UC San Diego

Lecture 3 - Fundamental Technologies III: Genomics and Transcriptomics

BENG168

Instructor: Adam M. Feist, Assistant Professor, Shu Chien - Gene Lay Department of Bioengineering

In-Class Announcements & Follow Up

- Update to previous lecture slides
- In-class announcements

Update of Pace and Class Module Structure

- The syllabus has been updated based on the class vote to hold the **Final Exam on the last day of class** and to add a Final Exam review lecture the class period before.

Discussion Section Feedback

- Thank you for attending the Discussion sections and for providing feedback! The Discussion sections are the ideal venue for sharing your thoughts.
- A central theme in the feedback was that the current class structure (3 modules per lecture) is content-heavy and that discussing modules is most effective in active learning sessions.
- **Actions Based on Feedback:** Beginning in Week 2, we will shift to **2 modules per lecture and have interactive discussions in the Modules**. It is very important to the instructional team that you understand the fundamental aspects of the course. As a result, later lectures focused on case studies will be integrated into earlier lectures or condensed.
- Please continue to provide feedback to your TAs during Discussion sections **this week** regarding whether you prefer the 2-module or 3-module lecture format

Update of Pace and Class Module Structure

Homework # 1 is posted!

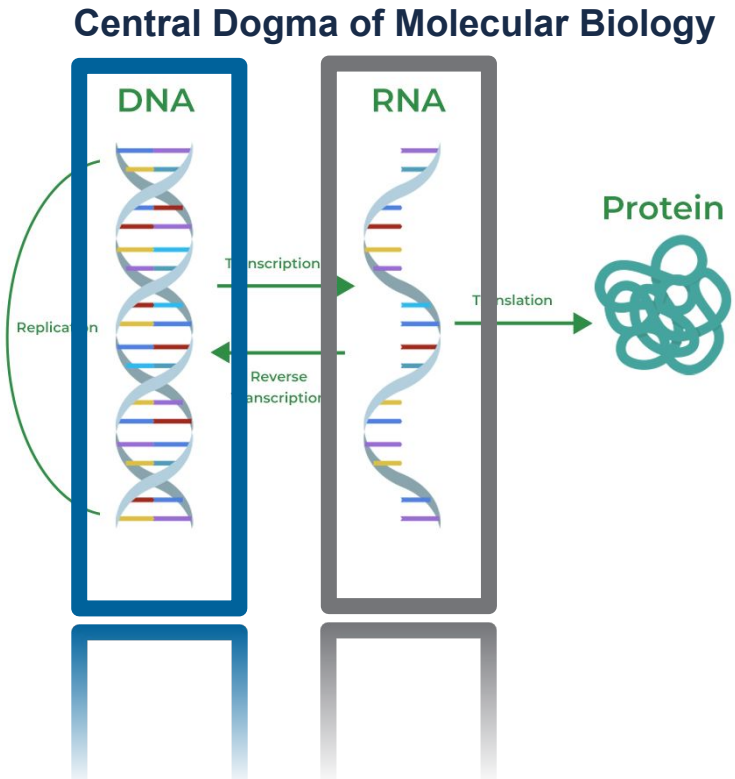
- Due on January 19 on Gradescope.
- Discussion Sections are also good places to go over homework questions.

Module 1: Genomics (Source Pages: Chapter 2: 62–65)

- Mapping structural features of genomes.
- Identifying biological functions using bioinformatics.
- Comparing related genomes for disease.

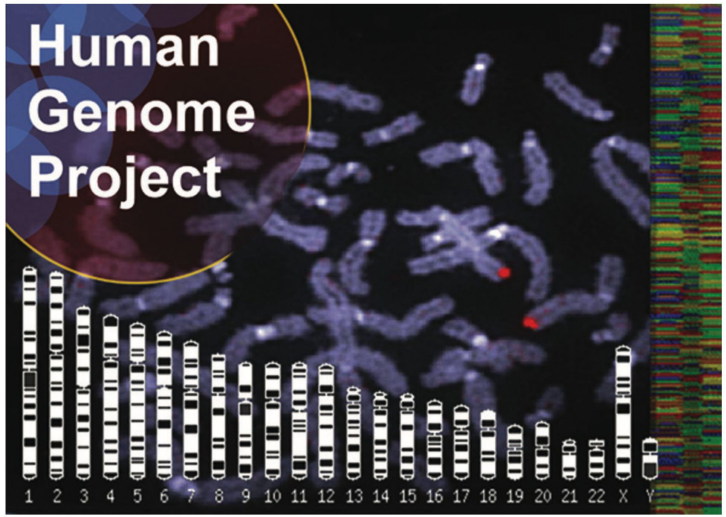
Prerequisite: The Central Dogma

- **Core Concept:** DNA is transcribed into RNA, which is then translated into functional protein machines.
- **Genomics** analyzes the DNA blueprint while **Transcriptomics** measures the total RNA readout.
- **Bioengineering Tool:** Mastering these stages is essential for modern bioengineering.
- These technologies drive **Systems Biology** - research to understanding the larger picture by putting pieces together. It's in stark contrast to decades of reductionist biology, which involves taking the pieces apart. (Christopher Wanjek, NIH)



