that introduces you to all the kinds of biology and medicine problems that bioinformatics addresses.

This class neither assumes nor teaches biology

- Focus in this class is on bioinformatics theory and problem solving skills, not biological applications (for that I offer a *Genomics & Computational Biology* course, Chem. 100, in Spring).
- Today we will quickly summarize some basic genetics principles and vocabulary, to get you comfortable with the types of data and problems we'll be working on.
- We're also giving you a short glossary of relevant bio terms, to ensure that no one feels uncertain about their meanings. We will also provide this glossary during exams.
- Jones & Pevzner chapters 1-3 give useful bio background.

Weekly Projects and "Mission Training"

Most weeks you'll have a **real-world problem** to solve, typically in the form of a **programming project**, and occasionally an in-class **quiz** (3).

- we give you **mission training** for solving each real-world problem, due Friday (noon). Do it online (Courselets.org). Full credit for fully answering each question (*no points off for errors*). Use this to sort out anything that's confusing you.
- **Project** due the following Monday (noon). Once you've done the mission training, the project should be straightforward.
- on three Mondays, instead a half hour **quiz**, applying the skills you learned in the project + mission training to a different real-world problem. (And there is no midterm).

Foundations: Conceptual Understanding

- Solving problems in the real world requires above all understanding the fundamental concepts and how to apply them correctly.
- But this is easier said than done. First of all, what does "understanding the concepts" actually mean?

What It Isn't

- **memorization**: "remembering what the textbook says"
- **plug-and-chug**: remembering how to do one kind of calculation, by practicing it over and over.
- **solving equations**: given a pre-specified set of conditions, solve for the unknown(s).

We may need to do some of these things, but by themselves they are *not enough* to figure out real-world problems.

Understanding = Correct Usage

- it's one thing to "know" Newton's three Laws of Motion, and another thing altogether to be able to use them to solve all problems in classical mechanics.
- Analogy: think of basic concepts as "axioms", vs. all their implications for real-world problem-solving as "theorems". I.e. there are many ways you could try to use a given concept, some of which are right (correct theorems) and many of which are wrong.
- Understanding means not only knowing which are right vs. wrong, but also being able to explain from first principles why the wrong ways are incorrect.
- To learn this requires doing the mental work to prove it to yourself, i.e. by actually *making* the error, and seeing it leads to obviously wrong conclusions.

In-class Exercises

- **Implicational thinking** is a muscle that you can only build by **exercising it**.
- We'll take time each class to solve conceptual problems;
- Submit answers to online server so we can categorize, share and discuss answers.
- Our goal is to capture the common mistakes we make, to clear up confusions for everyone in class.
- Together we'll filter out *error models* vs the solution.
- Ungraded; purely for learning (but you get points for participating).

Learning How to Learn

- this actually a research project that we've invested a lot of effort in building active learning tools for you...
- occasionally research observers will be in class taking notes on how we spend our time.
- We've built an online platform for the active-learning exercises which will be using in class (Courselets.org).
- if you run into any problems tell us, either via the class Discussion forum or our Github issue tracker (click the About link on Courselets.org).

Expect Different

This is not going to be a typical lecture class; much of it may seem surprising at first because it won't be what you're used to.

- in-class concept questions will probe your understanding of how to *use* a concept.
- making mistakes on these is a *good thing*: exposes the various ways we misunderstand and misuse that concept, so we can learn how to use it right in all situations.
- Expect to get at most half the questions in-class completely right.
- "Why are you asking us questions you didn't already show us how to do?" Because this is about *thinking*, not *memorization*.
- 100% correct = zero learning opportunity.
- similar conceptual questions will constitute about half the exams.

Learning = Collaboration NOT Competition

- "Grading on the curve" enforces strict competition: "no good deed goes unpunished"
- Having taught this class many years, I already know the exact score thresholds for "A performance", "B performance" etc.
- I've locked those grade thresholds, so if this year's class learns more (higher average score) then **everyone will get a higher grade**.
- You are **not** in competition with each other!
- Help each other learn. Everybody wins!
- On individually graded assignments, give & get help on conceptual understanding, but do your own work. After all, you need those skills for the exams.

My Goal: 100% of you understand 100% of the concepts

- on your own you'll understand about 50% of the concept tests.
- for each concept test that confused you, identify what **error model** you fell into, see what review and exercises help you clear it up (**Resolutions**).
- **Update your status** on each such problem: if the materials available for it aren't clearing it up for you, set your status as "Still confused, need help!"
- To address those help requests, we'll try to provide additional error explanations and resolutions.
- ask your classmates for help on concepts that confused you, and help them on anything that confuses *them*.
- Use these to clear up *all* your misunderstandings of every concept we work on that week!
- Remember: you are not in competition with each other!

