1 Introduction to Genomics

``That the fundamental aspects of heredity should have turned out to be so extraordinarily simple supports us in the hope that nature may, after all, be entirely approachable. … This is encouraging, for if the world in which we live were as complicated as some of our friends would have us believe, we might well despair that biology could ever become an exact science.`` Thomas Morgan, *The Physical Basis of Heredity (1919)*

1.1 Get Ready for Final Exam

I tried to emphasize some important aspects through the inserts labeled **Get Ready for Final Exam**. At least 70% of the questions that will be included in my part of the final exam will come from these **Get Ready for Final Exam** questions listed throughout the course notes. The vast majority of the other questions will come from topics discussed in class such as ongoing research projects in regional companies and research laboratories. All questions to be included in my part of the final exam will be short answer questions. Multiple choice questions are to be expected only for the online quizzes.

I should add that class attendance is probably the best way to get prepared and confident for the final exam.

Historical perspectives are particularly important for this introductory lecture dedicated to genomics as the contribution of some scientists who founded this field of research, genomics, will be reviewed.

2 On the Concept of Genomics

There is no all-encompassing definition for genomics and the word is often used with many meanings. The suffix "-ome" comes from the Greek for *all*, *every*, or *complete*. It was originally used in "genome," which refers to all the genes in a person or other organism. The suffix "-ics" refers to a field of study. Genomics therefore means the study of all genes within one or several organisms of interest.

McKusick and Ruddle first used the term Genomics to name their brand new scientific journal created in 1987 (see Genomics). At that time McKusick and Ruddle understood genomics to be **the mapping and sequencing of DNA to analyze the structure and organization of genomes**. Note that the Genomics journal was founded only three years after the invention of automated DNA sequencers. Also, the Genomics Option of the Biopharmaceutical Science Program was created in 1989, which is two years after the first mention of the concept of genomics – Dr. Durst behaved like a visionary leader when he put in place the BPS Program. The Genomics journal is still published and its current scope of investigation ``focuses on **the development and application of cutting-edge methods, addressing fundamental questions with potential interest to a wide audience.**`` Articles published in Genomics cover topics such as genome sequencing, sequencing technologies, functional genomics, evolutionary and comparative genomics (phylogenomics), bioinformatics, epigenomics, and medical genomics (don't worry if you don't understand the meaning of these fancy terms as we will have opportunities to talk about these in upcoming sections). Note that, over the years, the research in genomics has shifted from fundamental structural and functional organisation of the genomes towards a broader perspective including applications and technological advances.

The term **omics** has been recently created to describe complete datasets of biomolecules originating from one organism such as **metabolomics** (the whole set of metabolites), **proteomics** (whole set of proteins), and **transcriptomics** (whole set of transcription products or RNAs). In this context, genomics can be considered as the first omics discipline and most others fit in what can be referred as the post-genomic era. The human genome was first characterized and this opened the door to further characterization of the transcription products (transcriptomics), translation products (proteomics) and metabolites (metabolomics). Just remember that genomics is commonly understood as a broad concept encompassing classical genomics as well as post-genomic disciplines.

BPS2110 is an introductory course and you should not expect an in depth coverage of complex principles, but rather an exposure to some basic concepts that will help you understanding the original contribution of some of the modern disciplines related to genomics. Some of the research opportunities in genomics will also be discussed. The main learning outcome is to allow you to make an informed decision whenever you will be required to choose an option for your 3rd and 4th year of studies, if you have to do so.

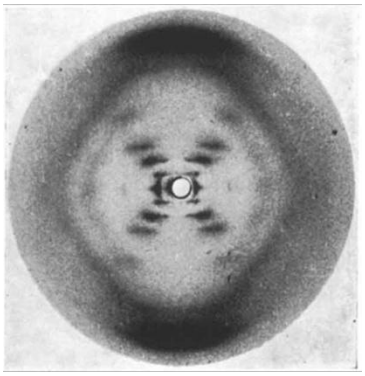
Genomics is still a relatively recent discipline - it was created in 1987 – and it is hoped that the few historical anecdotes included throughout the notes will help you appreciate the impressive progress that was experienced over the last 25 years.

The specific topics to be covered in the Genomics section of BPS2110 are:

- Genomics and the Human Genome Sequencing Project
- DNA Sequencing Technologies
- Bioinformatics
- The ENCODE Project
- Biologics
- Personalised medicine

2.1 Watson and Crick: Double Helix Model of DNA and the Underlying Principle of Gene Expression and Heredity

It`s been 60 years since Watson and Crick published their report in Nature proposing the 3D structural arrangement of DNA, that is the double stranded helix (Nature, 1953). Rosalind Franklin also contributed to the elucidation of the DNA structure, though her contribution is often overlooked. Francis Crick indeed acknowledged that Franklin's images were "the data we actually used" to formulate their 1953 hypothesis regarding the structure of DNA. The most significant of those images is shown below and the helical structure of DNA was deduced from it (Wikipedia, Rosalind Franklin). Crick and Watson had worked on a structural model with the phosphate backbone located on the inside of the helix and Franklin apparently told them that the phosphate groups are located toward the outside. It is interesting to note that Watson and Crick's proposal is more or less a hypothetical structural model of the DNA helix –something you do yourself in organic chemistry with your ball and stick model–, though Franklin was a crystallographer who generated and analyzed experimental data supporting the helical structure of DNA.