The Scope of Genomics

- **Genome-wide Analysis:** This field involves the study of information content and structural features of an organism's complete DNA sequence.
- **Functional Identification:** Includes investigating the biological roles of encoded proteins and regulatory elements that determine an organism's unique physiology, i.e., annotation.
- **Data Management:** Genomics generates massive amounts of information that must be analyzed and managed using high-throughput sequencing and specialized software.



Biotechnology for Beginners, 3rd Edition, Reinhard Renneberg, eBook ISBN: 9780323855709

Bioinformatics and Data Repositories

- **Search Algorithms:** Bioinformatics tools like **BLAST** (Basic Local Alignment Search Tool) are engines for "data mining," allowing search of databases for sequence matches.
- **Sequence Retrieval:** Complex computer algorithms enable retrieval of specific nucleotide or protein information from shared international repositories like GenBank.
- **Global Databases:** Sequence data are stored in these specialized repositories, which are searchable to facilitate the comparison of genetic info across all domains of life.

Web BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Genome Annotation Workflow

- **Functional Homology:** The annotation process relies heavily on **BLAST** to predict the potential biological function of new sequences by finding "homology," or significant similarity, to known, well-characterized genes.
- **Descriptive Labeling:** Annotation is the descriptive process of identifying structural features like **open reading frames (ORFs)** and functional RNA molecules (rRNA/tRNA) within a newly determined genome sequence.
- **Experimental Verification:** Computer-predicted annotations are *preliminary* and require lab testing to confirm biological roles.
    - Errors can get propagated in this way.

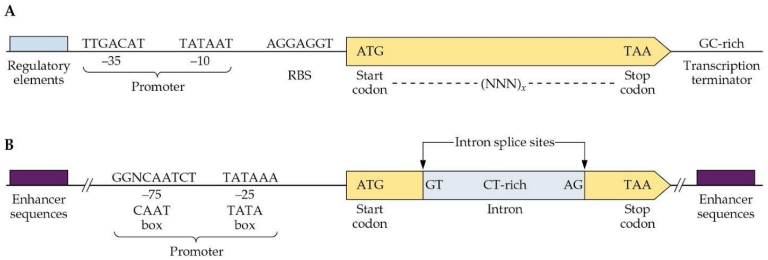


Figure 2.39 Genome annotation utilizes conserved sequence features. Predicting protein-coding sequences (open reading frames) in prokaryotes **(A)** and eukaryotes **(B)** requires identification of sequences that correspond to potential translation start (ATG or, more rarely, GTG or TTG) and stop (TAA, shown; also TAG or TGA) codons in mRNA. The number of nucleotides between the start and stop codons must be a multiple of three (i.e., triplet codons) and must be a reasonable size to encode a protein. In prokaryotes, a conserved ribosome-binding site (RBS) is often present 4 to 8 nucleotides upstream of the start codon **(A)**. Prokaryotic transcription regulatory sequences such as an RNA polymerase recognition (promoter) sequence and binding sites for regulatory proteins can often be predicted based on similarity to known consensus sequences. Transcription termination sequences are not as readily identifiable but are often GC-rich regions downstream of a predicted translation stop codon. In eukaryotes, protein coding genes typically have several intron sequences in primary RNA that are delineated by GU and AG and contain a pyrimidine-rich tract **(B)**. Introns are spliced from the primary transcript to produce mRNA. Transcription regulatory elements such as the TATA and CAAT boxes that are present in the promoters of many eukaryotic protein-coding genes can sometimes be predicted. Sequences that are important for regulation of transcription are often difficult to predict in eukaryotic genome sequences; for example, enhancer elements can be thousands of nucleotides upstream and/or downstream from the coding sequence that they regulate.