Learning = FAIL NOW rather than LATER

find out you can't solve real world problems...

- **A.** after you graduate;
- **B.** on the final exam;
- **C.** the first day of class.

In this class, we're going to try to press fast-forward to hit you with the kind of puzzlements that real-world problems will pose you.
Such as...

ABORT Errors: Never Got Off the Launch Pad

- *missing a basic definition*: you can't afford to sleep in this class. Anything I tell you, you're going to use 5 minutes later.
- *didn't really read the question*: if you don't understand the question, please ask!
- *got hung up on a detail and didn't think through the main question*
- *couldn't figure out how to get started*: failed to apply concept(s) that you *know*. Classic error: "Trying to remember the answer"... too big of a leap, often leads to a blank.

The starting line: THINK about what the question is really asking, and about what concept(s) that you know are relevant to it. You're sure to come up with something, and that's better than not even getting to the starting line. The key activity: THINK instead of REMEMBER.

Never Get Stuck Again: Use Three Superpowers

- three mental powertools: human language; pictures; math. **Use all three!**
- **Words**: speak the language! What words are relevant to your problem?
- draw a **picture**! Is there a picture that will help you make sense of the problem?
- **Equations** are powerful and essential, but they aren't the only way to think about these ideas, or always the best.
- If you get stuck using one, try another superpower!

Questions++

- **Please ask questions in class, any time, all the time**.
- I want everyone to ask all their questions, and I want everyone to hear the answer.
- Remember, you are *not* in competition with each other.
- If you have a question, chances are lots of other people in class are puzzled about it too. So **please ask it**; they will thank you!
- Questions are the only way to learn: you can listen to me without thinking whatsoever, but if you ask a question it shows you're thinking. And if you're thinking, you're learning.

Reading

- assigned textbook: Jones & Pevzner, *Introduction to Bioinformatics Algorithms*. Focuses mainly on algorithms.
- assigned chapters (Lee; download from web site). Will cover additional material e.g. probabilistic modeling.
- **Today's Assignment**: biology background for CS students. Jones & Pevzner, Chapter 3; Lee, Genetics & molecular biology basic terminology glossary.
- Provided for your learning and reference -- no need to memorize it! E.g. the glossary will be included with your exams.

Class Website

- We are using UCLA CCLE (aka Moodle): `http://ccle.ucla.edu`
- Login with your UCLA BOL account, and click on this course.
- For a start, see the posts in Week 1.

Points Breakdown

- Undergrad: 15% homework (mission training), 15% projects, 25% midterm, 35% final, 10% class participation
- Grad: 15% homework, 15% projects, 20% midterm, 30% final, 10% term project, 10% class participation
- Graduate term project should be a small research project, e.g. some kind of improved algorithm, or a real data analysis project. You'll receive a handout on detailed guidelines for the term project. Keep it focused and be realistic about what's doable in the available time.

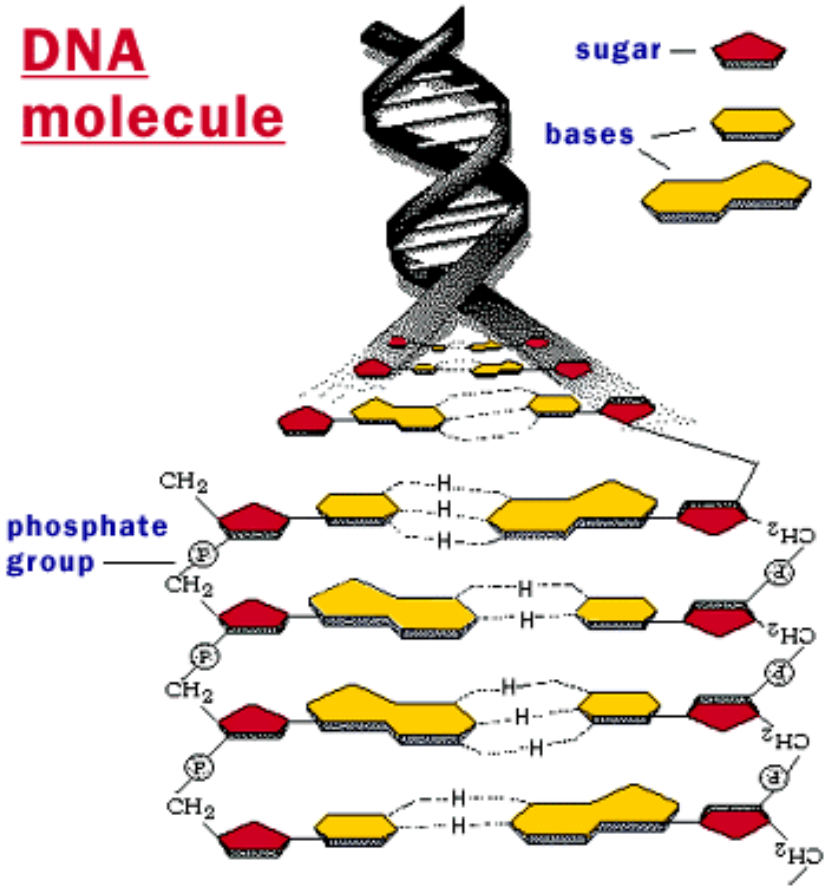
An Introduction to What?

