A short introduction to ‘–omics’

Gianluca Campanella

Contents

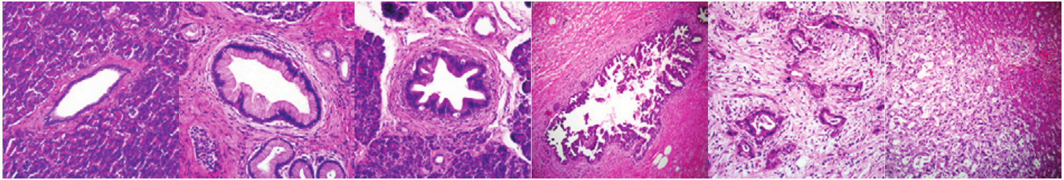
Molecular epidemiology

Technologies

Lessons learned

Molecular epidemiology

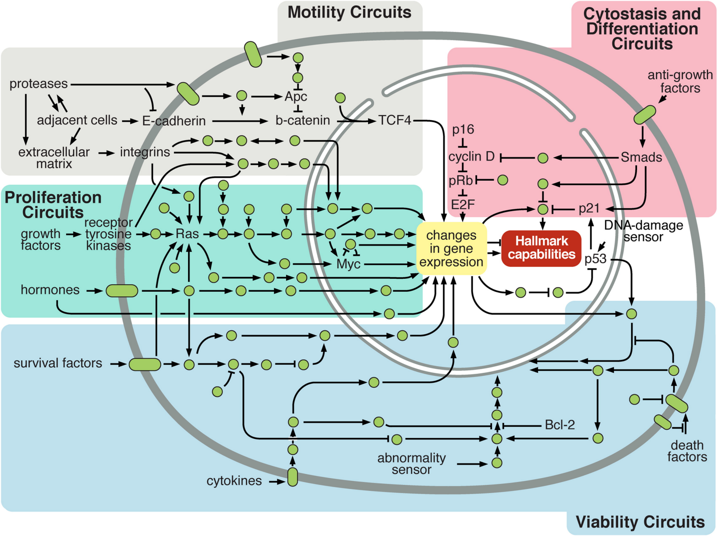
‘Hallmarks of cancer’ (Hanahan and Weinberg, 2011)



Normal PanIN-1 PanIN-2 PanIN-3 Cancer Metastasis

- Inducing angiogenesis
- Resisting cell death
- Enabling replicative immortality
- Sustaining proliferative signalling
- Evading growth suppressors
- ...

We haven't figured it all out…



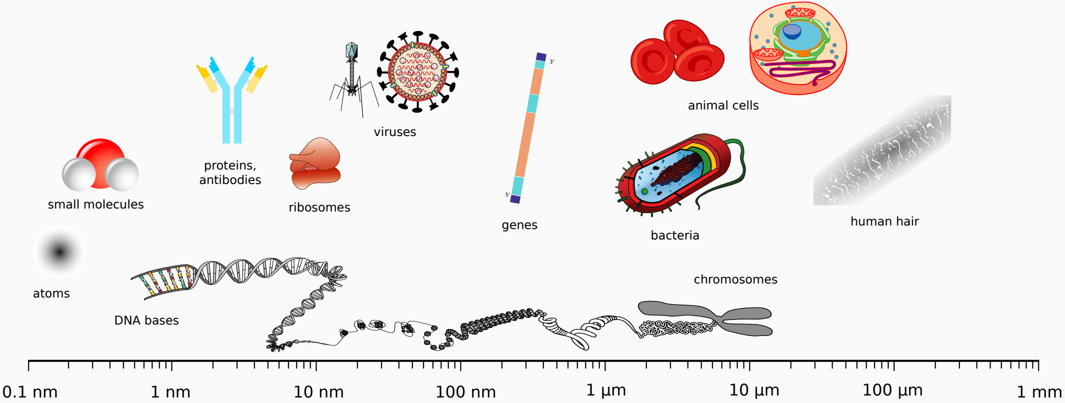
Complex (or multifactorial) diseases

- Do not have a single genetic cause
- Likely associated with the effects of:
  - Multiple genes
  - Lifestyle and environmental factors
  - Foetal programming?