This is the first X-ray diffraction image of a DNA sample. The analysis of this diffraction pattern, especially the X, is considered as the first experimental evidence of the helical structure of DNA. Photo from Rosalind Franklin: The Dark lady of DNA (2002)

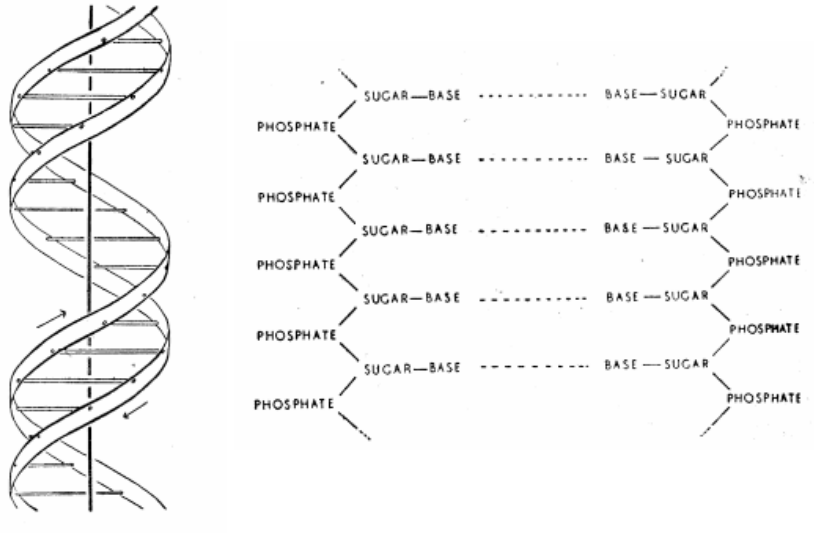


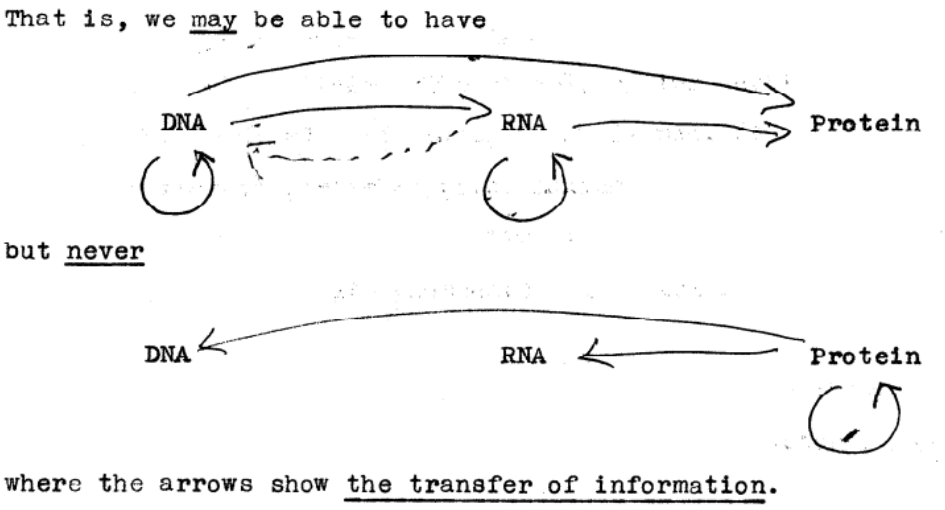
Figure from Nature 1953

The discovery of the double helix of DNA yielded ground-breaking insights into the genetic code and protein synthesis. One major, if not the most important, contribution of Watson and Crick is to have proposed a molecular model for heredity: DNA encodes for RNA which encodes translation or protein synthesis. Watson and Crick also predicted a model for the replication of DNA: the two strands of DNA can be pulled apart then each strand serves as template for the production of its complementary strand, a process referred to as semiconservative replication.

Before the discovery of the double helix, the term genetic code had no meaning; afterwards, deciphering the code (putting together the dictionary by which the three-letter nucleic acid language is translated into the twenty-letter protein language) became the most urgent and ambitious undertaking of biologists throughout the world, an effort that defined the classical age of molecular biology. Francis Crick was the intellectual leader in this effort, distributing, critiquing, and connecting experimental results from many sources, mediating between scientists of differing opinions, and proposing new experiments and lines of investigation. This opened up new avenues of research into the genetic control of essential biological processes, most importantly the synthesis of proteins. Watson and Crick were the first to realize that the seemingly random sequence of the four bases in DNA form a code which specifies the order of the twenty amino acids that make up most proteins.

"Crick summarized his ideas about the genetic code in a paper entitled *On Protein Synthesis* and presented it at University College London in September 1957. The paper … permanently altered the logic of biology. In it Crick proposed the sequence hypothesis, which held that genetic information was encoded in the sequence of the bases in DNA. The base sequence was to be read in linear fashion, from a fixed starting point and in one direction.

Secondly, Crick propounded the Central Dogma of molecular biology, which stated that genetic information which had been transcribed from DNA to messenger RNA and used to build a protein, could not again flow in the reverse direction from protein to RNA. The Central Dogma thus implied that acquired changes in a protein could not be inherited, an implication that conformed to Charles Darwin's theory of evolution. The Central Dogma further implied that DNA contained all the information necessary for specifying the sequence of amino acids in a protein, and thus its shape and function; no external information was needed. Finally, Crick asserted that the genetic code was universal to all higher forms of life, as in fact it has proved to be. " (Profiles in Science, National Library of Medicine).



Draft of a presentation by Crick in 1956. The same model was restated in 1970 (Nature 1970).

Get ready for exams

What are the two major scientific contributions of Watson and Crick?

What is the scientific contribution of Franklin toward the elucidation of the 3D structure of the DNA helix? How does it complement Watson and Crick`s discovery?

The flow of information from DNA to RNA to proteins is said to be sequential and unidirectional. What do sequential and unidirectional specifically mean and entail?

Use a diagram to illustrate and explain the process of semiconservative replication of DNA. Who first proposed this process and when?

What is exactly the Central Dogma of molecular biology? Who proposed it and when?

If you have time, look at the video created by Nature and displaying a modern version of the Central Dogma based on current molecular biology knowledge. It seems like a Star Trek movie and the music is quite futuristic too.

2.2 Historical and Technological Perspectives of the Human Genome Project

"The Human Genome Project (HGP) refers to the international 13-year effort, formally begun in October 1990 and completed in 2003, to discover all the estimated 20,000-25,000 human genes and make them accessible for further biological study" (HGP Information).

2.2.1 Phase I: Conceptualization

2.2.1.1 Molecular Nature of Heredity and Planning of the Human Genome Project

In the early 70`s the double-stranded DNA was formally recognized as responsible for the transmission of hereditary traits among generations. The elucidation of the molecular details of heredity fostered research in "DNA technology" to permit the characterisation and, eventually, manipulation of DNA fragments (see movie DNA Dreams).

