Gene expression technologies

Bioinformatics · Universitat Politècnica de Catalunya

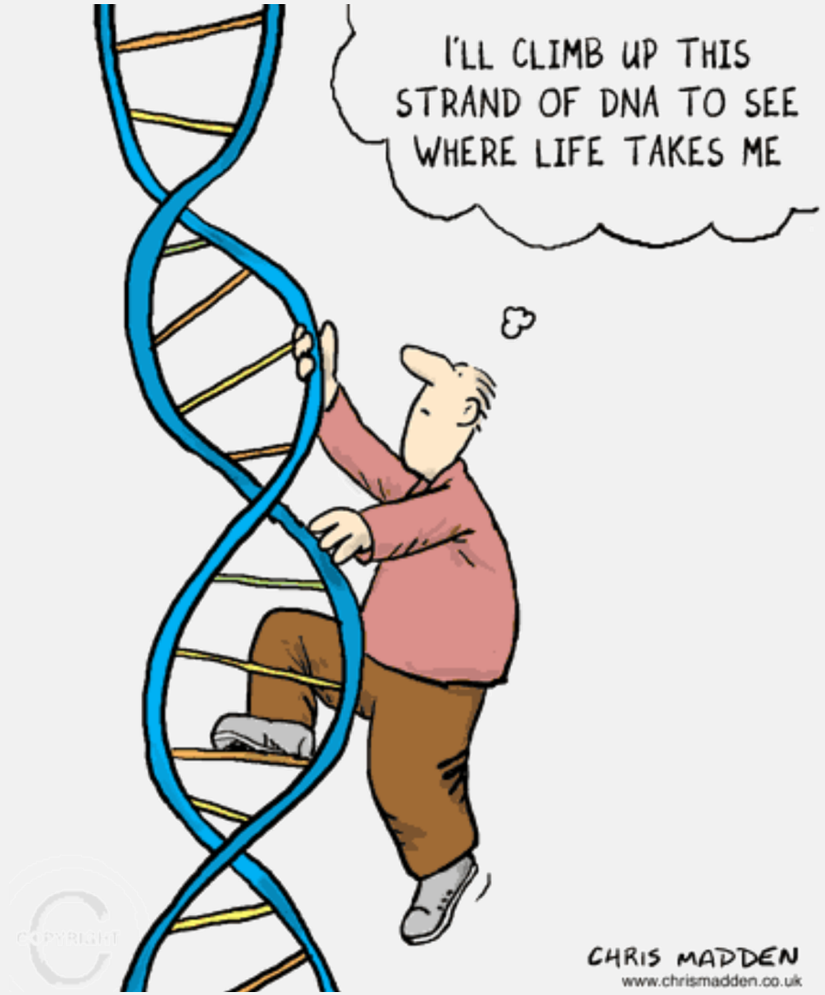
Alexandre Perera-Lluna

From transcripts to measurements

1. Gene regulation and alternative splicing
2. Microarrays: hybridisation, fluorescence and probe manufacture
3. PCR → library preparation → RNA sequencing
4. From reads to biological conclusions

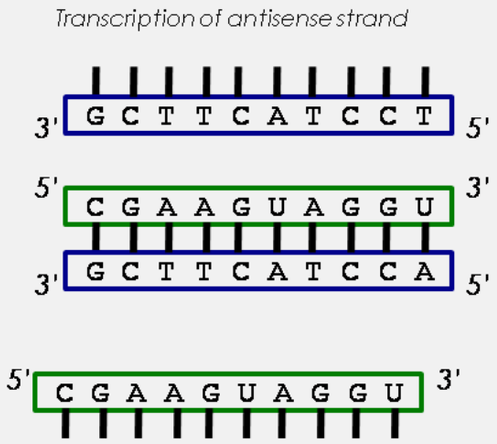
Learning goal: explain what each technology measures—and what it cannot tell us on its own.

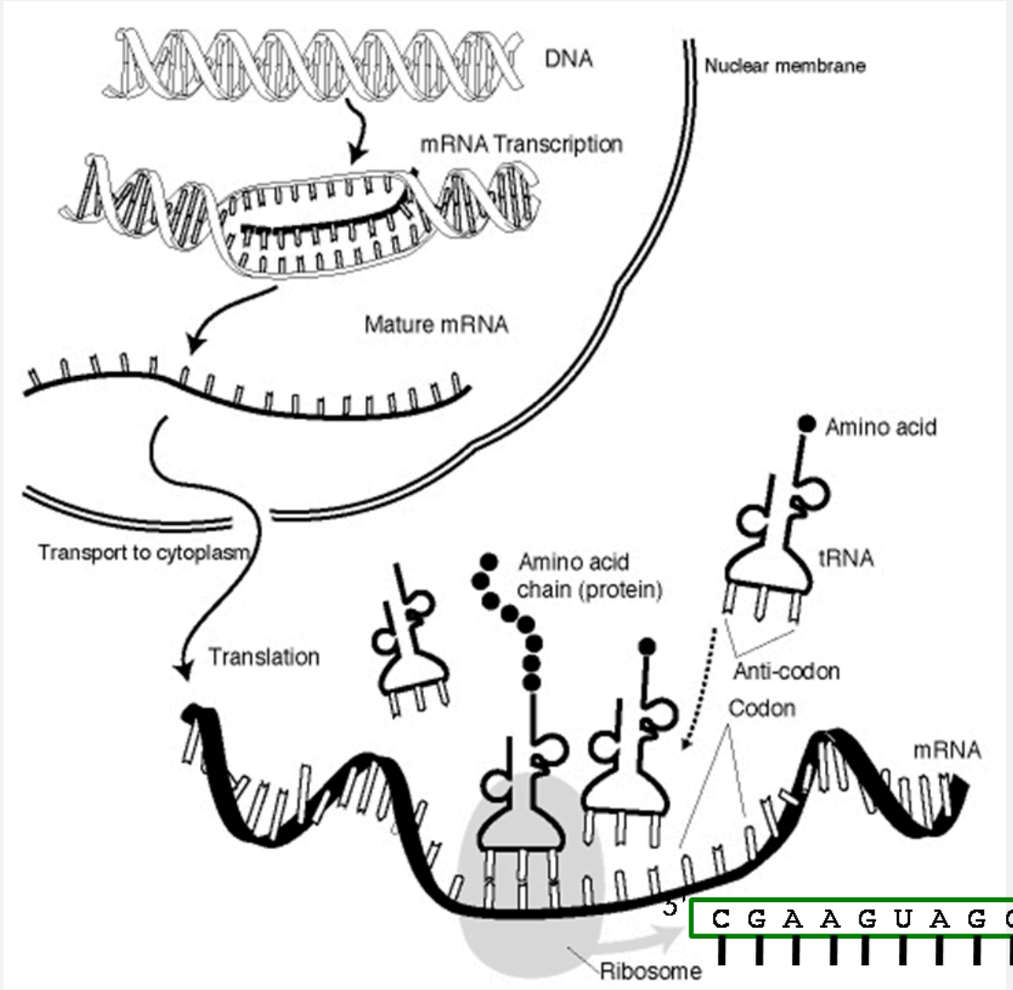
We still have to find a way to answer



from http://chrismadden.co.uk

Notes on Gene Transcription



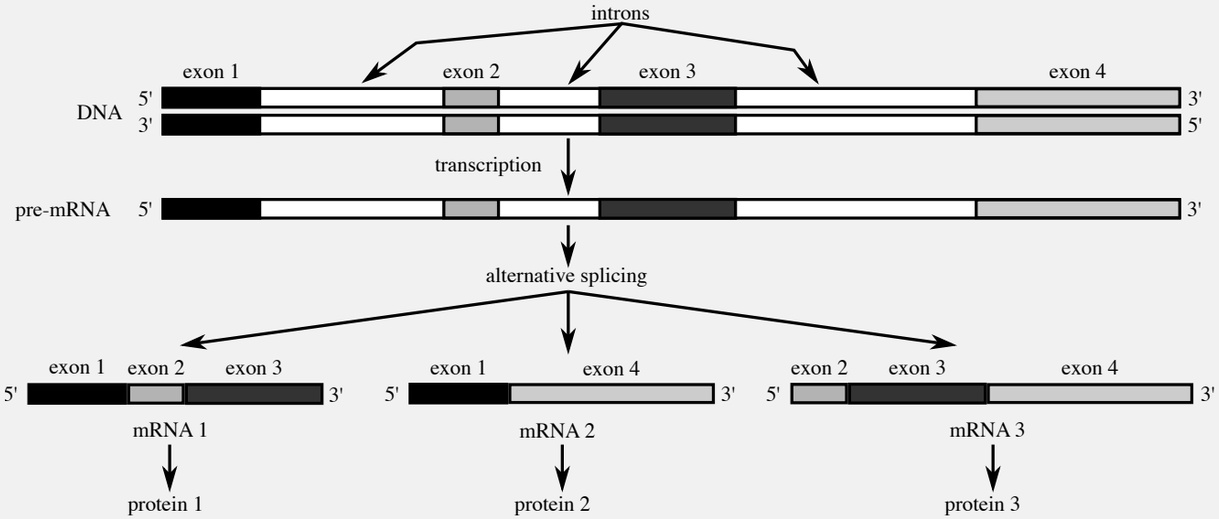


Cytosine
Thymine
Adenine
Guanine
Uracil

One gene can produce several RNA isoforms

- Transcription produces a **pre-mRNA** containing exons and introns.
- Splicing joins selected exon segments and removes intervening sequences.
- Different splice choices create different **transcript isoforms**.

Exon does not mean protein-coding: exons can contain untranslated regions.



