

Integrative Genomics, Mapping and Functional Explanation

Monday 28.11.11

Wednesday 30.11.11

G - Genome



T - Transcriptome



P - Proteome



M - Metabolome



F - Phenome

G → T



G → F



The Cell, the Central Dogma and the Multicellular Organism

The Cell – ignoring shape and compartmentalisation (10^{-5} m):

DNA – string over 4 letters/nucleotides {A,C,G,T}

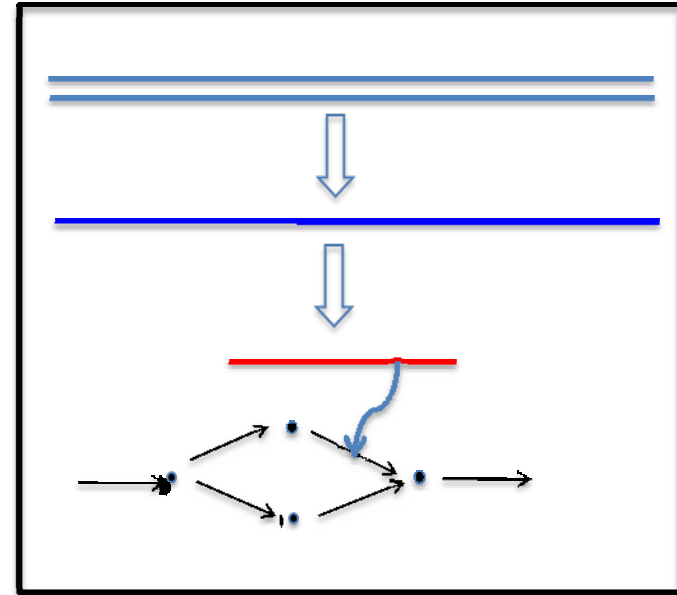
Transcribed by base pairing (A-T(U), C-G) into:

RNA – string over 4 letters/nucleotides {A,C,G,U}

Nucleotides in groups of 3 (codons) translated into amino acids:

