

RNA-seq (1)

**Peter N.
Robinson**

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

RNA-seq

Read mapping and Quantification

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

Genomics: Lecture #12

Today

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

- Gene Expression
- Previous gold standard: Microarrays
- Basic RNA-seq protocol and transcript quantification
- RNA-seq read mapping

Eukaryotic Gene Expression: Overview

RNA-seq (1)

Peter N.
Robinson

Microarrays

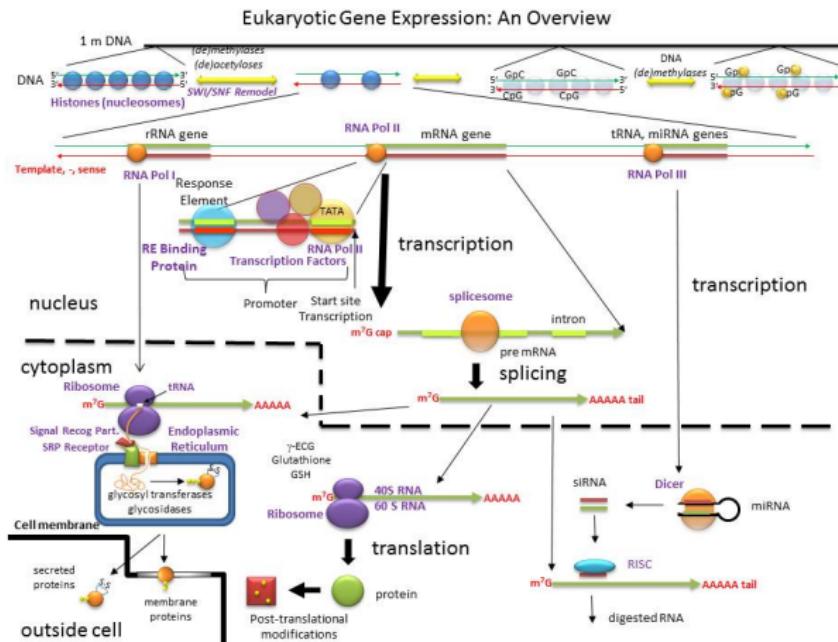
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



Graphics credit: CSBCJU; Biochemistry, Dr Jakubowski

<http://employees.csbsju.edu/hjakubowski/classes/ch331/bind/olbindtranscription.html>



Microarrays

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

- Hybridization of samples to thousands of probes on a slide simultaneously
- Many applications:
 - ① Transcriptional profiling (e.g., search for DE genes)
 - ② Copy-number variation
 - ③ SNP genotyping
 - ④ DNA protein interaction (Chip-on-Chip)
 - ⑤ many others
- Likely to be gradually replaced by next-generation sequencing, but the technology will probably remain relevant in the near future

Affymetrix Technology

RNA-seq (1)

Peter N.
Robinson

Microarrays

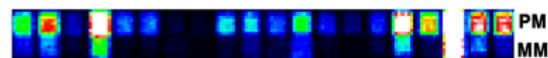
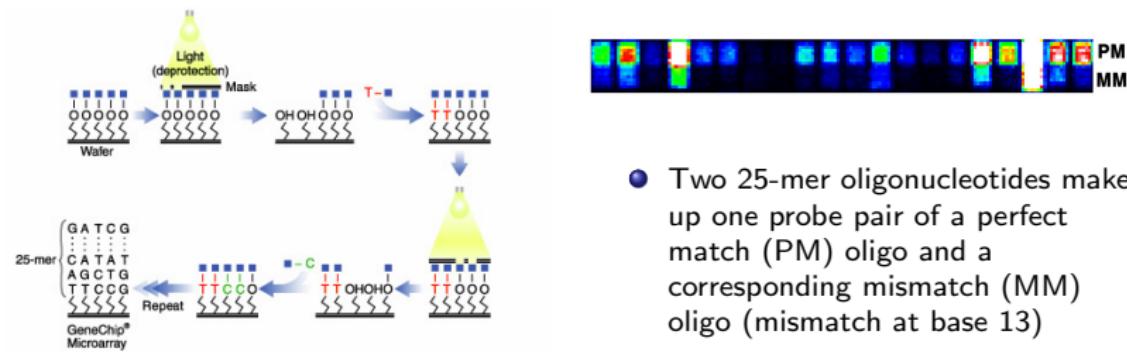
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



- The Affymetrix technology uses photolithographic synthesis of oligonucleotides on microarrays.
- The chip can hold up to 1.6 million features
- Two 25-mer oligonucleotides make up one probe pair of a perfect match (PM) oligo and a corresponding mismatch (MM) oligo (mismatch at base 13)
- The probe pairs allow the quantization and subtraction of signals caused by non-specific cross-hybridization.
- $PM - MM \Rightarrow$ indicators of specific target abundance.

Affymetrix Technology

RNA-seq (1)

Peter N.
Robinson

Microarrays

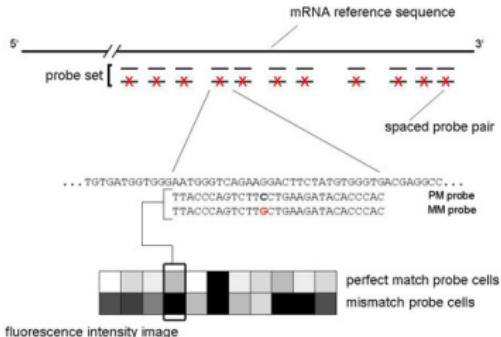
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



- The presence of messenger RNA (mRNA) is detected by a series of probes that differ in only one nucleotide.
- Hybridization of fluorescent mRNA to these probes on the chip is detected by laser scanning of the chip surface.
- A **probe set** consists 11 PM, MM pairs – the expression level is calculated by synthesizing information from all such PM/MM probes

Affymetrix Technology

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

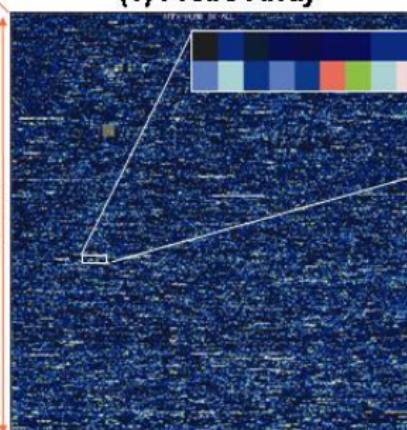
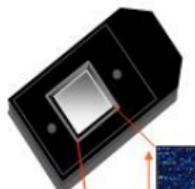
Alternative
splicing

mapping

cufflinks

Bipartite

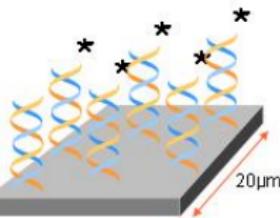
Human Genome U133A GeneChip® Array



(1) Probe Array

(4) Probe Cell

Each Probe Cell contains ~ 40×10^3 copies of a specific probe complementary to genetic information of interest
probe: single stranded, sense, fluorescently labeled oligonucleotide (25 mers)



(2) Probe Set

Each Probe Set contains 11 Probe Pairs (PM:MM) of different probes

(3) Probe Pair

Each Perfect Match (PM) and Mismatch (MM) Probe Cells are associated by pairs

The Human Genome U133 A GeneChip® array represents more than 22,000 full-length genes and EST clusters.