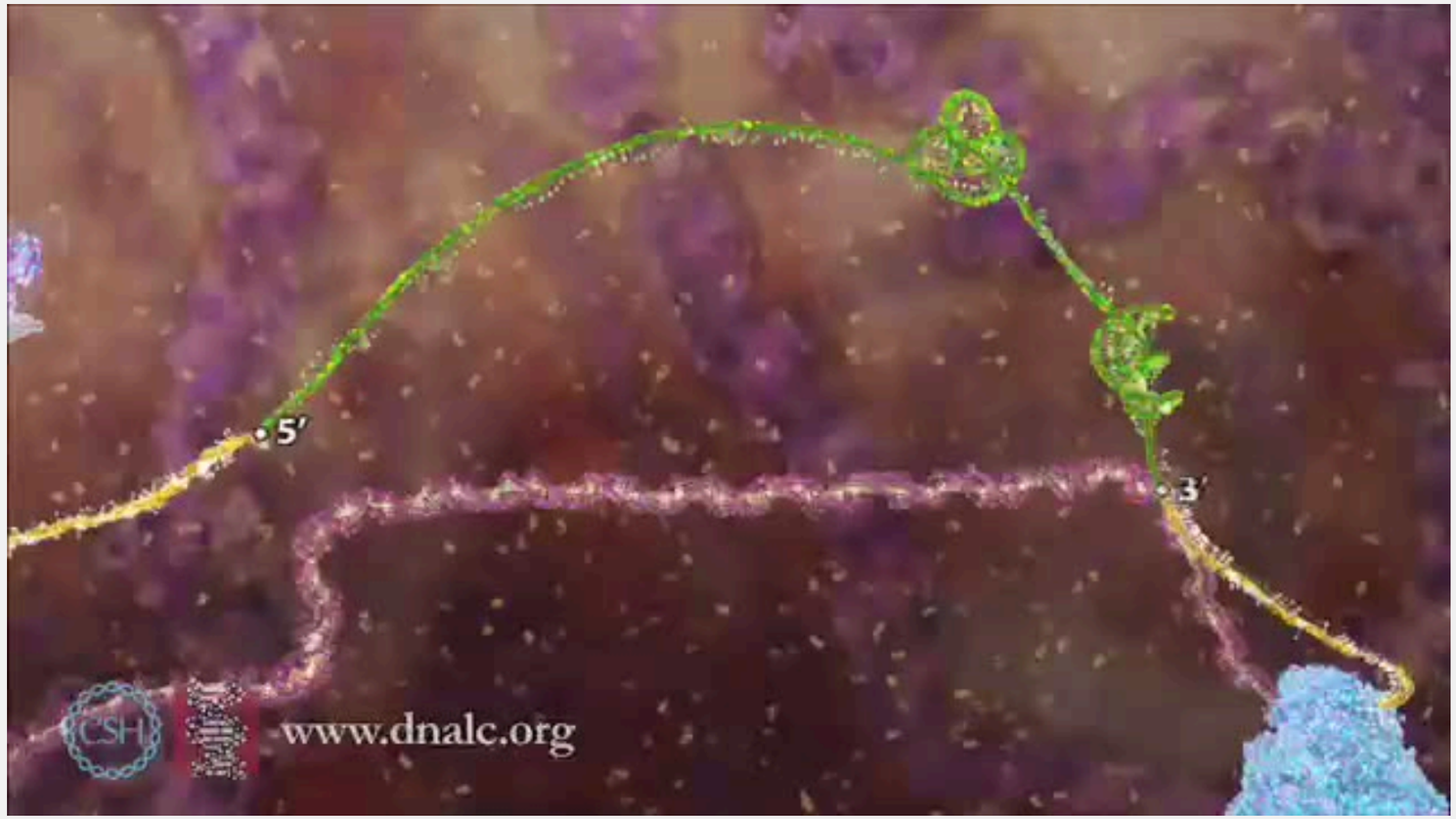
Philippe Hupé (2012), CC BY-SA 3.0 · original SVG

The spliceosome makes a precise connection

- The spliceosome recognises splice sites and a branch point in the pre-mRNA.
- The intron forms a **lariat** intermediate; the adjacent exons are ligated.
- Regulatory proteins and RNA sequence/structure influence which sites are used.
- Most major-spliceosome introns have **GU…AG** boundaries in RNA; these two dinucleotides alone do not define a splice site.

Alternative splicing changes the RNA product without changing the genomic DNA.

RNA splicing in three dimensions



Animation: Cold Spring Harbor Laboratory, DNA Learning Center

Watch for this

How do the exon ends meet while the intron is removed?

The film simplifies exons as coding sequence. Remember that exons also include UTRs.

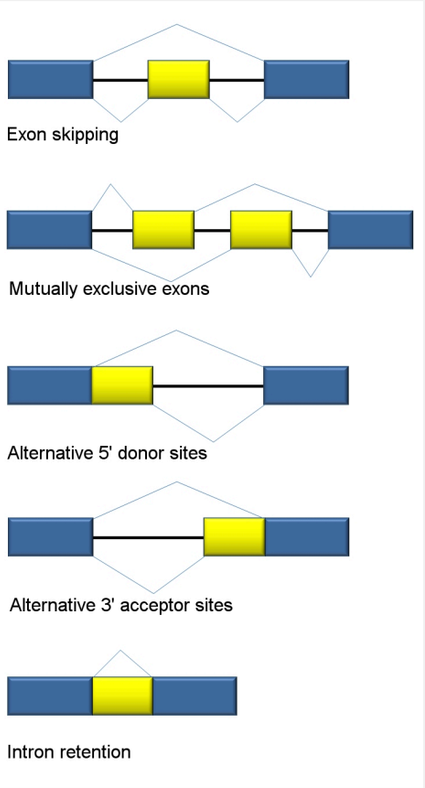
Then explain

Explain which molecular interaction makes this step selective.

Five common patterns of alternative splicing

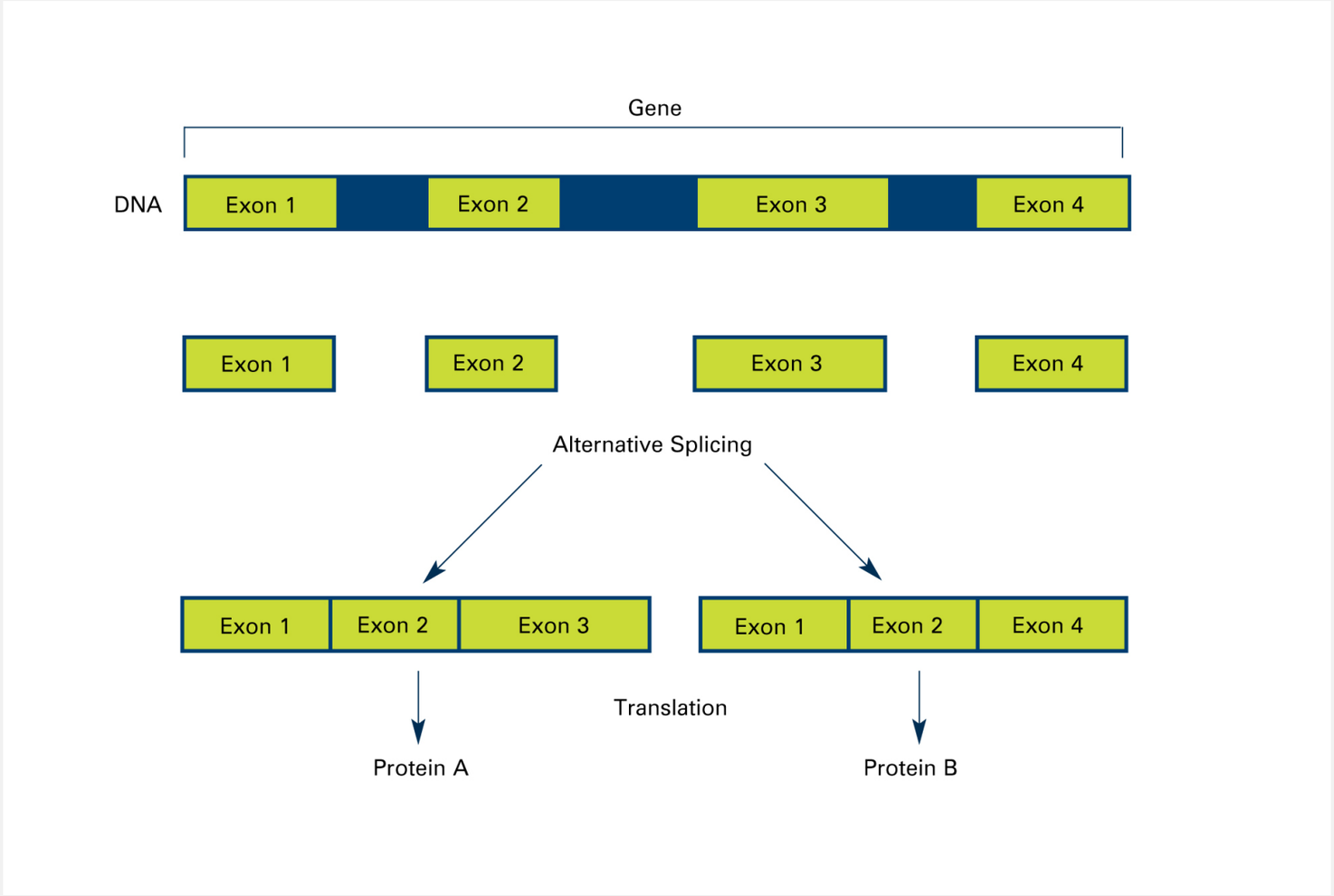
Pattern	What changes?
Exon skipping	An exon is included or omitted
Alternative 5′ splice site	One exon boundary changes
Alternative 3′ splice site	The other exon boundary changes
Mutually exclusive exons	One of two alternatives is selected
Intron retention	An intron remains in the RNA

5′/3′ refer to the **intron's splice sites**, not the left/right of every drawing.



Agathman (2009), CC BY-SA 3.0 · downloaded original

Alternative splicing: different exon combinations



Alternative splicing produces transcripts with different exon combinations that can encode different proteins.

Figure: Crabtree + Company / NIH National Institute of General Medical Sciences, *The New Genetics*. Original full-resolution image.

Different RNA does not always mean different protein

- A coding-region change can add or remove a protein domain.
- A splice change may shift the reading frame or introduce a premature stop.
- Some transcripts are targeted by **nonsense-mediated decay**.
- Changes confined to UTRs can affect RNA stability, localisation or translation.

An expressed transcript is not proof of a stable, functional protein.

A concrete example: SMN exon 7

Two related genes

SMN1 and **SMN2** are paralogous genes, not two isoforms of one gene.

A sequence difference in SMN2 exon 7 changes splicing regulation. SMN2 transcripts frequently skip exon 7.

Two splice outcomes



Exon 7 included → full-length SMN

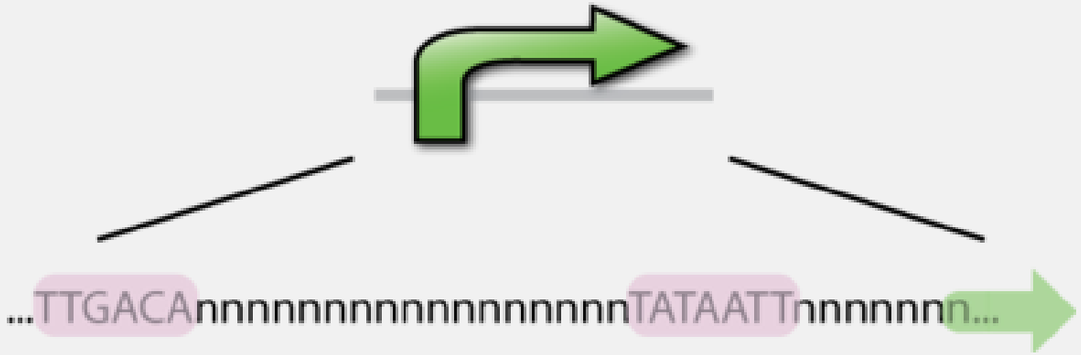


Exon 7 skipped → shortened isoform

Both outcomes can occur: this is a change in proportions.

Promoters and regulatory sequences

- A **promoter** helps recruit the machinery that initiates transcription.
- Initiation depends on regulatory proteins, DNA context and, in eukaryotes, chromatin.
- In bacterial operons, an **operator** is a regulatory protein-binding site.
- Eukaryotic enhancers can act at a distance; they are not simply bacterial operators.