Prokaryotic Annotation Features

- **Open Reading Frames (ORFs):** Software scans for potential translation start and stop codons to encode functional bacterial proteins.
- **Ribosome Binding:** In prokaryotes, a conserved ribosome-binding site (RBS) is typically located 4 to 8 nucleotides upstream of the start codon.
- **Regulatory Signals:** Transcription sites, such as RNA polymerase recognition (promoter) sequences, are predicted based on their similarity to known consensus sequences.
- **Operon Organization:** Unlike eukaryotes, prokaryotic genes are frequently clustered into operons, which are groups of related genes transcribed under the control of a single promoter.

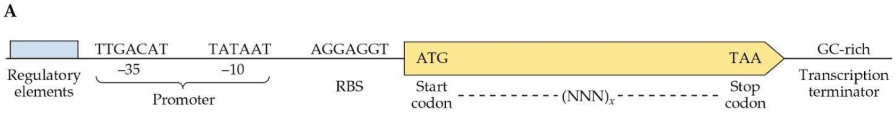


Figure 2.39 Genome annotation utilizes conserved sequence features. Predicting protein-coding sequences (open reading frames) in prokaryotes **(A)** and eukaryotes **(B)** requires identification of sequences that correspond to potential translation start (ATG or, more rarely, GTG or TTG) and stop (TAA, shown; also TAG or TGA) codons in mRNA. The number of nucleotides between the start and stop codons must be a multiple of three (i.e., triplet codons) and must be a reasonable size to encode a protein. In prokaryotes, a conserved ribosome-binding site (RBS) is often present 4 to 8 nucleotides upstream of the start codon **(A)**. Prokaryotic transcription regulatory sequences such as an RNA polymerase recognition (promoter) sequence and binding sites for regulatory proteins can often be predicted based on similarity to known consensus sequences. Transcription termination sequences are not as readily identifiable but are often GC-rich regions downstream of a predicted translation stop codon. In eukaryotes, protein coding genes typically have several intron sequences in primary RNA that are delineated by GU and AG and contain a pyrimidine-rich tract **(B)**. Introns are spliced from the primary transcript to produce

Eukaryotic Annotation Complexity

- **Exon-Intron Boundaries:** Primary RNA transcripts must be mapped to identify introns, which are typically delineated by GU and AG splice sites.
- **Promoter Motifs:** Transcription regulatory elements, such as the TATA and CAAT boxes, can sometimes be predicted in many eukaryotic protein-coding genes.
- **Distant Regulation:** Enhancer elements are difficult to predict because they can be located thousands of nucleotides away from the sequence they regulate.

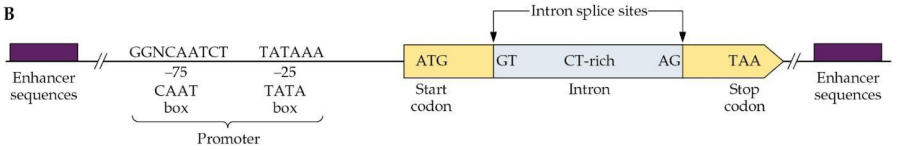


Figure 2.39 Genome annotation utilizes conserved sequence features. Predicting protein-coding sequences (open reading frames) in prokaryotes **(A)** and eukaryotes **(B)** requires identification of sequences that correspond to potential translation start (ATG or, more rarely, GTG or TTG) and stop (TAA, shown; also TAG or TGA) codons in mRNA. The number of nucleotides between the start and stop codons must be a multiple of three (i.e., triplet codons) and must be a reasonable size to encode a protein. In prokaryotes, a conserved ribosome-binding site (RBS) is often present 4 to 8 nucleotides upstream of the start codon **(A)**. Prokaryotic transcription regulatory sequences such as an RNA polymerase recognition (promoter) sequence and binding sites for regulatory proteins can often be predicted based on similarity to known consensus sequences. Transcription termination sequences are not as readily identifiable but are often GC-rich regions downstream of a predicted translation stop codon. In eukaryotes, protein coding genes typically have several intron sequences in primary RNA that are delineated by GU and AG and contain a pyrimidine-rich tract **(B)**. Introns are spliced from the primary transcript to produce mRNA. Transcription regulatory elements such as the TATA and CAAT boxes that are present in the promoters of many eukaryotic protein-coding genes can sometimes be predicted. Sequences that are important for regulation of transcription are often difficult to predict in eukaryotic genome sequences; for example, enhancer elements can be thousands of nucleotides upstream and/or downstream from the coding sequence that they regulate.