- This course is "intro" in the sense that it focuses on basics.
- Fundamentals that come up in *every* bioinformatics problem.
- Learn these fundamentals like you're really going to need to use them throughout your life! That means: know how to think using these powertools.
- "Why's he asking another probabilistic modeling question?" Because if you have a basic misunderstanding of these key concepts, you are almost sure to go wrong (get stuck) when you try to solve real world problems, whereas if you have a basic understanding, you will keep coming back to the right path no matter how many mistakes you make along the way.
- Other courses: CSB 184, *Intro Computational & Systems Biology*; CS124, *Computational Genetics*; Chem C100, *Genomics & Computational Biology*.

Basic Genetics & Genomics Overview

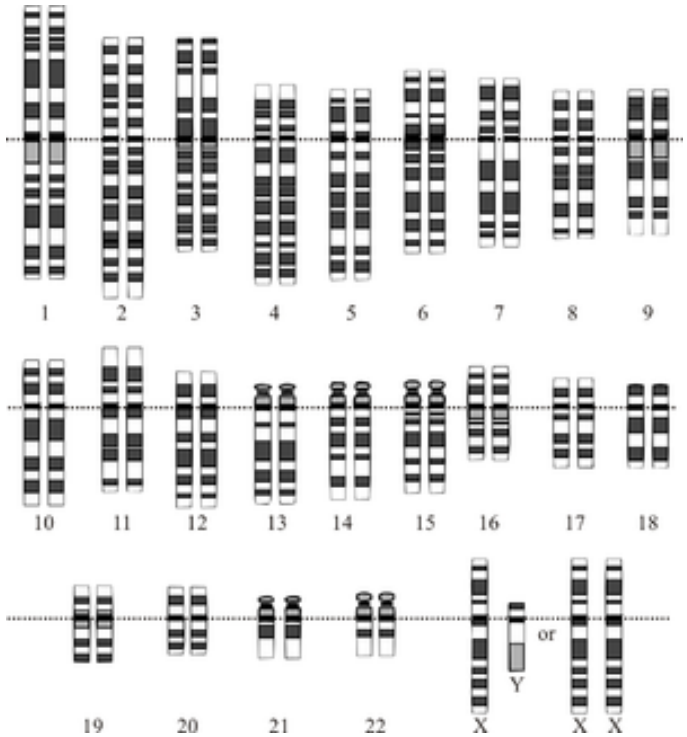
- Biology has its share of specialized jargon just like CS...
- Don't worry: it may sound complicated at first, but it's actually quite simple.
- Concretely, we are going to restrict ourselves to a very simple subset of biology problems throughout this class, which will gradually become very familiar to you.
- Let's start with a quick overview of basic genetics concepts and the kinds of biological data ("genomics") we'll be working with.
- Purpose is to start getting you comfortable with these, not to memorize them (remember, you'll always have the glossary...)

DNA: the Text of Life



- *DNA*: A DNA sequence is a string consisting of a four letter alphabet (A, C, G, T). The four "letters" are called nucleotides or bases.