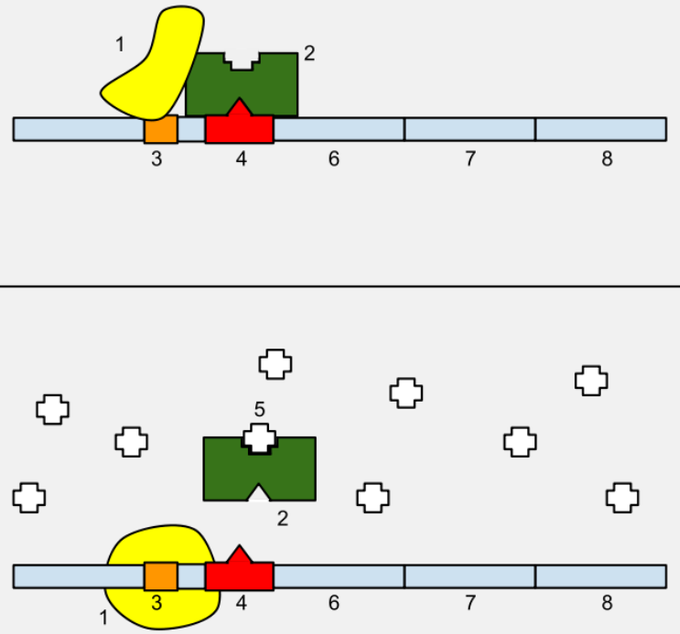


The lac operon: a bacterial example

lacZ encodes β-galactosidase, not lactase.

- LacI represses transcription by binding the operator.
- Allolactose, derived from lactose, reduces repressor binding.
- Strong expression also depends on glucose availability through cAMP–CAP.

The figure simplifies induction by showing lactose.



Promoter Choice

- A gene may have multiple promoters and transcription start sites.
- Different promoters can produce different first exons or 5′ UTRs.
- Coding sequences may also differ, depending on transcript structure.

Alternative promoter use and alternative splicing are distinct mechanisms that can occur together.

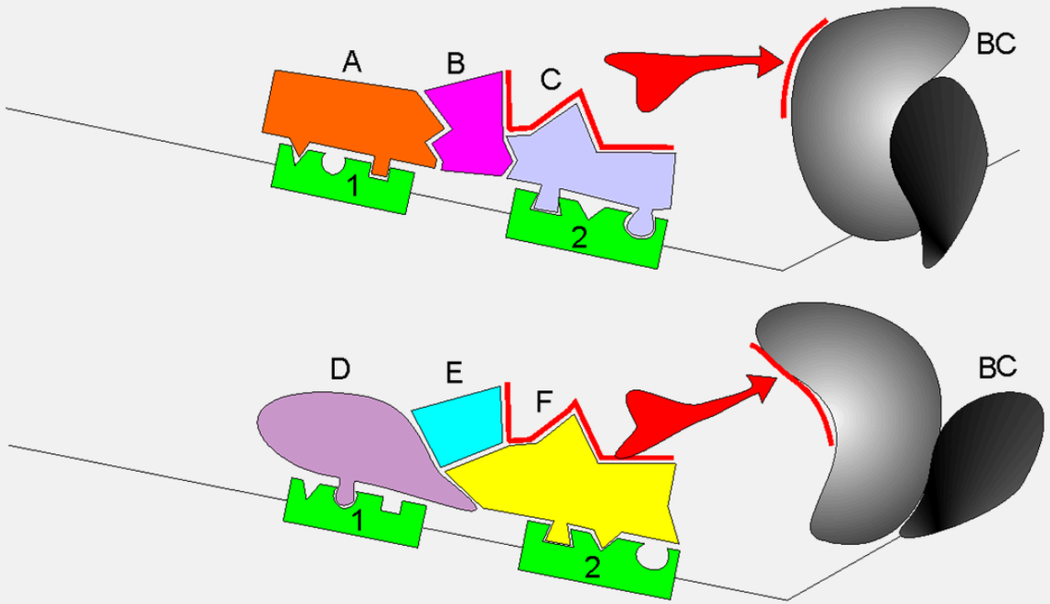


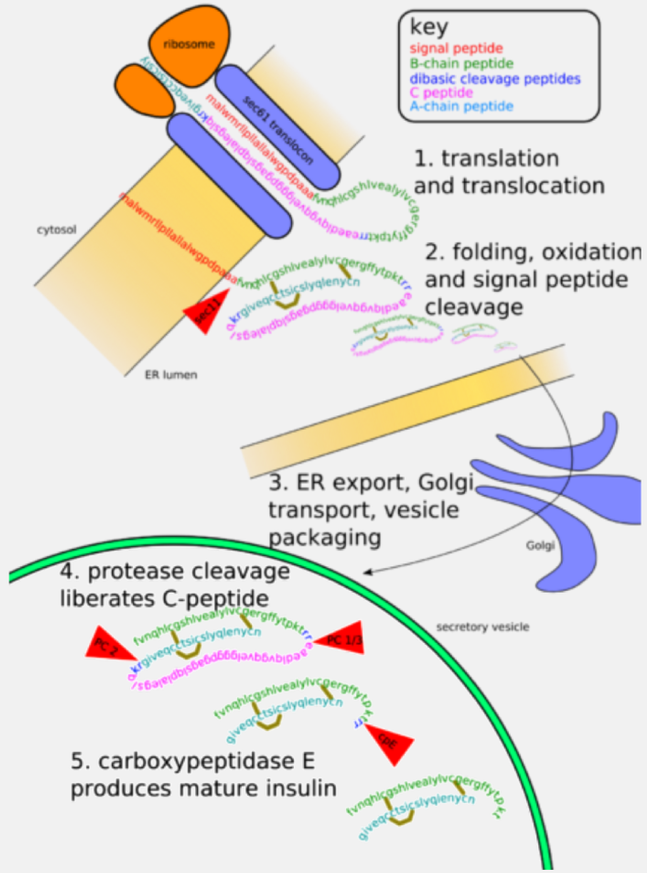
Image from Kel A.E., Ratner V.A.

Post Translational Modifications

Posttranslational modification (PTM) is the chemical modification of a protein after its translation.

PTM extends the range of functions of a protein:

- Attaching to it other biochemical functional groups.
- Changing the chemical nature of an amino acid (e.g. citrullination).
- Making structural changes, like the formation of disulfide bridges.
- Removing amino acids from the amino end of the protein
- Cutting the peptide chain in the middle.

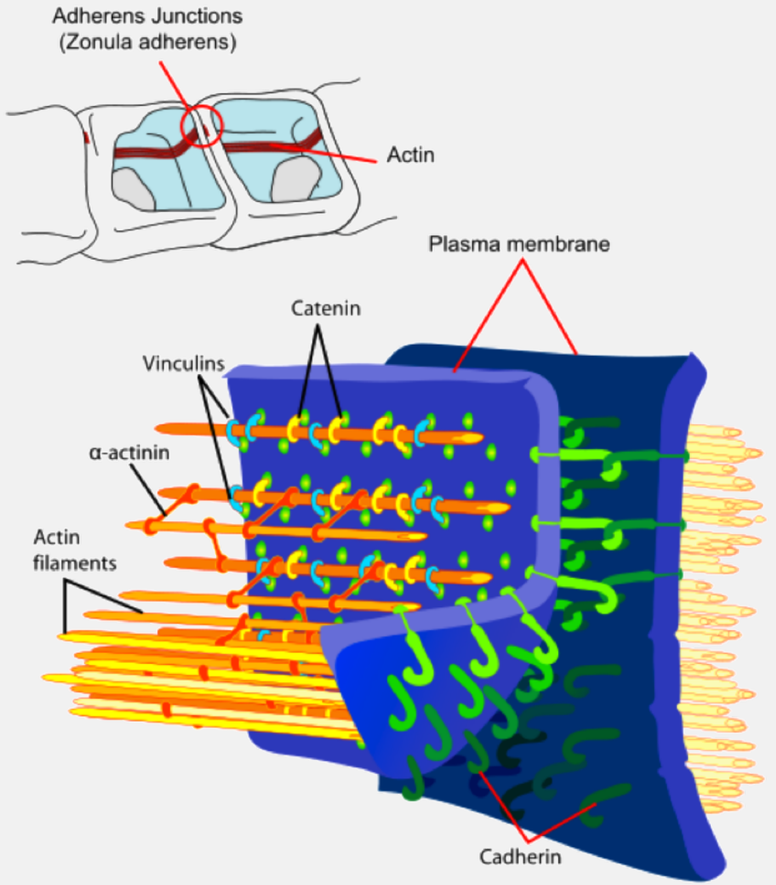


The proteome

The **proteome** is the set of proteins present in a biological system under specified conditions.

Proteins have structural, catalytic, transport, regulatory, and immune roles. Protein abundance and modification depend on cell type and state.

Some genes produce functional RNAs rather than proteins.



What are we actually measuring?

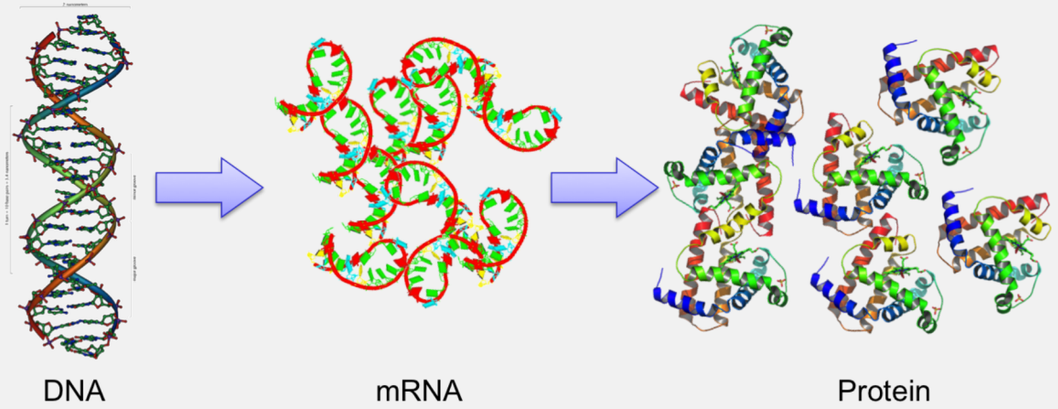
Biological quantity	A suitable measurement	Main caution
RNA abundance	Microarray or RNA-seq	Not a direct protein measurement
Protein abundance	Mass spectrometry or immunoassay	Isoforms and detection biases matter
Protein activity	Functional assay	Abundance alone is insufficient

RNA abundance reflects both synthesis and degradation.

Translation and protein turnover further separate RNA and protein levels.

Measuring mRNA abundance

Expression microarrays estimate the abundance of selected RNA targets. They do not directly measure protein abundance or post-translational modifications.



How do DNA arrays work?

