Genetics and biochemistry for OMICs:

cheat sheet

This document is for people who are studying OMICs in a computational, big-data context and who haven’t got a background in biology or genetics. It is intended to explain some key concepts, and how those concepts link to OMICs, in a basic way, to give you the minimum viable amount of understanding to start working with OMICs data.

# The ‘central dogma of molecular biology’

You’ll see a lot of slides with this phrase on it. The ‘central dogma’ describes how information contained in DNA flows into proteins. First, DNA is *transcribed* into RNA, then RNA is *translated* into proteins. Proteins are the chemical worker bees in cells. They execute the instructions contained in DNA.

**The central dogma of molecular biology**

(Image from Khan Academy)

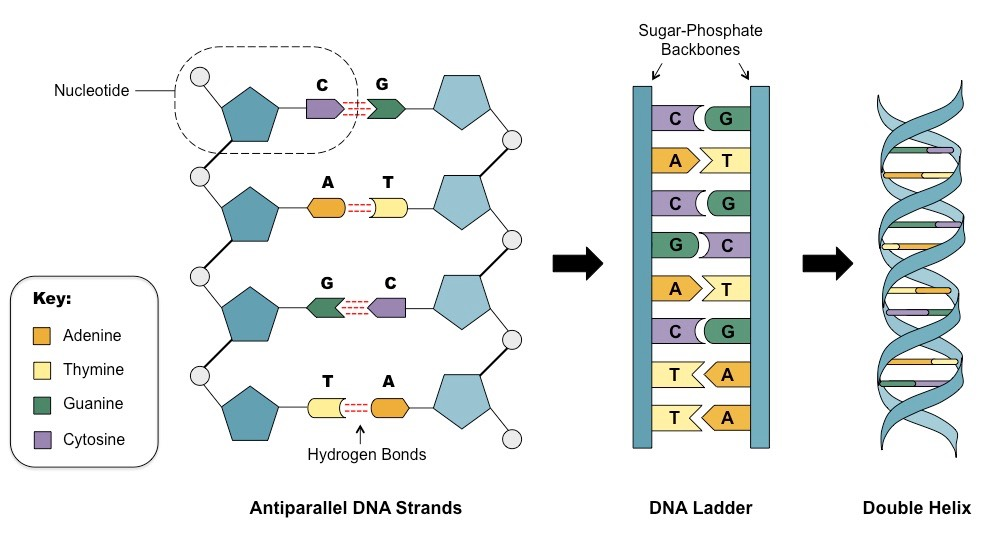
Understanding the chemical processes is not especially important. It’s just crucial to grasp that there is a flow of information going from DNA into other molecules in cells, which then go on to perform functions.

Each ‘level’ of OMICs involves studying a particular group of molecules at a single stage on this pathway. There are a lot of commonalities between the methods of data gathering and analysis for each OMIC, but there are also some important differences. Below is a brief explanation of each of the main OMICs, with the key terms explained. Afterwards there is a table with some information about the data you will encounter in each field of OMICs.

# Genetics

## Nucleotides, SNPs, DNA, genes, and chromosomes

The genome is the complete set of DNA in an organism. Each cell in an organism contains a copy of the genome. A simple, if a little reductive, analogy is to think of the genome like a book.

**Nucleotides** are the letters on the page – the smallest meaningful unit of variation. They can only take four forms (G,A,T,C), and each possible variant is called an allele. Because of the double helix nature of DNA strands, nucleotides are always ‘paired’ with another nucleotide on the opposing DNA strand.   
Ts are always paired with As, and Cs are always paired with Gs.

**Genes** are paragraphs – groups of nucleotides that can be meaningfully interpreted when grouped together.

**DNA** is just a collective term for all the genes / nucleotides.

And **chromosomes** are chapters.

So chromosomes are made of genes and genes are made of DNA (nucleotides). Understanding of the chemical properties of DNA is not essential to conducting OMICs analysis.

**SNPs** are “single nucleotide polymorphisms”. It’s just a way of describing a nucleotide base pair that has been shown to vary in at least 1% of individuals within a given population. The majority of our genomes are identical and very rarely vary between individuals, so knowing which places to look for variation substantially speeds up genetic studies.

**The double helix structure of DNA.**

Nucleotides are paired together to form base pairs. Cs always pair with Gs. Ts always pair with As.

Image from https://socratic.org/questions/5a5c59b4b72cff5dc9754fa8

Gene sequencing and genotyping

These are both ways of ‘reading’ the book.

**Gene sequencing** involves reading every single letter (approximately 3.2 billion nucleotide base pairs in the human genome), and is costly and time consuming.

**Genotyping** is faster, and involves only looking at a subset of SNPs with so-called microarrays / bead chips. More than 335 million SNPs are known in the human genome, and generally a subset of between 500k and 10 million are selected for study, on the basis of those which are known to vary the most in the population, or those that are most relevant to known diseases or phenotypes.

Genotyping is what 23andMe and other similar companies do, and is how most genetic data for OMICs studies is gathered.