The Human Genome Consists of 23 Chromosomes

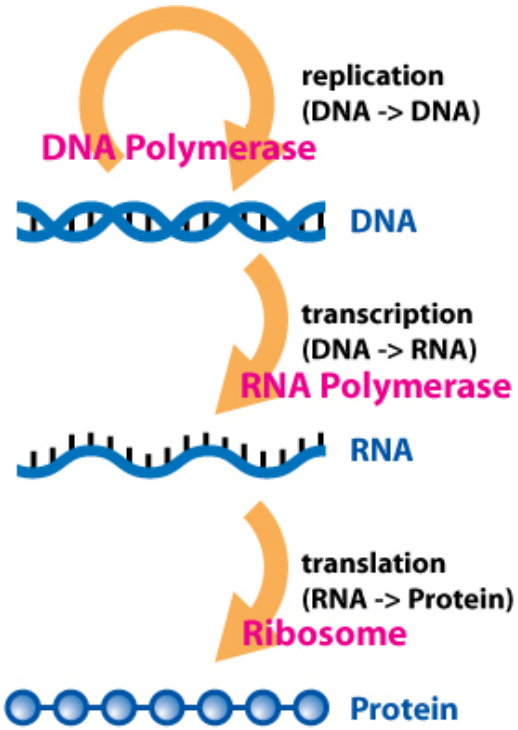


- *genome*: The total DNA sequence of an organism. The genome of a small bacterium is about 4 million letters long and contains about 4000 genes. The human genome is about 3 billion letters long, and contains about 25,000 genes. The length of a DNA string is given in "base-pairs" (bp; just means a single letter); kilobases (Kb, 1000 letters) or megabases (Mb, a million letters).

Basic Genetics

- *genome copy number*: Some organisms keep a single copy of the genome in each individual, e.g. bacteria. Animals and plants generally have two distinct copies of the genome in each individual. These two copies can have (somewhat) different sequences, see *polymorphism*.
- *gene*: a specific substring of the DNA that encodes a specific functional unit called a protein.
- *mutation*: an alteration of the DNA sequence that changes one or more letters in the DNA string. Mutation occurs at a low rate over time (roughly 10^{-8} letters change per generation). Some mutations change only a single letter, while other mutations might insert or delete a substring of the DNA.
- *phenotype*: A specific, observable change in function due to a change in the DNA sequence.

From Blueprint to Function: DNA to RNA to Protein



- *transcription*: the process of copying the DNA sequence representing a gene, to an RNA sequence, often referred to as the "messenger RNA" or "mRNA" for that gene.
- *gene expression*: regulation of the activity (or "expression") of a gene by increasing or reducing the amount of mRNA for that gene.

RNA Makes Proteins, Which Do the Work of Life

- *protein*: a string composed of a 20 letter alphabet (the amino acids) encoded by translating the DNA sequence of a gene. Proteins perform most of the biochemical activities in a cell, and a specific protein sequence carries out a specific activity, thanks to its unique physical structure determined by its sequence.
- *codon*: a group of three consecutive nucleotides that encode a single amino acid according to the universal genetic code.
- *translation*: the process of making a protein, by making the string of amino acids specified by a string of codons in a gene.

Let's try some concept tests

- A chance for you to have some fun with these ideas, by starting to think about them.
- Not graded and not on the exams -- this is not a biology class!
- Connect to wifi: **UCLA_WIFI** (UCLA password) or **UCLA_WEB**
- Login to **ccle.ucla.edu**, click on our class site, go to **Week 1**, and click link to **our problem-solving exercise platform (courselets)**.
- Bookmark **https://courselets.org/ct/** for easy access in future classes!

Types of Mutational Effects

Given the basic flow of information in molecular biology, propose two specific molecular explanations for how a mutation can impact gene function (for simplicity, assume an organism with only one copy of the genome).

Types of Mutational Effects Answer

- a mutation that alters a codon in the gene can change the amino acid sequence of the resulting protein. That could inactivate or alter the function of that protein.
- a mutation that alters the transcriptional regulation (expression) of the gene could block its expression or cause it to be incorrectly regulated (expressed at the wrong time or place).
- a large deletion could completely remove multiple genes, eliminating production of the proteins they encode.
- a mutation could alter an RNA splice site or splicing regulatory sequence, resulting in abnormal RNA splicing and typically a large change in the resulting protein.

(Other specific mechanisms are also possible; any two specific mechanisms would be acceptable).

Types of Mutational Effects Errors

- Some people didn't give a concrete molecular mechanism, although the question asks you to propose something specific and molecular.
- Some people were unsure what "gene function" means.

Attack of the Clones

- *clone*: If two individuals are genetically identical, we say they are *clones* (of each other). Asexual organisms (such as bacteria) typically have only one copy of the genome, and reproduce clonally, i.e. the "daughter" cell is an exact genetic copy of the "mother" cell. Biologists often separate a complex population of cells by "plating" them at low density (say 1 cell / cm^2) and letting them grow; each cell will make a colony; each colony is clonal. Note that human clones (i.e. exact twins) are not very rare.

Sequence Variation

- *polymorphism*: Mutations in a population that have not yet been lost or "fixed" are called "polymorphisms", which simply means that some individuals have the mutation while others do not.
- *SNP*: A single nucleotide polymorphism (SNP) is a single letter in the DNA sequence that differs between individuals.
- *allele frequency*: the probability of finding a specific polymorphism on any given copy of the genome in a population.