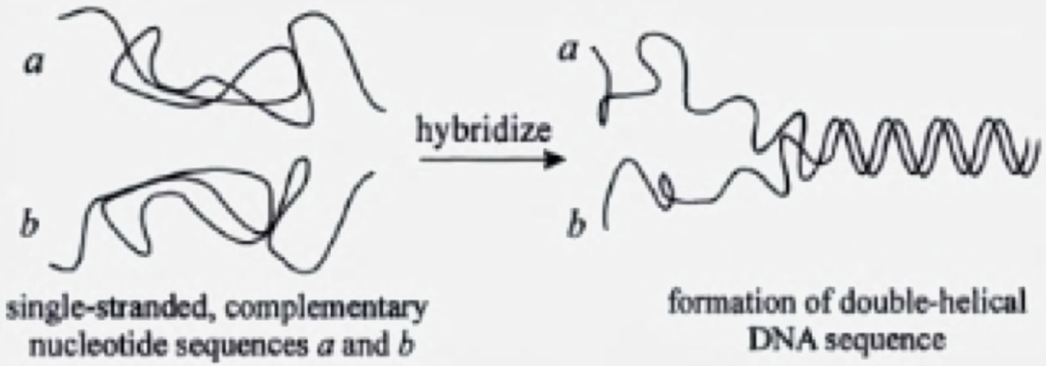
Hybridization

Complementary single-stranded DNA or RNA molecules associate through base pairing to form a duplex.

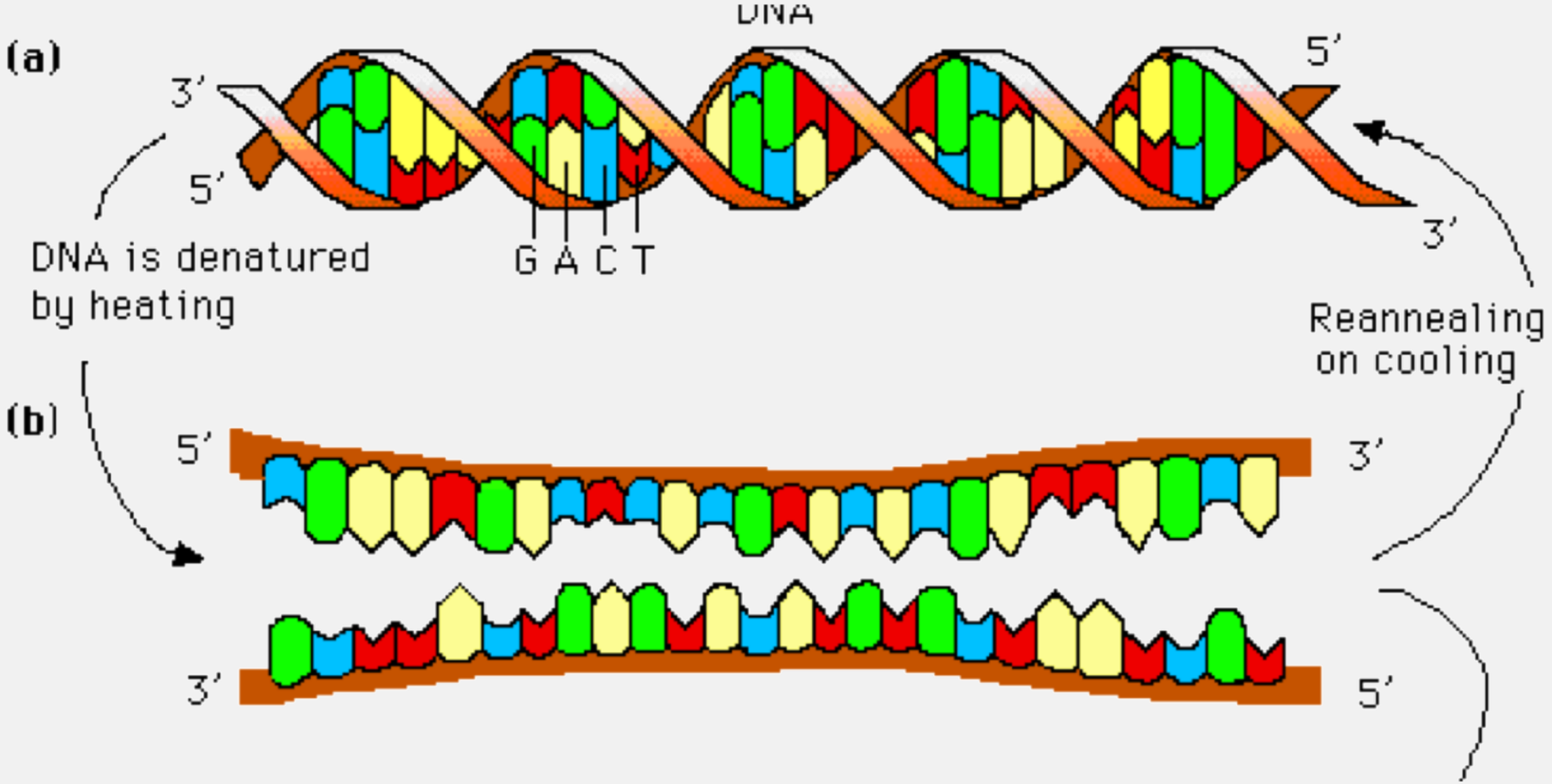
Yield

Depends on:

1. The concentration of the original single-stranded polymers,
2. Sequence complementarity, temperature, salt concentration and time.



DNA Hybridization

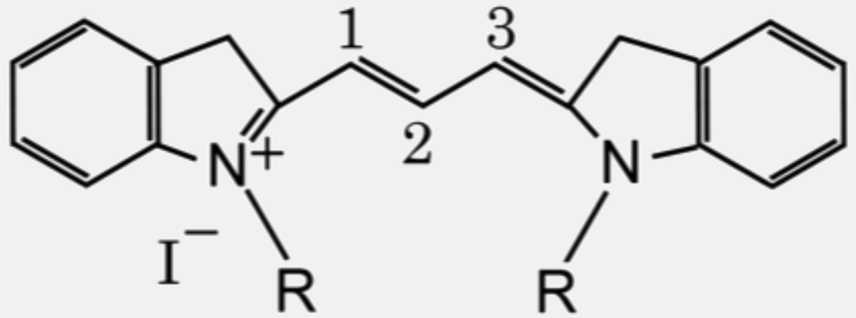


Fluorescent labels turn binding into a signal

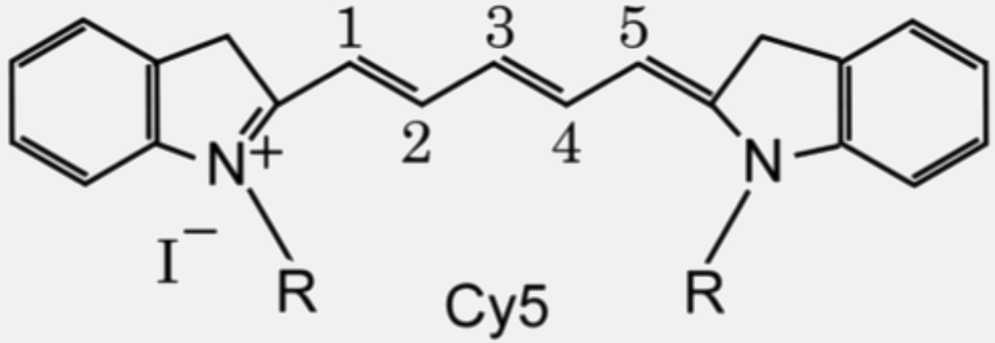
Two-colour arrays

- Label one sample with **Cy3** and the other with **Cy5**.
- Hybridise both to the same array; scan the channels separately.
- Displayed green/red colours represent channel intensities.
- Yellow in the composite means signal in both channels—not automatically equal biological expression.

Dye bias and background must be corrected before interpreting ratios.



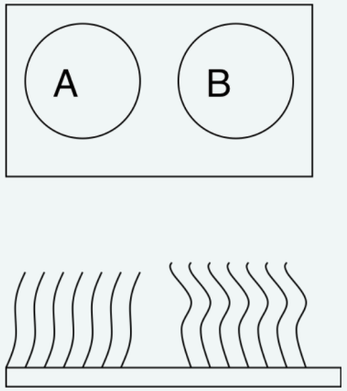
Cy3



Cy5

Hybridisation: from solution to a surface signal

1 · Immobilise probes



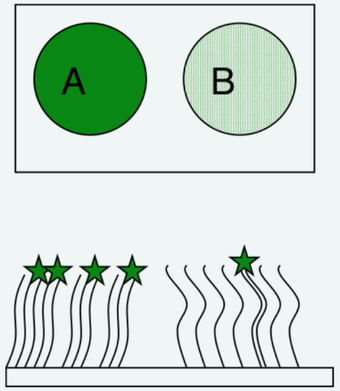
Known sequences at known positions.

2 · Add targets



Labelled sample molecules encounter probes.