Compare with:

- Genetic disorders
- Infectious diseases

Biological scale



The ‘central dogma’ of molecular biology

‘DNA makes RNA makes proteins’

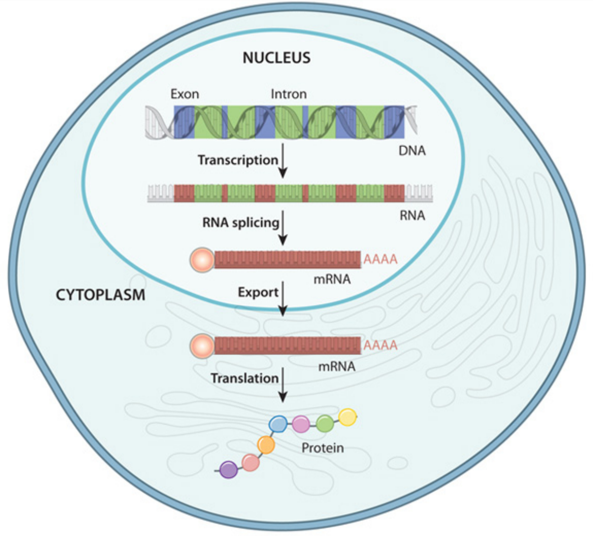
General transfers

1. Replication (DNA → DNA)
2. Transcription (DNA → RNA)
3. Translation (RNA → proteins)

Special transfers

1. Reverse transcription (RNA → DNA)
2. RNA replication (RNA → RNA)

Information flow: DNA → RNA → proteins



Regulation of gene expression

Transcriptional regulation

- *Cis*/*trans* regulation
- Epigenetics (DNA methylation and histone modifications)

Post-transcriptional regulation

- Co-transcriptional modification
- miRNAs

Post-translational regulation

- Modification (reversible)
- Degradation (irreversible)

Epigenetics

DNA methylation

- Methyl groups (–CH_3) attached to cytosines
- Usually (but not exclusively) at C followed by G (CpG loci)
- Most CpG loci clustered in dense ‘CpG islands’
- Effect on transcription dependent on location

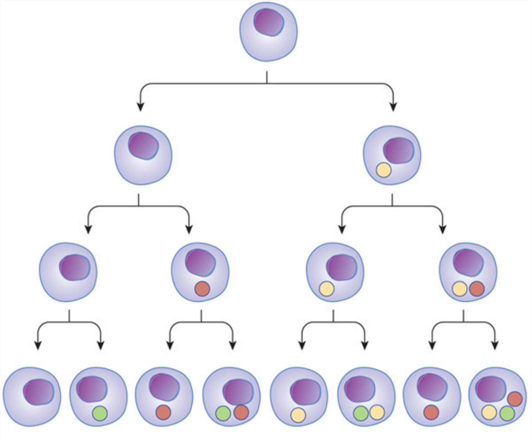
Histone methylation and acetylation

- Methyl/acetyl groups (–$COCH_3$) attached to histone tails
- Very complex (combinatorial) effects on transcription

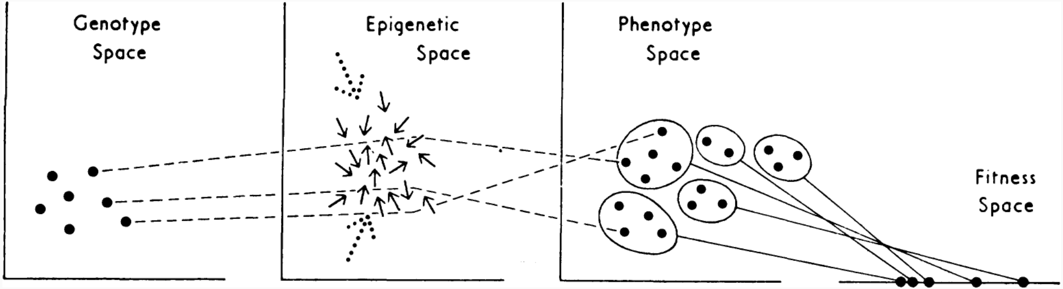
Why regulate gene expression?

