**Learning Goals**

Week 1

* List and explain the twelve principles for data organization in spreadsheets from Broman & Woo (2018).

1. **Be consistent**: whatever you do, do it consistently. Even throughout multiple files. Use consistent file names. Consistent format for all dates YYYY-MM-DD.
2. **Choose good names for things**: do not use spaces, use underscores or hyphens, avoid special characters. The names should be short, but meaningful. Finally, do not use “final” in file name.
3. **Write Dates as YYYY-MM-DD**: be aware of Excel’s treatment of dates, this can cause problems in data. Sometimes it even turns other thins into dates. To prevent problems, do this: select column 🡪 in menu bar, select format > cells 🡪 choose “Text” on the left. Or begin date with apostrophe. Or as YYYYMMDD
4. **No empty cells***:* fill in all cells and use common code for missing data like “NA” or even a hyphen.
5. **Put just one thing in a cell**: the cells in spreadsheet should each contain one piece of data. Do not include the units in the cell. Do not include notes within a cell. Do also not merge cells.
6. **Make it a rectangle**: best layout is a single big rectangle with rows corresponding to subjects and columns corresponding to variables. Or if necessary, as a set of rectangles).
7. **Create a data dictionary**: This might contain the exact variable names as in the data file, a version of the variable name that might be used in data visualizations, a longer explanation of what the variable names means, the measurements units, expected min and max values.

This is part of the metadata. You will also want a “ReadMe” file that includes an overview of the project and data.

1. **No calculations in the raw data files:** just the data, nothing else. Write-protect it, back it up, and do not touch it. For analysis, copy the file and do calculations and graphs in the copy.
2. **Do not use font color or highlighting as data:** better to include an additional column.
3. **Make backups:** regularly and in multiple locations. “Read only” for when you are not actively entering the data.
4. **Use data validation to avoid errors:** validation criteria in excel, or set as text.
5. **Save the data in plain text files:** CSV format

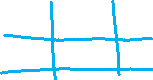
* Apply the principles for data organization in creating spreadsheets.
* List the three properties (or “rules”) of tidy data according to Wickham (2014).

This paper defines tidy data sets as a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).

* facilitate analysis, as you spend less time on data cleaning and wrangling;
* make it reproducible and easy for others to understand;
* format your data to easily apply tools used in R (tidyverse) or other languages.
* Observations: rows



* Variables: columns



* One type of observational unit per table

Column headers are headers not variable names



A dataset is a collection of values, either numbers (quantitative) or strings (qualitative). `

It is easier to describe functional relationships between variables than between rows, and it is easier to make comparisons between groups of observations than between groups of columns.

Fixed should come first, than the measured values

* List and recognize the five most common problems with messy data sets and supply appropriate solutions to solve these issues.

Messy data is any other arrangement of the data then described by Codd’s 3rd normal form (see last goal).

1. Column headers are values, not variable names.

*Solution: melt or stack, turn columns into rows.*

1. Multiple variables are stored in one column.

*Solution: splitting*

1. Variables are stored in both rows and columns.

*Solutions: melting, cast or unstack operation (inverse of melting by rotating the element variable back out into the columns)*

1. Multiple types of observational units are stored in the same table.

*Solution: normalization, useful for tidying and eliminating inconsistencies. Divide this table with 2 types in two tables.*

1. A single observational unit is stored in multiple tables.

*Solution: read the files into a list of tables 🡪 add a new column that records the original file name for each table 🡪 combine all tables into a single table*

*Plyr package makes this a simple task in R. The following code generates a*

*vector of file names in a directory (data/) which match a regular expression (ends in .csv).*

*Next we name each element of the vector with the name of the file. We do this because plyr*

*will preserve the names in the following step, ensuring that each row in the final data frame*

*is labeled with its source. Finally, ldply() loops over each path, reading in the CSV file and*

*combining the results into a single data frame.*

*R> paths <- dir("data", pattern = "\\.csv$", full.names = TRUE)*

*R> names(paths) <- basename(paths)*

*R> ldply(paths, read.csv, stringsAsFactors = FALSE)*

* Create data sets in a tidy format.
* Reformat non-tidy ("messy") data sets to make them tidy.

Messy datasets can be tidied with the tools: melting, string splitting and casting.

Question 1

|  |  |  |
| --- | --- | --- |
| **Table 1** |  |  |
| **Factor** | **Day** | **Freq** |
| 1 | 10 | 1 |
| 1 | 16 | 3 |
| 2 | 10 | 6 |
| 2 | 16 | 28 |
| 3 | 10 | 0 |
| 3 | 16 | 10 |
| 4 | 10 | 0 |
| 4 | 16 | 5 |

|  |  |  |
| --- | --- | --- |
| **Table 3** |  |  |
| **Colony\_number** | **Factor** | **Day** |
| 1 | 1 | 10 |
| 3 | 1 | 16 |
| 6 | 2 | 10 |
| 28 | 2 | 16 |

Question 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Taxon** | **Habitat** | **Change** | **Bf influence** | **Cw influence** | **Date** |
| Coleoptera | A | 1 | 1 | 1 | 2010 |
| Coleoptera | D | 1 | 0 | 1 | 2010 |
| Coleoptera | U | 1 | 0 | 1 | 2010 |
| Coleoptera (Carabidae) | A | 1 | 0 | -1 | 2012 |
| Coleoptera (Coccinellidae) | A | 1 | 0 | -1 | 2014 |
| Coleoptera (Coccinellidae) | A | 1 | 0 | 0 | 2007 |
| Coleoptera (Coccinellidae) | D | 1 | 0 | 0 | 2007 |
| Hymenoptera | A | 1 | 0 | 0 | 2009 |
| Hymenoptera | U | 1 | 0 | 0 | 2009 |
| Lepidoptera (moths\_ | A | 1 | 0 | -1 | 2006 |
| Lepidoptera (moths\_ | D | -1 | 0 | -1 | 2006 |
| Odonata | U | 1 | 0 | 0 | 2010 |
| Odonata | WR | 1 | 0 | 0 | 2010 |

* Describe why gene names can get corrupted when using Excel

Some gene symbols resemble dates, such as SEPT2 and MARCH1. Excel has the tendency to convert these gene names into those dates. This causes a lot of error in gene data files. There is no way to permanently deactivate automatic conversion to dates in many spreadsheet

Week 2

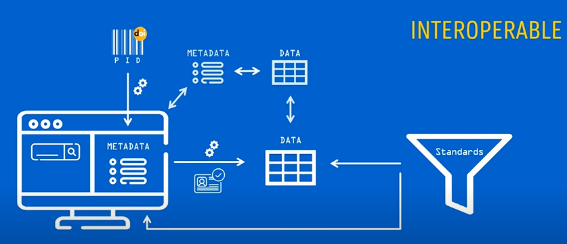
* List the principles of FAIR data and describe what these principles mean.

Findable: Data should be easily discoverable by humans or machines through unique identifiers and descriptive metadata.

Accessible: Data should be easily accessible for use by humans or machines, without undue barriers. Accessible is not the same as open. A protocol is allowed.

Interoperable: Data should be usable across different systems and be accessible to a wide range of users. Understandable.

Reusable: Data should be usable and accessible for subsequent use, without restriction or technical barriers.

These principles aim to ensure that data is well-documented, standardized, and readily available to researchers, scientists, and other users, so that it can be used, reused, and built upon in a manner that maximizes its value. The goal is to promote transparency and collaboration and make it easier for researchers to integrate data into their own work and build upon Afbeelding met tekst

Automatisch gegenereerde beschrijvingthe work of others.