RNA-seq

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

RNA-seq can be used for many different types of experiment

- Measuring gene expression
- Differential expression
- Detecting novel transcripts
- Splice junction analysis
- De novo assembly
- SNP analysis
- Allele specific expression
- RNA-editing
- Studying small/microRNAs

blue: (Nearly) impossible with microarrays

green: Requires special chip

RNA-seq

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

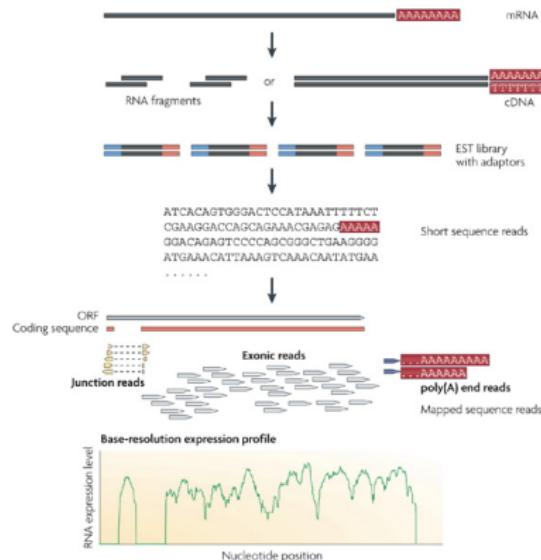
Alternative
splicing

mapping

cufflinks

Bipartite

General RNA-seq experiment



Nature Reviews | Genetics

Wang Z et al. (2009) RNA-Seq: a revolutionary tool for transcriptomics *Nature Reviews Genetics* 10:57–63

RNA-seq

RNA-seq (1)

Peter N.
Robinson

Microarrays

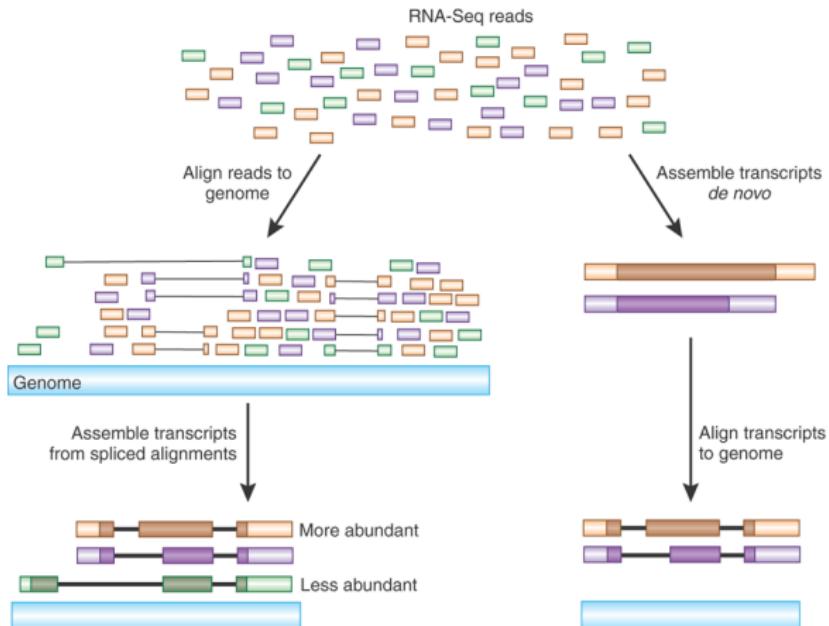
RNA-seq

Alternative splicing

mapping

cufflinks

Bipartite



RNA-seq

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

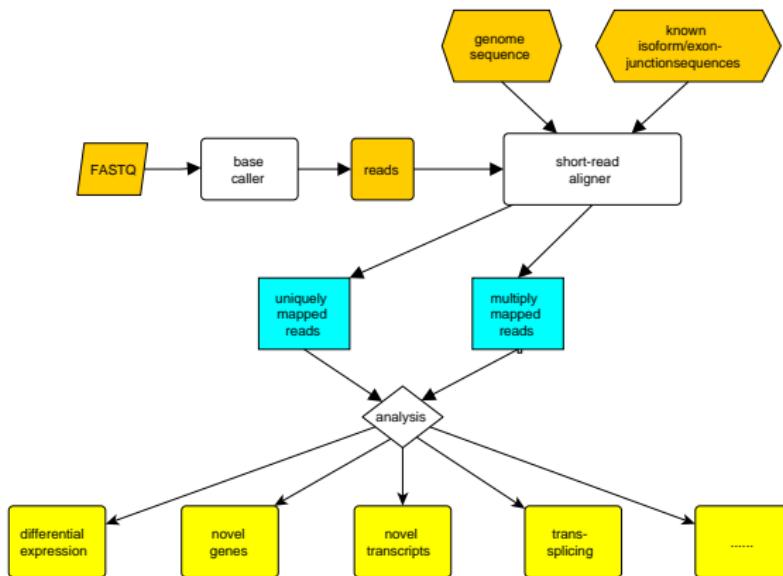
Alternative
splicing

mapping

cufflinks

Bipartite

Multiple downstream applications...



- Today: mapped reads \Rightarrow genes/transcript models
- Next time, we will talk about analyzing differential expression

Mapping Reads to Transcriptome

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

One of the critical steps in an RNA-Seq experiment is that of mapping the NGS reads to the reference transcriptome. However, we still do not know all transcripts even for well studied species such as our own.

- RNA-Seq analyses are thus forced to map to the reference genome as a proxy for the transcriptome.
- Mapping to the genome achieves two major objectives of RNA-Seq experiments:
 - ① Identification of novel transcripts from the locations of regions covered in the mapping.
 - ② Estimation of the abundance of the transcripts from their depth of coverage in the mapping.