Differentiation
into different cell types

Response to
acute and chronic stress

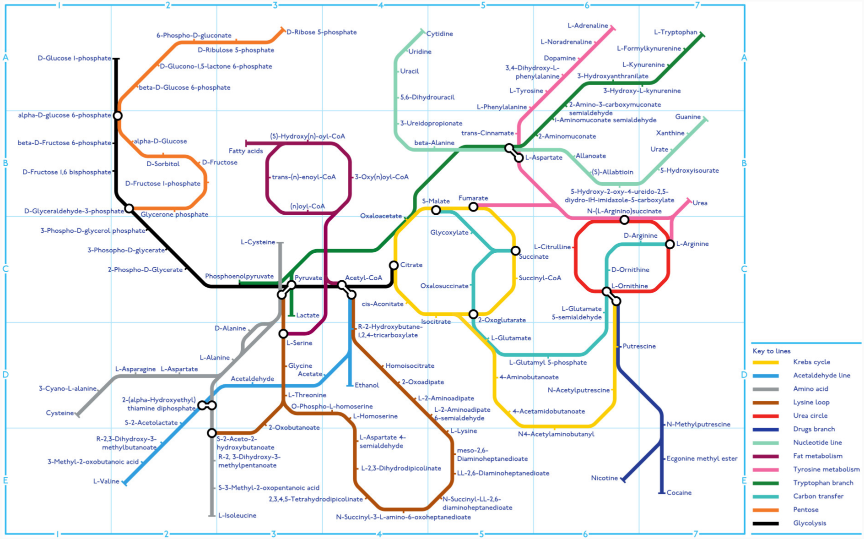


Complex diseases: deregulation of information flow?



From Scarr and McCartney (1983)

Complex diseases: there is more...



Of ‘–omes’ and ‘–omics’…

‘–ome’

Forming nouns with the sense ‘all of the specified constituents of a cell, considered collectively or in total’

‘–omics’: the study of ‘–omes’?

- Collective characterization of the building blocks of structure, function, and dynamics of organisms
- ~~Hypothesis-free~~ Agnostic

Technologies

Overview

	Supporting Structure	Platforms (log$_{10}$ order of magnitude)	Features
Genome	DNA	Microarrays (6) Sequencing (9)	Categorical data Distance-driven correlation Extremely stable over time
Epigenome	DNA methylation Histone modifications Non-coding RNA	Microarrays (5) Bisulfite sequencing (1)	Continuous data Affected by time and exposures (with reduced plasticity)
Transcriptome	mRNA	Microarrays (5) RNA sequencing (9)	Continuous data Affected by time and exposures Strong measurement noise
Proteome	Proteins	Microarrays (5) Mass spectrometry (5)	Continuous data Affected by time and exposures
Metabolome	Small molecules	Mass spectrometry (5) NMR spectroscopy (4)	Continuous data Structured correlation Strongly affected by exposures

Genomics

Methods

- Targeted:
  - Single-nucleotide polymorphisms (SNPs)
  - Copy-number variations (CNVs)
- Partly targeted: exome sequencing
- Untargeted: whole genome sequencing

Genomics

Outputs

- Targeted: alleles or copy number
  → Statistical analysis is straightforward
- Partly targeted and untargeted: sequence reads
  → Must map to reference genome

Epigenomics: DNA methylation

Method

1. Create polymorphisms at methylated cytosines using bisulphite conversion (C $\rightarrow$ U/T, me-C $\rightarrow$ C)
2. Use genomic methods