* Define the following concepts open data, metadata, open science.

**Open data** refers to data that is freely available to anyone to access, use, and share without legal, financial, or technical barriers. The goal of open data is to promote transparency, collaboration, and innovation by making data accessible to the public and enabling reuse and integration of data into new projects and research.

**Metadata** is information that describes the characteristics of data. It provides contextual information about data, such as the date it was created, the author, the format, and the subject matter. Metadata is important for finding, accessing, and using data effectively.

**Open Science** is a movement that advocates for making the scientific process more transparent and accessible by making research data, methods, and results publicly available. The goal of Open Science is to promote collaboration, encourage innovation, and increase the credibility and reproducibility of scientific research by making it easier for others to build upon and validate existing work. Open Science encompasses open data, open access publishing, open-source software, and open collaboration practices.

* Define the following terms and describe how to relate to each other: reproducibility, replicability and repeatability.

**Reproducibility** refers to the ability of obtaining the same results by independently conducting the same experiment or analysis using the same data and methods. Some allowance for differences in machine precision.

**Replicability** refers to the ability of obtaining the same results by independently conducting a similar experiment or analysis with a new dataset and/or new methods. Allows for differences in results.

**Repeatability** refers to the ability of obtaining the same results by conducting the same experiment or analysis multiple times under the same conditions.

In general, reproducibility is considered the gold standard in scientific research, and replicability and repeatability are seen as weaker forms of evidence for the validity of results. However, replicability and repeatability can still provide important information about the robustness and generalizability of results.

* List advantages and disadvantages of open data and open science.

See table:

|  |  |
| --- | --- |
| Advantages of open data: | Disadvantages of open data: |
| Increased transparency and accountability.  Improved quality control and error correction.  Greater accessibility for researchers and the public.  Enhanced collaboration and interdisciplinary research.  Opportunities for innovation and problem-solving. | Privacy and security concerns.  Intellectual property and commercialization issues.  Lack of incentives for data sharing.  Inadequate infrastructure for data management and preservation.  Burden on researchers to properly document and format data. |
| Advantages of open science: | Disadvantages of open science: |
| Enhanced transparency and reproducibility.  Increased collaboration and interdisciplinary research.  Faster progress and discovery in scientific research.  Greater accessibility of scientific results and methods.  Opportunities for citizen science and public engagement. | Privacy and security concerns.  Intellectual property and commercialization issues.  Lack of incentives for openness.  Competition for funding and recognition.  Resistance to change and resistance from some traditional scientific communities. |

* Form a well-reasoned opinion on open science, open data and reproducibility.
* *In the genomics field, scientists work with big amounts of data. How does this relate to the reproducibility crisis?*

Not the same outcome, p value only valid for a simple measure: p hacking is a big problem. Increasing likelihood of false positives. Data does not speak for it’s self. If only one research group does its own statistical analysis, replication studies boring: not in scientific journals.

* *What’s the importance of reproducibility (in the Genomics field)?*

To increase the correctness of scientific findings

* *What is Open science and what is its goal?*

Making publications available and reusable via open access and making research data accessible. To allow others to use and reuse your results. Before, during or after. Everyone can benefit from it.

* *What are characteristics of Open science?*

The four R’s of open science: Reliable, Reproducible, Reusable, Relevant

Research Culture is Broken, open science can fix it. There is pressure for scientists to publish paper. Publish or perish.

Let’s stop rating research on quantity, but on quality. Time to adapt. Exclusive 🡪 Inclusive.

* *What are the six pillars of open science and what do these pillars entail?*

1. **Open data:** the process of releasing both raw and processed data, enabling other to analyze it without restriction.
2. **open access:** the model under which papers are available for anyone to read without having to pay, and that license allows secondary use such as text-mining.
3. **open methodology:** one which has been described in sufficient detail to allow other researchers to repeat the work and apply it elsewhere
4. **open source:** refers to open and free access to the blueprint of a product; applied to software, it refers to the source code.
5. **open peer review:** transforming the peer review process; it is about making peer review a collaborative process between authors and reviewers; it is about constructive criticism, but with the goal of helping the authors to get published.
6. **open education:** the open and free availability of educational resources.

* *What are the definitions of repeatability, replicability, and reproducibility? How do they differ?*
* *What are characteristics of reproducible research?*
* *What are the opportunities reproducible research offers concerning data analysis and future studies?*

Time can be saved, data and code can be maintained.

* *What are the challenges that come with reproducible research conducts like sharing data and code, and what are best practices to overcome these challenges?*

Not supported yet.

* What are some doubts and concerns researchers may have related to open science and open data?

Data uit context 🡪 misinterpretatie.

* What are possible solutions to remove these doubts and concerns?

Software te updaten, spreadsheet regels

Weighing up these advantages and disadvantages, what is your opinion on open science, open data and reproducibility?

While open data focuses on accessibility, FAIR data takes a more comprehensive approach to ensuring data quality and usability.

What was the authors’ goal?

To asses AI systems

What was their (general) approach?

Datasets representative of UK and US were curated. Outcomes derived. AI system trained to identify presence of breast cancer. And was evaluated.

What were their conclusions?

This system might be integrated into screening workflows. The system can generalize across continents.

What are the major points that Haibe-Kains et al. have, concerning the reproducibility of the McKinney et al. study based on the data and code provided?

* Absence of sufficiently documented methods and comuter code underlying the study effectively undermindes its scientific value.
* Textual descriptions of deep-learning models can hide their high level of complexity. 🡪 actual computer code.
* Definition of several hyperparameters are missing 🡪 frame works excist to to make AI more transparent and reproducible.
* Sharing code and data
* The lack of access to code and data in prominent scientific publications may lead to unwarranted and even potentially harmful clinical trials1

What are the major points in the reply to the concerns of Haibe-Kains et al.?