Finding the genetic cause of a mutant phenotype

Say a bacterium has about 4000 genes (each mapped to a specific interval of its known genome sequence), and is subjected to a chemical treatment that induces about 50 random mutations (single nucleotide substitutions) per genome. A large number of resulting mutant bacteria are screened for a useful phenotype (e.g. greatly increased production of a biofuel), and the screen successfully finds many different mutant strains with the desired phenotype (whereas the parental strain lacks this phenotype). A researcher sequences **one** such strain (clone) and compares it with the parental (unmutated) sequence.

- Say the researcher knows there is only one gene in the genome where mutations can cause this phenotype, but does not initially know *which* gene it is. Can the researcher identify which gene this is, directly from this sequencing data? If so, explain how the researcher could do this. If not, explain whether or not there is a simple extension of this analysis that will do so.

Finding the genetic cause of a mutant phenotype Answer

- No, sequencing one mutant strain will identify approximately 50 mutated genes, so the researcher will still have no clear idea of which gene actually causes the phenotype.
- But simply sequencing more (independent) mutant strains with the phenotype should quickly identify the gene. It should be mutated at least once in every strain, whereas other (non-causal) genes have only a probability of 50/4000 of being mutated in each strain.

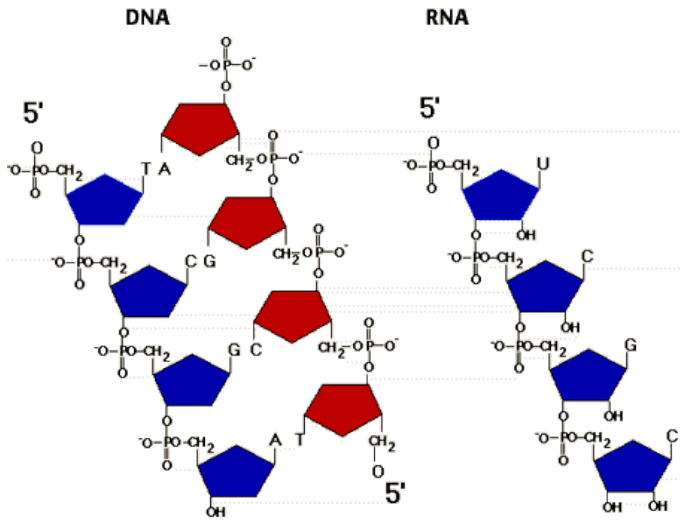
Finding the genetic cause of a mutant phenotype Errors

- Some people seemed to assume the mutant strain will only contain one mutation (contrary to what the question said).
- Some people did not consider that 50 random mutations (in a genome containing 4000 genes) will result in (approximately) 50 mutated genes, making it very uncertain which mutated gene is the one that causes the phenotype.
- Some people didn't consider that the huge difference in expected mutations per strain (one per strain in the causal gene, vs. approx. 50/4000 per strain in any non-causal gene) should make it easy to distinguish the causal gene, if we just sequence more independent mutant strains.

Basic Mendelian Genetics (Plants, Animals, etc.)

- *Mendelian inheritance*: the standard pattern of gene transmission in plants and animals, in which each individual has two copies of a gene (one copy from each of its two parents), and passes on one of its copies (chosen at random) to each of its children.
- *recessive mutation*: Because of this, most mutations that damage a gene function are "recessive", which means that *both* copies of the gene must have a damaging mutation, in order to produce the mutation's phenotype (i.e. disease symptoms). An individual with one mutated copy plus one normal copy would simply have a normal phenotype (i.e. no disease symptoms).
- *dominant mutation*: A mutation that can cause its phenotype even if it is present on only one of the two copies is said to be dominant.