**GWAS** studies involve genotyping a cohort of people to look at how their alleles vary across a subset of SNPs. The most common set-up for a GWAS is a case-control, where you have a group of people with a phenotype/disease, and a group without, and you run a huge sequence of – fairly simple – identical statistical tests on all the SNPs to see if any of them are statistically significantly different between the case group and the control group. GWAS is explained in more detail in the table below.

# Epigenetics

To continue the book metaphor, epigenetics refers to a process by which bits of text get tippexed over, usually via a process called **methylation*.*** Again, understanding the chemistry is not essential for understanding the principles and the data.

This methylation process happens in parts of the genome called **CpG sites,** where a C nucleotide is followed by a G nucleotide. For chemical reasons, only CpGs can be methylated, and not other nucleotide base pairs. There are approximately 28 million CpG sites in the human genome, and they usually group together in **CpG islands**. Genes can be ‘silenced’ – meaning that they don’t perform their usual function – if they have large areas of methylated CpG islands, which is why it is of interest to study.

**Epigenome wide association studies (EWAS)** use similar methodology to GWAS, but rather than looking at SNPs, they are measuring the levels of methylation at CpG sites. Methylation levels are measured as proportions, between 0 and 1.

# Transcriptomics

**CpG sites on a CpG island**

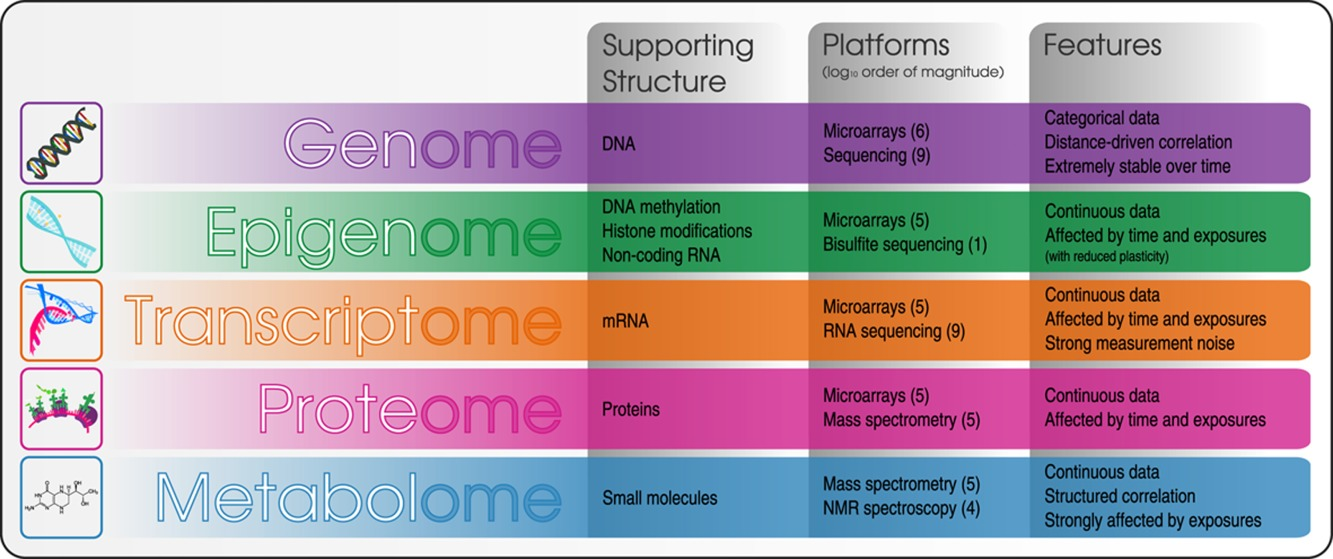
(Image from Wikipedia)

Transcriptomics is the study of gene expression – the extent to which a given gene is being transcribed into RNA at any given moment in a cell. There are different types of RNA, but **messenger RNA** (mRNA) is the type that is largely responsible for making proteins, so mRNA is generally studied in transcriptomics.

# Proteomics

The study of all the proteins in cells.

# Metabolomics

The study of all the metabolites in cells / blood. 