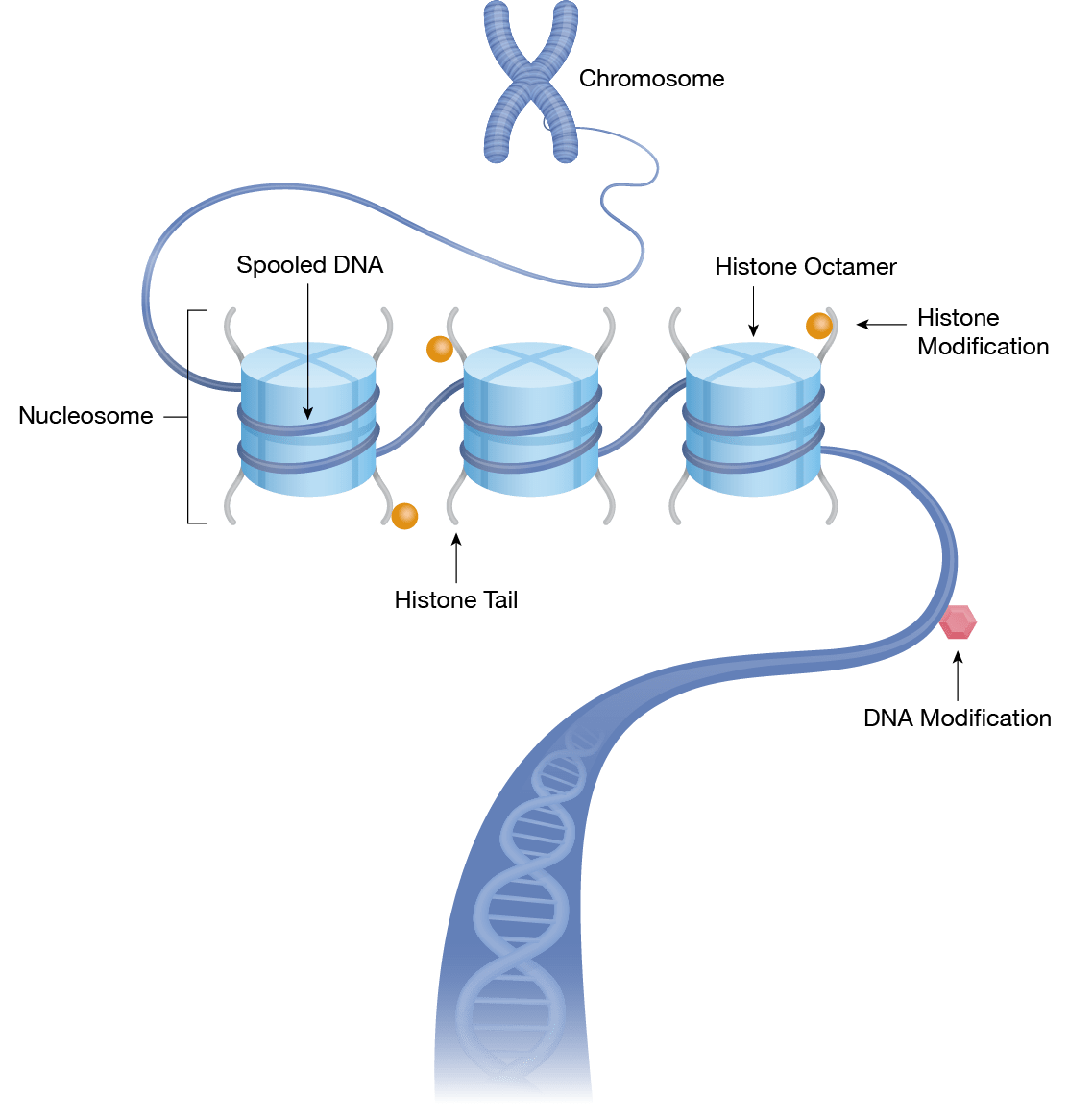
Week 3

* List key characteristics of the human genome, such as size, number of chromosomes and number of annotated genes.

The key characteristics are:

* Size: 2.096 billion basepairs (haploid)
* Number of chromosomes: 23 times 2
* 20,441 protein coding genes
* 198,002 coding transcripts
* Define and explain the terms histone, histone modification, nucleosome and chromatin, and explain how these terms relate to each other and their role in genome organization.

Histones are small basic proteins that serve as the building blocks for chromatin, the material that makes up chromosomes. There are five main types of histones (H1, H2A, H2B, H3, and H4), which form a tight and regular structure around DNA, providing the physical basis for DNA packaging in the nucleus.

Histone modification refers to the covalent alteration of histones, such as acetylation, methylation, or phosphorylation, which changes the charge, shape, or interaction properties of the histones. These modifications play a critical role in regulating gene expression, DNA replication, and other cellular processes.

A nucleosome is the basic unit of chromatin, consisting of a DNA strand wrapped around an octamer of histones (two copies each of H2A, H2B, H3, and H4). The DNA within the nucleosome is organized into 147 base pair "beads" with "linker" DNA between the beads. The regulation of gene expression often depends on the organization of the nucleosomes and the regulation of the linker DNA.

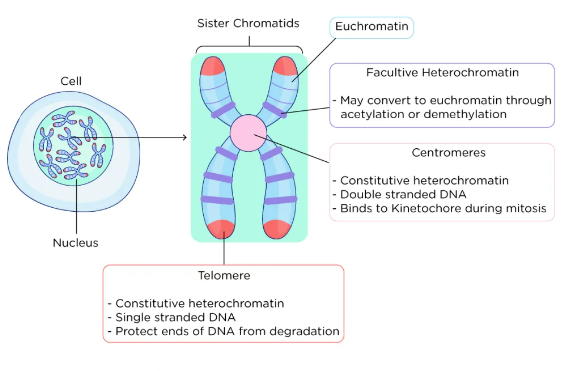
Chromatin is the material that makes up chromosomes and is composed of DNA and proteins, including histones and non-histone proteins. Chromatin plays a critical role in the regulation of gene expression and the maintenance of genomic stability. In its compact and inaccessible state, chromatin effectively "silences" genes and prevents their expression. Conversely, when chromatin is more open and accessible, genes are more likely to be transcribed and expressed.

In conclusion, histones, histone modifications, nucleosomes, and chromatin are all critical components of genome organization, with histones and their modifications playing a key role in regulating gene expression and chromosomal structure. The precise arrangement of nucleosomes and the modulation of histone modifications through various enzymes, including histone acetyltransferases and histone deacetylases, provides a flexible and dynamic mechanism for regulating gene expression and other cellular processes.

* Define and explain the terms euchromatin and heterochromatin and compare and contrast these two types of chromatin and list the differences.

Euchromatin and heterochromatin are two distinct types of chromatin that differ in their degree of compaction and gene expression.

Euchromatin refers to the more open and accessible form of chromatin that is often associated with actively transcribed genes and a more open, gene-rich genome. The DNA within euchromatin is generally less tightly packed, allowing for the transcriptional machinery to access the genes for expression. Euchromatin is often composed of mostly euchromatic histones, which are generally less highly modified and more susceptible to acetylation, a modification that contributes to the more relaxed, open chromatin structure.

Heterochromatin, on the other hand, is a more tightly packed form of chromatin that is often associated with silencing of genes and the repression of gene expression. Heterochromatin is generally composed of more compact histones, which are highly modified and often contain histone modifications that are associated with gene repression. Heterochromatin is often found near the centromeres and telomeres of chromosomes, where it helps to ensure proper segregation of chromosomes during cell division.

In summary, euchromatin and heterochromatin differ in their degree of compaction and gene expression, with euchromatin being more open and gene-rich, and heterochromatin being more tightly packed and gene-poor. Euchromatin is often composed of euchromatic histones, which are less highly modified and more susceptible to acetylation, while heterochromatin is often composed of compact histones, which are highly modified and associated with gene repression.

* Describe how chromatin modifications can regulate transcription and list modifications associated with either active transcription or inactive transcription.

Chromatin modifications play a critical role in regulating transcription by affecting the accessibility of DNA to the transcriptional machinery. Acetylation and methylation of histone H3 at lysine 4 are associated with active transcription, while methylation of histone H3 at lysine 9 and deacetylation of histones H3 and H4 are associated with inactive transcription.

* List and describe the different types of repetitive DNA and compare and contrast their unique properties and characteristics.