Comparative Genomics in Industrial Applications

- **Industrial Optimization:** Comparing industrial strains to wild-types helps identify "genomic instability." Deleting non-essential elements like prophages in *C. glutamicum* can increase growth rates and improve yields.
- **Environmental Bioremediation:** Analysis of diverse *Pseudomonas* and *Burkholderia* genomes allows bioengineers to combine degradative plasmids to create new capabilities (e.g., capable of cleaning oil spills or toxic herbicides).

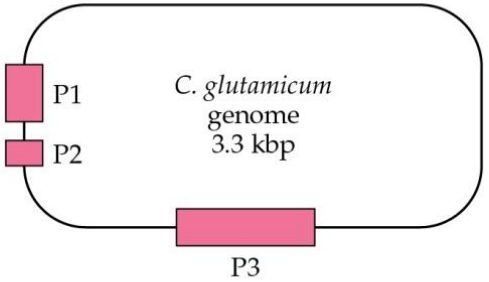


Figure 8.10 The genome of a strain of *Corynebacterium glutamicum* containing three prophages (not drawn to scale). The prophages are labeled P1 to P3 for prophages 1 to 3.

The book can have errors!

The genome size is 3.3 **Mbp**!

Comparative Genomics in Medicine

- **Mutation Identification:** Comparing related genomes reveals mutations associated with the development of specific human diseases.
- **Cancer Analysis:** Sequencing tumor genomes and comparing them to normal cells reveals cancer-specific mutations.
- **Pathogen Virulence:** Comparing genomes of bacterial pathogens with non-pathogens helps identify virulence genes for drug targets.
- **What is Precision Medicine?** Precision medicine is health care that is based on you as an individual. It takes into account factors like where you live, what you do, and your family health history… and your genome!

All of Us RESEARCH PROGRAM

NIH National Institutes of Health

https://www.joinallofus.org/

Large-Scale Disease Identification

- **100,000 Genomes Project:** Initiatives like this use whole-genome sequencing to identify the genetic basis of rare diseases and common cancers.
- **Precision Medicine:** Genomic data are used to predict risk, diagnose conditions, and tailor medical treatments to an individual's specific genotype.
- **Pathogen Detection:** Unique sequences derived from comparative studies are utilized for rapid pathogen detection and the development of more effective vaccines.

Video - Concepts in the module or a demonstration

- BLAST Search: How homology is determined (1min58s)

More videos

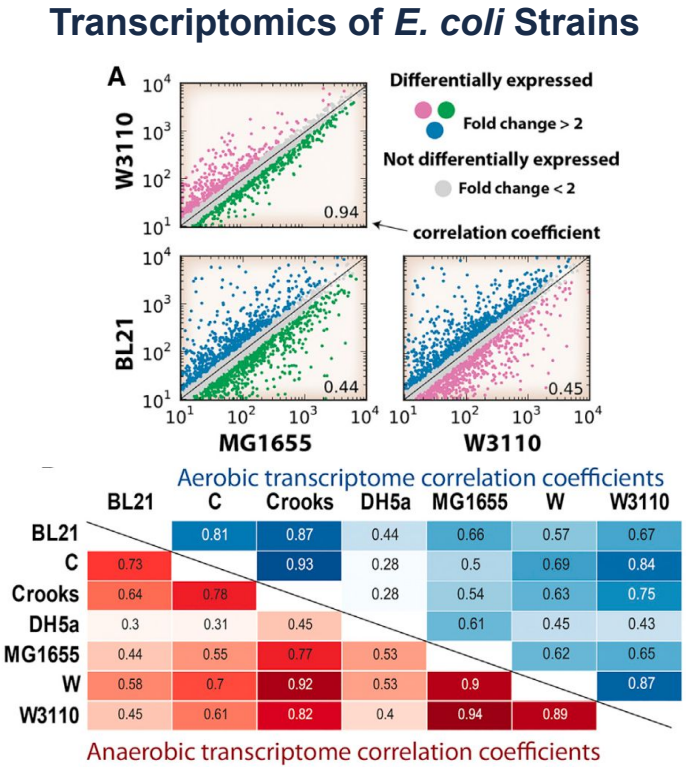
- Trey Ideker, UCSD: Big Data in Biomedicine (Stanford Medicine Precision Health 2016)
- Genomics work in FAH - Siavash Mirarab, Turakhia Lab, Vineet Bafna

Module 2: Transcriptomics and Expression Profiling (Source Pages: Chapter 2: 65–70)

- Moving from DNA code to measuring total RNA expression.
- Quantitative analysis of the whole-genome transcriptome
- Comparing profiles across different conditions.
- Analyzing single-cell level gene activity.