Protein – string over 20 letters/amino acids

Proteins governs (among other things) Metabolism

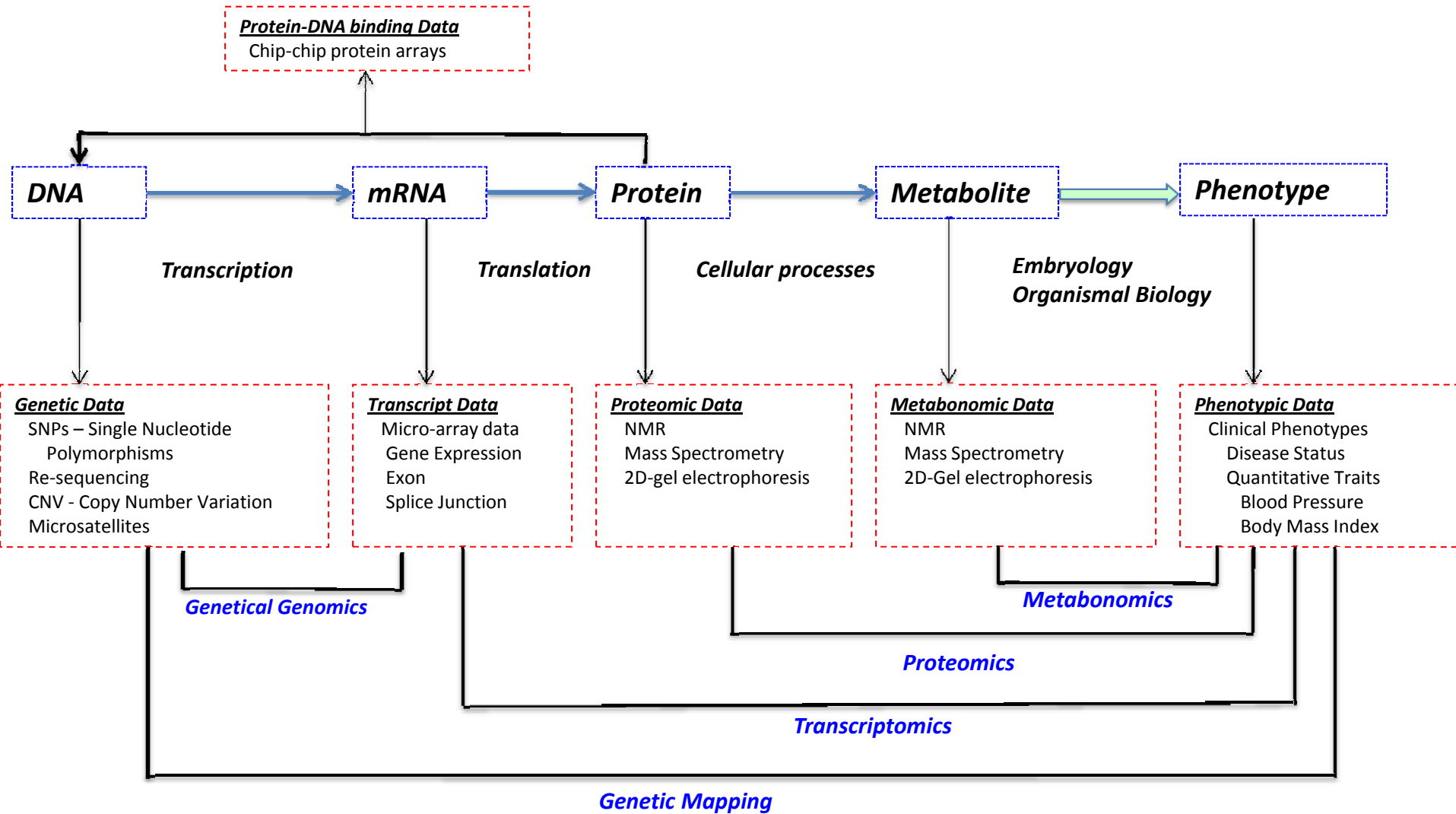


Epigenetics – DNA and chromosome is modified as part of governing regulation.

Data: *highthroughput*-collected without reference to a hypothesis, *experiment* – data collected relative to hypothesis

The Cell creates the individual through ~40 duplications

The Central Dogma & Data



Structure of Integrative Genomics

Classes

DNA

mRNA

Protein

Metabolite

Phenotype

Parts

.....

.....

.....

.....

.....

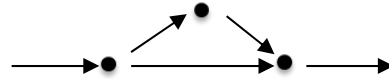
Concepts

G → F Mapping



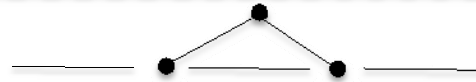
Models: Networks

Physical models:



Systems Biology

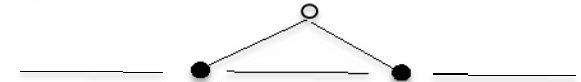
Phenomenological models:



Integrative Genomics

Hidden Structures/ Processes

○ Unobserved/unobservable

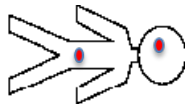


Knowledge:

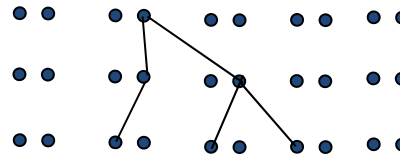
Externally Derived Constraints on which Models are acceptable

Evolution:

Cells in Ontogeny



Individuals/Sequences in a Population



Species



Analysis:

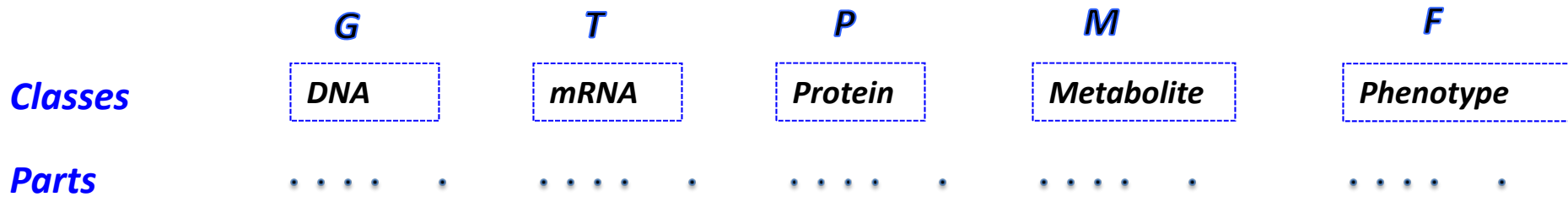
Data + Models + Inference



Model Selection

Functional Explanation

The key questions for any data type(s)