Double vs. Single Stranded Structures

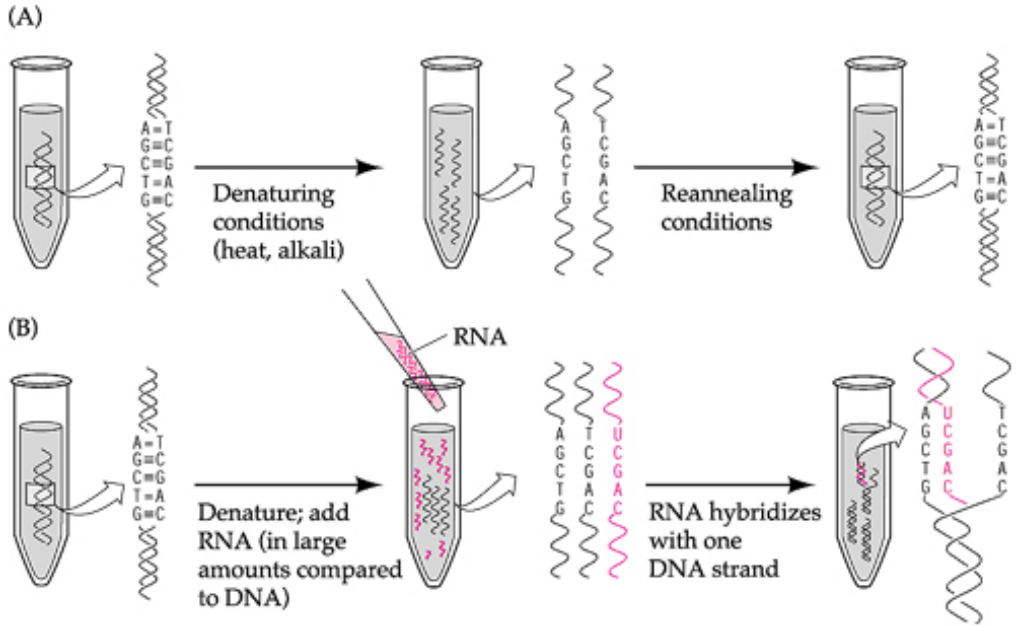


- *base pairing*: the physical attraction and binding of two DNA chain segments that are reverse-complements of each other. This can occur between two separate DNA molecules, or (due to the physical flexibility of the DNA chain) between two separate segments of a single DNA molecule.
- *RNA*: a variant of DNA whose chemical structure is just slightly different. Whereas DNA is the "permanent storage" of the genetic code, RNA usually operates as an "active form" of the genetic code e.g. for producing proteins coded by the gene sequence.

Complementarity of the Double Helix

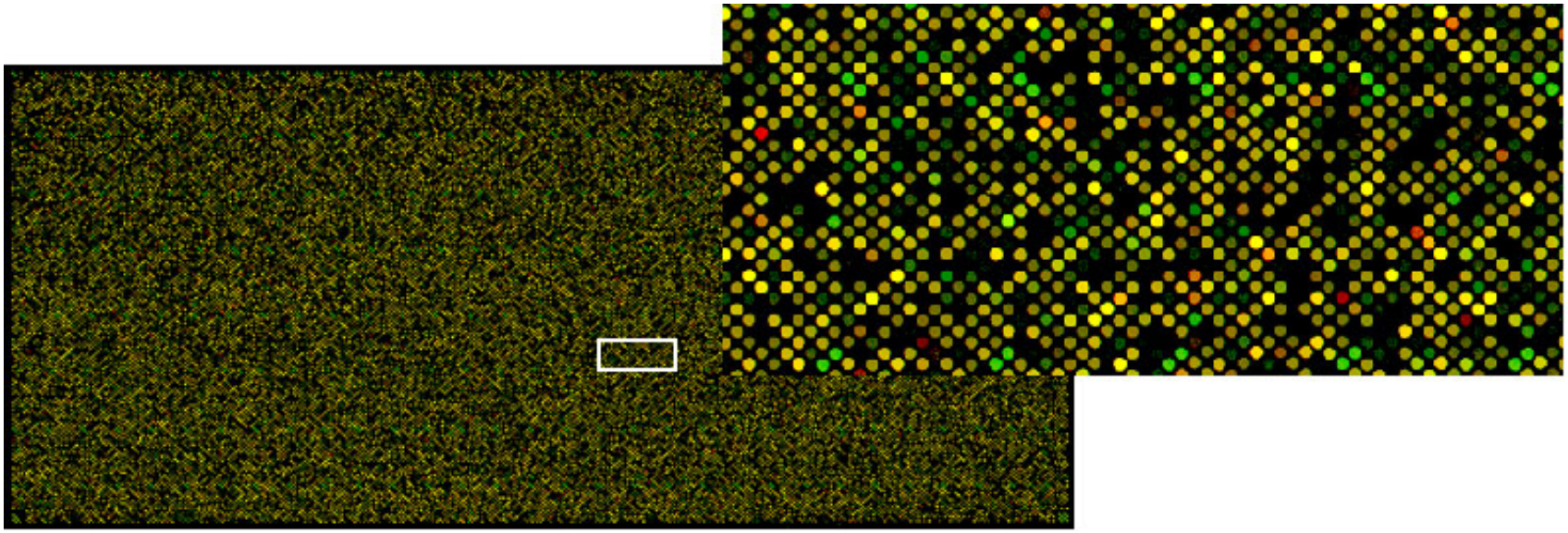
- *reverse complement*: the specific DNA string that will naturally base-pair with a given DNA sequence: it is formed by the following rule: A pairs with T; G pairs with C; and the paired DNA strings run in opposite directions, i.e. the reverse complement of ATTGC is GCAAT, with the A in the first string base-paired with the T in the second string, and the C in the first string base-paired with the G in the second string.
- *5' end, 3' end*: A DNA chain has inherent asymmetry, and its two different ends are referred to as 5' (pronounced "five-prime") vs. 3'. By convention, DNA sequences are always written from 5' to 3'.