Differential Expression: Identifying Response

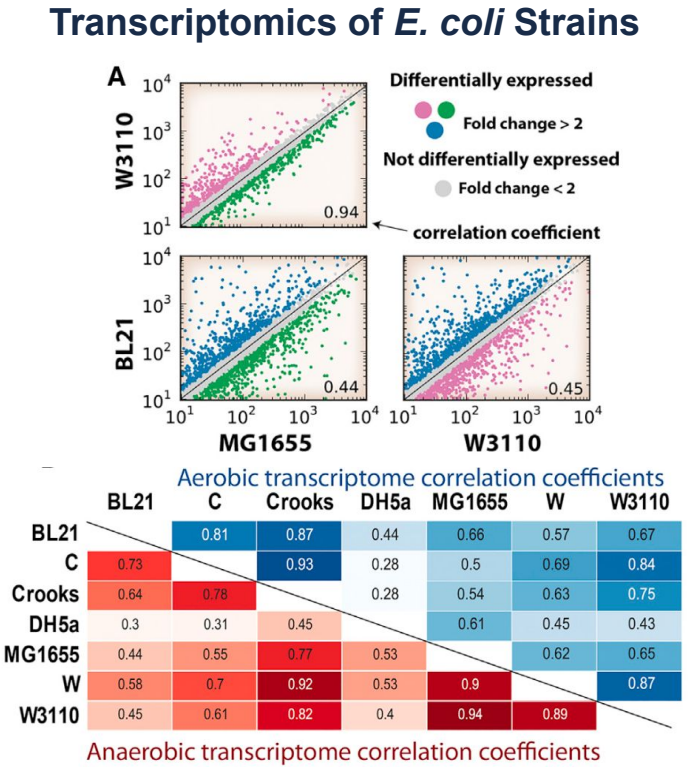
- **Comparative Strategy:** Identifies genes that are up- or downregulated by comparing test samples to a reference control (e.g., healthy vs. diseased).
- **Fold Change Analysis:** Uses ratios to measure the magnitude of a biological response to stimuli like toxins or stress.
- **Clustering:** Groups genes with similar patterns in heat maps to predict cooperative function in biological pathways.
- **Functional Inference:** Allows bioengineers to infer a cell's "current priorities" during different developmental stages.



Monk et al. & Feist, 2016, Cell Systems 3, 238–251, http://dx.doi.org/10.1016/j.cels.2016.08.013

Quantitative Transcriptomics & Calibration

- **Absolute Abundance:** RNAseq counts individual sequence reads to measure the exact quantity of transcripts in a sample.
- **Internal Controls:** Uses "housekeeping genes", i.e., genes with stable expression, as a biological baseline.
- **Spiked Controls:** Adds external RNA sequences of a known amount to calibrate and normalize data between different sequencing runs.
- **Data Normalization:** Scales raw counts to total sample size (e.g., total reads) to ensure comparisons are fair and accurate.



Monk et al. & Feist, 2016, Cell Systems 3, 238–251, http://dx.doi.org/10.1016/j.cels.2016.08.013

RNA Sequencing (RNAseq) Workflow

- **RNA Isolation:** Total RNA is extracted from a sample, and ribosomal RNA (rRNA) is often removed because it constitutes over 80% of cellular RNA.
- **cDNA Synthesis:** Reverse transcriptase and random hexamer primers convert the isolated RNA into stable complementary DNA (cDNA) libraries for sequencing.
- **Adaptor Ligation:** The cDNA fragments are ligated to adaptors that provide specific binding sites for high-throughput sequencing primers.