The goal was to unravel the mysteries of our genome, but it was impossible to anticipate the Human Genome Project until the necessary molecular tools could be developed. It is the development of suitable DNA technologies that finally convinced the members of the US Congress to adopt a dedicated research budget of $200 million annually for 15 years to determine the sequence of the 3 billion nucleotides of the human genome and to map and identify all human genes. The purpose of the HGP was not only to get a sequential series of nucleotides – A's, C's, G's and T's - but to also position genes onto chromosomes. Genome

mapping is the creation of a genetic map listing the locations of different genes and other genetic elements or distinctive features.

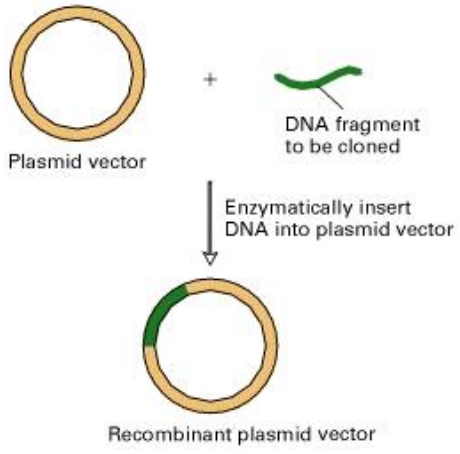
"By 1985, progress in genetic and DNA technologies led to serious discussions in the scientific community about initiating a major project to analyze the structure of the human genome. After concluding that a DNA sequence would offer the most useful approach for detecting inherited mutations, the US Department of Energy announced in 1986 its Human Genome Initiative. The project emphasized development of resources and technologies for genome mapping, sequencing, computation, and infrastructure support that would culminate in a complete sequence of the human genome" (HGP Report).

2.2.1.2 Cloning Vectors and Cloning Libraries

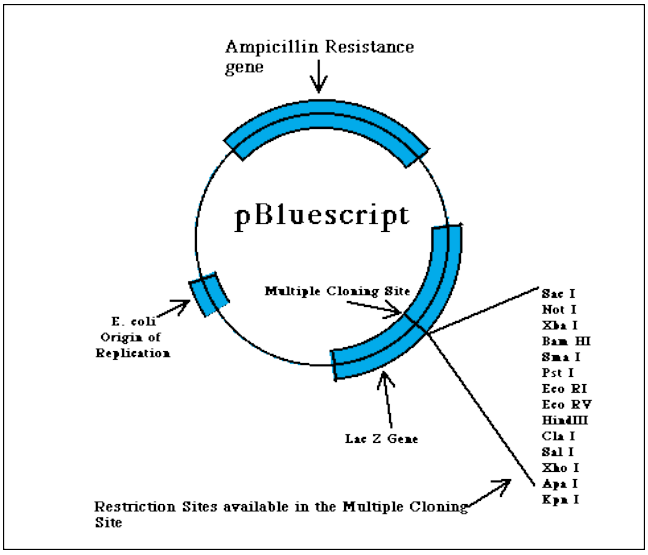
Some of the key DNA technologies whose development made it possible to anticipate a plan for sequencing the human genome are listed below. The molecular principles behind these DNA technologies are important as well as their respective contribution to the feasibility of the Human Genome Project.

DNA cloning refers to the ability of manipulating and amplifying by several times a given DNA fragment, usually in the range of a very few thousand base pairs or kbps. **Cloning vectors** are the molecular "vehicles" used to carry a DNA fragment of interest. Without proper vectors DNA fragments would be difficult to manipulate. Most, if not all, cloning vectors share basic features that facilitate the manipulation of DNA. These features are:

- Can be digested or "opened" at specific positions to facilitate the insertion or release of DNA fragments. A vector that contains an exogenous DNA fragment, the DNA insert to be cloned, is referred as a **recombinant vector**. There are several vectors and each one is suited to integrate a given range of DNA size into a specific host of category of hosts. The plasmid pBluescript, for instance, can effectively integrate cell fragments in the range of 100bp to 5000bp into bacterial cells such as *Escherichia coli*.



- An **origin or replication** that will permit the self-replication of the cloning vector along with its inserted DNA fragment, if present. The origin or replication is a genetic element recognized by replication enzymes within the genetically transformed host cell – a cell with a recombinant plasmid - to initiate the copying of the recombinant plasmid. The tern **copy number** refers to the number of identical plasmids that are found within a transformed host cell. pBluescript, for instance, can typically produce up to 1000 copies per host cell. High copy numbers are essential to favor effective amplification of a given DNA fragment.
- A marker of resistance to avoid contamination of transformed cells with non transformed cells. The **selection marker** found in pBluescript is the gene that codes for the resistance to ampicillin. A selection marker is essential as the transformation of a host cell, that is the integration of a plasmid through its cell membrane, is a process that is poorly effective. For a typical transformation protocol, one should expect to obtain a couple of transformed cells out of an initial pool of 1000 cells. The presence of a selection marker makes it possible to selectively eliminate the vast majority of cells that failed to be transformed and to permit the very few transformed cells to grow due to the presence of the recombinant plasmid containing the gene coding for resistance to ampicillin.
- The **multiple cloning site** (MCS) is the part of a vector where DNA fragments can be inserted. A MCS is characterized by the presence of several recognition sites that can be recognized and cut by restrictions enzymes. A restriction enzyme recognizes a short and unique DNA sequence of 4 to 6 bp long and specifically cuts or “digests” DNA at this unique recognition site. Restriction enzymes are used to cut within the MCS and permit the insertion of a DNA fragment to be cloned.
- The overall structural stability of a cloning vector facilitates its manipulation (digest, purification, long term storage, …).



Get ready for final exam

Use a diagram to identify the important components of a plasmid for effective cloning of a DNA fragment of interest.