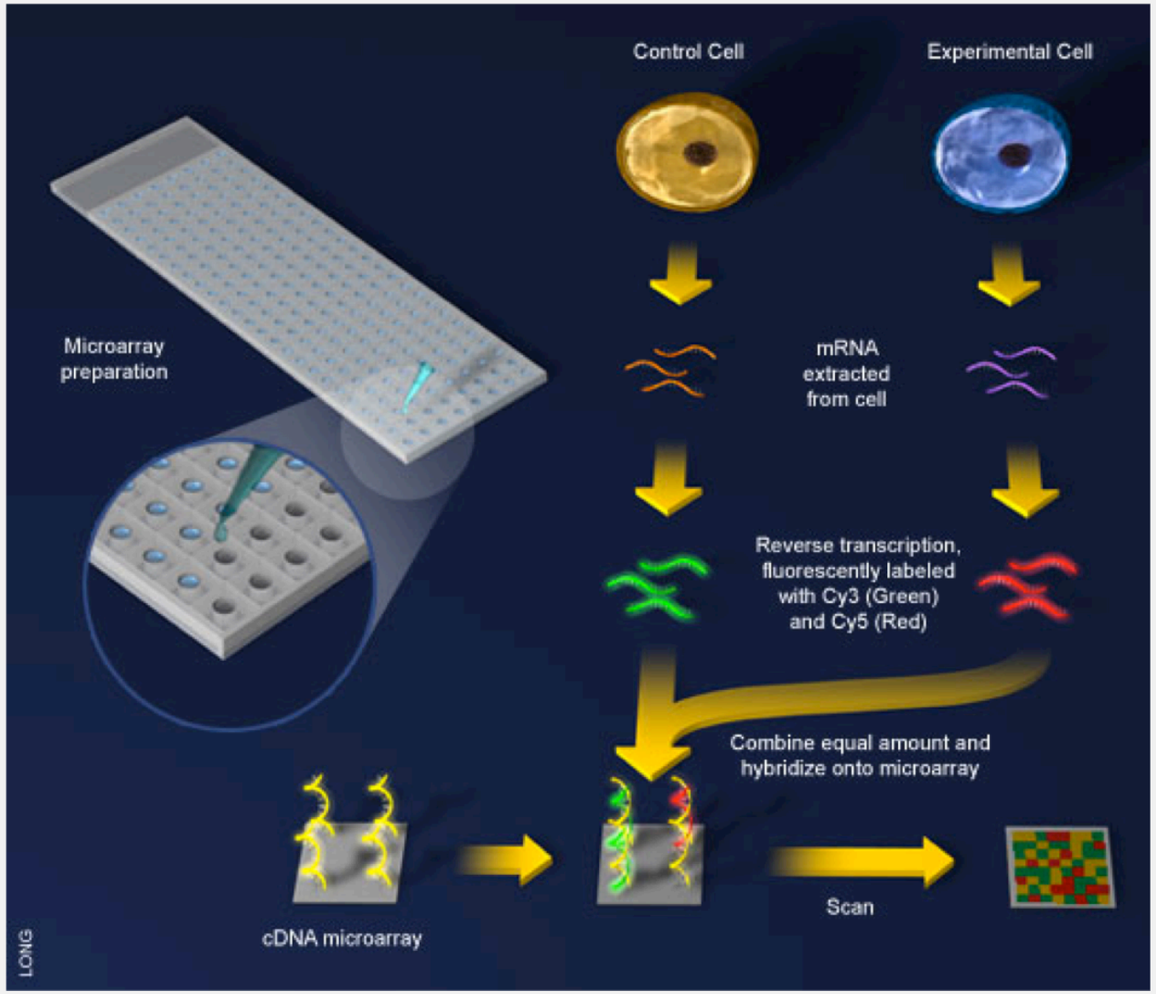
3 · Wash and scan



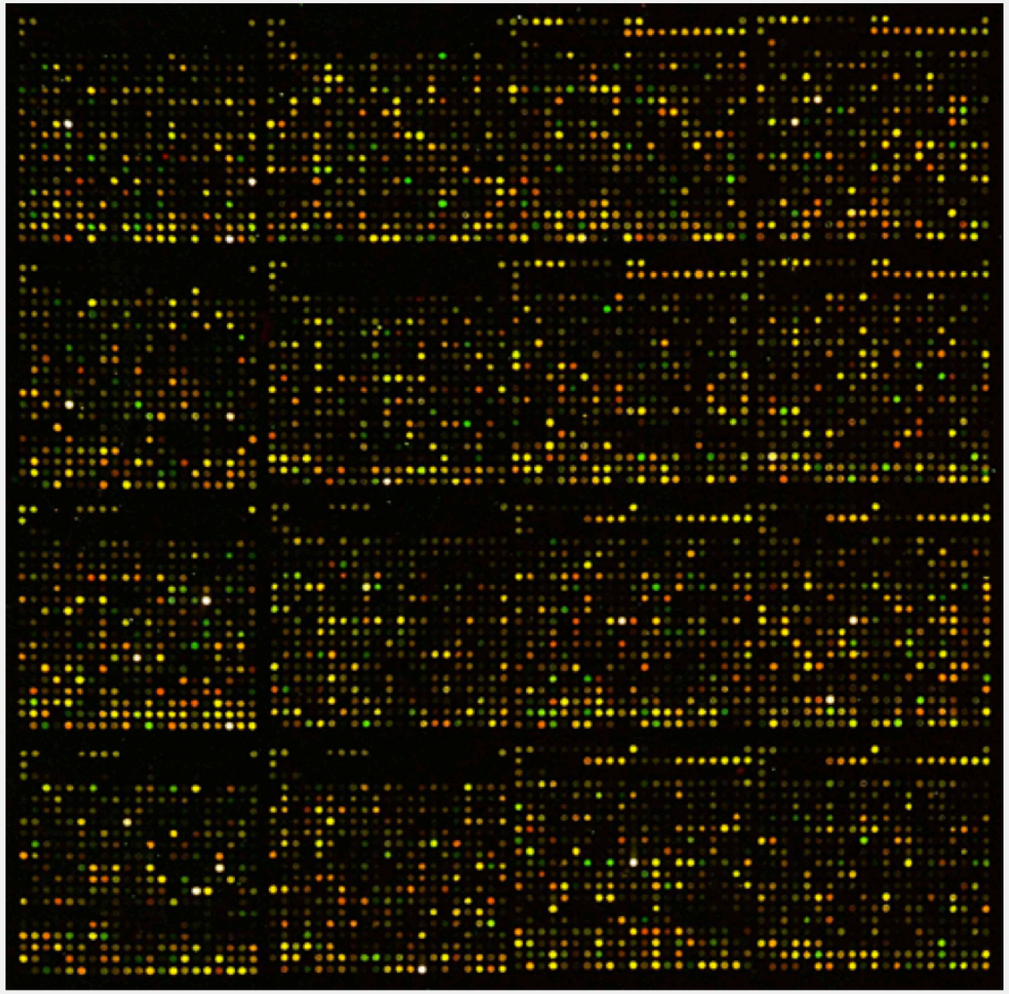
Retained fluorescence provides an abundance-related signal.

Signal also depends on affinity, background and saturation.

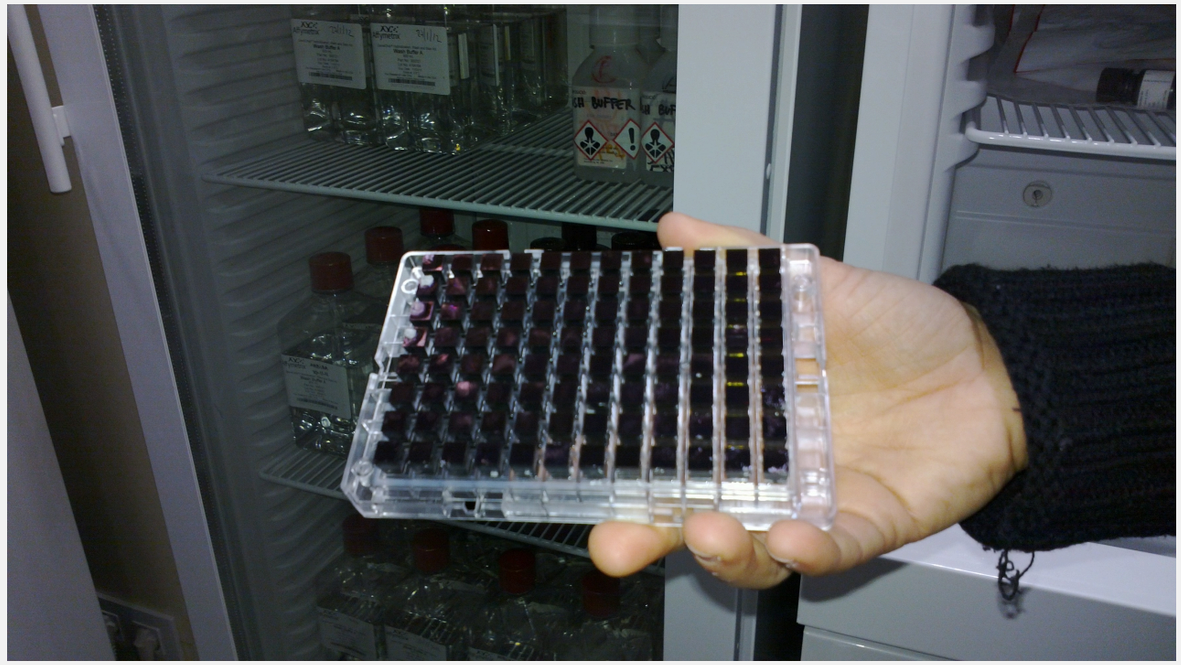
Competitive hybridization



Probes



From a chip to an automated assay





A dense probe array is one component of a workflow: sample preparation → hybridisation → washing → scanning.

One channel and two channels answer comparisons differently

	One-colour array	Two-colour array
Samples per array	One	Usually two
Labelling	Same label across samples	Different dyes
Comparison	Across arrays	Within an array, then across arrays
Essential correction	Background and between-array effects	Background, dye and array effects
Result	Relative expression after processing	Relative expression after processing

Neither produces absolute transcript counts directly.

Two-colour microarray: a paired measurement



The samples compete for the same probes. Consider dye swaps and biological replication when designing the comparison.

One-colour microarray: one sample per chip



Background correction → normalisation → compare corresponding probes

A brighter raw spot on array B is **not enough** to conclude higher expression. Processing and replicate variation matter.

How are the probes put on a chip?

Manufacturing approach	How probes are made	Main idea
Spotted arrays	Prepare DNA/oligos, then deposit them	Place existing probes at chosen positions
Light-directed in situ synthesis	Add bases on the surface using patterned light	Light selects which features can grow
Inkjet in situ synthesis	Deliver reagents to selected positions	Local reagent delivery selects growth

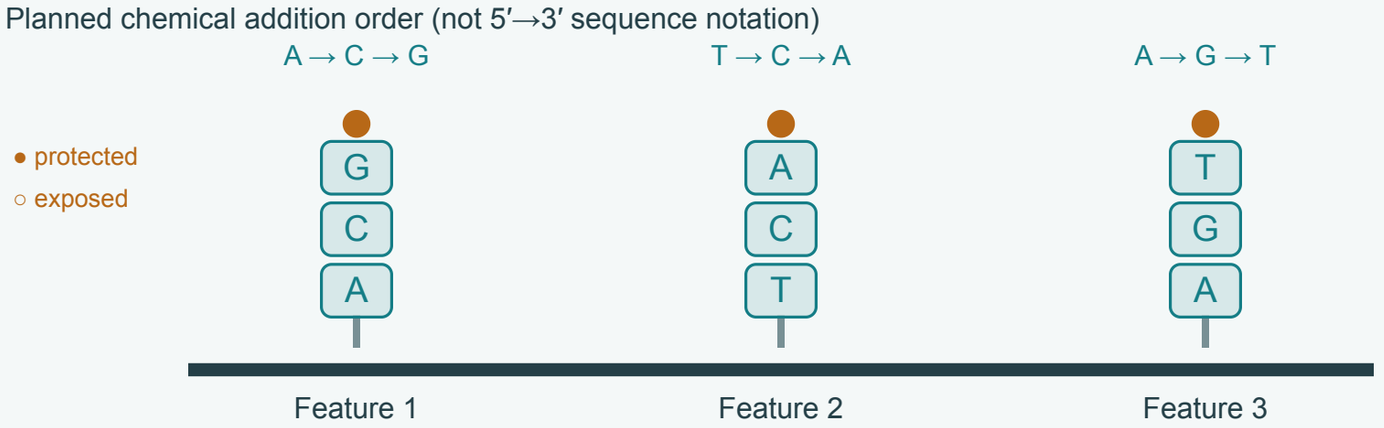
A **feature** contains many copies of a designed probe sequence.

The next slides explain classical light-directed synthesis.

Light selects where the next base is added

Position 3 · reagent T · wash

Step-controlled teaching model



Wash away excess reagent. The new ends are protected again; choose the next mask/base.

Protected surface → patterned light removes selected protecting groups → one protected nucleotide couples at those sites → wash → repeat with other masks and bases. Each feature builds many identical probes; different features follow different growth schedules.