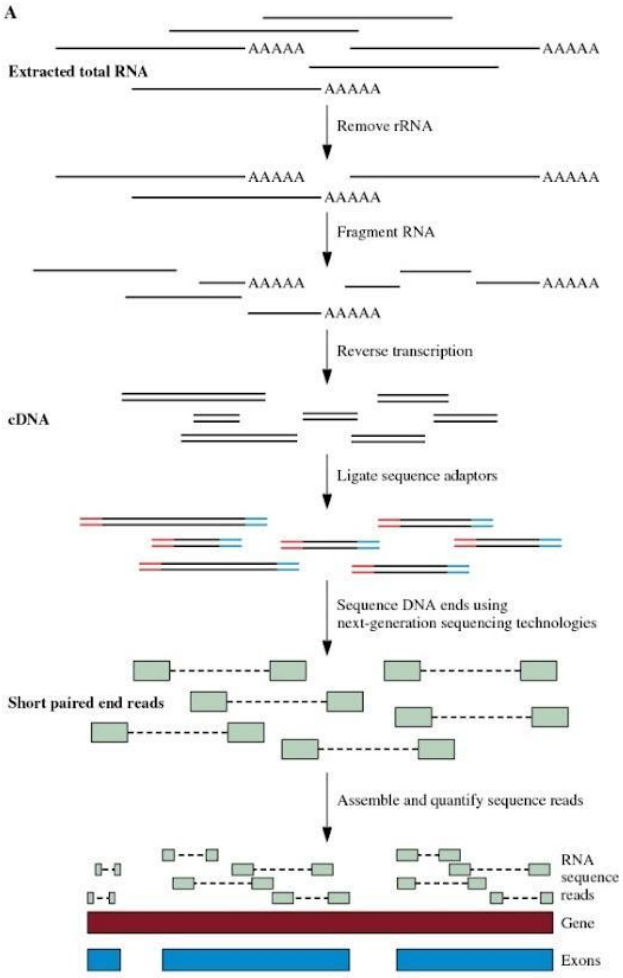


Figure 2.40 High-throughput RNA sequencing. **(A)** Total RNA is extracted from a sample, and rRNA may be removed. The RNA is fragmented and then converted to cDNA using reverse transcriptase. Adaptors are added to the ends of the cDNA to provide binding sites for sequencing primers. High-throughput next-generation sequencing technologies are used to determine the sequences at the ends of the cDNA molecules (paired-end reads). The sequence reads are aligned to a reference genome or assembled into contigs using the overlapping sequences. Shown is the alignment of paired-end reads to a gene containing two introns. **(B)** RNA expression levels are determined by counting the reads that correspond to a gene.

Quantifying Gene Expression

- **Read Alignment:** Sequence reads are aligned to a reference genome or assembled *de novo* into contiguous "contigs" when no reference is available.
- **Counting Reads:** Gene expression levels are determined by counting the sequence reads that correspond to each specific nucleotide position in a gene.
- **Data Normalization:** Expression levels are normalized between different samples by scaling the results to the total number of reads per sample or by quantifying them against internal housekeeping genes or spiked standards.

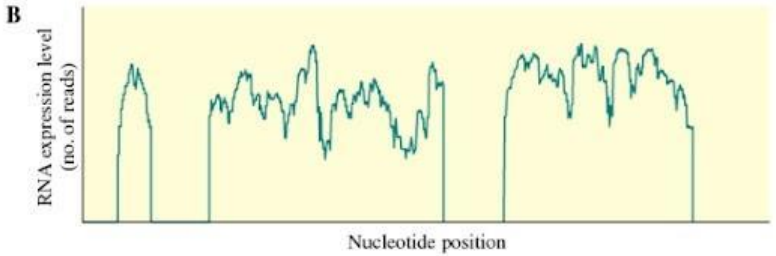


Figure 2.40 High-throughput RNA sequencing. **(A)** Total RNA is extracted from a sample, and rRNA may be removed. The RNA is fragmented and then converted to cDNA using reverse transcriptase. Adaptors are added to the ends of the cDNA to provide binding sites for sequencing primers. High-throughput next-generation sequencing technologies are used to determine the sequences at the ends of the cDNA molecules (paired-end reads). The sequence reads are aligned to a reference genome or assembled into contigs using the overlapping sequences. Shown is the alignment of paired-end reads to a gene containing two introns. **(B)** RNA expression levels are determined by counting the reads that correspond to a gene.

Single-Cell RNA Sequencing (scRNA-seq)

- **Cellular Heterogeneity:** This technique assesses gene expression in individual cells to reveal diversity within a population that "bulk" RNA-seq might miss.
- **Microfluidic Encapsulation:** Individual cells are encapsulated within oil droplets with gel beads containing enzymes and primers for cDNA synthesis.
- **Molecular Barcoding:** Unique identifier sequences (barcodes) are assigned to each cell's transcripts, allowing sequence reads to be mapped back to individual cells.

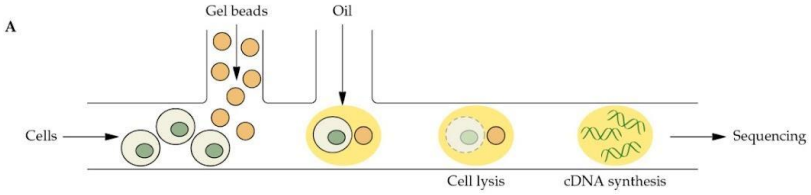


Figure 2.41 Single-cell RNA sequencing. **(A)** Cells in a tissue are dissociated, and a single cell is encapsulated in a droplet of oil with a gel bead containing primers and reverse transcriptase for cDNA synthesis, often within a microfluidic device as depicted here. In addition to sequences to prime cDNA synthesis, such as poly(T) that binds to the poly(A) tails of eukaryotic mRNA, the primers may contain unique identifier sequences, for example, to indicate the origin of the cell. Within the oil droplet, the cells are lysed and cDNA is synthesized from the released RNA. The cDNA can then be extracted from the oil, amplified, and sequenced as described in Fig. 2.40. **(B)** Heat map of gene expression in individual cells in regenerating earth-