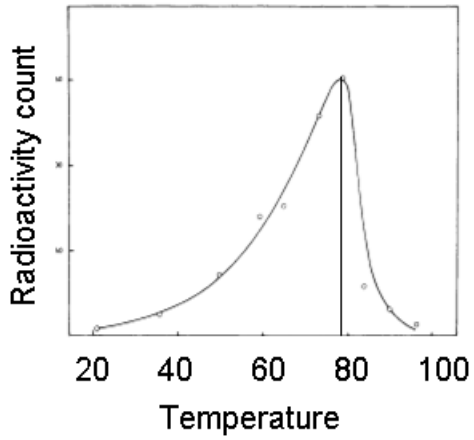
Explain the likely difficulty that someone should expect when cloning a gene with a vector lacking (A) a selection marker and (B) an origin of replication.

Provide a short definition of the following terms : copy number, origin of replication, selection marker.

2.2.1.3 DNA Polymerase as the Tool for In Vitro DNA Synthesis

The first DNA polymerase was discovered by Arthur Kornberg in 1956, just 3 years after Watson and Crick had proposed their double helix model (see discussion in Nature, 2006). This finding was truly remarkable as many researchers, at that time, believed that it was impossible to duplicate DNA replication outside the cell. Many researchers remained unconvinced that DNA was indeed the hereditary material in cells. In 1959, Kornberg was awarded the Nobel Prize in Physiology or Medicine for this work.

Another major step in DNA polymerase research was achieved by Chien who discovered, in 1976, an enzyme that achieves DNA synthesis at high temperatures. The enzyme, which had optimal activity at 80°C, was purified from *Thermus aquaticus* which can be found within hot springs *(*J of Bacteriology, 1976). Further research sponsored by Stratagene led to the characterisation of a DNA polymerase from *Pyroccocus furiosus,* which is a hyperthermophilic bacterium growing optimally at 100°C (Gene, 1991). We'll see in next lecture that the ability of DNA polymerases to catalyse DNA synthesis at high temperatures is essential for the PCR reaction that is used in Sanger method of DNA sequencing.



Effect of temperature on the activity of the DNA polymerase from *Thermus aquaticus*. Reactions were incubated for 30 min at different temperatures. Fractions were then taken, and the conversion of [^{3}H]dTTP into acid-insoluble material (DNA) was measured. Figure at the left adapted from J of Bacteriology, 1976). Figure at the right shows a thermal spring such as the one in which *T. aquaticus* lives.

Get ready for exams

How would you design an experimental protocol to measure the activity of a DNA polymerase enzyme at different temperatures? How could you specifically quantitatively measure incorporation of nucleotides? Please refer to the figure above (the figure and caption would be supplied if this question were incorporated within the final exam).

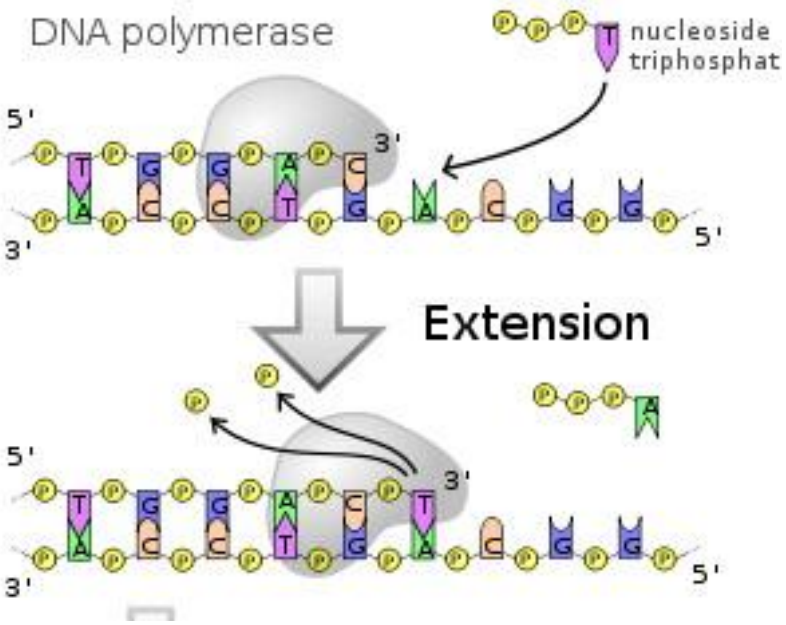
Never stop thinking

How can living organisms survive in boiling water? What kind of biochemical adaptations could you predict within hyperthermophilic cells? How can the integrity of the cell membrane be preserved at 100°C? Are there some specific commercial applications you can imagine for enzymes that are stable and active at very high temperatures?

In its first steps, DNA polymerase research was quite competitive as researchers knew the importance of that enzyme to duplicate and eventually amplify DNA fragments through PCR. The dream was therefore to find, and later engineer, more effective DNA polymerases that could facilitate the synthesis and manipulation or study of DNA, the molecule coding for heredity.

Most DNA polymerases have unique features that make possible the synthesis of DNA. These are :

- DNA synthesis requires a **single-stranded template** to guide the synthesis of a new complementary DNA strand.
- DNA elongation should be **primed**; DNA polymerase can only incorporate new nucleotides at the free 3' end of a primer or an elongating fragment. The double helix is antiparallel with its two strands oriented in opposite direction (by convention a DNA fragment is read from its 5' end towards its 3' end). DNA elongation mediated by DNA polymerase occurs in the opposite direction to the template strand.
- Additional essential reagents necessary for proper activity of DNA polymerases are Mg++ and a mix of all four deoxynucleotide triphosphates (dATP, dCTP, dGTP, and dTTP).



Get ready for exams

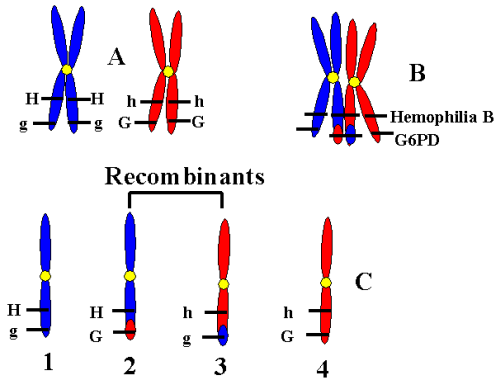
DNA polymerase is a fundamental enzyme with several applications in genomics research. Make sure that you know very well how DNA extension occurs (incorporation of free nucleotide triphosphates, release of two phosphates, sense of elongation, template reading). Yes, the final exam will contain at least one short answer question on DNA polymerase. For instance, could you design a primer to be used to synthesize the complementary strand to the sequence below? A primer being a short oligonucleotide of 10-20 nucleotides that anneals onto its complementary sequence to prime DNA synthesis. Don`t forget to indicate the 5`and 3`ends!