Epigenomics: DNA methylation

Output

- Percentage of methylated cytosines at each CpG locus
  → Statistical analysis is (more or less) straightforward
- Average over many cells, possibly of different types
- Sequence reads must again be mapped to reference genome after *in silico* ‘bisulphite conversion’

Transcriptomics (and miRNAs)

Methods

- Targeted: micro-arrays
- Untargeted: RNA sequencing (RNA-seq)

Transcriptomics (and miRNAs)

Outputs

- Targeted: intensities proportional to RNA abundances
  → Statistical analysis is straightforward
- Untargeted: sequence reads
  → Must map to reference transcriptome
  → Must take into account splicing

Proteomics and metabolomics

Methods

- Targeted: mass spectrometry assays
- Untargeted: mass spectrometry and NMR spectroscopy

Proteomics and metabolomics

Outputs

- Targeted: quantified proteins/metabolites
- Untargeted: mass and retention times, or spectra

→ Statistical analysis is straightforward, but unknown compounds from untargeted studies may be very difficult to identify

Lessons learned

Know your biology

You need some knowledge of the biological process
if you are to model it meaningfully

- Aim to grasp the subject decently: get a good biology textbook if needed, and ask questions
- Find out which questions are still unanswered: they make great hypotheses to test in your dataset

Know your technology

You need some knowledge of the measurement procedure
if you are to model it meaningfully

- Read the manuals, possibly several times
- Understand what is being measured, and how
- Be aware of quirks in the design!

The plague of batch effects

Protocols are tedious and involve many complex (and often complicated) steps that will introduce nuisance variation

1. Record as much information as possible
2. Identify influential factors (QC)
3. Attenuate by means of preprocessing
4. Model any residual confounding

Know your statistics

You need some knowledge of statistical modelling
if you are to write down a model

- What is your question?
- What assumptions can you reasonably make (and verify)?
- What type of data do you have at hand?
- Explore different options, but be careful when borrowing methods from other fields

Trust, but verify

- Women with Y chromosome
- Controls with date of diagnosis
- ‘Matched’ pairs with huge age differences
- Secondary instead of primary cancer
- Technical replicates with different genotype

Always check: it takes little time, and saves future headaches

Know your computer science

You need some knowledge of programming
if you are working with ‘-omics’ data

Given the sheer amount of data, we must standardise and automate statistical analysis as much as possible

Validation and replication

No matter how stringent your QC and preprocessing, and how accurate your models, false positive results will still occur

Validation

Are results reliable? Repeat the experiment using the same samples, but a different lab technique

Replication

Are results generalisable? Reproduce the findings using different samples, and possibly a different lab technique

Summary

- Complex diseases as deregulation of information flow
- The ‘-omics’ paradigm: a holistic point of view
- Multidisciplinarity:
    - Biological processes
    - Measurement procedures
    - Statistical modelling
    - Computer science

Opportunities

1. Identification of novel biomarkers for:
    - Disease risk
    - Exposures
    - ‘Meet-in-the-middle’ approach
2. Understanding at a molecular level of:
    - Disease states
    - Exposures
3. Characterisation of dynamic molecular environment
4. Development of new treatments

Biomedical challenges

- **Holistic view**
  What is the effect of multiple '-omics' markers?
- **Tissue heterogeneity**
  What is the value of '-omics' measurements in samples that contain multiple, heterogeneous cell types?
- **Surrogate tissues**
  What is the value of '-omics' measurements in surrogate tissues, e.g. in blood, for localised diseases?
- **Effect sizes**
  What is the magnitude of clinically significant changes?

Statistical challenges

- **Multiple comparisons**
  What significance threshold should be used when performing millions of tests simultaneously?
- **Nuisance variation**
  How can we distinguish between biological and technical variation?
- **Combined effects**
  How can we model the combined effect of multiple ‘-omics’ markers?
- **‘Crossomics’**
  How can we analyse multiple ‘-omics’ datasets jointly?