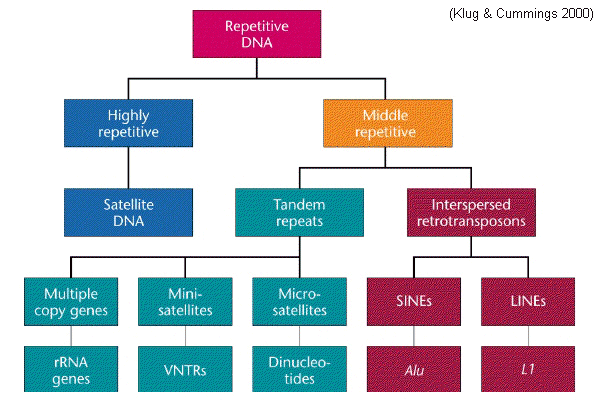
Repetitive DNA refers to DNA sequences that are repeated several times within the genome. Here are the main types of repetitive DNA, each with its unique properties and characteristics:

Satellite DNA: This type of repetitive DNA consists of short, tandemly repeated DNA sequences that are usually several hundred base pairs long. Satellite DNA is often found in the centromeres or telomeres of chromosomes and is associated with heterochromatin, which is a tightly packed, transcriptionally inactive form of chromatin.

Microsatellites: Also known as simple sequence repeats (SSRs), microsatellites are short DNA sequences (1-6 base pairs long) that are repeated several times in a tandem manner. They are commonly used as genetic markers for DNA fingerprinting, linkage mapping, and population genetics studies.

Minisatellites: Minisatellites are similar to microsatellites, but they are longer, usually several tens to a few hundred base pairs long. They also tend to have more complex repeat units, with variations in the number of repeats within the same individual. Minisatellites are also used as genetic markers, but they have lower variability than microsatellites.

Transposable Elements: This type of repetitive DNA consists of sequences that can move (or "transpose") from one location in the genome to another. There are two main types of transposable elements: class I and class II. Class I transposable elements, also known as retrotransposons, replicate via an RNA intermediate, while class II transposons replicate directly via DNA intermediates. Transposable elements are found throughout the genome and can impact gene regulation and contribute to genome evolution and diversity.

Ribosomal DNA (rDNA): This type of repetitive DNA consists of the genes that encode ribosomal RNA, which is a component of ribosomes, the cellular structures responsible for protein synthesis. rDNA is usually organized into tandem arrays and is found in a specific chromosomal locus. The number of rDNA copies in a genome can impact cellular growth and protein synthesis.

In general, repetitive DNA sequences play a crucial role in genome structure, function, and evolution. However, their repetitive nature can also cause problems such as genome instability and the formation of diseases, such as trinucleotide repeat disorders.

* Explain the mechanism of transposition of 1) DNA transposons and 2) retrotransposons.

Transposons are genetic elements that can move or "jump" from one location to another within a genome. There are two main mechanisms of transposition: transposition of DNA transposons and retrotransposition.

Transposition of DNA transposons involves the cut-and-paste mechanism, in which the transposon is excised from its original location in the genome and inserted into a new site. This process is usually mediated by the action of transposase enzymes, which recognize specific sequences at the ends of the transposon and catalyze the excision and integration events.

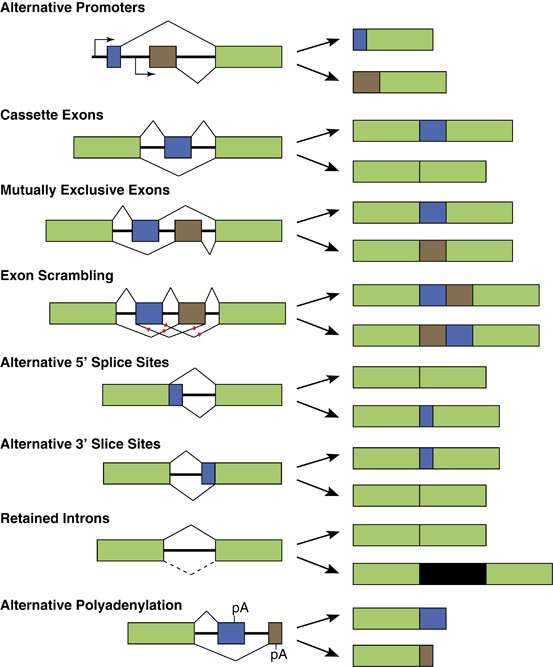
Retrotransposition, on the other hand, is a copy-and-paste mechanism that involves reverse transcription of RNA intermediates. In this process, the transposon is transcribed into RNA, which then reverse-transcribed back into DNA by reverse transcriptase enzymes. The newly synthesized DNA copy is then inserted into a new site in the genome.

In both mechanisms of transposition, the movement of transposons can have various effects on the host genome, including the creation of new mutations, rearrangements, and changes in gene expression. These events can contribute to genetic diversity and evolution, but they can also have negative consequences for the organism, such as the activation of oncogenes or the disruption of essential genes.

* Describe the differences between prokaryotic and eukaryotic gene structure and organization.

|  |  |  |
| --- | --- | --- |
|  | prokaryotes | eukaryotes |
| Chromosome structure | One circular | Multiple linear |
| Gene organization | Operons (clusters of genes that are transcribed and translated together) | genes are dispersed along the chromosomes and are generally transcribed and translated individually |
| Introns and exons | Only exons, no introns | Complex with introns and exons |
| Regulatory elements | Simpler, promotor regions that control transcription initiation | Complex with enhancers and silencers |
| size | smaller | Larger and complexer |

* List, define, draw and explain the following forms of alternative splicing: alternative promoters, cassette exons, mutually exclusive exons, exon scrambling, alternative splice sites, retained introns, alternative polyadenylation.

Alternative promoters: use of different promoter regions to initiate transcription of the same gene.

Cassette exons: exons that are included or excluded from the final mRNA depending on the specific cell type or tissue.

Mutually exclusive exons: exons that are included or excluded from the final mRNA depending on the specific cell type or tissue.

Exon scrambling: rearrangement of exons within a transcript.

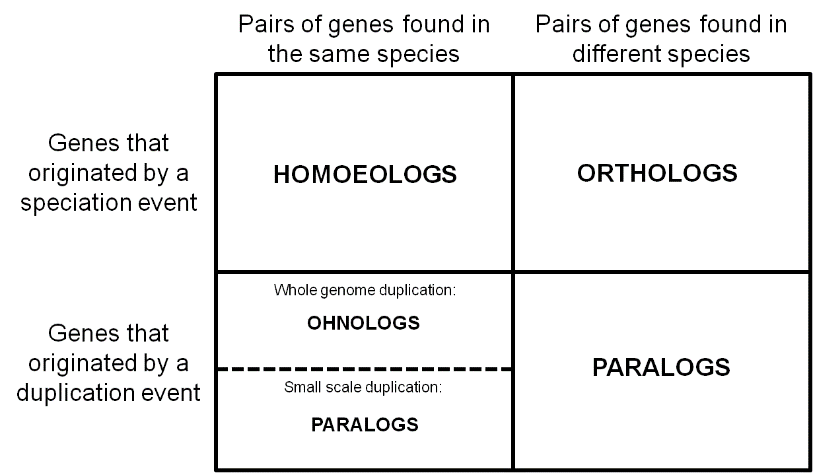
Alternative splice sites: the use of different splice sites to remove introns from the transcript.

Retained introns: introns that are not properly spliced from the transcript.

Alternative polyadenylation: different polyadenylation sites to add a poly(A) tail to the 3' end of the transcript.

* Define and explain the following terms: ortholog, homolog, paralog and describe how these terms relate to each other.

Orthologs are genes or proteins in different species that have evolved from a common ancestral gene and have similar functions. For example, the human gene that codes for insulin is considered an ortholog to the gene that codes for insulin in other mammals, such as mice.