5`ACTGACCTTACTGCATGAATTCCGGAATATGCGCATACGTACGTTAACCGTATT3`

2.2.2 Phase II: Chromosomal Mapping

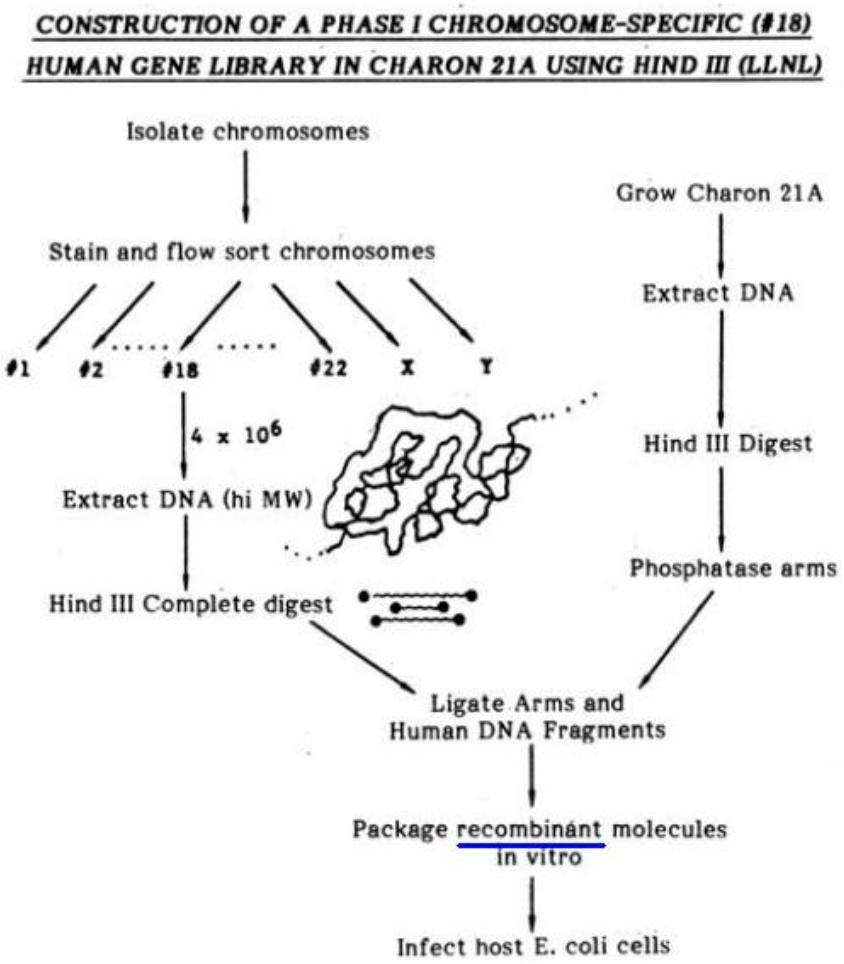
The human genome contains approximately $3x10^9$ base pairs (bps) of nucleotides that are divided into 23 pairs of chromosomes. The average size of a chromosome is therefore $125X10^6$ bps. At the time the HGP was initiated there was no cloning vector able to integrate DNA fragments as large as human individual chromosomes, though this is not the case anymore (see PNAS, 1992). Scientists were therefore aiming at initially mapping chromosomes with reference marks or tags that may allow them to locate DNA sequences onto chromosomes and proceed with the HGP in an orderly fashion. This orderly sequencing approach was considered as more effective at that time, but we`ll see later that it`s not true anymore – random shotgun sequencing is currently used. At the beginning of the HGP, it was assumed that a reference genome map would facilitate the overlapping and fitting together of short, sequenced fragments, ``… in the same way that having a picture of a finished jigsaw puzzle helps in its assembly even if pieces are missing`` (Craig Venter, A Life Decoded). But by the time the HGP was initiated DNA landmarks were relatively few and unspecific.

Genetic linkage mapping: During meiosis recombination between homologous chromosomes can occur. The farther apart the genes are from a known locus and the less likely that two traits will co-segregate. By investigating the rates of co-segregation between different traits geneticists have been able to establish a reference genetic linkage map. This strategy can only provide relative positioning of a gene of interest along its host chromosome. The first chromosome mapped this way dates to pioneering work on the fruit fly in the early 1900s by Thomas Morgan (see the quotation at the beginning of this section). Another form of genetic cartography is **physical mapping** based on finding the physical location of a given gene, that is its position onto a given chromosome (see below). Figure from http://www.uic.edu/classes/bms/bms655/lesson12.html .



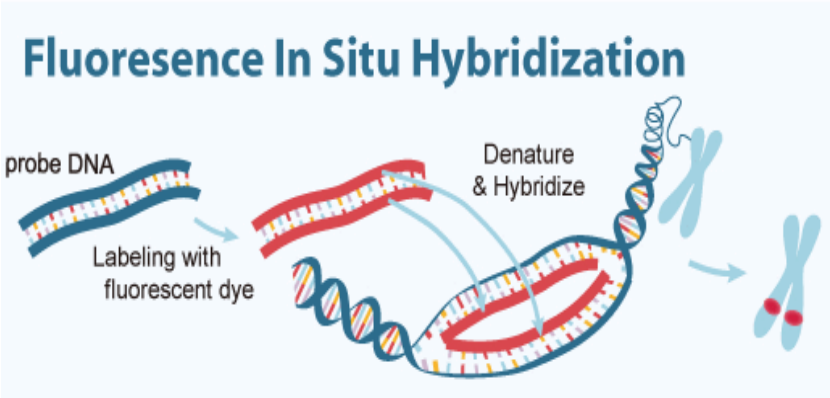
An early step towards the HGP was the characterization of **chromosome physical maps**. Vectors such as pBluescript allowed the cloning and sequencing of fragments up to 2000-3000bps. It would have been challenging, if not impossible, to identify the specific originating chromosome of a short DNA fragment without an access to a reference map of the different chromosomes. “Critical to this effort were the libraries of individual human chromosomes

produced at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL). These libraries allowed the huge task of mapping and sequencing the entire 3 billion bases in the human genome to be broken down into 24 much smaller single-chromosome units" (HGP Report).