Visualizing Data: Heat Maps

- **Expression Visualization:** Gene expression levels are depicted as variations in color intensity, with yellow often indicating high and pink/blue indicating low activity.
- **Clustering Analysis:** Genes with similar expression patterns are clustered to facilitate predictions of their cooperative function in a biological pathway.
- **Cell Identification:** Clustering data from single cells helps marker genes identify dominant cell types, such as stem cells or muscle cells.

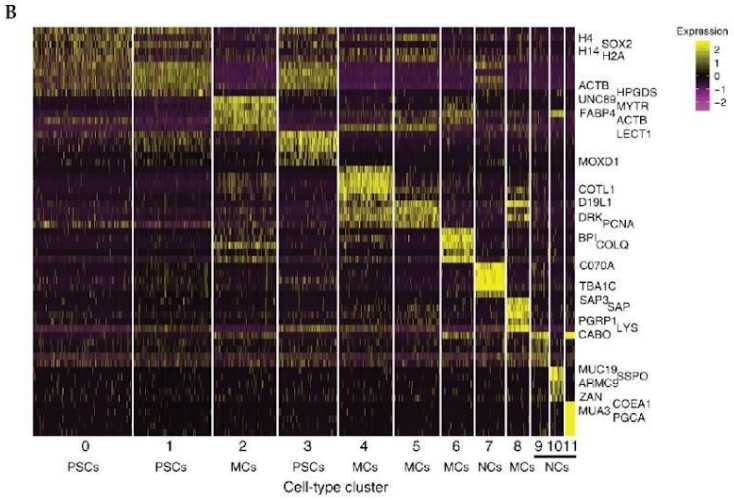


Fig. 2.40. **(B)** Heat map of gene expression in individual cells in regenerating earthworm tissue. Each row shows the expression level of a specific gene (listed on the right) in the cells. Highly expressed genes are indicated in yellow, and genes with low levels of expression are indicated in pink. Five genes with the greatest fold change in expression for each cell type are shown. PSCs, pluripotent stem cells; MCs, muscle cells; NCs, neuronal cells. Reprinted from Shao et al., *Nat. Commun.* **11**:2656, 2020, under license CC BY 4.0.

DNA Microarray Technology

- **Probe Hybridization: cDNA** Samples are derived from mRNAs and hybridized to single-stranded DNA sequences (probes) arrayed on a solid platform.
- **Oligonucleotide Design:** Microarrays can contain more than 500,000 probes representing up to 30,000 different genes synthesized directly on the array surface.
- **Diagnostic Utility:** This technology is used in clinical tests assess the risk of cancer metastasis based on gene expression.

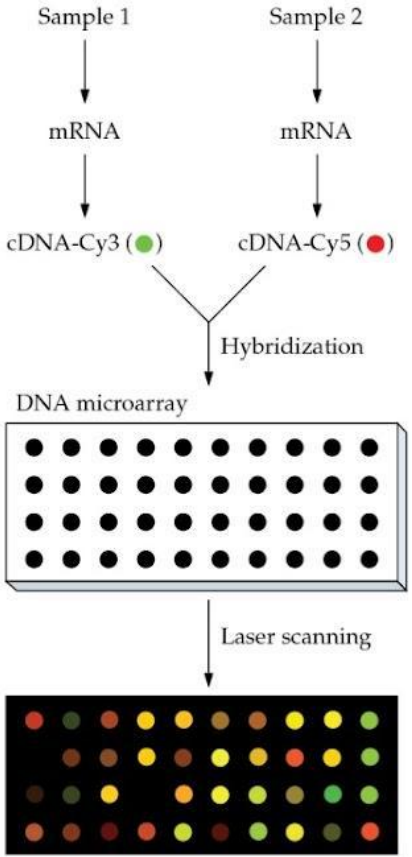


Figure 2.42 Gene expression profiling with a DNA microarray. The mRNA is extracted from two samples (sample 1 and sample 2), and during reverse transcription, the cDNA strands are labeled with the fluorescent dyes Cy3 and Cy5, respectively. The cDNA samples are mixed and hybridized to an ordered array of either gene sequences or gene-specific oligonucleotides. After the hybridization reaction, each probe cell is scanned for both fluorescent dyes and the separate emissions are recorded. Probe cells that produce only a green or red emission represent genes that are transcribed only in sample 1 or 2, respectively; yellow emissions indicate genes that are active in both samples, and the absence of emissions (black) represents genes that are not transcribed in either sample.