- *What is the state space of a single of observable and its (unobservable) biological state ?*
- *What is the dimension of the observation vector at each level?*
- *What is the distribution of an individual observable*
- *Are there correlation **within** a level? Statistical? Mechanistic?*
- *Are there correlation **between** levels? Statistical? Mechanistic?*
- *Are there conditional independencies? Say T and M are conditionally independent given P ?*
- *How does a level evolve between species? How does it vary within a population?*
- *Does it vary between tissues or diseases states?*

G: Assembly and Hybridisation

Target genome

3×10^9 bp

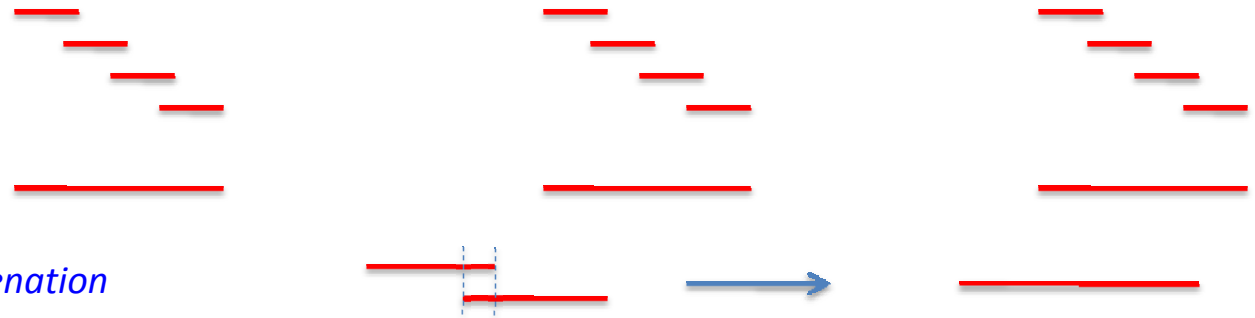
(unobservable)

Reads

3-400 bp

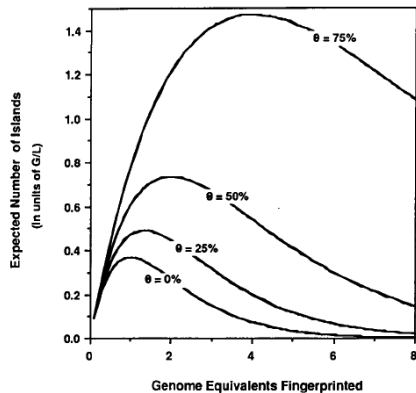
(observable)

Contigs



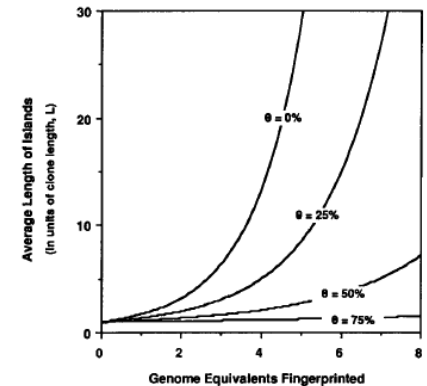
Sufficient overlap allows concatenation

Contigs and Contig Sizes as function of Coverage, Genome Size (G), Read Size (L) and overlap (θ):

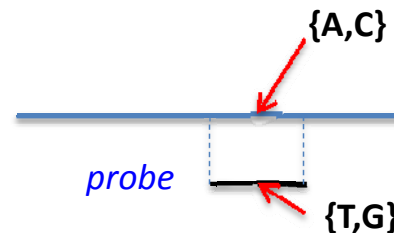


Approximate value of G/L

	Phage (15kb)	Cosmid (40kb)	Yeast (1Mb)
<i>E. coli</i>	267	100	4
<i>S. cerevisiae</i>	1333	500	20
<i>C. elegans</i>	5,667	2,125	85
Human	200,000	75,000	3,000