Splicing (review)

RNA-seq (1)

Peter N.
Robinson

Microarrays

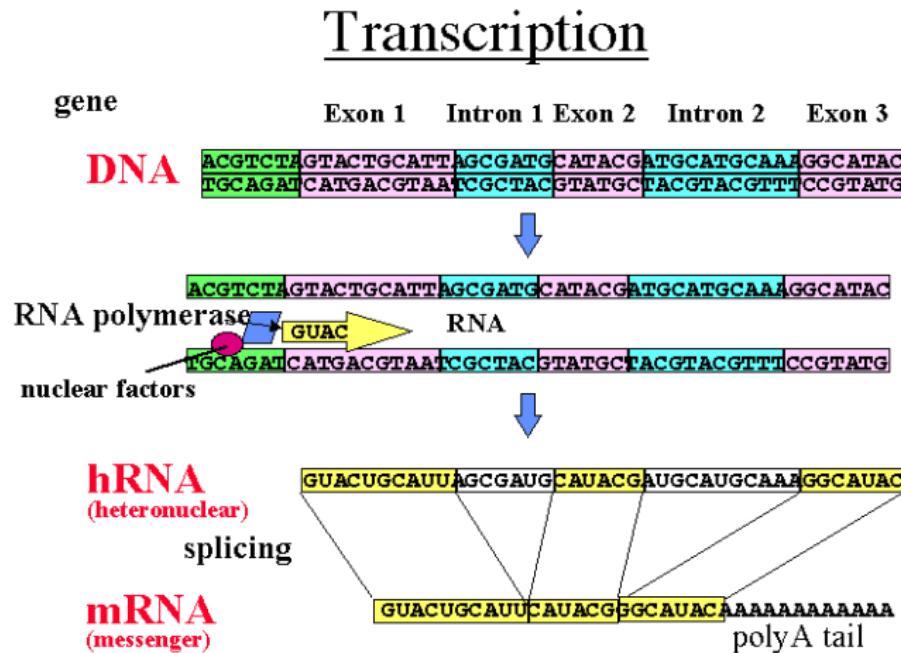
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



You should know (or review) general concepts of transcription, pre-RNA (near synonym to "heteronuclear RNA"), spliceosome, splicing

Splicing (review)

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

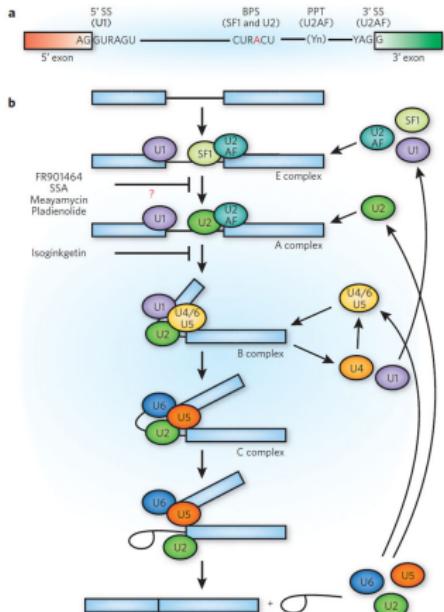
Alternative
splicing

mapping

cufflinks

Bipartite

- A spliceosome is a complex of snRNA and protein subunits
- A spliceosome removes introns from a transcribed pre-mRNA (hnRNA) segment.



Schneider-Poetsch et al (2010) *Nature Chemical Biology* 6:189–198

Alternative splicing

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

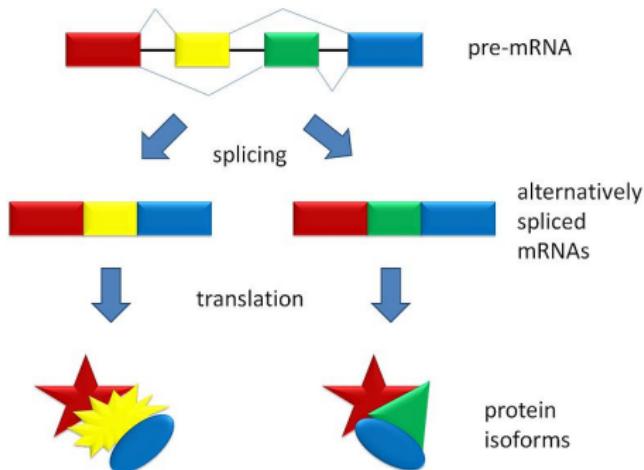
Alternative
splicing

mapping

cufflinks

Bipartite

Single gene coding for multiple proteins. Each distinct splicing is known as an **isoform** or transcript of the gene.



graphic credit: wikipedia

Alternative splicing

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

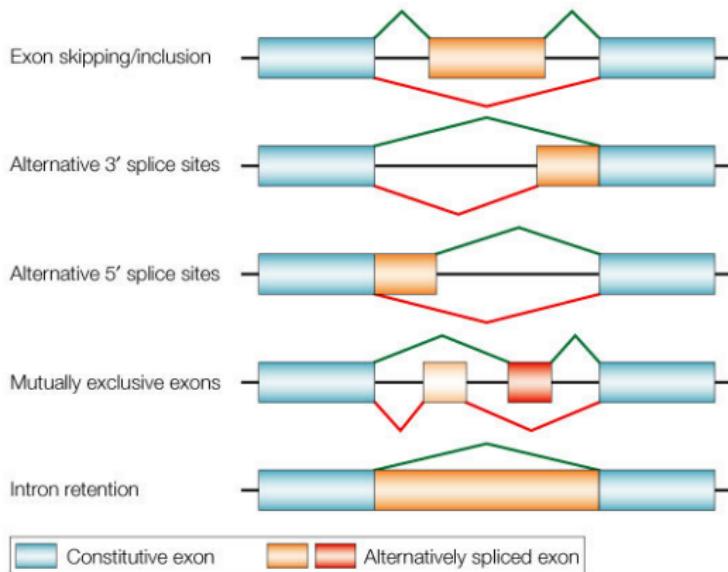
Alternative
splicing

mapping

cufflinks

Bipartite

Several different classes of alternative splicing events



Cartegni L et al. (2002) *Nature Reviews Genetics* 3:285–298

Alternative splicing: Biological roles

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

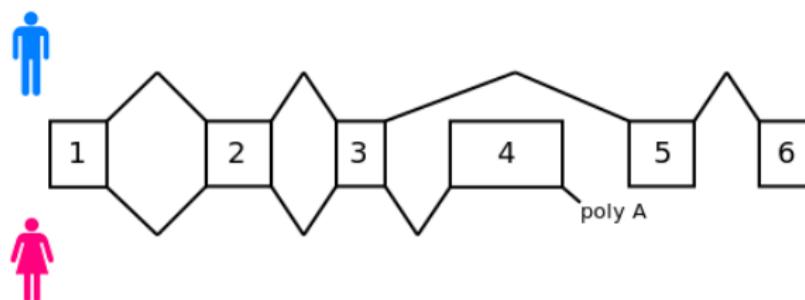
Alternative
splicing

mapping

cufflinks

Bipartite

The different isoforms of a gene can have quite distinct functional roles. Here we see the *Drosophila dsx* gene.