Homologs, on the other hand, are genes or proteins in different organisms that have evolved from a common ancestor and have similar sequences, structures, and/or functions. This can include orthologs as well as other genes that have similar properties, such as those that play a role in the same biological process but have different functions.

Paralogs are genes or proteins that have evolved from a duplication event within the same genome and may have similar or distinct functions. For example, the human genome contains multiple paralogs of the gene coding for hemoglobin, each with a distinct function.

In summary, orthologs are homologs that have a specific function, while paralogs are homologs that have arisen from a duplication event within the same genome.

* Describe how gene families evolve and describe the role of gene and genome duplication in evolution.

Gene duplication is the process by which a single gene is duplicated, creating two identical copies within a genome. This can happen through a variety of mechanisms, including errors in DNA replication, unequal crossing over during meiosis, or transposition of mobile genetic elements. The duplicate genes can then diverge over time, evolving into distinct genes with different functions.

Gene and genome duplication play a key role in the evolution of gene families by providing a mechanism for the creation of new genes and new functions, which can provide a selective advantage in changing environments.

* Define and explain the concept of non-coding RNA.

Non-coding RNA (ncRNA) refers to RNA molecules that are transcribed from DNA but do not code for proteins. Non-coding RNA molecules play a variety of important roles in cellular processes such as regulation of gene expression, RNA splicing, and maintenance of genomic stability.

Non-coding RNA molecules play a variety of important roles in cellular processes and have been shown to be involved in a range of diseases and conditions, including cancer and neurodegenerative disorders. The study of non-coding RNA continues to be an important area of research in the field of molecular biology.

* List different types of non-coding RNA and describe different functions of non-coding RNA.
* Transfer RNA (tRNA): These molecules are involved in translating the genetic information in messenger RNA (mRNA) into proteins. They bring the appropriate amino acids to the ribosome during protein synthesis.
* Ribosomal RNA (rRNA): These molecules are the structural components of ribosomes, which are the molecular machines that carry out protein synthesis. rRNA molecules provide the scaffold for protein synthesis and catalyze the formation of peptide bonds between amino acids.
* Small nuclear RNA (snRNA): These molecules play a role in splicing mRNA, removing non-coding introns and joining exons to form mature mRNA for protein translation.
* Small nucleolar RNA (snoRNA): These molecules play a role in guiding chemical modifications of other RNAs, including ribosomal RNA and transfer RNA.
* MicroRNA (miRNA): These small non-coding RNA molecules regulate gene expression by binding to complementary sites on mRNA, leading to degradation of the mRNA or repression of its translation into protein.
* Long non-coding RNA (lncRNA): These long RNA molecules do not code for proteins, but play a role in regulation of gene expression and other cellular processes. lncRNAs can act as molecular sponges, binding to specific proteins and regulating their activity, or they can act as decoys, sequestering specific RNA-binding proteins to prevent them from binding to other RNAs.
* Piwi-interacting RNA (piRNA): These small non-coding RNA molecules play a role in the defense against transposable elements and the regulation of gene expression.
* Define and explain the terms GC%, CpG islands and DNA methylation and describe how these terms relate to each other.

CpG islands are regions of DNA that are rich in CpG dinucleotides, which are pairs of cytosine and guanine separated by a single phosphate group. CpG islands are typically found in the promoter regions of genes and are often associated with the regulation of gene expression.

DNA methylation is the addition of a methyl group (-CH3) to the cytosine base in a DNA molecule. DNA methylation is an epigenetic modification that affects gene expression and can be heritable. CpG islands are often targets of DNA methylation, with methylation of CpG islands in promoter regions often leading to repression of gene expression.

In conclusion, GC%, CpG islands, and DNA methylation are related in that they are all aspects of DNA that can affect gene expression and cellular function. GC% affects the stability of DNA, CpG islands are often associated with gene regulation, and DNA methylation can modify gene expression by adding a methyl group to specific cytosine residues. These concepts are important for understanding the regulation of gene expression and the molecular basis of cellular function.

* Describe the cause of the underrepresentation of the CpG dinucleotide in the human genome.

The underrepresentation of CpG dinucleotides in the human genome is due to a combination of spontaneous deamination and DNA methylation, which can lead to mutations and changes in gene expression and function. These processes have been shaped by selective pressures over time, leading to the depletion of CpG dinucleotides in the human genome.

* Describe the purpose and possibilities of the following on-line resources and databases: NCBI BLAST, UCSC Genome Browser, OMIM, NCBI taxonomy browser, NCBI Genome, RefSeq, GenBank and Uniprot.

These online resources and databases are important tools for researchers in the fields of genomics, genetics, and biology. They provide access to a wealth of information on a wide range of topics, including genomic sequences, gene function, and evolutionary relationships, and they are essential for a variety of research and analysis applications.

* NCBI BLAST (Basic Local Alignment Search Tool) = sequence alignment, compare a query sequence to a database of known sequences, such as nucleotide or protein sequences, to find similar or identical regions. This tool is useful for identifying the source or origin of a particular sequence, or for finding related sequences for further analysis.
* UCSC Genome Browser is a web-based tool for exploring and visualizing genomic data. The browser provides access to a variety of genomic data sets, including reference genomes, epigenetic marks, and gene expression data, as well as tools for analyzing and annotating the data.
* OMIM (Online Mendelian Inheritance in Man) is a database of information on human genes and genetic disorders. The database includes information on the genetic basis of a wide range of inherited diseases, including information on gene mutations, associated disorders, and treatment options.
* NCBI Taxonomy Browser is a database that provides information on the classification and relationships of living organisms. This database is used to identify and classify species and to explore evolutionary relationships between different organisms.
* NCBI Genome is a database that provides access to genomic information, including complete genomes and associated annotations, for a wide range of organisms. This database is useful for exploring the genetic content of different species and for identifying specific genes and genetic features.
* RefSeq is a database of reference sequences for a variety of organisms, including bacteria, viruses, fungi, plants, and animals. The database provides a standardized and curated set of sequences for use in a variety of genomic and bioinformatics applications.
* GenBank is a public database that provides access to DNA sequences from a wide range of organisms, including bacteria, viruses, fungi, plants, and animals. This database is useful for identifying and accessing DNA sequences for research and analysis.
* Uniprot is a database of protein information that provides information on the function, structure, and evolution of proteins from a wide range of organisms. The database includes information on protein sequences, functions, and interactions, as well as tools for exploring and analyzing protein information.
* Use web-based NCBI BLAST to identify related DNA or protein sequences and interpret (the significance of) the results.
* Explain the concept of the BLAST e-value.

The BLAST e-value is an important measure of the significance of sequence alignments in BLAST search results. It provides a way to rank and filter matches based on the degree of similarity and the likelihood of a match being due to chance, and is an essential tool for analyzing sequence data and identifying meaningful relationships between sequences.

* Use the UCSC Genome Browser to navigate to a gene of interest and to identify gene properties of interest.
* Explain the purpose of BLAT in the UCSC Genome Browser.

BLAT in the UCSC Genome Browser is a valuable tool for researchers in the fields of genomics, genetics, and biology. It provides a fast and efficient way to align sequences to a reference genome, and the results can be easily visualized and analyzed in the context of the genome assembly. This makes it an essential tool for exploring and understanding the relationships between sequences and the genomic regions they occupy.