Protocol used to generate a library of recombinant DNA vectors with different fragments of the human chromosome #18.

The appropriate technology for cloning entire human chromosomes was made available by the early 90's (see PNAS, 1992). The **Bacterial Artificial Chromosome** (BAC) system facilitates the high-resolution of physical map of each of the human chromosomes. BAC chromosomal clones in combination with Fluorescent In Situ Hybridization (see figure below) make it possible to directly visualize the specific position of a gene onto its host chromosome. BAC chromosomes were superior to **Yeast Artificial Chromosomes** (YACs) as these later tend to be unstable due to recombination when incorporated within yeast cells – the repository reference chromosomes must remain stable over time if one needs to determine their DNA sequence.



"*The FISH technique is dependent upon hybridizing a probe with a fluorescent tag, complementary in sequence, to a short section of DNA on a target gene. The tag and probe are applied to a sample of interest under conditions that allow for the probe to attach itself to the complementary sequence in the specimen if it is present. After the specimen has been treated, excess fluorophore is washed away and the sample can be visualized under a fluorescent microscope. By quantifying the amount of fluorescence with the scope it can be determined if the type of cell the probe was designed for is present, and if so, how much of it is present in a sample.*" (Carleton University) *The picture at the right displays the in situ localization of a given clone on two distinct chromosomes – the clone is duplicated, at least in part* (ChromBIOS). *Access to a gallery of results obtained by FISH and to a video illustrating the step-by-step FISH protocol.*

.

Get ready for final exam

In the HGP, what was the purpose for mapping chromosomes prior to initiate sequencing of DNA?

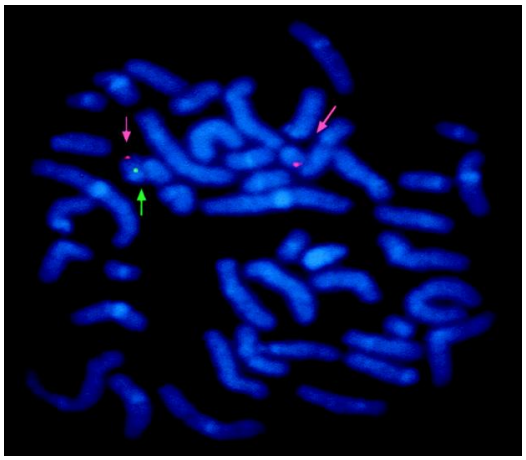
Use a diagram to explain the FISH protocol. What`s the main purpose for using the FISH protocol in genomics research?

In 2013, is it possible to directly visualize the chromosomal position of a gene of interest? If yes, explain how. (If you say no, you have the wrong answer and you should try again!)

Name three types of cloning vectors and, for each one, identify the size of DNA fragments that can be inserted and cloned.

Application of FISH in modern genomic research

FISH can be used to confirm the deletion of DNA fragments from a target chromosome. This approach is commonly used to confirm the successful targeted deletion of a gene or chromosomal fragment or to diagnose genetic diseases characterized by a partial chromosomal deletion. A metaphase spread, which is the display of all chromosomes at the mitotic stage of a cell during which chromosomes are condensed and easily visible, from a patient with 22q11.2 deletion syndrome. FISH mapping was done with a positive control probe to 22q13.3 (red arrow; found on each of the two copies of chromosomes 9) and a plasmid including *GPIb-β* to 22q11.2 (green arrow; found on only one of the two copies). On the normal chromosome, a green signal (green arrow) can be seen at 22q11.2 and a red signal (red arrow) can be seen at 22q13.3. On the deleted chromatic chromosome, only a red signal (red arrow) can be seen at 22q13.3. (Genetics in Medicine, 2003).



2.3.1 Phase III: Sequencing Strategies and Genome Assembly

At the time the HGP was initiated the technology allowed to sequence roughly 500-700 nucleotides per reaction. Chromosomes were broken down into smaller fragments that could be completely sequenced one at a time. Taking into consideration that sequencing could be initiated from either side of a DNA insert, DNA fragments of 1000bp were generated and cloned into bacterial vectors very similar to pBluescript. This systematic clone-by-clone approach aiming at the systematic and orderly coverage of the human genome had been used by the NIH-funded laboratories. “A set of 33,221 BAC clones forming the minimal set covering the maximum region of each chromosome was first selected. Each clone was then broken down into smaller subclones of ≈1000bp each for direct DNA sequencing. The plan was to put the fragments to be sequenced in order, followed by complete sequence determination of each fragment in a systematic manner so that the entire human DNA sequence of each BAC (and, therefore, the corresponding region on the chromosome) was known. This process of building the sequence-ready maps, subclone library construction, and directed gap filling is costly, time consuming, and, therefore, often rate-limiting.” (Tsui and Scherer, The Human Genome Project, Essentials of Genomics and Bioinformatics, 2002)