- Males: exons 1–3,5–6 ⇒ transcriptional regulatory protein required for male development.
- Females: exons 1–4 ⇒ transcriptional regulatory protein required for female development

Alternative splicing: Regulation

RNA-seq (1)

Peter N.
Robinson

Microarrays

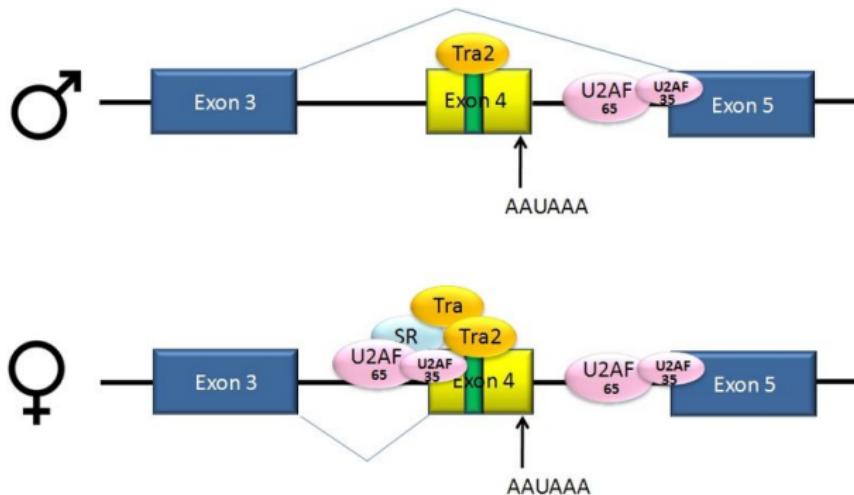
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



graphics credit: wikipedia

- The intron upstream from exon 4 has a polypyrimidine tract that doesn't match the consensus sequence well, so that U2AF proteins bind poorly to it without assistance from splicing activators. This 3' splice acceptor site is therefore not used in males.
- In general, we are just beginning to understand the regulatory mechanisms responsible for alternative splicing

Alternative splicing: Regulation

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

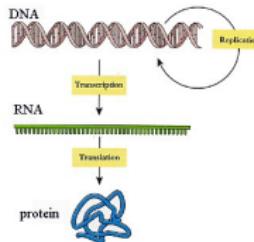
Alternative
splicing

mapping

cufflinks

Bipartite

The **central dogma** of molecular biology...is thus slightly dodgy



- Instead: One gene – many polypeptides
- Several proteins can be encoded by a single gene, rather than requiring a separate gene for each, and thus allowing a more varied proteome from a genome of limited size.
- Evolutionary flexibility. (“change just one isoform at a time”)

Alternative splicing and RNA-seq

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

- In the rest of this lecture, we will therefore discuss how one might investigate alternative splicing with RNA-seq
- There are by now a multitude of methods and algorithms, each with particular focuses, strengths, and weaknesses.
- Today, we will concentrate on one particular algorithm that uses some concepts from graph theory to infer the presence of known and novel isoforms of individual genes in RNA-seq data

The big picture

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

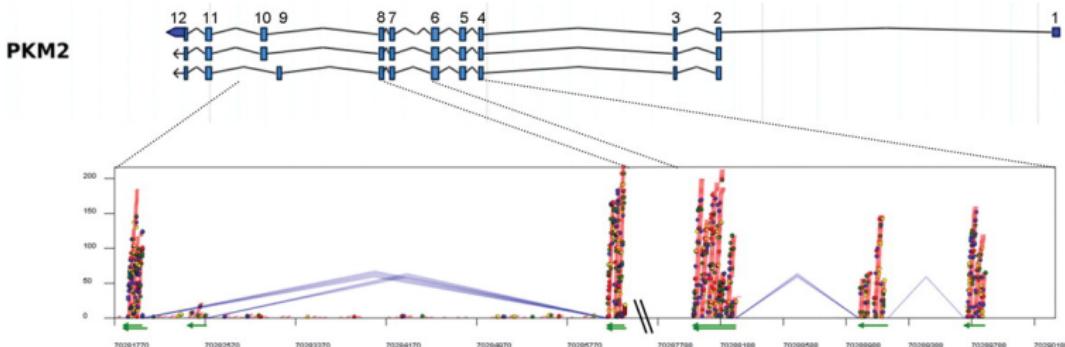
Alternative
splicing

mapping

cufflinks

Bipartite

Assuming we can map all reads correctly, we will find that there are some reads that map within exons, and some that span two or more exons.



Sultan M, Schulz MH et al. (2008) *Science* 321:956–960

- Two different splice junctions (blue lines) connect either exon 9 or exon 10 and identify alternative *PKM2* transcripts with mutually exclusive exons.

The big picture: Reference-based transcriptome assembly

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

There are two major classes of RNA-seq assembly algorithms

- ① Reference-based transcriptome assembly (We will talk about this today)
- ② de novo transcriptome assembly

Major steps:

- Map reads to genome
- Use annotation of locations and transcripts and their exons to identify and count reads that
 - ① map within single exons
 - ② span two or more exons
- Use this information to reconstruct an isoform distribution for each gene that appears likely given the patterns of reads (many different algorithms)

The big picture: Reference-based transcriptome assembly

RNA-seq (1)