The Key to Biotechnology: DNA Hybridization



- *hybridization*: The process of single-stranded nucleotide sequence encountering a reverse-complement nucleotide sequence, and binding to it via base pairing (resulting in a double-stranded complex). Note that even in a mixture of a huge number of different sequences, those that are reverse-complementary will quickly find and bind to each other (i.e. "hybridize").

Example: DNA Microarrays

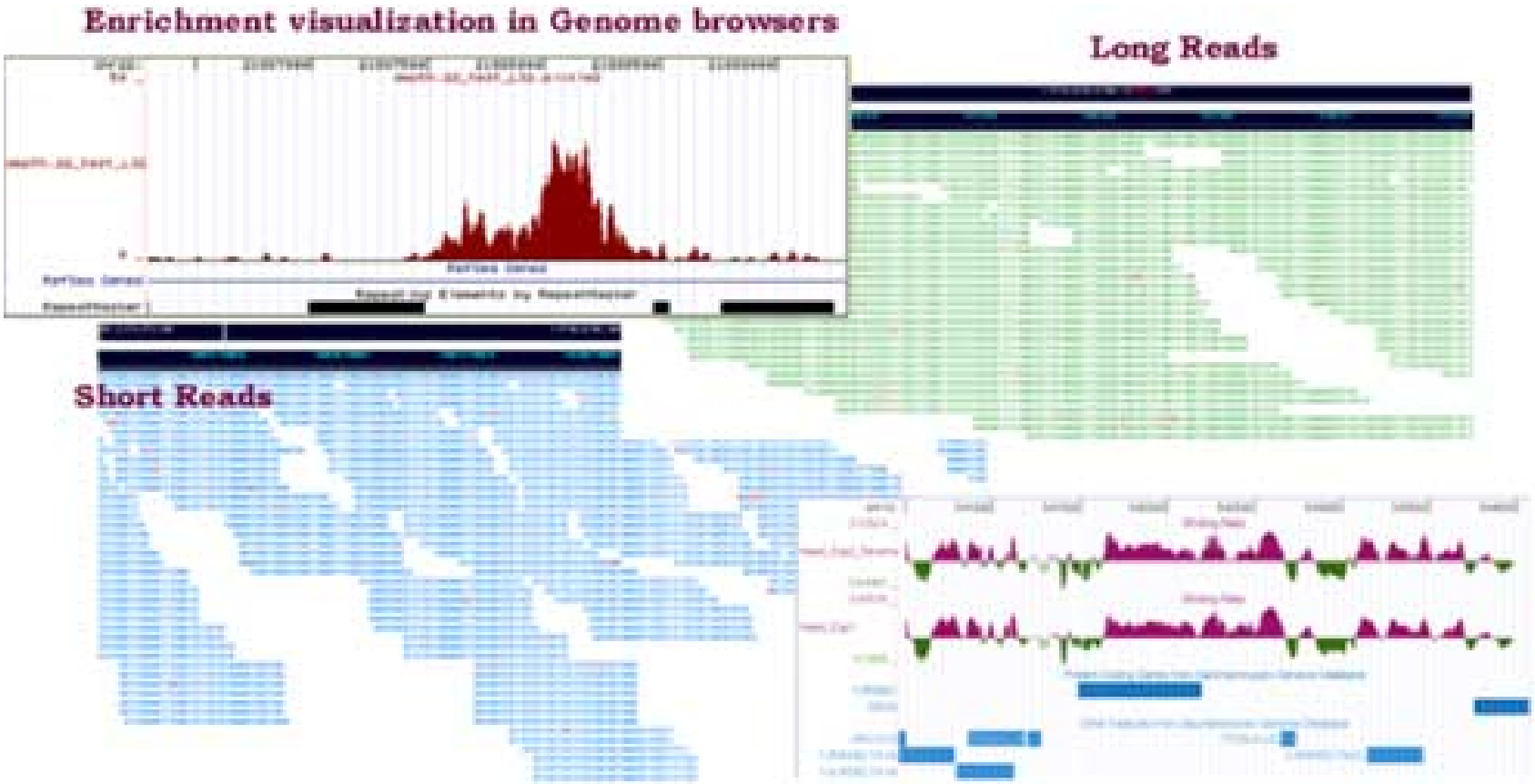


- *DNA microarray*: To measure the individual amounts of many specific sequences in a sample, a microscopic dot of a specified reverse-complement ("probe") sequence is attached to the surface of a glass slide; millions of such dots are arrayed microscopically. A fluorescently labeled DNA sample is then allowed to hybridize to the array; each dot will only bind its specific target sequence. A laser scans the array to measure the fluorescence bound to each dot.