* Interpret the gene annotation track of the UCSC Genome Browser.
* Use OMIM to identify genetic phenotypes associated with human genes.
* Use the NCBI taxonomy browser to identify the taxonomy of a species of interest.
* Use NCBI Genome to locate genome assembly statistics, such as number and size of chromosomes, of a species of interest.
* Use the RefSeq, GenBank and Uniprot databases to retrieve information about DNA or protein sequences.

Week 5

* List and describe the main principles, advantages, disadvantages and applications of the following sequencing techniques: Illumina, Nanopore and PacBio SMRT Sequencing.

*Illumina Sequencing:*

Illumina sequencing is a high-throughput next-generation sequencing (NGS) technology that generates millions of short DNA sequences simultaneously. The main principles, advantages, disadvantages and applications of Illumina sequencing are:

* Principles: Illumina sequencing works by creating a DNA library, which is then fragmented and amplified using PCR to create clusters of DNA fragments. The clusters are then sequenced by synthesis, where fluorescently-labeled nucleotides are incorporated into the growing DNA strands and detected by a camera.

Afbeelding met tekst, binnen

Automatisch gegenereerde beschrijving

90% of sequencing market

Imaging-based method

Many reads: millions to billions per run

300 to 600 bases per read

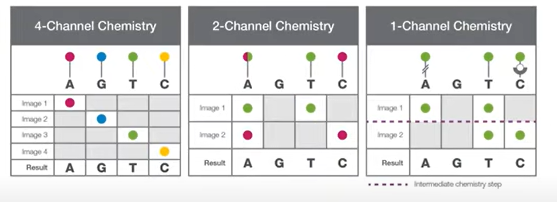
High fidelity: > 99.9% accuracy

$1,000 human genome in 48 hours

Uses flow cells

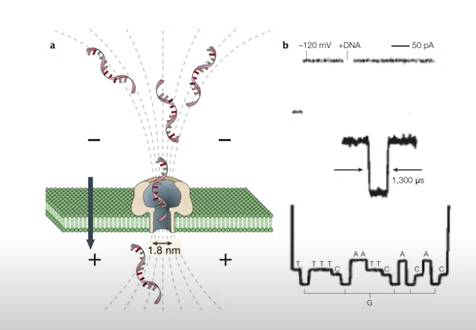
Afbeelding met muur

Automatisch gegenereerde beschrijving

The four colors overlap: so instead of 4 colour sequencing, 2 colour sequencing. Or even one color sequencing.

* Advantages: Illumina sequencing is fast, accurate, and relatively inexpensive, making it a popular choice for large-scale genomics projects. It can generate high coverage of the genome, which is useful for detecting rare mutations and variants. It is also highly scalable, allowing researchers to sequence hundreds of samples simultaneously.
* Disadvantages: Illumina sequencing produces short reads, typically between 50-300 base pairs in length. This can make it difficult to assemble genomes and identify structural variations. It is also prone to errors caused by base-calling inaccuracies, PCR biases, and sequencing artifacts.
* Applications: Illumina sequencing is widely used in a variety of genomics applications, including genome sequencing, transcriptome sequencing, metagenomics, epigenomics, and single-cell sequencing.

*Nanopore Sequencing:*

Nanopore sequencing is a third-generation sequencing technology that detects changes in electrical current as DNA strands pass through a nanopore. The main principles, advantages, disadvantages, and applications of nanopore sequencing are:

* Principles: Nanopore sequencing works by passing a single strand of DNA through a tiny nanopore, which causes changes in the electrical current. The changes are then recorded and used to generate a DNA sequence.
* Advantages: Nanopore sequencing produces **long reads**, typically up to tens of kilobases in length, which allows for the detection of structural variations, including large insertions, deletions, and inversions. It is also **portable** and can be used in field settings, making it useful for on-site sequencing applications. Also **directly sequence RNA**, maybe **protein in the future.**
* Disadvantages: Nanopore sequencing is currently less accurate than Illumina sequencing, with **a higher error rate** caused by both sequencing and sample preparation artifacts. The technology is also **more expensive** than Illumina sequencing.
* Applications: Nanopore sequencing is used in a variety of applications, including genome sequencing, transcriptome sequencing, metagenomics, epigenomics, and real-time monitoring of DNA modifications.

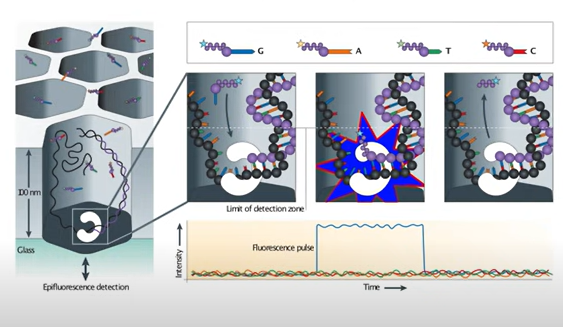
Unlike the Illumina system, this approach does not attempt gapless, end-to-end coverage of a single DNA fragment. Instead it relies on linked reads, in which dispersed, small fragments that are derived from a single long molecule share a communal barcode. Although these fragments leave segments of the original large molecule without any coverage, the gaps are overcome by ensuring that there are many long fragments from the same genomic region in the initial preparation, thus generating a read cloud wherein linked reads from each long fragment can be stacked, combining their individual coverage into an overall map

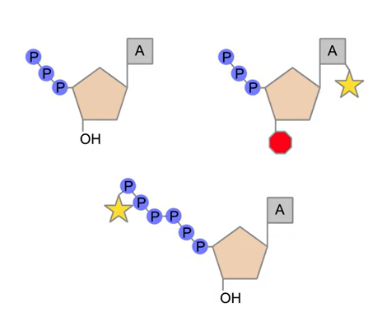
*PacBio SMRT Sequencing:*

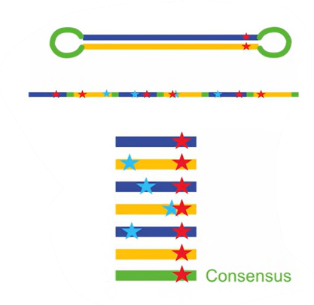
PacBio SMRT sequencing is another third-generation sequencing technology that detects changes in light scattering as DNA strands are sequenced. The main principles, advantages, disadvantages, and applications of PacBio SMRT sequencing are:

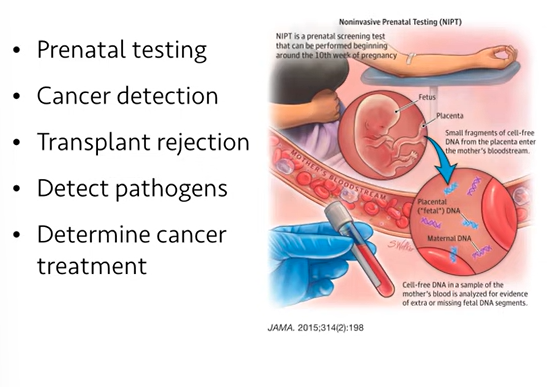
* Principles: PacBio SMRT sequencing works by sequencing a single strand of DNA using a process called single-molecule real-time (SMRT) sequencing. The DNA polymerase used in SMRT sequencing emits a light signal as each nucleotide is incorporated, which is used to generate a DNA sequence.