Peter N.
Robinson

Microarrays

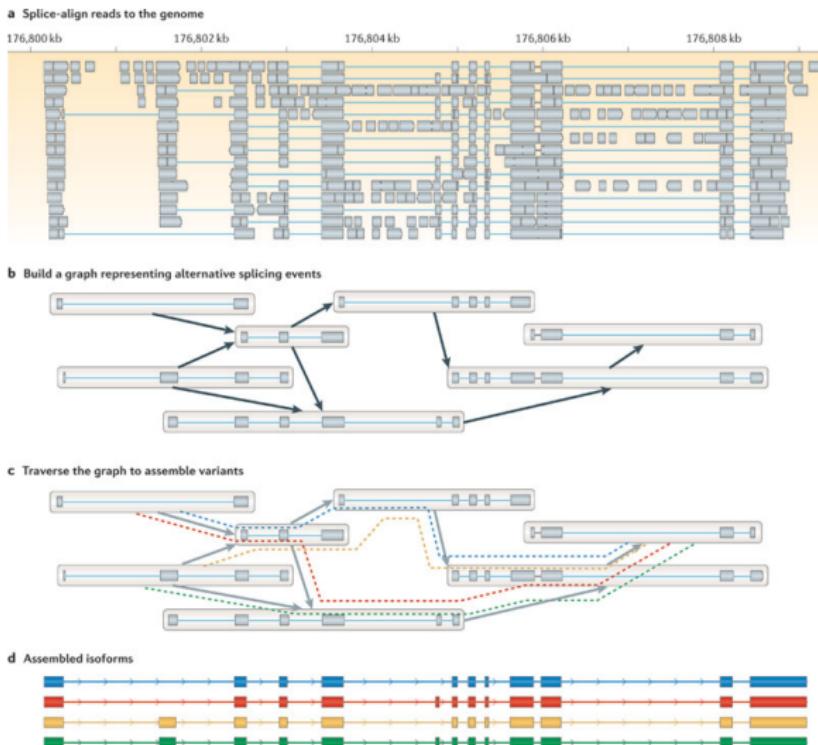
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



Martin JA, Wang Z (2011) *Nature Reviews Genetics* 12:671–682

Nature Reviews | Genetics



RNA-seq read mapping

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

RNA-seq read mapping uses the algorithms that you have learned about in the read-mapping lectures of this course. However, we additionally must take some particularities of RNA-seq data into account, including especially the fact that some reads might not map well to the genome because they “skip” one or more introns

- We will talk about tophat
- Trapnell C et al. (2009) *Bioinformatics* **25**:1105-11.
- Extension of original algorithm in supplementary material of Trapnell C et al. (2010) *Nat Biotechnol* **28**:511-5.

Tophat

RNA-seq (1)

Peter N.
Robinson

Microarrays

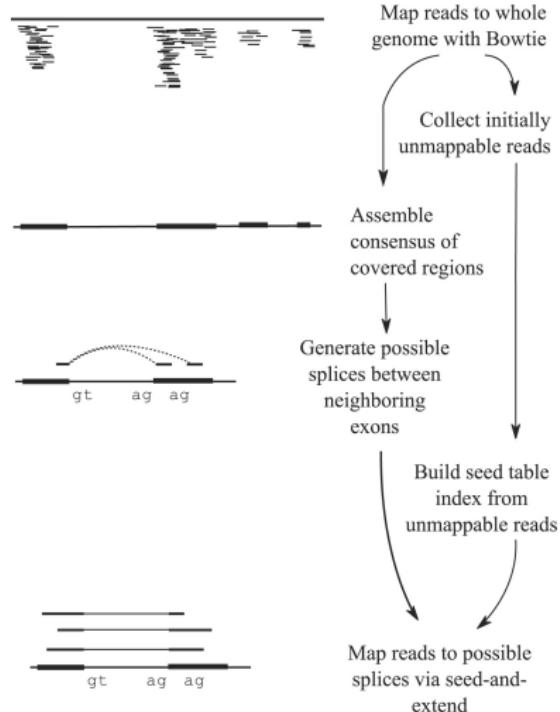
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



Tophat

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

We will talk here about the latest version of tophat (version 1.0.7 and above).

Algorithm 1 Find intron-spanning reads

- 1: Split read S (of length ℓ nucleotides) into $n = \lfloor \frac{\ell}{k} \rfloor$ segments (default: $k = 25$).
 - 2: Map each of the s_1, s_2, \dots, s_n reads to the genome separately with bowtie
 - 3: **if** s_1, s_2, \dots, s_n cannot be mapped contiguously **then**
 - 4: Mark S as a possibly intron-spanning read
 - 5: **end if**
-

Tophat

RNA-seq (1)

Peter N.
Robinson

Microarrays

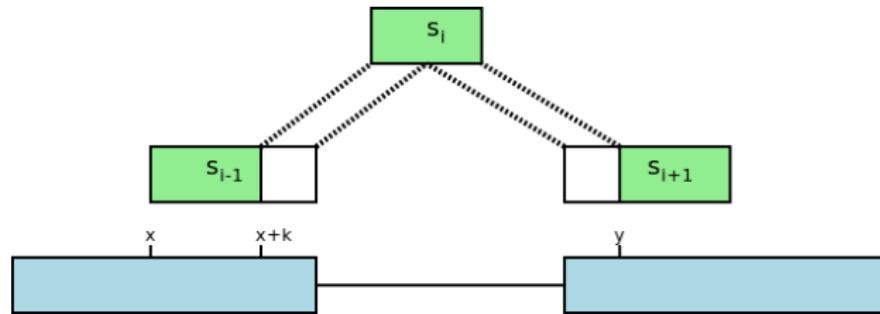
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



- When a segment s_i fails to align because it crosses a splice junction, but s_{i-1} and s_{i+1} are aligned (at positions x and y), Tophat looks for the donor and acceptor sites for the junction near x and y .
- Must be within k bases downstream of $x+k$ and within k bases upstream of y : There are thus k **possible exon-exon splice junctions**.

k is the segment size (25 nt)

Tophat

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Algorithm 2 Identify splice junctions

- 1: **for each** unmappable segment s_i of possibly intron-spanning
read S **do**
 - 2: concatenate k bp upstream of s_{i-1} and k bp
downstream s_{i+1}
 - 3: Align segment s_i to the concatenated
sequences with Bowtie.
 - 4: Merge contiguous and spliced segment
alignments for s_{i-1}, s_i, s_{i+1}
 - 5: **end for**
-

Tophat

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

There are many heuristics, bells, and whistles that tophat uses to perform the final alignment, that also take advantage from signals from readpairs, and wind up ranking candidate alignments according to some biological assumptions, such as for instance that really long introns are rare. Additionally, in cases where there are multiple plausible candidate alignments, the reads are assigned to each of n such alignments with a probability of $\frac{1}{n}$. We will not look at these details further.