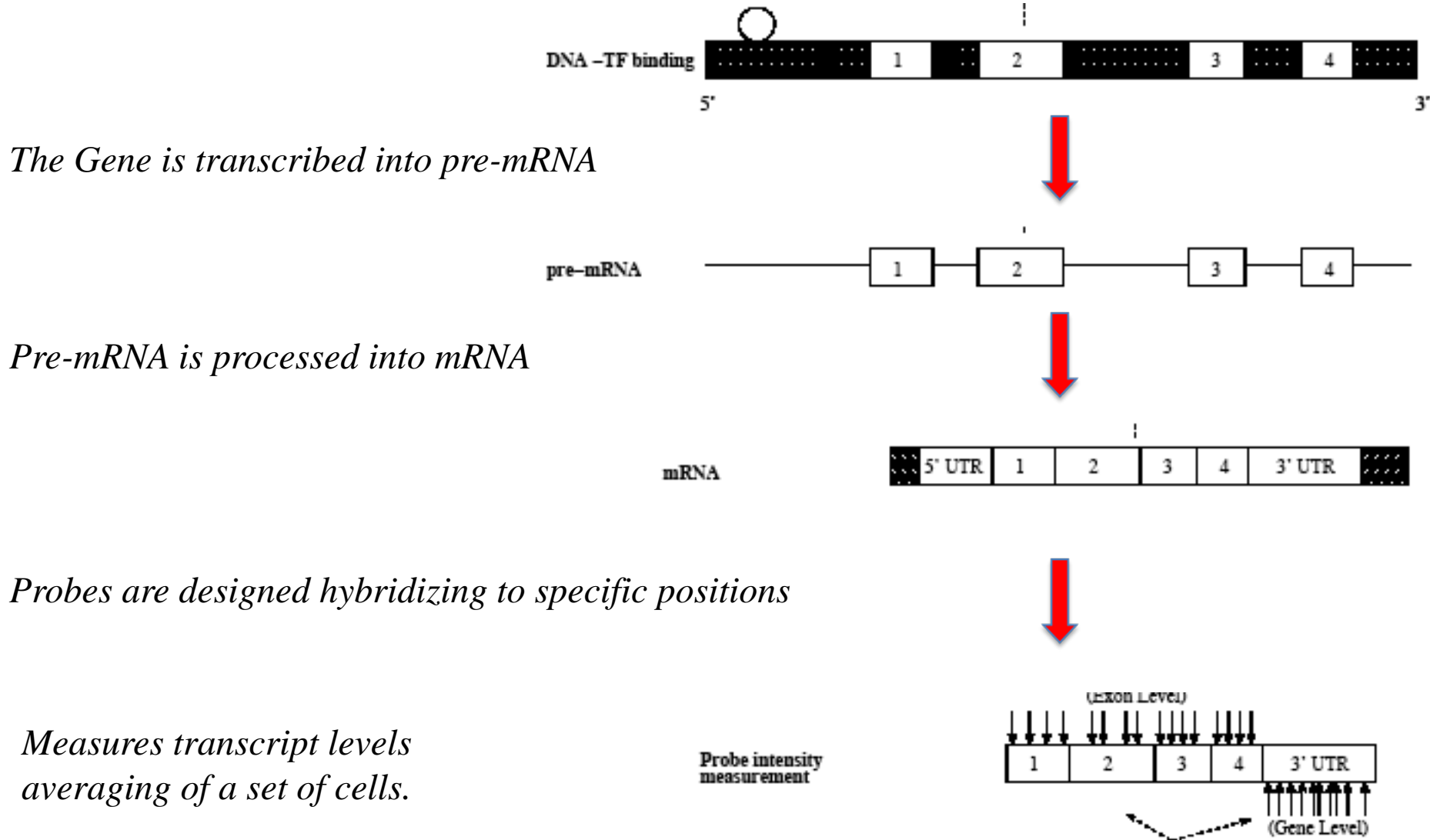


Complementary or almost complementary strings allow interrogation.