Manufacturing reference: Affymetrix / Thermo Fisher educational guide

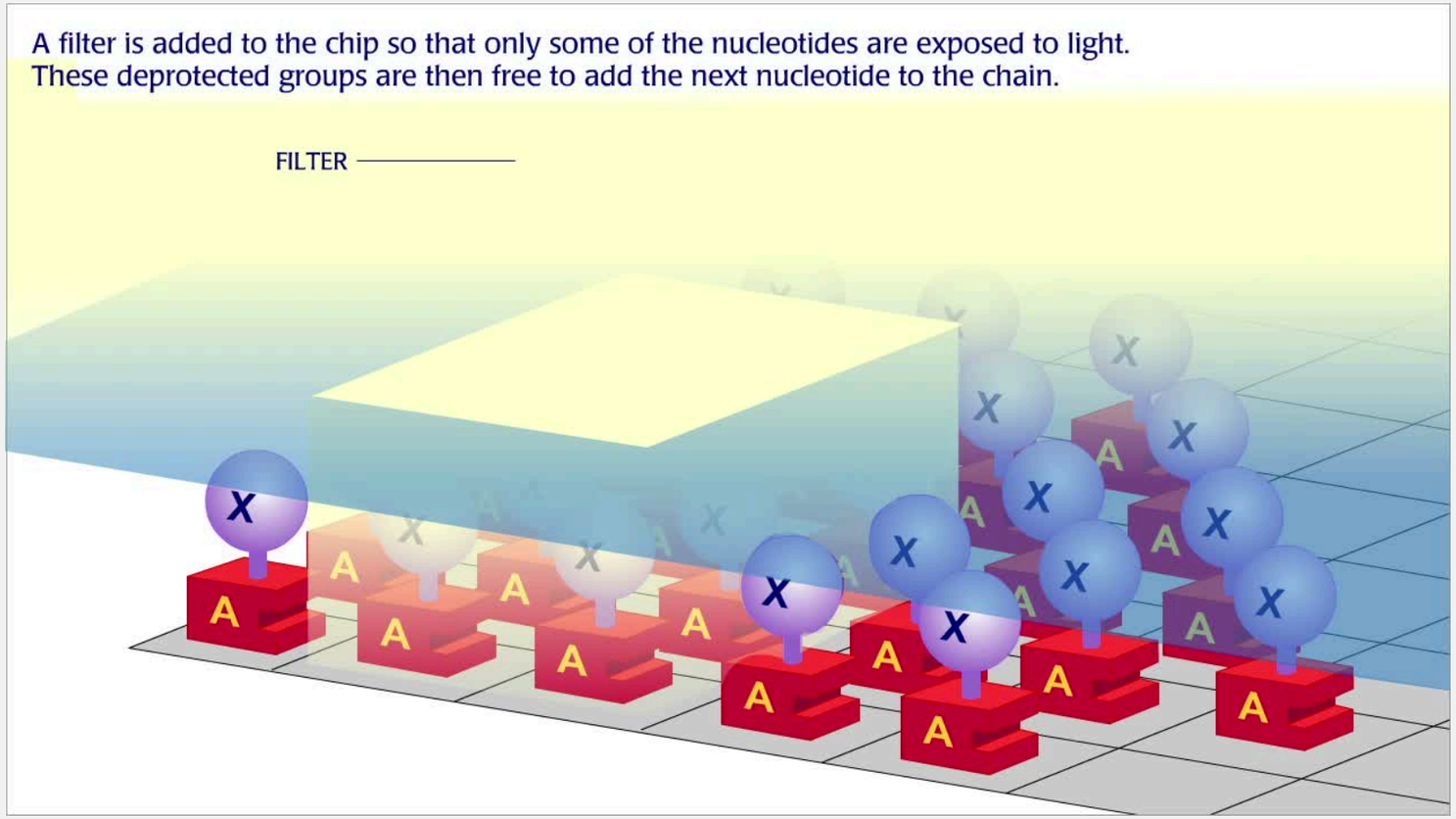
Probe synthesis: sequence space versus manufacturing cycles

Probe length N	Possible sequences: 4^N	Base-specific cycles: at most 4N*
1	4	4
2	16	8
3	64	12
10	1,048,576	40
25	1,125,899,906,842,624	100

Parallel synthesis saves cycles; it does not create unlimited chip area.

*Idealised A/C/G/T schedule at each position. A cycle includes exposure, coupling and washing; actual chemistry has additional steps. Only the selected probe designs fit on the chip.

Watch a GeneChip being built



Animation: Cold Spring Harbor Laboratory, DNA Learning Center

Watch for this

Which step provides spatial selectivity: the light pattern or the nucleotide reagent?

Why can many different probes grow in parallel?

Then explain

Explain which molecular interaction makes this step selective.

Microarrays: powerful, but constrained by their probes

- You mainly measure sequences represented by the probe design.
- Related sequences may cross-hybridise.
- Background and saturation limit the usable signal range.
- Transcript isoforms are distinguishable only if the probes target informative regions or junctions.

Transition: can we read sample-derived sequences instead of measuring binding to a fixed probe set?

RNA-seq: sample RNA, read DNA, infer expression

Biological material	Sequencing material	Observed data
RNA molecules in the sample	A library of RNA-derived DNA fragments	Sequence reads and quality scores

In conventional short-read RNA-seq, **reverse transcription** makes cDNA. The sequencer reads DNA; analysis relates those reads back to RNA abundance and structure.

Direct RNA sequencing is a different technology.

Before the flow cell: what does PCR do?

The ingredients

- A DNA template
- Two primers that flank a target
- A thermostable DNA polymerase
- dNTPs, buffer and suitable ions

The principle

1. Separate the DNA strands.
2. Let primers bind complementary sites.
3. Extend from each primer's **3′ end**.

   New DNA is always synthesised **5′ → 3′**.

Reverse transcription makes cDNA; PCR amplifies DNA.

PCR: predict the next step

Repeat the cycle

Step-controlled teaching model

Blue / teal: original templates · orange / pink: newly synthesised DNA

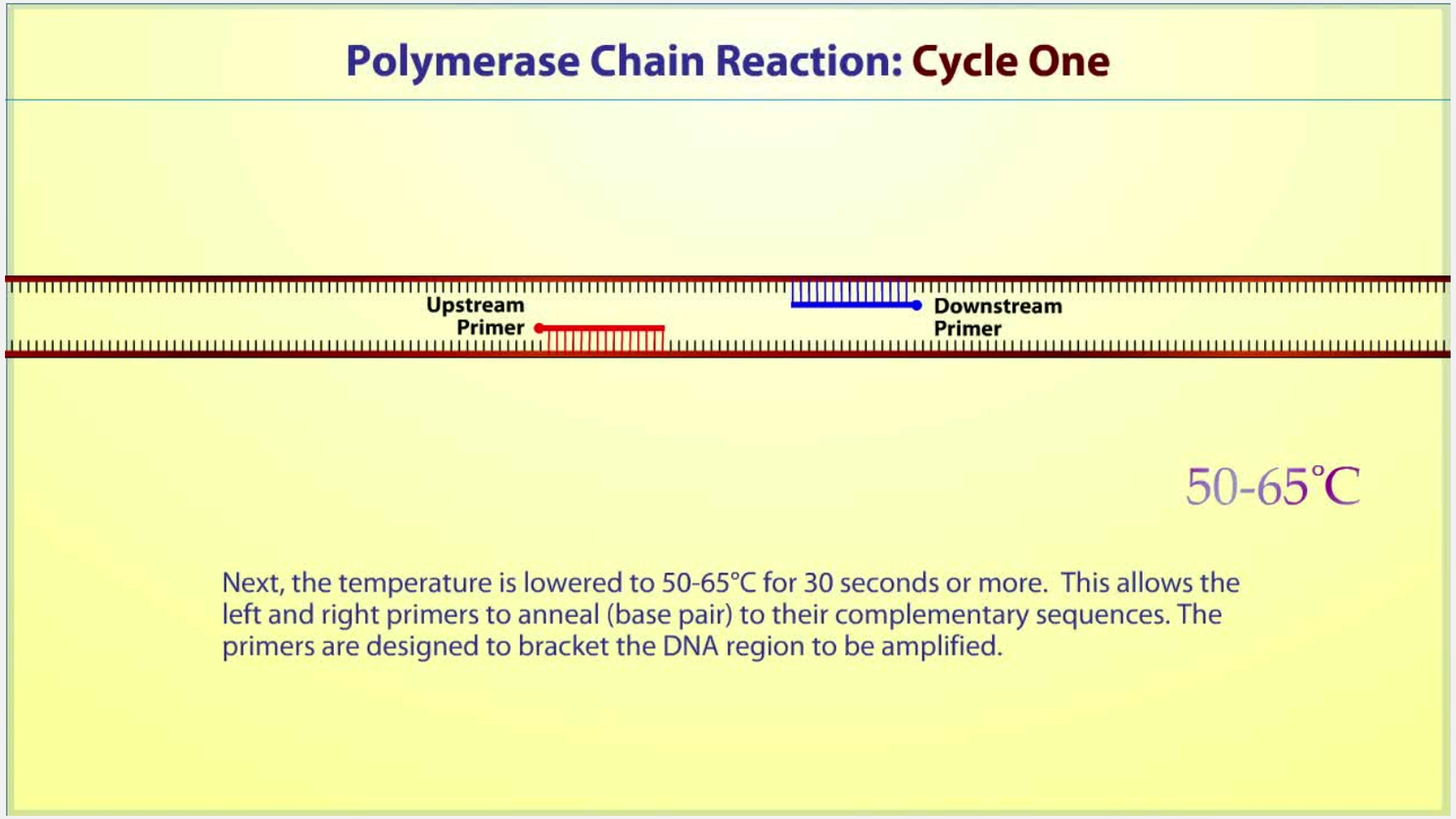


Each daughter duplex: one original strand + one new strand

Semiconservative copying preserves one original strand per daughter duplex. Ideal amplification gives 2^n copies; real reactions plateau.

Denature (~95 °C) → anneal primers (~50–65 °C) → extend (~72 °C for Taq). Repeat. One ideal double-stranded template becomes 2, 4, 8… copies. Exact temperatures and efficiency depend on the protocol.

PCR in motion



Animation: Cold Spring Harbor Laboratory, DNA Learning Center

Watch for this

Locate the primers. Which ends can the polymerase extend?

Why is one primer not sufficient for ordinary exponential PCR?

Then explain

Explain which molecular interaction makes this step selective.

Amplification creates copies—and can introduce bias

Completed cycles	Ideal copies from one template
0	1
1	2
3	8
10	1,024
20	1,048,576

The ideal model is $\mathbf{2^n}$. Real reactions have variable efficiency and eventually plateau.

Copies from one molecule are **not independent biological replicates**. Sequence-dependent amplification can distort representation.

RNA-seq starts with an experimental design

- Define the contrast: tissue, condition, time point or treatment.
- Collect **independent biological replicates**; technical repeats answer a different question.
- Balance conditions across extraction, library preparation and sequencing batches.
- Assess RNA quality and choose a protocol suited to the material.

A perfect sequencer cannot rescue a confounded experiment.

Which RNA population will enter the library?

Strategy	What is enriched or retained?	Important consequence
Poly(A) selection	Polyadenylated RNA, mainly eukaryotic mRNA	Misses many non-polyadenylated RNAs
rRNA depletion	RNA remaining after removing targeted rRNAs	Can include non-coding RNA and pre-mRNA
Targeted capture	Selected transcripts/regions	Limited by the target design

Choose for the organism, RNA integrity and biological question.

The library is a selected representation of the RNA sample.