Tophat

RNA-seq (1)

Peter N.
Robinson

Microarrays

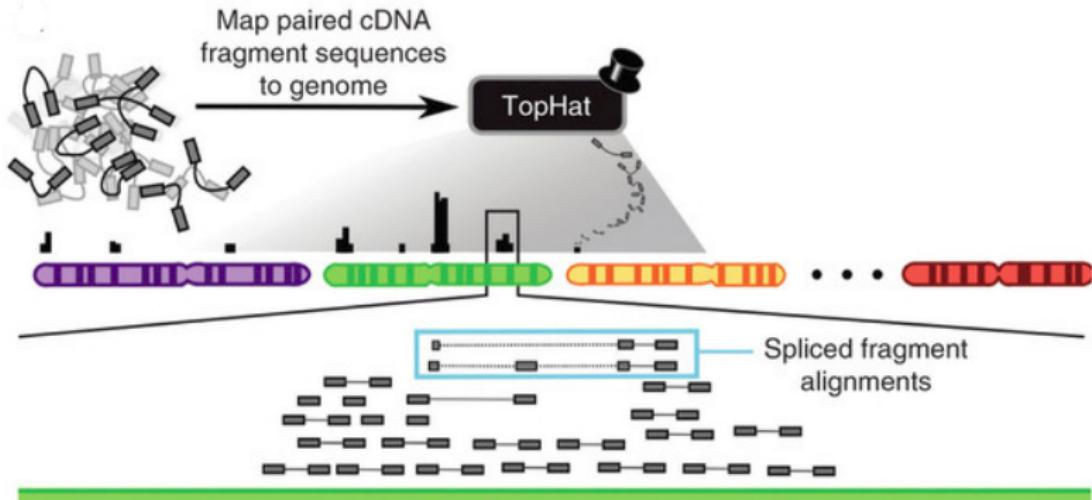
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



- cufflinks uses the alignments of tophat (or any alignment, i.e., samfile) to estimate the isoform distribution in a sample
- In the rest of this lecture, we will examine the graph algorithms used by cufflinks to do all of this

Cufflinks

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Trapnell C et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**:511-5.

- Probably the best known algorithm for reference-guided transcriptome assembly

Cufflinks: Typical data following tophat analysis

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Sample	Sequenced fragments	Aligned fragments	Singleton fragments	Spliced fragments	Multi-mapping fragments	Total alignments
-24 hours	42,184,539	35,852,366	11,031,886	8,824,825	1,768,041	41,663,170
60 hours	70,192,031	57,071,494	18,104,211	15,778,114	2,265,378	64,637,511
120 hours	41,069,106	27,914,989	14,431,734	7,711,026	1,881,772	33,929,133
168 hours	61,787,833	50,705,080	20,396,250	14,585,287	2,458,292	58,797,912
Total	215,233,509	171,543,929	63,964,081	46,899,252	8,373,483	199,027,726

- In a typical experiment, there were 215 million fragments, of which 171 million (79%) mapped to the genome.
- 46 million of these spanned at least one putative splice junction ($\approx 22\%$)
- In 63 million, only one end of the read could be mapped (singleton: $\approx 30\%$)
- 8 million reads mapped to multiple locations (multi-mapping fragments: $\approx 4\%$)

Cufflinks: Goals of transcript assembly

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

The assembly algorithm is designed to aim for a parsimonious explanation of the fragments from the RNA-seq experiment, i.e.:

- ① Every fragment is consistent with at least one assembled transcript.
- ② Every transcript is tiled by reads.
- ③ The number of transcripts is the smallest required to satisfy requirement (1)
- ④ The resulting RNA-Seq models display some desirable qualities

Cufflinks: Compatible Reads

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

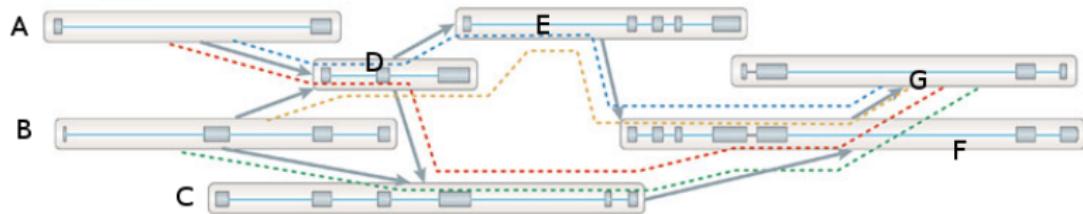
Alternative
splicing

mapping

cufflinks

Bipartite

Two reads are **compatible** if their overlap contains the exact same implied introns (or none). If two reads are not compatible they are **incompatible**.



- Read A is *incompatible* with reads B and C
- Read B is *compatible* with read C

Cufflinks: Compatible Reads

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

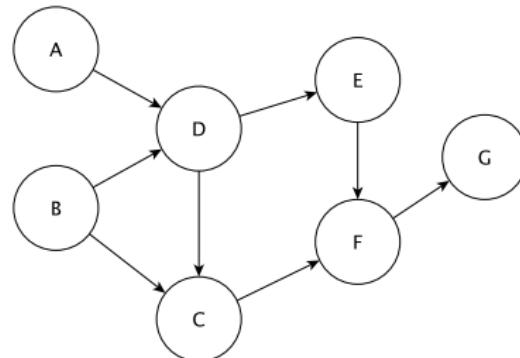
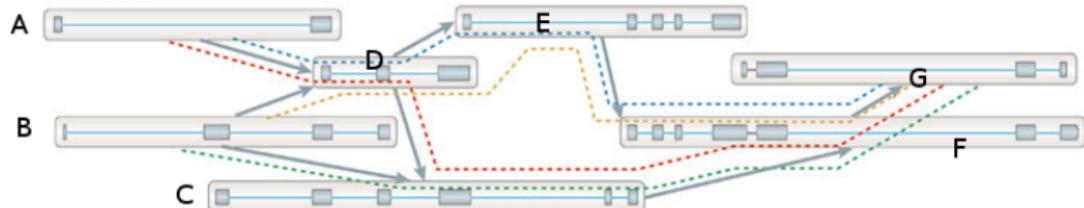
Alternative
splicing

mapping

cufflinks

Bipartite

We will now view this set of reads as a directed acyclic graph, which will first require some explanations.



Partial Order

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Definition

A relation \preccurlyeq on a set S is called a **partial order** if it is reflexive ($x \preccurlyeq x$), antisymmetric (if $x \preccurlyeq y$ and $y \preccurlyeq x$ then $x = y$) and transitive (if $x \preccurlyeq y$ and $y \preccurlyeq z$ then $x \preccurlyeq z$). A set S together with a partial ordering \preccurlyeq is called a partially ordered set or **poset** for short and is denoted (S, \preccurlyeq) .