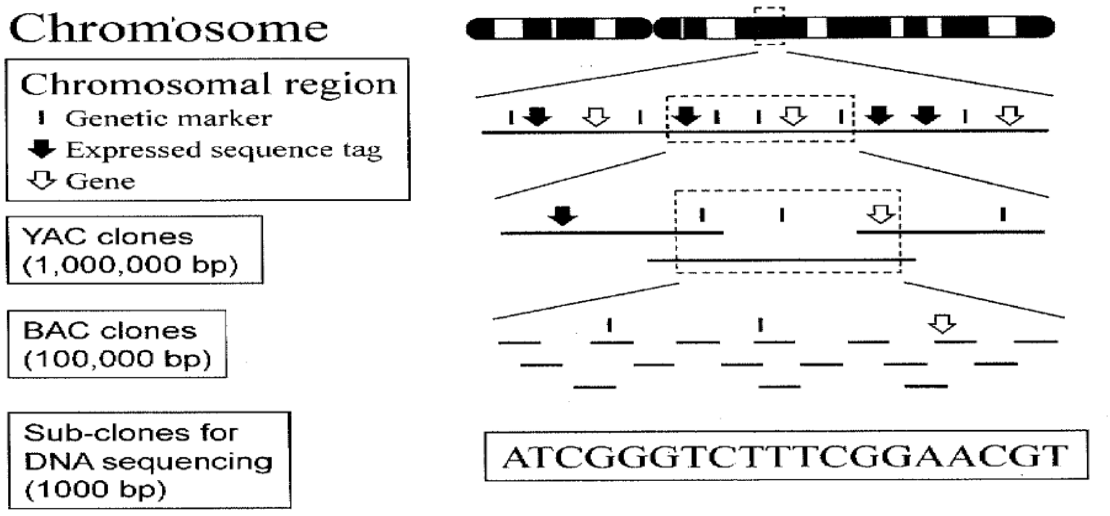
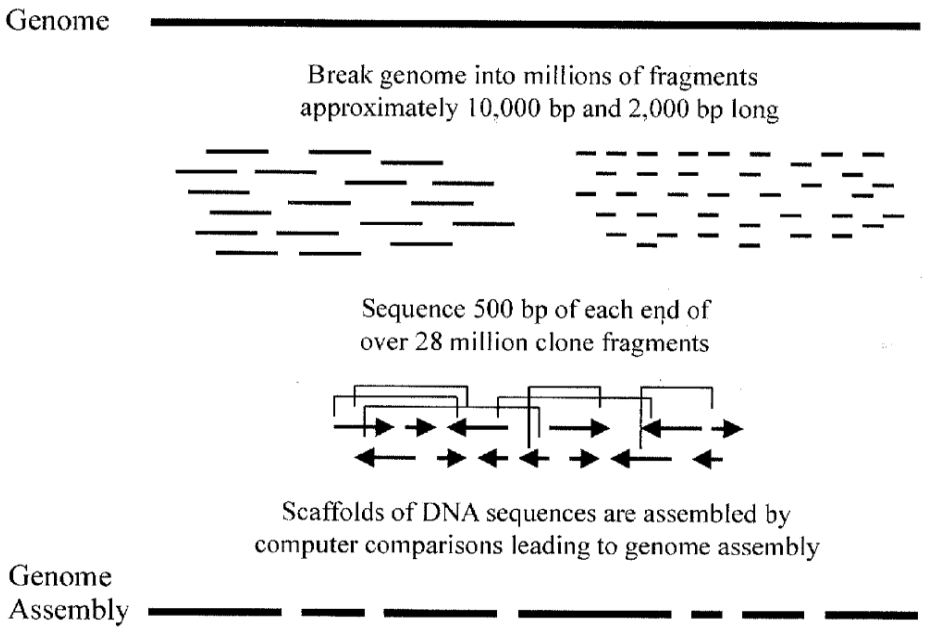


Figure from Tsui and Scherer, The Human Genome Project, a chapter in Essentials of Genomics and Bioinformatics, 2002.

The second sequencing strategy is referred as whole genome shotgun (WGS) and it was first used by Craig Venter, the founder and president of Celera Genomics. `` While this shotgun approach was widely criticized at the time, it has subsequently become a standard method for sequencing complex organisms that is now broadly accepted and routinely used by many of the same scientists who originally scorned the approach.`` (Celera.com) Venter didn`t want to make either a linkage or a physical map a prerequisite of sequencing, as had the NIH-funded laboratories. His reasoning was that ``As anyone who has assembled a jigsaw puzzle knows, you can proceed without knowing what the bigger picture is if you take advantage of edges and other recognizable features to reconstruct the jigsaw from the bottom up.``

All that is required to proceed with WGS is:

- A library of large DNA fragments of the genome to be sequenced, Venter used a library of lamda clones with each one containing a DNA fragment of 25,000bp to 50,000bp.
- A mixture of the DNA to be sequenced that is randomly broken down through sonication – use of ultrasound waves - into easy-to-handle and easy-to-sequence fragments. This fragmented mixture of DNA can then be easily separated by chromatography to collect fragments of given sizes (for the *Haemophilus* genome, fragments of 2,000, 10,000, and 50,000bp were used by Venter and his colleagues to sequence the human genome, though the figure below illustrates the situation for fragments of 2,000 and 10,000 bp).
- Computing power … For sequencing the human genome, Venter and his colleagues negotiated a contract of $50-$100 million with Compaq to build one of the biggest computers ever. An overview of the computating demand for the sequencing of three genome projects is listed in the table below.



Overview of the whole shotgun approach whose distinctive feature is to first assemble short DNA fragments based on the sequences of their two ends. Figure from Tsui and Scherer, The Human Genome Project, a chapter in Essentials of Genomics and Bioinformatics, 2002.

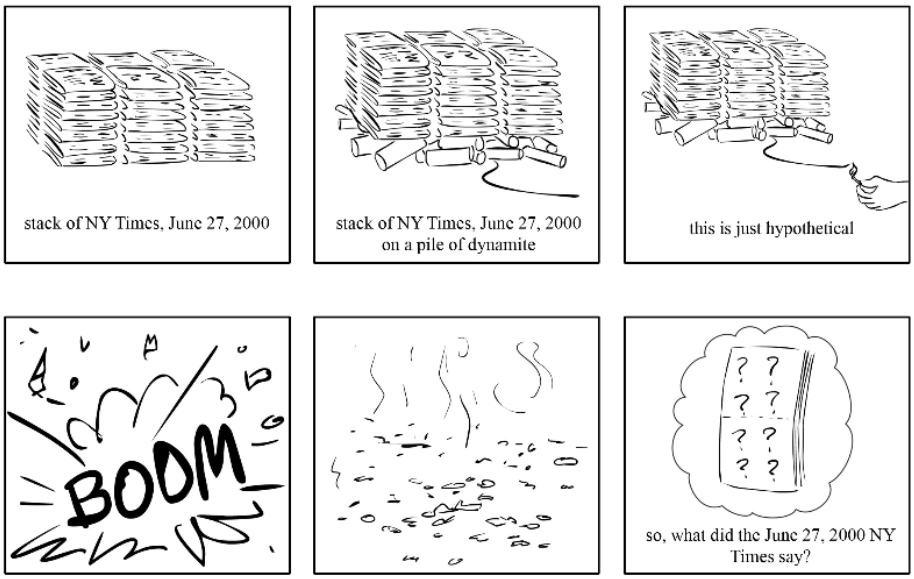
Overview of the number of DNA reads and computing power that was necessary for three genome sequencing projects completed by Celera Genomics (values obtained from A Life Decoded).