**Chart from:**

https://onlinelibrary.wiley.com/doi/full/10.1002/em.21797

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **OMIC** | **Molecules studied** | **Method of data gathering** | **What does the data look like?** | **Statistical methods** | **Environmental influence?** |
| **Genome**  GWAS:  Genome Wide Association Studies | DNA: SNPs – nucleotide base pairs that are known to vary within a population. Usually around 500k to 10million are studied in a GWAS | Chip-based microarrays (often machines from Illumina or Affymetrix) | A wide data set with 500k–10million variables (SNPs).  The SNPs are really categorical variables, but they are generally treated as continuous variables, taking values of 0 (non-carrier), 1 (heterozygous) and 2 (homozygous).  This represents the number of alternate alleles for this specific SNP (“mutation”): 0 means both chromosomes carry the reference allele, 1 means one reference and one alternate allele, 2 means both chromosomes carry the alternate allele. | Generally, univariate linear mixed models are run for each SNP individually, with the outcome variable being the phenotype (case or control, in the most common cases). Either logistic regression of Chi2 for allele count can be used.  SnPs are then matched back to their chromosomal position so that the gene in which they appear can be identified.  The results are commonly presented in a manhattan plot, where the -log10 p value for each individual test is plotted on the y-axis, with the SNP on the x-axis, organised by its chromosomal position. Each point on the plot is a SNP.    Multivariate methods also applied when the combined effects of multiple SnPs / genes is under investigation. | DNA is heritable and very stable. However, it can be changed for example through cancerogenic environmental influences like smoking, UV-light or radiation.  Epigenetic alterations can, however, change the way that DNA is transcribed into mRNA (see below). |
| **Epigenome**  EWAS: Epigenome Wide Association Studies | CpG sites (usually) | Usually the Illumina Infinium 450k / EPIC (850k) array | A wide data set with around 450,000 / 850,000 variables (each one is a CpG site). This is still only around 2% of the total human epigenome.  Values range between 0 and 1, representing the proportion of methylation at a given CpG site. | Similar to GWAS. Univariate results may also be presented in a Manhattan plot.  Approaches will depend on whether you are examining the epigenomic ‘response’ to environmental factors, or whether you are looking to identify CpGs that are markers of a particular disease or phenotype. Multivariate approaches may be more appropriate for the latter.  For a full run-down of statistical approaches, see <https://onlinelibrary.wiley.com/doi/full/10.1002/em.21797> | Methylation can be influenced by multiple environmental factors – stress, lifestyle, nutrient intake, etc.  Methylation levels also change over a person’s life course, and some CpGs are known to follow consistent trajectories, meaning that an ‘epigenetic clock’ can be used to measure age.  Some epigenetic modifications can be transmitted to offspring, meaning that, contrary to what was once thought, there is a mechanism for heritability beyond the raw genetic code of DNA. |
| **Transcriptome**  Transcriptomic studies | mRNA | Chip-based microarrays | A wide data set with ~40,000 variables (each one measures levels of a particular transcript).  Genes can encode more than one transcript, hence humans have 20–25k genes, and 40k+ transcripts. | Similar to above. Manhattan plots not generally used for transcriptomic studies as a single gene can produce >1 transcript.  Results of univariate analysis often presented in volcano plots (below), where the x-axis is the regression coefficient and the y axis is the pvalue (-log10, so higher value = more significant). | As with epigenetics, the transcriptome is altered by exposures and changes over time.Also, different cell types and organs are differentiated and while sharing they same genome, they have distinct transcription profiles. |
| **Proteomics**  Proteomic studies | Proteins | Mass specrometry / Microarrays | A wide data set with ~10k proteins.  Each data point is a continuous numerical value representing the mass-to-charge ratio of the given protein, which is a way of quantifying the protein molecules.  Values are scaled normalised so they should be roughly standard normally distributed  (mean = 0, SD=1)  Easily accessible data set available here:  <https://www.kaggle.com/piotrgrabo/breastcancerproteomes> | Similar to above. | The proteome is altered by exposure and over time. |
| **Metabolomics** | Metabolites | Mass spec /NMR spectroscopy | A wide data set of 15–20k metabolites.  As with proteins, data points usually represent normalised mass/charge ratio measurements. Expected to be standard normal.  Data sets available here: <http://www.metabolomexchange.org/site/> | Similar to above | The metabolome is altered by exposure and over time. |

## Useful links

**Explanations of concepts related to 1) DNA**

1. Basic explanation of the genome:

<https://www.genome.gov/About-Genomics/Introduction-to-Genomics>

1. DNA Sequencing:

<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet>

1. DNA Microarray Technology:

<https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology>

1. Basic explanation of GWAS:

<https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet>

1. More in-depth explanation of GWAS

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531285/>

1. List of all GWAS published

<https://www.ebi.ac.uk/gwas/>

1. Polygenic risk scores:

<https://www.nature.com/articles/d42473-019-00270-w>

1. Linkage disequilibrium:

<https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-16483-5_3368#:~:text=Linkage%20disequilibrium%20refers%20to%20the,occur%20at%20the%20expected%20frequencies>

1. Types of genetic variation:

<https://www.ebi.ac.uk/training-beta/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/types-of-genetic-variation/>

1. Solving Hardy-Weinberg problems:

<https://www.youtube.com/watch?v=xPkOAnK20kw>

**Explanations of concepts related to 2) RNA**

Transcriptome fact sheet:

<https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet>

**Explanations of concepts related to 3) Protein**

What is proteomics?

<https://www.ebi.ac.uk/training/online/course/proteomics-introduction-ebi-resources/what-proteomics>

**Glossary:**

<https://www.genome.gov/genetics-glossary>

**A few useful videos:**

<https://www.youtube.com/playlist?list=PL1wN5MzqlxD3PxdkjbAH6Vc7zEP6t45MY>

**Methodological overview of all OMICs:**

<https://onlinelibrary.wiley.com/doi/full/10.1002/em.21797>

**For further reading see chapters 1 to 7 of**

<https://link-springer-com.iclibezp1.cc.ic.ac.uk/book/10.1007%2F978-1-4419-7338-2>