- Partial orderings are used to give an order to sets that may not have a natural one.
- We use the notation $a \preccurlyeq b$ for $a, b \in S$ if a comes before b .
- If $a \neq b$, then we can also write $a \prec b$.
- \prec is not necessarily “less than”, rather it denotes the partial ordering

Partial Order and Comparability

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Definition

The elements a and b of a poset (S, \preccurlyeq) are called **comparable** if either $a \preccurlyeq b$ or $b \preccurlyeq a$. When $a, b \in S$ such that neither are comparable, we say that they are **incomparable**.

- (\mathbb{R}, \leq) real numbers and the less-than-equal-to relation:
All pairs of elements are compatible (this is a totally ordered set)
- $(\mathbb{Z}, \text{divisibility})$: natural numbers and the relation of “divisibility”, i.e., $m|n$: Only some pairs of elements of this set are comparable

Partial Order for mapped reads

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

We now define a partial ordering for the reads. Here, we will consider only the simple case of single reads and neglect the more complicated case of paired end reads.

- We define compatibility of two reads as mentioned above based on whether their overlap contains the exact same implied introns (or none)
- If two reads are compatible, they are considered comparable by our relation \preccurlyeq , otherwise not
- If we denote the starting mapped coordinate of a read x as $pos(x)$, then $x \preccurlyeq y$ iff $pos(x) \leq pos(y)$ and x and y are compatible with one another.

Partial Order for mapped reads

RNA-seq (1)

Peter N.
Robinson

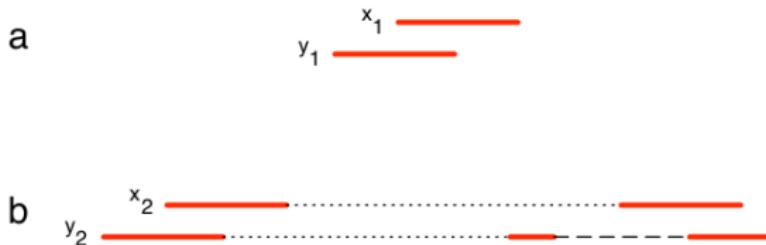
Microarrays

RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite



- x_1 and y_1 are **comparable** because they are compatible (they both contain no introns): $y_1 \preceq x_1$ because $\text{pos}(y_1) \leq \text{pos}(x_1)$
- x_2 and y_2 are **incomparable** because their overlap implies different introns. Thus, we cannot use the relation \preceq for this pair of reads

Chains and Antichains

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Definition

A **chain** is a set of elements in $C \subseteq S$ such that for every $x, y \in C$ either $x \preccurlyeq y$ or $y \preccurlyeq x$. An **antichain** is a set of elements that are pairwise incomparable.

Posets and DAGs

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

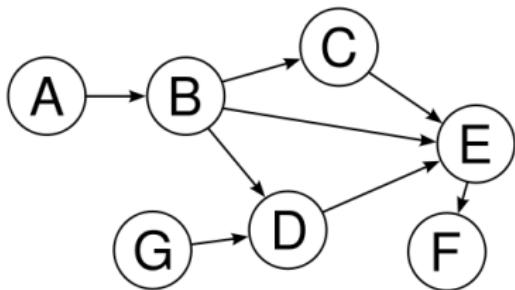
Alternative
splicing

mapping

cufflinks

Bipartite

It is easy to see that posets are equivalent to directed acyclic graphs (DAGs).



- For instance, $A \preccurlyeq B$ and $B \preccurlyeq C$, but A and G are incomparable with one another.

Dilworth's theorem

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

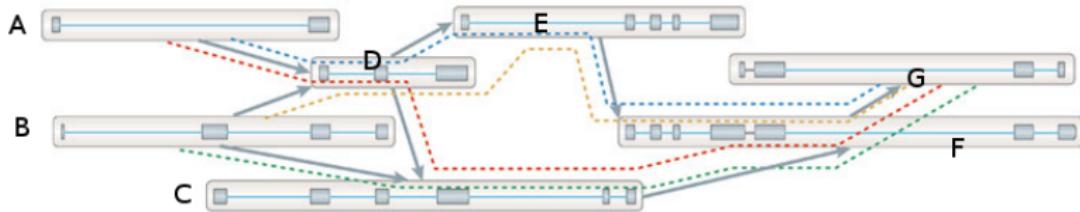
cufflinks

Bipartite

Theorem (Dilworth)

Let P be a finite partially ordered set. The maximum number of elements in any antichain of P equals the minimum number of chains in any partition of P into chains.

- In the setting of RNA-seq, this essentially means that the maximum cardinality of a set of fragments that are pairwise incompatible is the same as the minimum number of isoforms needed to explain the reads.



Let's check this with an example. Keep transitivity in mind!

Dilworth's theorem

RNA-seq (1)

Peter N.
Robinson

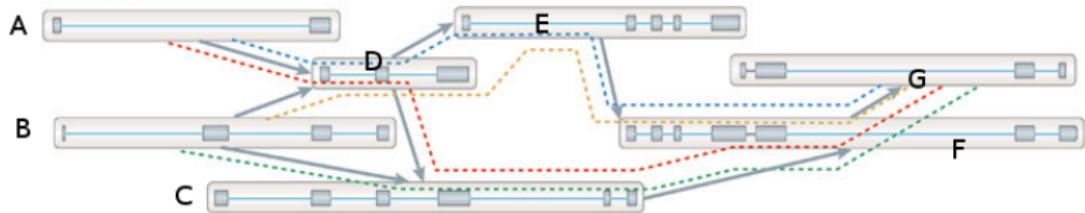
Microarrays

RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite



- Maximum cardinality of a set of pairwise incompatible fragments: 2, e.g., $\{A, B\}$ or $\{A, C\}$ or $\{C, E\}$. The set $\{A, B, C\}$ is no longer pairwise incompatible because $B \preccurlyeq C$. The set $\{A, C, E\}$ is no longer pairwise incompatible because $A \preccurlyeq C$
- Minimum number of chains that partition all reads: 2, e.g., $\{A \preccurlyeq D \preccurlyeq E \preccurlyeq F \preccurlyeq G, B \preccurlyeq C\}$. or $\{A \preccurlyeq D \preccurlyeq E, B \preccurlyeq C \preccurlyeq F \preccurlyeq G\}$.

RNA-seq assembly: Reformulating the problem

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

A partition of P into chains yields an assembly because every chain is a totally ordered set of compatible fragments x_1, x_2, \dots, x_l and therefore there is a set of overlapping fragments that connects them.