The Key to Genomics Today: Short Read Sequencing

- *short read sequencing*: a recent technology for rapid sequencing of entire genomes. Current systems can read up to 500 million short sequence fragments each approximately 150 nucleotides in length, in one machine run. Due to the intrinsic error rates in the sequencing process, many fragment reads must be compared to arrive at a reliable "base call" at each letter position in the genome. Experimentally, DNA is randomly fragmented, and the resulting "short read" sequence strings must be aligned to each other (computationally, based on matching strings from overlapping reads) as the starting point for all analysis.

Short Read Sequencing



- *sequencing coverage*: the average number of independent reads covering any position in the genome (or region) being sequenced. Typically, sequencing projects aim for coverage of 70x or higher.

Detecting rare SNPs by short read sequencing

Say we are looking for a rare Single Nucleotide Polymorphism (found in 1% of people). We collect DNA samples from many individual people, and generate short read sequencing data separately for each person. Now assume that short read sequencing has a 1% random error rate, i.e. 1 in 100 bases are randomly mis-called. Will it be possible to distinguish a real case where an individual has the rare SNP, from cases of sequencing error? Assume that each person's short read dataset is properly aligned and has approximately 50x coverage (i.e. each nucleotide in the genome is covered by 50 reads on average). What would you expect to see in the aligned short read data for a real SNP? For a sequencing error?

Detecting rare SNPs by short read sequencing Answer

- Humans have two copies of each chromosome; a rare SNP would be expected to be found on one of the two chromosomes in an individual person. Therefore about half the reads covering that genome position (i.e. about 25 reads) should show the SNP, vs. about half the "consensus" base.
- By contrast, a sequencing error should occur in about 1% of the reads covering that position (i.e. at most 1 or 2 reads).
- So real SNP observations will be easily distinguishable from sequencing errors.

Detecting rare SNPs by short read sequencing Errors

- Some people made the (relatively unimportant) error of forgetting that people have two copies of each chromosome, and thus expected that the SNP would be seen in 100% of reads instead of 50%.
- Some people didn't understand the meaning of "50x coverage". Suggestion: review the description of short read sequencing coverage in the lecture and glossary.
- Some people thought a sequencing error will never occur at 50x coverage since $Np = 50(0.01) < 1$. But Np is just the *mean*. By the binomial, the probability of getting exactly one miscall out of 50 trials is $p(n = 1|N = 50, p = 0.01) = \binom{50}{1} p^1 (1-p)^{49} \approx 0.31$.

SNP Detection via Microarray

- whereas short-read sequencing is the best tool for *discovering* new SNPs, DNA microarrays are a cheap way of detecting all *known* SNPs in a sample.
- A set of common SNPs is collated from existing databases.
- Probe sequences are designed to detect the major vs. minor alleles of each SNP, and a microarray manufactured with the full set of probes.
- Fluorescently labeled DNA sample hybridized to the microarray, then scanned: fluourescence signal measures amount of each allele of each SNP.
- Minor allele will be either 0%, 50% or 100% of the signal.
- Anybody can order the microarray and analyze any DNA sample: genetic counseling for a patient; disease mapping ("Genome Wide Association Study"); paternity testing; forensics...

Mendelian Inheritance

- Say we have a "normal" gene (+), and a mutation that knocks out that gene's function (-).
- A person with one normal copy and one mutant copy (+-) will show no symptoms (the mutation is "recessive").
- If two such people have a child, the child gets one copy of the gene (chosen at random) from each parent, so the probabilities are 25% ++, 50% +-, 25% --.
- Exactly the same as flipping two coins...
- Only if the child is -- will he show symptoms.

Sex Chromosomes & the Disease of Kings

- Women have two copies of the X chromosome (XX), whereas men have one copy of X and one copy of the Y chromosome (XY).
- Each child randomly receives one of the two chromosomes from each parent, i.e. one X from mom and either X or Y from dad, giving either XX (girl) or XY (boy).
- So if mom is X^+X^-, her son has 50% chance of being X^-Y
- Disease example: hemophilia (blood fails to clot) is a recessive disease mutation on the X chromosome, so it is far more common (actual disease symptoms) among men than women. And historically it was associated with European royal families (due to inbreeding)!