T - Transcriptomics

Classical Expression Experiment:



T - Transcriptomics

RNA-Seq Expression Experiment:

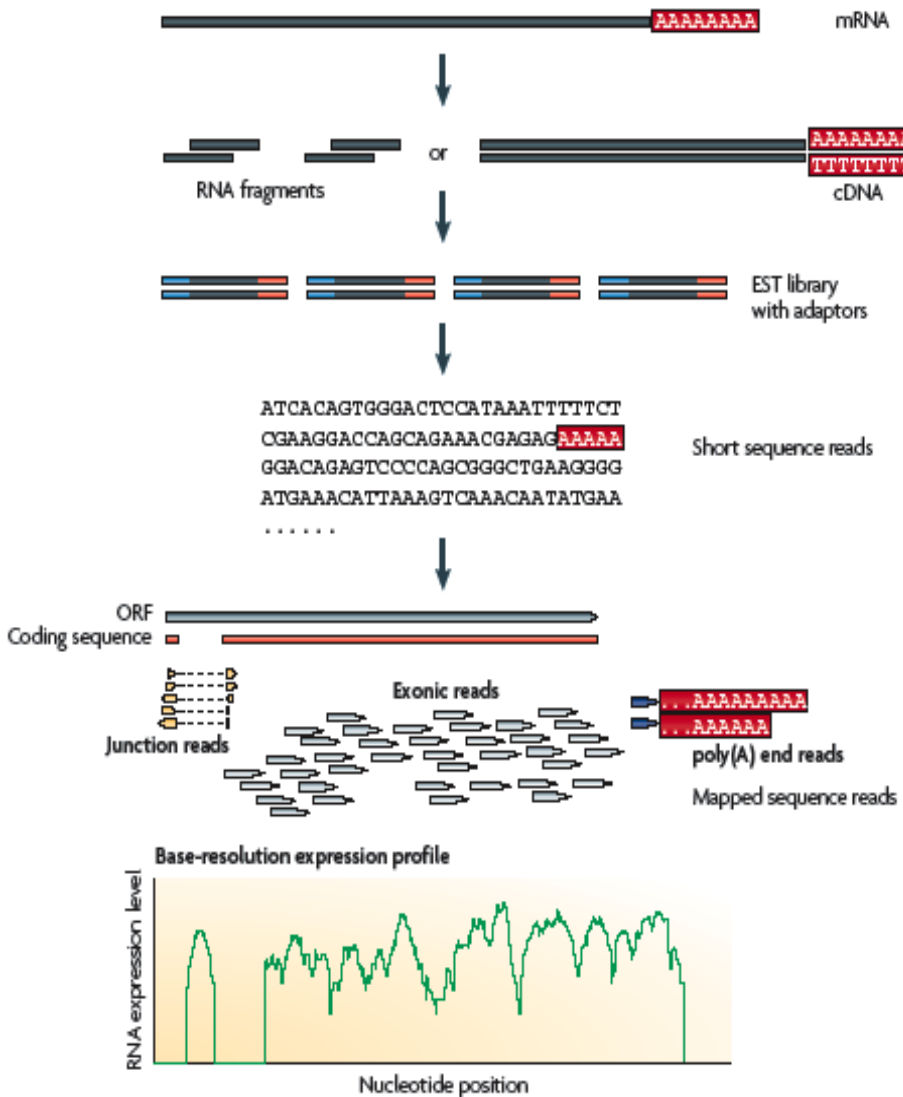
Advantages - Discoveries

More quantitative in evaluating expression levels

More precise in positioning

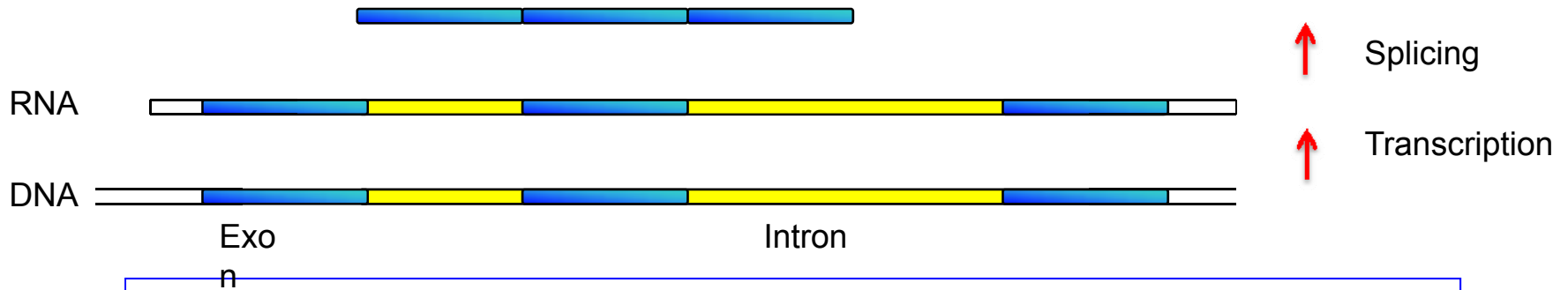
Much more is transcribed than expected.

Transcription of genes very imprecise



Genomics → Transcriptomics: Alternative Splicing

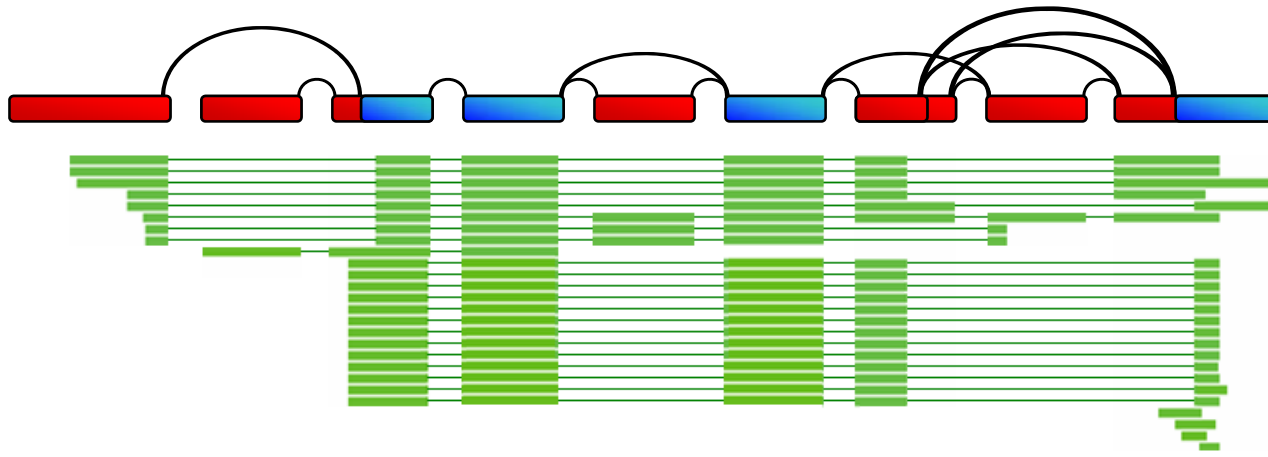
- *AS: one genomic segment can create different transcripts by skipping exons (sequence intervals)*



Problem: Describe the set of possible transcripts and their probabilities.

Define the alternative splicing graph (ASG) –

- *Vertices are exon fragments*
- *Edges connect exon fragments observed to be consecutive in at least one transcript*
- *This defines a directed, acyclic graph*
- *A putative transcript is any path through the graph*



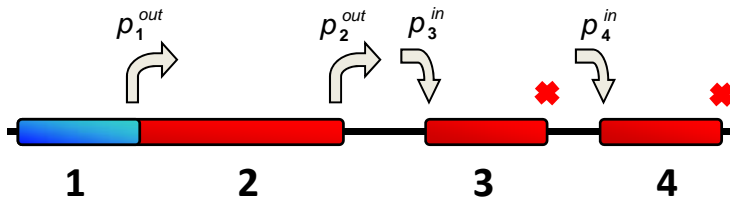
$G \rightarrow T$: Alternative Splicing

Problem: Inferring the ASG from transcripts

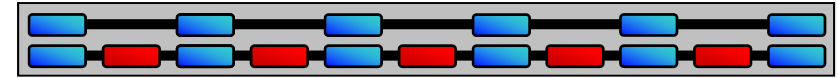
- *Maximally informative transcripts*
- *Minimally informative transcripts*
- *Random transcripts*