Afbeelding met tekst

Automatisch gegenereerde beschrijving



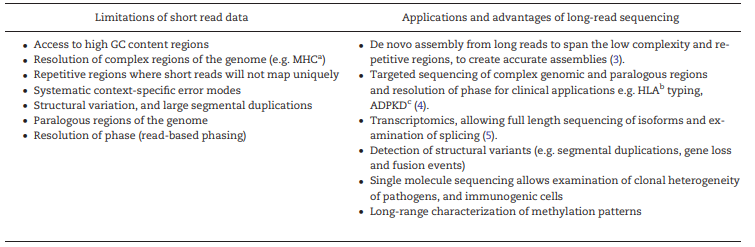
* Advantages: PacBio SMRT sequencing produces **very long reads**, up to tens of kilobases in length, which allows for the detection of structural variations, including large insertions, deletions, and inversions. It is also **highly accurate** and produces low error rates compared to other long-read sequencing technologies.
* Disadvantages: PacBio SMRT sequencing is currently more expensive than Illumina sequencing and nanopore sequencing. It also requires high input DNA quantities and can be sensitive to DNA damage and degradation. **High error rates (10-15%), but random error rates**
* Applications: PacBio SMRT sequencing is used in a variety of applications, including genome sequencing, transcriptome sequencing, metagenomics, epigenomics, and detection of DNA modifications. It is particularly useful for studying complex regions of the genome, such as centromeres.



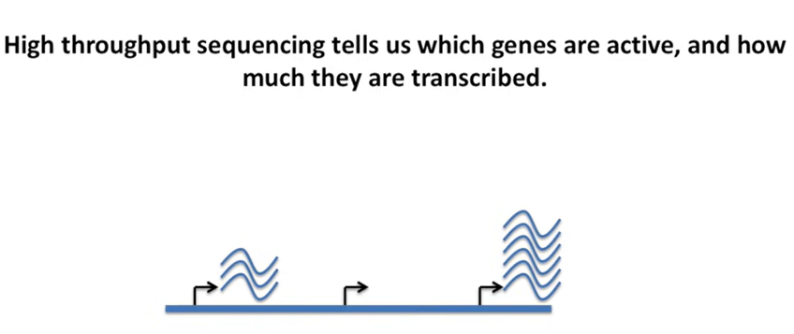


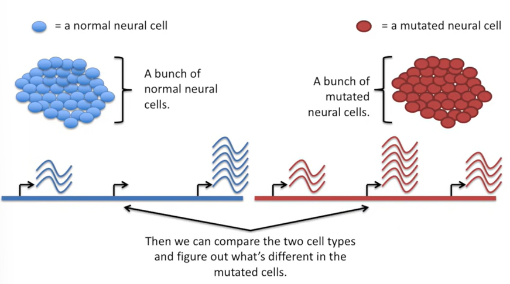
Afbeelding met tafel

Automatisch gegenereerde beschrijving



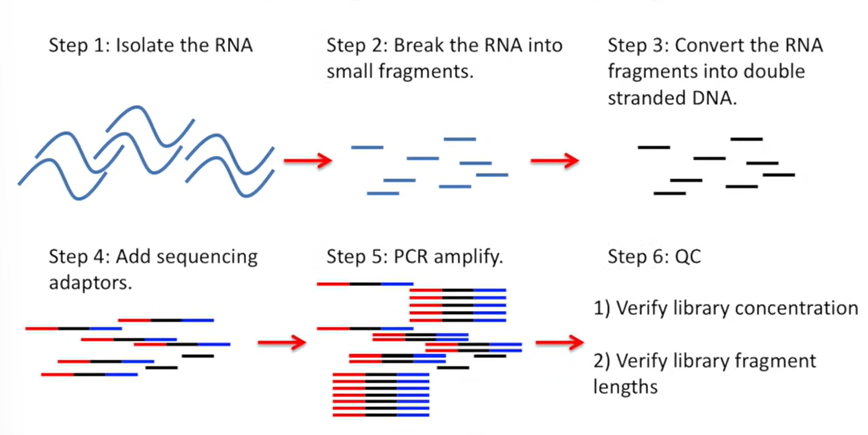
* Describe the main principles of the following applications based on high-throughput sequencing: RNA-seq, allele-specific RNA-seq, metagenomics and whole genome sequencing (WGS).



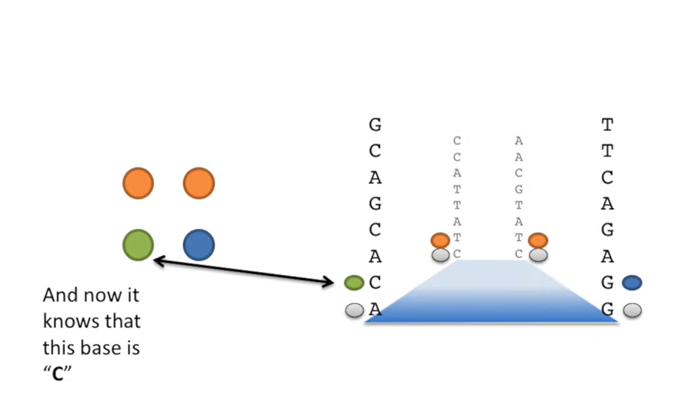


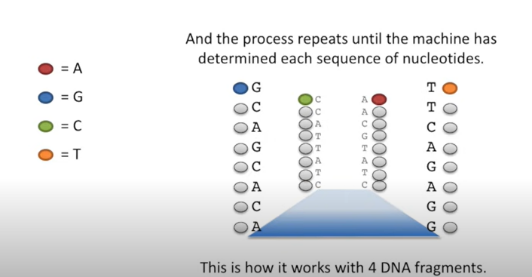
Steps for RNA seq

1. Preparing an RNA-seq library



1. Sequence the library





If there is a low diversity or if the probe doesn’t shine as bright, it will get a low quality score. Together with artifacts of the chemistry these lead to garbage reads, which need to be filtered out.

Reads per gene need to be counted.

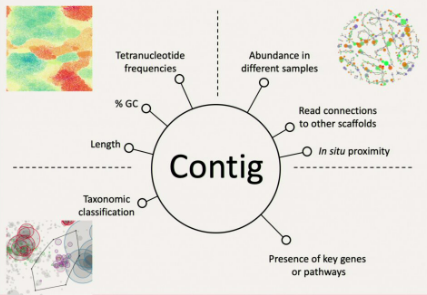
And we need to adjust the read counts per gene to reflect differences in how many reads were assigned to each sample: divide the read counts per gene by the total mapped to each sample.

1. Analyze

Plot the data!!! We use PCA, differences on the x-axis are the most important.

Plotting data tells us if we can expect to find interesting differences and tells us if we should exclude some samples from any downstream analysis.

* Choose the most appropriate sequencing application for a biological question and substantiate your choice.
* Define the term contig.

a contig refers to a contiguous sequence of DNA that has been assembled from overlapping DNA fragments. These fragments are typically generated by breaking down a larger DNA molecule into smaller pieces, which are then sequenced and subsequently aligned to generate a contiguous sequence. Contigs are important for understanding the structure and organization of a genome, as well as for identifying genes and other functional elements within the genome. The term "contig" is derived from the word "contiguous," which means "touching or adjacent in space or time."

* Explain the concept of binning in metagenomics and list and explain different properties that can be used for binning.

Binning is a computational technique used in metagenomics to cluster DNA sequences from mixed microbial communities into groups, known as bins, based on their genomic properties. The goal of binning is to separate the genetic material of different microorganisms and to assign each group to a specific taxonomic identity, thereby enabling the study of the community structure, function, and diversity.