Species	Size of the genome	# of reads or DNA sequences generated throughout sequencing project	# of attempted alignments or computing steps
Haemophilis influenza	1.8×10^6 bp (1.8Mbp)	26,000	$\approx 10^9$
Drosophila melanogaster	140×10^6 bp (140 Mbp)	3,200,000	$\approx 10^{13}$
Homo sapiens	3×10^9 bp (3Gbp or 3000Mbp)	26,000,000	$\approx 10^{15}$

"Genome assembly refers to the process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated. A genome assembly algorithm works by taking all the pieces and aligning them to one another, and detecting all places where two of the short

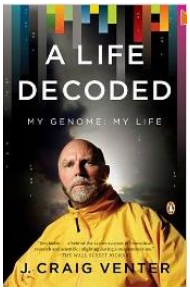
sequences, or *reads*, overlap. These overlapping reads can be merged, and the process continues." (Wikipedia) DNA fragments that are sequenced must overlap, at least partially, to permit alignment of sequences and their assembly into longer stretches of DNA. The coverage value is the amount of coverage of sequences that is covered for complete assembly of the whole DNA fragment of genome to be sequenced. For example, a onefold coverage of a 100,000 bp clone means that 100,00bp of DNA sequence are generated. For the HGP, the NIH-funded approach used a tenfold coverage to complete the human genome. Genome assembly requires significant computating power; for the HGP, roughly $40x10^6$ sequencing results or reads of 500-100 nucleotides each one were generated.

You should read the first five pages of the intriguing article written by Compeau and Pevzner to illustrate, in a simple way, the nature and complexity of genome assembly. Their analogy of the exploding newspapers is something you will likely remember forever! (reference)



Never stop thinking

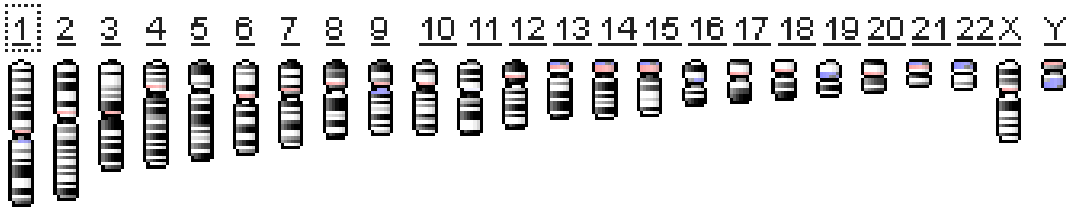
The rise and fall of Celera Genomics as an ambitious competitor of the Human Genome Project is an interesting story that is presented in the book *The Genome War*. Craig Venter, the founding president of Celera Genomics, has indeed an interesting personality; he created the not-profit Craig Venter Institute in 2006. This genomics research institute released in 2007 the first complete genome (both chromosomal copies were sequenced for a total of 46 chromosomes or $6x10^9$bp) of an individual – Venter's personal genome. Venter has further released a book relating his adventure: A Life Decoded: My Genome, My Life.



2.3.2 Phase IV: Identification and Annotation

A complete understanding of the biology and function of the genome is the ultimate goal of the HGP. This is still an ongoing project even if the HGP was considered as officially completed in April 2003 (Science and Nature). Gene identification specifically refers to the delineation of genes along their host chromosomes while annotation refers to the addition of complementary biological information such as sequence variations (Single Nucleotide Polymorphism or SNP, …), important regulatory motifs, and disease mutations.

The amount of genomic information that is currently available is enormous and it has become an overwhelming task to centralize all that information within one single repository centre. This is why several specialized databases have been created with each one covering some specific genomics aspects. We will have an in class discussion about some of the information that can be easily accessed and retrieved from the National Center for Biotechnology Information (NCBI) that houses a series of databases relevant to biotechnology and biomedicine. Major databases accessible from NCBI include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. NCBI is located in Bethesda Maryland since 1988.



Click on chromosome name to open MapViewer

NCBI link to access annotated version of the human genome: http://www.ncbi.nlm.nih.gov/genome/51

We will have other opportunities to talk more about annotation options when covering bioinformatics and the ENCODE Project.

DNA annotation jamboree as an opportunity to combine business with pleasure

The sequencing team at Celera Genomics had promised to release the genome of the fruit fly at a conference held September 17, 1999. After two intense weeks the chief programmer, Arthur Delcher, fixed a bug about one of the 150,000 programming steps so a draft of the fruit fly genome could be completed by September 7. But the team didn`t have the opportunity to analyze and annotate the genome information to figure out its meaning. ``Annotating and describing the fly genome could take well over a year. … In discussions with colleagues, we came up with a novel way to solve the problem, one that would involve the *Drosophila* scientific community, be exciting science, and move things forward rapidly. We decided to hold what we termed an `annotation jamboree`, inviting top scientists from around the world … to analyze the fly genome over the course of a week to ten days. We would then write up our results and publish a series of papers on the genome.`` And the results of this impromptu meeting were beyond expectations: ``The community of thousands of scientists devoted to *Drosophila* research had spent decades hunting down every one of the 2,500 known genes, one study at a time, and now they had all 13,600 [genes] laid out before them on a computer. After eleven days we had found out more than enough to provide an initial analysis of the genome.`` (Venter, A Life Decoded)

.

Get ready for final exam

List the four main phases of the HGP and, for each one, briefly summarize a new technology that had to be implemented to make possible its completion.

Craig Venter designed a new strategy to hasten the completion of the HGP. What are the specific advantages of his sequencing strategy compared to the one that was put in place by the NIH-funded laboratories?