A Hierarchy of Models can be envisaged

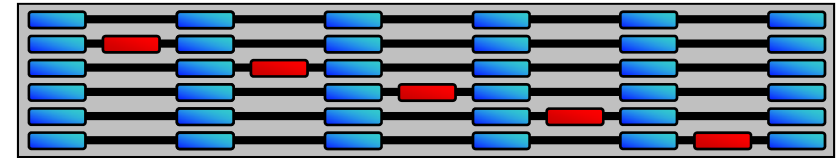
*Simpler still: model ‘donation’ and ‘acceptance’ separately
Jump ‘in’ or ‘out’ of transcript with well-defined probabilities
Isolated exons are included independently, based only on the
strength of its acceptor site*



This ASG could have been obtained from as few as two ‘informative’ transcripts...



...or as many as six. There are 32 putative transcripts.

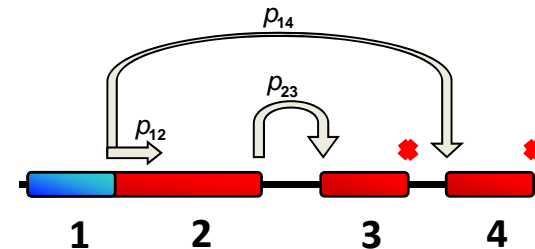


Enrich the ASG to a Markov chain

Pairwise probabilities

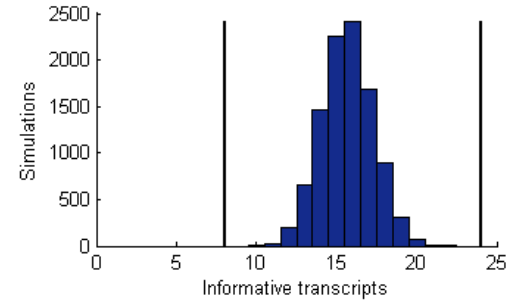
Transcripts generated by a ‘walk’ along the ASG

A natural model for dependencies between donors and acceptors

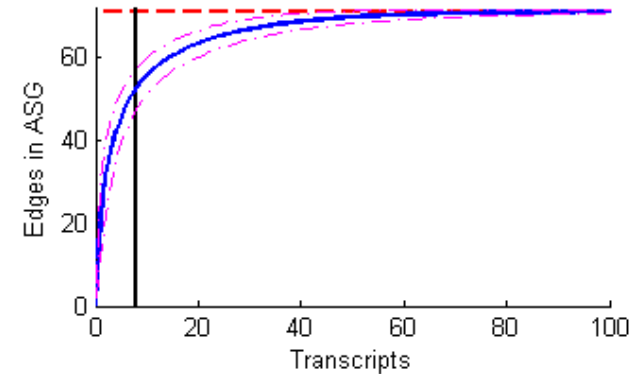


$G \rightarrow T$: Alternative Splicing

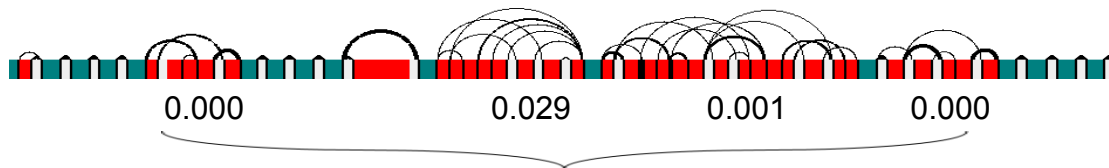
- *The distribution of necessary distinct transcripts*



- *The size of the inferred ASG*



- *Testing nested ASG modes*



Pairwise model: V^2 parameters

In-out model: V parameters

Models can be nested:

$In-out \subseteq pairwise \subseteq non-parametric$

Hence, given sufficient observations, likelihood ratio tests can determine the most appropriate model for transcript generation

The pairwise model was accepted, In-Out rejected

$G \rightarrow F$

- *Mechanistically predicting relationships between different data types is very difficult*
- *Empirical mappings are important*
- *Functions from Genome to Phenotype stands out in importance*

G is the most abundant data form - heritable and precise. F is of greatest interest.