Build a sequenceable library

RNA selection / depletion

↓

Fragment RNA · reverse transcribe · make double-stranded cDNA

↓

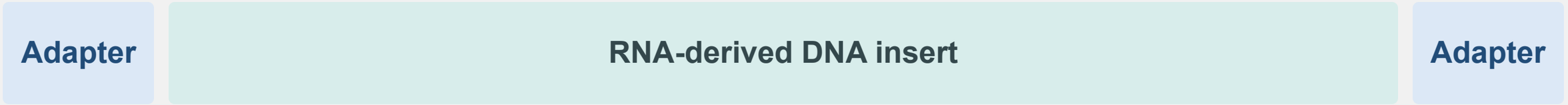
Prepare ends · attach adapters and sample indexes

↓

Optional / protocol-dependent PCR · library QC · pool

A representative workflow: some protocols fragment cDNA instead, combine steps, or use different adapter-addition strategies. Strand-specific chemistry preserves information about the original RNA strand.

Adapters give each fragment an interface



- Flow-cell binding sequences tether the library to the surface.
- Sequencing-primer sites let reads begin at defined positions.
- A **sample index** identifies the pooled library.
- A **UMI**, when included, tags an original molecule before amplification.

Index ≠ UMI: sample identity and molecule identity answer different questions.

Why make a cluster?

A single library molecule

One labelled base is too dim for the classical ensemble imaging strategy.

Local amplification creates a **clonal cluster** of many copies.

A brighter local signal

Copies ideally incorporate the same base in the same cycle.

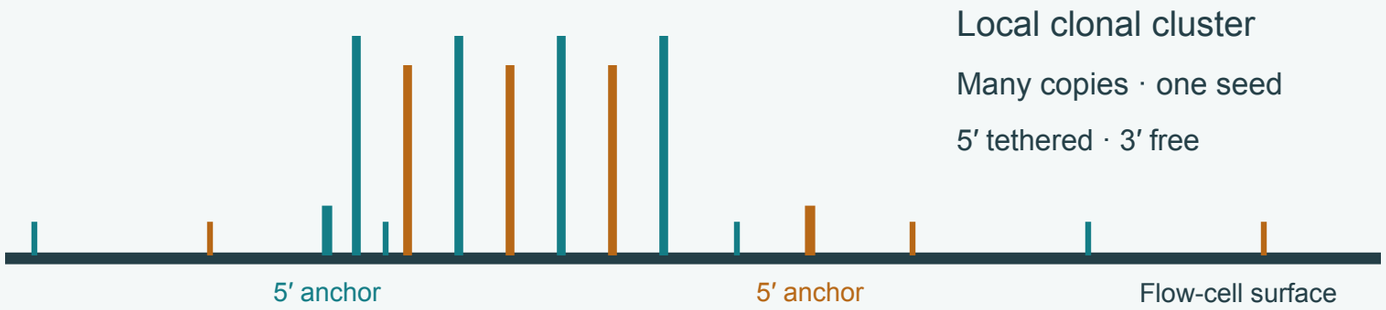
The imager reads the combined signal at that position.

A cluster is a technical copy population, not a biological replicate.

Bridge amplification: tether, bend, copy, separate

7 · Repeat locally

Step-controlled teaching model



Repetition creates a clonal cluster. Sequencing preparation selects suitably oriented templates.

A library strand binds a complementary surface oligo and is copied. The free end of the tethered copy binds a nearby compatible oligo, forming a bridge. Polymerase makes a complementary strand; denaturation leaves two tethered strands. Repetition makes a local clonal cluster.

Solution PCR versus surface amplification

	PCR in solution	Classical bridge amplification
Primers	Free in the liquid	Tethered to the surface
Products	Dispersed through the tube	Remain near the seed molecule
Main teaching purpose here	Enrich/amplify a library	Make individually imaged clusters
Shared principle	Primer binding → extension → separation	Primer binding → extension → separation

The library contains **DNA derived from RNA**, not RNA being copied by DNA polymerase.

Bridge amplification is a teaching model—not every flow cell

Classical flow cells

Randomly positioned clusters can be generated by bridge amplification.

Patterned flow cells

Defined nanowells control cluster positions; Illumina platforms can use **exclusion amplification (ExAmp)** and platform-specific clustering chemistry.

Keep the invariant idea: many local copies produce an interpretable sequencing signal. Check the instrument's chemistry before describing its exact clustering method.

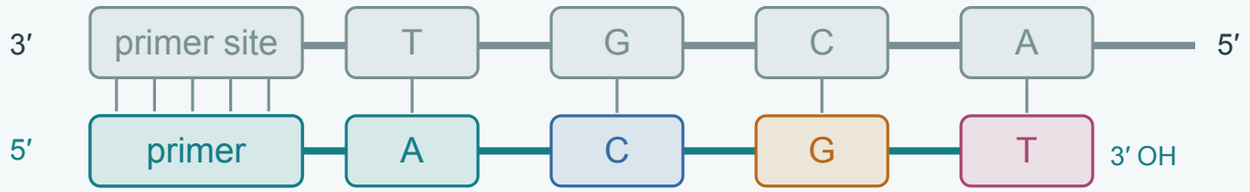
Illumina · patterned flow-cell technology

Sequencing by synthesis: one base, one image cycle

Completed teaching read

Step-controlled teaching model

A primer is already paired to a complementary template region



Called read: ACGT

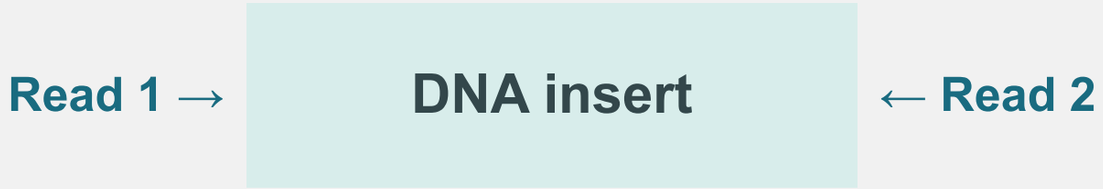
Four cycles → ACGT

The read is the newly synthesised strand, complementary to the template. Four colours illustrate the principle, not every instrument chemistry.

A primed template receives fluorescent reversible-terminator nucleotides. One base is incorporated, imaging identifies it, then the dye/block is removed. Repeating the cycle generates a read. Teaching read: ACGT; template: 3′-TGCA-5′.

Illumina · sequencing by synthesis

Paired-end sequencing reads both ends of a fragment



- The two reads originate from the **same library fragment**.
- They can improve placement across repeats and splice junctions.
- The middle may remain unsequenced if the insert is longer than both reads combined.
- Paired reads are two observations of one fragment, not two independent RNA molecules.

Index reads identify the sample; they are separate from the biological insert reads.

What comes off the sequencer?


```
@fragment_001/1
ACGTACGTACGT
+
IIIIIIIIIIII
```

- A read identifier, a nucleotide sequence, and per-base quality scores.
- **Demultiplexing** assigns reads to samples using their indexes.
- Base qualities describe confidence in calls—not biological importance.
- Inspect adapters, quality, contamination and library composition before interpretation.

Spliced reads connect RNA-seq to alternative splicing



Read evidence	What does it support?
Inside shared E1	Expression, but not a unique isoform
Across E1–E2 or E2–E3	Exon 2 inclusion
Across E1–E3	Exon 2 skipping

A splice-aware genome alignment can jump across introns. Transcriptome-based quantification uses a different reference representation.

Short reads may not uniquely resolve complete isoforms.

Gene expression and isoform usage are different questions

Illustrative molecule counts	Condition A	Condition B
Isoform with exon 2	80	20
Isoform without exon 2	20	80
Total for the gene	**100**	**100**

Gene-level abundance is unchanged, while exon inclusion falls from **80% to 20%**.

Real RNA-seq estimates must account for ambiguity, effective length, sampling and replicate variation. These are idealised molecule counts, not raw read counts.

From reads to a gene-by-sample matrix

Quality control and sample assignment

↓

Splice-aware alignment OR transcript-level quantification

↓

Summarise expression with an appropriate gene/transcript model

↓

Normalisation · replicate-aware modelling · biological interpretation

Raw counts depend on sequencing depth and composition. A count is not automatically an absolute number of cellular RNA molecules.

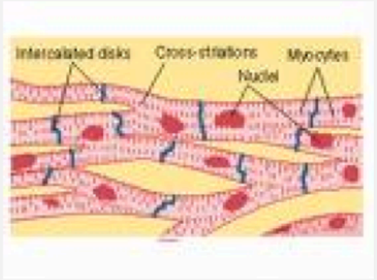
Choose a technology for the biological question

Question or property	Microarray	Short-read RNA-seq
Known genes	Probe-based intensity	Sequence-derived abundance
Unanticipated transcripts	Limited by probe design	Can provide new sequence evidence
Alternative splicing	Needs informative probes	Junctions and isoform evidence
Main limitations	Cross-hybridisation, saturation	Sampling, mapping ambiguity, library bias
Protein abundance?	Not directly	Not directly

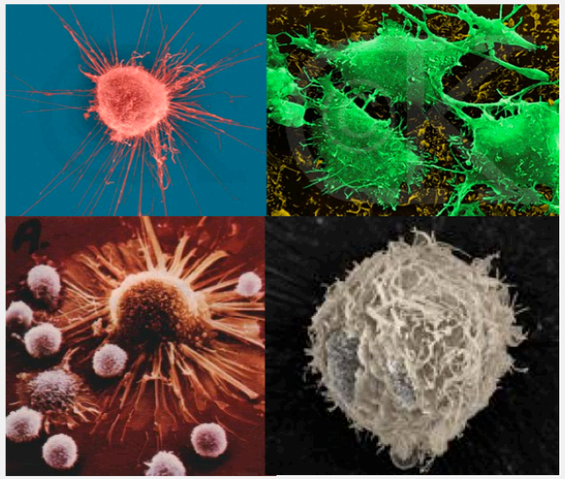
Both need good design, controls, replication and validation.

Applications

- Compare RNA abundance in cardiac versus skeletal muscle, or across cell differentiation.
- Cells exposed to different conditions.
  - Physical (temperature, radiation).
  - Chemical (toxins).
  - Changing nutrient availability.
- Compare tumour with matched non-tumour samples.
- Study disease-associated expression changes alongside other molecular and clinical evidence.
- Drug response.







Historical case study: measurement is not validation

The following original figures document early biomarker enthusiasm and reproducibility concerns.

- The ovarian-cancer example concerns **serum proteomics**, not RNA-seq.
- Historical performance estimates apply to the reported study, not to general clinical use.
- The shared lesson is experimental design, independent validation and reproducible analysis.

The New England
Journal of Medicine

Copyright © 2002 by the Massachusetts Medical Society

VOLUME 347 DECEMBER 19, 2002 NUMBER 25

A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T VEER, PH.D., HONGYUE DAI, PH.D., AUGUSTINUS A.M. HART, M.SC., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER, M.SC., JOHANNES L. PETERSE, M.D., CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOUWE ATSMA, ANKE WITTEVEEN, ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY BARTELINK, M.D., PH.D., SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPHEN H. FRIEND, M.D., PH.D., AND RENÉ BERNARDS, PH.D.