There are several properties that can be used for binning, including:

Sequence composition: This involves analyzing the base composition, such as GC content, of the DNA sequences in order to distinguish between different organisms. This is because the base composition of DNA can vary significantly between different microbial groups.

Sequence coverage: This refers to the number of reads that map to a particular genomic region. Bins can be generated by grouping sequences with similar coverage patterns, as organisms with different abundances are likely to have different coverage depths.

Sequence similarity: This involves comparing DNA sequences to reference genomes or databases to identify similar sequences. Bins can be generated by grouping sequences that share a high degree of similarity, which can be indicative of the same organism or taxonomic group.

Gene content: This involves analyzing the presence and absence of specific genes or gene families within DNA sequences. Bins can be generated by grouping sequences that share similar gene content, as organisms with similar functions are likely to have similar gene repertoires.

Tetranucleotide frequency: This involves analyzing the frequency of four-letter DNA words, or tetranucleotides, within DNA sequences. Bins can be generated by grouping sequences with similar tetranucleotide frequencies, as this can be indicative of the same organism or taxonomic group.

Overall, binning is a powerful tool for analyzing metagenomic data and has enabled researchers to gain insights into the complex microbial communities that inhabit our world. However, it is important to note that binning is not a perfect technique, and there can be challenges in accurately assigning sequences to specific bins due to the high level of genetic diversity within microbial communities.

* Describe the concept of functional annotation in metagenomics

Functional annotation in metagenomics refers to the process of assigning functions or biological roles to the genetic sequences obtained from a microbial community or environmental sample. In metagenomics, researchers use high-throughput sequencing techniques to obtain a large number of DNA or RNA sequences from an environmental sample, without the need for isolating individual organisms. These sequences are then compared to existing databases to identify homologous sequences that have been previously characterized.

Functional annotation in metagenomics typically involves two steps: (1) predicting the open reading frames (ORFs) within the sequenced DNA or RNA, and (2) assigning a putative function to the predicted ORFs based on homology searches against reference databases such as the Gene Ontology (GO) or the Kyoto Encyclopedia of Genes and Genomes (KEGG). Additional tools like functional domain prediction, phylogenetic analyses and other bioinformatics tools can be used to further refine the functional annotation.

Functional annotation is important in metagenomics as it allows us to understand the metabolic potential of the microbial community, identify potential biotechnological applications, and study the role of microorganisms in various environmental processes. It can also help identify novel genes and pathways that may have important biological functions that were previously unknown.

* Define the concept of imprinting and explain how RNA-seq can be used to detect imprinting.

Imprinting is an epigenetic phenomenon where specific genes are expressed in a parent-of-origin specific manner. This means that the expression of these genes depends on whether they were inherited from the mother or the father. Imprinting is established during gametogenesis, and it is maintained throughout development.

RNA sequencing (RNA-seq) can be used to detect imprinting by comparing the expression levels of the same gene in cells or tissues of different parental origin. In practice, RNA-seq data from a biological sample are compared to a reference genome, and differences in expression between the maternal and paternal alleles of a given gene are quantified. These differences can be detected by analyzing the allelic expression ratios (AER) of the transcripts, which are the ratio of the number of reads mapping to each parental allele.

There are several methods for detecting imprinting using RNA-seq data. One common approach is to use single-nucleotide polymorphisms (SNPs) to differentiate between the maternal and paternal alleles. SNPs are positions in the genome where different individuals have different nucleotides. By identifying SNPs that are heterozygous in the parental genomes, it is possible to distinguish between the two alleles in the RNA-seq data.

Another approach for detecting imprinting using RNA-seq is to use allele-specific expression (ASE) analysis. In this method, the expression of each allele is quantified independently, without the need for SNPs. ASE analysis can be more sensitive than SNP-based methods, but it requires high-quality RNA-seq data and a large number of reads to achieve accurate results.

In conclusion, RNA-seq can be used to detect imprinting by analyzing the expression of genes in a parent-of-origin specific manner. This is important for understanding the molecular mechanisms underlying imprinting and for identifying genes that may play a role in developmental disorders and diseases.

* Define the terms genetic variance, SNPs.

Genetic variance refers to the total variation in a trait or phenotype that can be attributed to genetic factors. It is the variation in traits that is due to differences in DNA sequences among individuals. Genetic variance can be further divided into two types: additive genetic variance and non-additive genetic variance. Additive genetic variance refers to the contribution of individual alleles to the trait, while non-additive genetic variance refers to the interactions between alleles or between genes that affect the trait.

Single-nucleotide polymorphisms (SNPs) are one of the most common types of genetic variation. SNPs are single nucleotide substitutions that occur at a specific position in the DNA sequence, where one nucleotide is replaced by another (e.g., A to C, T to G). SNPs can occur in both coding and non-coding regions of the genome, and they can affect gene expression and protein function. SNPs are used extensively in genetic studies to identify associations between genetic variation and various phenotypes, including disease susceptibility, drug response, and behavioral traits.

SNPs are typically identified by comparing the DNA sequences of multiple individuals or populations using high-throughput sequencing technologies. In a typical SNP analysis, the DNA sequences of individuals are aligned to a reference genome, and differences in the nucleotide sequence at specific positions are identified. SNPs can be used to genotype individuals, which involves determining which specific alleles they carry at specific SNP loci. Genotyping is commonly used in association studies to identify genetic variants that are associated with a particular trait or disease. SNPs are also used in population genetics studies to infer the demographic history of populations and to identify genes that have undergone positive selection during evolution.

* Describe advantages and limitation of RNA-seq for diagnostic use, as compared to whole-genome or exome sequencing.

RNA-seq, whole-genome sequencing (WGS), and exome sequencing (ES) are all powerful tools for analyzing genetic variation and gene expression. However, they differ in several ways, including their advantages and limitations for diagnostic use.

Advantages of RNA-seq for diagnostic use:

1. Gene expression analysis: RNA-seq can provide information about the expression levels of thousands of genes in a single experiment, allowing for the identification of differentially expressed genes in disease samples compared to controls. This can provide insights into the molecular mechanisms underlying disease and help identify potential therapeutic targets.
2. Detection of alternative splicing and fusion genes: RNA-seq can identify alternative splicing events and fusion genes that can contribute to disease development and progression.
3. Cost-effectiveness: RNA-seq is generally less expensive than WGS or ES, as it only requires sequencing of the transcribed portion of the genome, rather than the entire genome.

Limitations of RNA-seq for diagnostic use:

1. Limited coverage of non-coding regions: RNA-seq primarily focuses on the transcribed portion of the genome, which means that non-coding regions may be underrepresented or missed entirely.
2. Differential gene expression in different tissues: Gene expression levels can vary between different tissues or even between different regions of the same tissue, which may lead to misinterpretation of results if the sample used for RNA-seq does not represent the tissue or region of interest.

Advantages of WGS and ES for diagnostic use:

1. Comprehensive coverage of the genome: WGS and ES provide a comprehensive view of the entire genome, including non-coding regions, which can help identify genetic variants that may be responsible for disease.
2. Detection of de novo mutations: WGS and ES can detect de novo mutations, which can provide important information for families with genetic diseases.

Limitations of WGS and ES for diagnostic use:

1. High cost: WGS and ES are generally more expensive than RNA-seq due to the need to sequence the entire genome or exome.
2. Difficulty in interpreting non-coding variants: While WGS and ES provide comprehensive coverage of the genome, the interpretation of non-coding variants is still a challenge, and their functional significance may be unclear.