Microarray Experimental Design

- **Dual-Color Labeling:** Test and reference samples are labeled with different fluorescent dyes, typically Cy3 (green) and Cy5 (red), for comparative analysis.
- **Ratio Measurement:** A laser scanner determines the intensity ratio of the dyes to indicate relative abundance of transcripts between two samples.
- **Fold Change:** The raw fluorescence data are converted to a "fold change" ratio to identify genes responding to specific biological conditions.

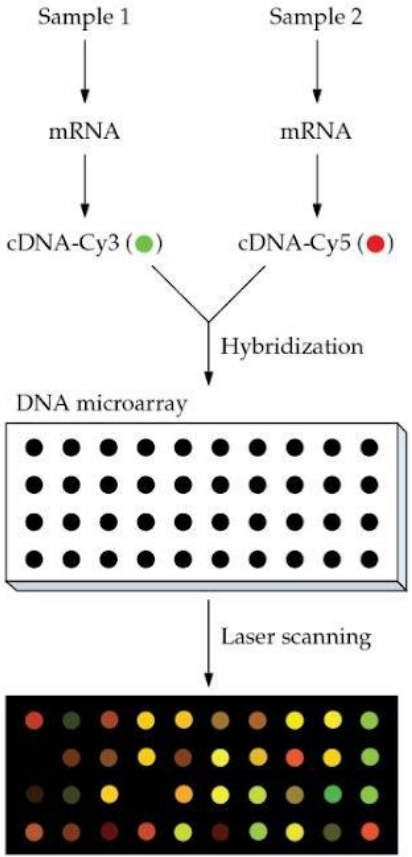
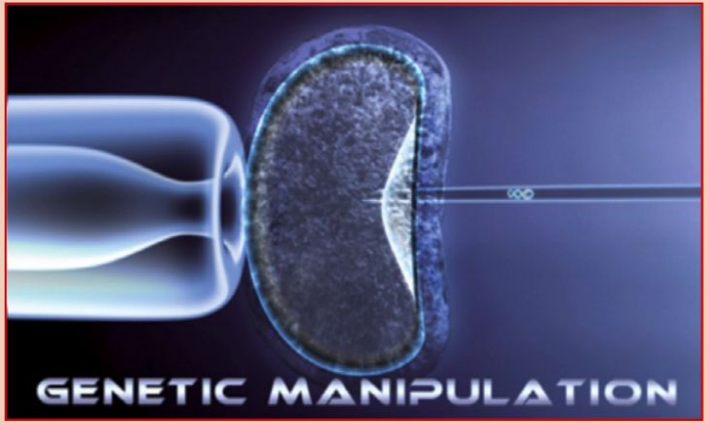


Figure 2.42 Gene expression profiling with a DNA microarray. The mRNA is extracted from two samples (sample 1 and sample 2), and during reverse transcription, the cDNA strands are labeled with the fluorescent dyes Cy3 and Cy5, respectively. The cDNA samples are mixed and hybridized to an ordered array of either gene sequences or gene-specific oligonucleotides. After the hybridization reaction, each probe cell is scanned for both fluorescent dyes and the separate emissions are recorded. Probe cells that produce only a green or red emission represent genes that are transcribed only in sample 1 or 2, respectively; yellow emissions indicate genes that are active in both samples, and the absence of emissions (black) represents genes that are not transcribed in either sample.

Ethical Boundaries of the "Postgenomic" Era

- **Justifiable Intervention:** Interventions are currently constrained by three factors: extraordinary suffering, highly penetrant genotypes, and safe technology.
- **The Postgenomic Manifesto:** As DNA reading costs fall below $1,000, society must redefine the boundary between "healing" and "perfection".
- **Data Privacy:** Genomic information is invaluable to patients but carries risks if accessed by insurance companies or employers.



Biotechnology for Beginners, 3rd Edition, Reinhard Renneberg, eBook ISBN: 9780323855709

Video - Concepts in the module or a demonstration

- RNA Sequencing (RNA-Seq) & DNA Microarrays (4min03s)
- Single-Cell Sequencing and Analysis Workflow (2min57s)

More videos

- DNA Microarray Technique (1min56s)

The End