ABSTRACT

Background A more accurate means of prognostication in breast cancer will improve the selection of patients for adjuvant systemic therapy.

Methods Using microarray analysis to evaluate our previously established 70-gene prognosis profile, we classified a series of 295 consecutive patients with primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis. All patients had stage I or II breast cancer and were younger than 53 years old; 151 had lymph-node–negative disease, and 144 had lymph-node–positive disease. We evaluated the predictive power of the prognosis profile using univariable and multivariable statistical analyses.

ADJUVANT systemic therapy substantially improves disease-free and overall survival in both premenopausal and postmenopausal women up to the age of 70 years with lymph-node–negative or lymph-node–positive breast cancer.[1,2] It is generally agreed that patients with poor prognostic features benefit the most from adjuvant therapy.[3,4] The main prognostic factors in breast cancer are age, tumor size, status of axillary lymph nodes, histologic type of the tumor, pathological grade, and hormone-receptor status. A large number of other factors have been investigated for their potential to predict the outcome of disease, but in general, they have only limited predictive power.[5]

The technology produced a breakthrough

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,[1] but to achieve this goal, specific and sensitive molecular markers are essential.[1–5] This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as *BRCA1* and *BRCA2*. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients,[1] and is associated with a 5-year survival of 35% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and most patients are cured of their disease by surgery alone.[1–6] Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or chemotherapeutic approaches.

Cancer antigen 125 (CA125) is the most widely used biomarker for ovarian cancer.[1–6] Although concentrations of CA125 are abnormal in about 80% of

Proteomics, Petricoin, Lancet 2.02

“The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non malignant disease, 63 were recognized as not cancer. This result yielded a sensitivity of 100% specificity of 95%”

``New cancer test stirs hope and concern'' Pollack A. New York Times, Feb. 3, 2004

"I've been in cancer research for 40 years and I think it's the most important breakthrough in those years," said Dr. John S. Kovach, director of…

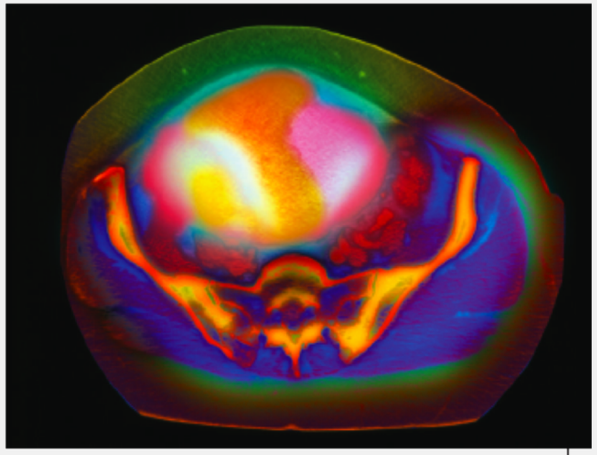
Correlogic Systems… Quest Diagnostics and LabCorp… say they expect to begin offering the test in the next few months.

Early biomarker claims: reproducibility concerns

news feature

Running before we can walk?

Two years ago, a new proteomic test was heralded as the future of cancer diagnostics. But since then, doubts about its effectiveness have begun to grow. Erika Check reports.



On target: can proteins in the blood reveal ovarian tumours (pink/yellow) before they reach this stage?

Seldom does a single piece of research prompt the US Congress to pass a resolution urging continued funding to drive a new diagnostic test towards the clinic. But that's what happened in 2002, when *The Lancet* published a paper[1] claiming a breakthrough in the diagnosis of ovarian cancer.

The paper described the use of mass spectrometry to analyse the pattern of proteins present in samples of blood serum. On the basis of these patterns, the test detected all the patients with ovarian cancers in a set of 50 samples, and falsely identified just three healthy patients as suffering from the disease from a total of 66 control samples.

Most encouragingly, the technique seemed to work well on patients with early-stage cancer — offering the prospect of earlier diagnosis, which improves the chances of successful treatment. The best current blood test, which relies on the detection of a single protein called CA125, misses at least half of patients in the earliest stages of the disease, and gives a high rate of false positives.

The researchers, led by Lance Liotta and Emanuel Petricoin, who co-direct a proteomics company called Correlogic in Rockville, Maryland, described their test as "virtually perfect". The *Lancet* paper was accompanied by an editorial[2] that declared: "The new millennium will be the era of proteomics, and ovarian cancer will be the first cancer to be conquered."

Early warning

Critics warn that the episode illustrates the dangers of moving rapidly to the clinic with immature technologies such as those of proteomics. They say that scientists and regulators need to develop standards to ensure that such tests really work before they hit the market, because early detection

> "Whether or not OvaCheck works, we will learn from this experience what rules of evidence we might apply in the future to find useful results more efficiently."

Lancet paper. In November 2002, Correlogic granted licences to two larger firms, Quest Diagnostics and the Laboratory Corporation of America, which are now hoping to market the test under the brand name OvaCheck.

But those plans could be thrown off track by reanalyses of Liotta and Petricoin's data by independent groups, which have raised serious doubts about OvaCheck's reliability.

These questions prompted the Society of Gynecologic Oncologists to review all of the published work about OvaCheck. On 7 February, the society declared that "more research is needed to validate the test's effectiveness before offering it to the public".

Bioinformatics[3]. They had reanalysed a data set that Liotta and Petricoin's team posted online in August 2002. Sorace and Zhan similarly found numerous differences in the protein patterns that discriminated between the cancer patients and the healthy controls. The trouble, according to Sorace and Zahn, was that these looked more like experimental artefacts than real biological differences.

The proteomics test relies on using gravity and electric fields to separate the proteins in a given sample. Each protein is then given a number that represents the ratio of its charge and mass — called its m/z value. The test identifies patterns in these numbers to give a diagnosis. Sorace and Zahn were concerned because the most marked differences between the cancer patients and unaffected controls occurred for proteins with m/z values of less than 500. Many spectrometry experts consider m/z values below 2,000 to be suspect, because they tend to include values generated by experimental artefacts or

Why independent validation matters

Lancet paper. In November 2002, Correlogic granted licences to two larger firms, Quest Diagnostics and the Laboratory Corporation of America, which are now hoping to market the test under the brand name OvaCheck.

But those plans could be thrown off track by reanalyses of Liotta and Petricoin's data by independent groups, which have raised serious doubts about OvaCheck's reliability.

These questions prompted the Society of Gynecologic Oncologists to review all of the published work about OvaCheck. On 7 February, the society declared that "more research is needed to validate the test's effectiveness before offering it to the public".

Early warning

Critics warn that the episode illustrates the dangers of moving rapidly to the clinic with immature technologies such as those of proteomics. They say that scientists and regulators need to develop standards to ensure that such tests really work before they hit the market, because early detection

"Whether or not OvaCheck works, we will learn from this experience what rules of evidence we might apply in the future to find useful results more efficiently."

Bioinformatics[2]. They had reanalysed a data set that Liotta and Petricoin's team posted online in August 2002. Sorace and Zhan similarly found numerous differences in the protein patterns that discriminated between the cancer patients and the healthy controls. The trouble, according to Sorace and Zahn, was that these looked more like experimental artefacts than real biological differences.

The proteomics test relies on using gravity and electric fields to separate the proteins in a given sample. Each protein is then given a number that represents the ratio of its charge and mass — called its m/z value. The test identifies patterns in these numbers to give a diagnosis. Sorace and Zahn were concerned because the most marked differences between the cancer patients and unaffected controls occurred for proteins with m/z values of less than 500. Many spectrometry experts consider m/z values below 2,000 to be suspect, because they tend to include values generated by experimental artefacts or

SPECIAL SECTION

GENES IN ACTION

NEWS

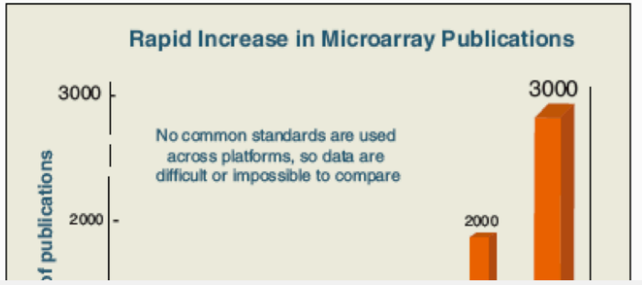
Getting the Noise Out of Gene Arrays

Thousands of papers have reported results obtained using gene arrays, which track the activity of multiple genes simultaneously. But are these results reproducible?

When Margaret Cam began hunting for genes that are turned up or down in stressed-out pancreas cells a couple of years ago, she wasn't looking for a scientific breakthrough. She was shopping. As director of a support lab at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), she wanted to test-drive manufactured devices called microarrays or gene arrays that measure gene expression; she had her eye on three different brands. These devices are hot, as they provide panoramic views of the genes that are active in a particular cell or tissue at a particular time.

being highly up- or down-regulated.

The disharmony appears in a striking illustration in Cam's 2003 paper in *Nucleic Acids Research*. It shows a Venn diagram of overlapping circles representing the number of genes that were the most or least active on each device. From a set of 185 common genes that Cam selected, only four behaved consistently on all three platforms—"very low concordance," she said at an August forum in Washington, D.C., run by the Cambridge Healthtech Institute, based in Newton Upper Falls, Massachusetts. Using less rigorous criteria, she found about 30% agreement—but never more than 52% between two brands. "It was nowhere near what we would expect if the probes were assaying for the same genes."



Cam's findings caused "one's jaw to drop," says Marc

he gathered on kidney tu
significant it seemed.

But those who have
gene expression arrays a
lems to early growing p
that experienced labs are
useful clinical inform
whether a breast cancer
require strong chemother
analytical methods will r
combine results from dif
and devices. Francis B
University's Weill Medic
York City insists that arr
one digs deeply into the u

Imperfections

Digging into the biology
did after her experiment
of discordant data. She
Marvin Gershengorn's
wanted to pick out a set c
in pancreatic tumor c
differentiation. From the
go on to examine how i
"We were very surpris
when they couldn't cros
from studies done w
Agilent, and Amershar

Check your understanding

1. Can two samples have equal gene expression but different splice-isoform proportions?
2. Why does one-colour microarray comparison require separate chips?
3. In chip manufacture, what determines **where** the next nucleotide attaches?
4. How are reverse transcription, PCR and bridge amplification different?
5. What does an RNA-seq read tell you—and what does it not tell you?

Next: preprocessing, quality control and differential expression.

Teaching resources and further reading

- **Animations:** CSHL DNA Learning Center — local copies of RNA splicing, PCR and GeneChip videos.
- **Splicing figures:** Philippe Hupé and Agathman, CC BY-SA 3.0.
- **Chip manufacture:** Affymetrix / Thermo Fisher educational guide.
- **Sequencing:** Illumina SBS and patterned flow cells.
- **Analysis:** Conesa et al. (2016), Genome Biology.

Detailed provenance and reuse information: MEDIA_SOURCES.md. Original interactive diagrams are editable teaching schematics; they simplify molecular scale and chemistry.