- By Dilworth's theorem, the problem of finding a minimum partition P into chains is equivalent to finding a maximum antichain in P^1
- In the following, we will show that this problem can be reformulated in the framework of bipartite graphs, which we will need to review first

¹Again, an antichain is a set of mutually incompatible fragments.

Matchings

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

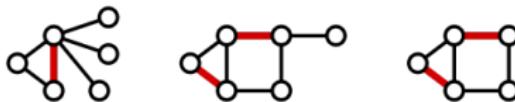
mapping

cufflinks

Bipartite

Given a graph $G = (V, E)$, a matching M in G is a set of pairwise non-adjacent edges; that is, **no two edges share a common vertex.**

- A **maximal matching** is a matching M of a graph G with the property that if any edge not in M is added to M , it is no longer a matching
- That is, M is maximal if it is not a proper subset of any other matching in graph G .



Maximum matchings

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

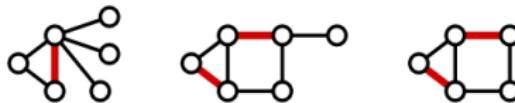
mapping

cufflinks

Bipartite

A **maximum matching** (also known as maximum-cardinality matching) is a matching that contains the largest possible number of edges.

- These matchings are maximal but two of them are not maximum



- These matchings are maximum (and therefore also maximal)



Vertex cover

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

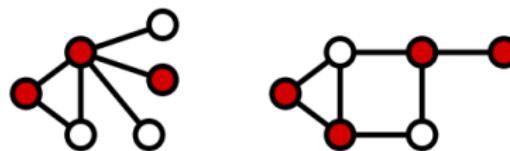
cufflinks

Bipartite

Definition

A **vertex cover** of a graph G is a set C of vertices such that each edge of G is incident to at least one vertex in C . The set C is said to cover the edges of G .

- Vertex covers in two graphs



Minimum Vertex cover

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

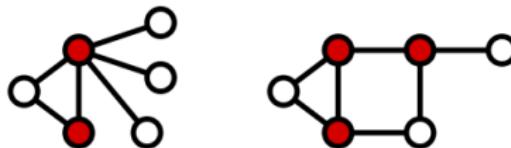
cufflinks

Bipartite

Definition

A **minimum vertex cover** is a vertex cover of smallest possible size.

- The vertex cover number τ is the size of a minimum vertex cover.
- For the left graph, $\tau(G) = 2$, for the right graph, $\tau(G) = 3$



König's theorem

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

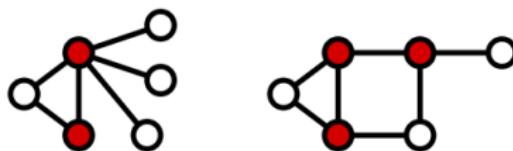
mapping

cufflinks

Bipartite

Theorem (König)

In a bipartite graph, the number of edges in a maximum matching equals the number of vertices in a minimum vertex cover.



Try it!

Cufflinks, König's theorem, and Dilworth's theorem

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Cufflinks exploits the equivalence of König's theorem and Dilworth's theorem to transform the problem of finding transcripts into a matching problem in a bipartite graph. We will explain this and then show how it works using our example graph from above.

bird's eye:

- A partition of P into chains yields an assembly because every chain is a totally ordered set of compatible fragments $x_1; \dots; x_l$ and therefore there is a set of overlapping fragments that connects them.
- The problem of finding such chains can be reduced to finding a maximum matching in an appropriate bipartite graph, which can be done at a complexity of $\mathcal{O}(VE)$ for a naive algorithm and $\mathcal{O}(\sqrt{VE})$ for a somewhat more sophisticated algorithm.

König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Theorem

Dilworth's theorem is equivalent to König's theorem

We need to show the the following are equivalent to justify the cufflinks algorithm:

Theorem (Dilworth)

Let P be a finite partially ordered set. The maximum number of elements in any antichain of P equals the minimum number of chains in any partition of P into chains.

Theorem (König)

In a bipartite graph, the number of edges in a maximum matching equals the number of vertices in a minimum vertex cover.

König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Theorem

Dilworth's theorem is equivalent to König's theorem

Proof (1): $K \rightarrow D$.

Let P be a poset with n elements. We define a bipartite graph $G = (U; V; E)$ where $U = V = P$, i.e. each partition in the bipartite graph is equal to P . Two nodes $u; v$ form an edge $(u; v) \in E$ in the graph G iff $u \prec v$ in P . □

see graph next page

König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

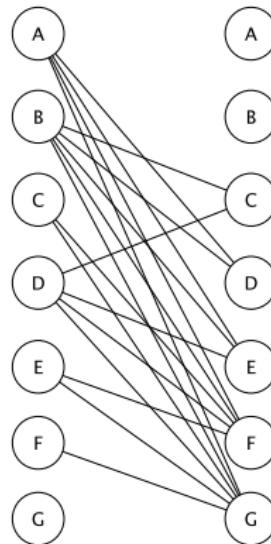
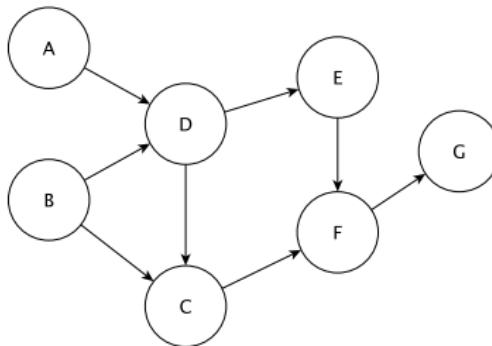
Microarrays

RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite



We want to prove: König (Number of edges in a maximum matching equals the number of vertices in a minimum vertex cover, see graph on right). \Rightarrow Dilworth (Minimum number of chains is equal to maximum number of elements in an antichain, see graph on left)

König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

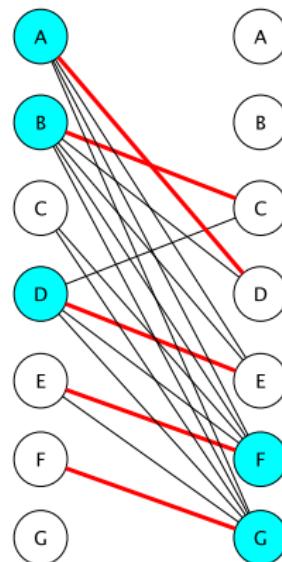
mapping

cufflinks

Bipartite

Proof (2): $K \rightarrow D$.

By König's theorem
there exist both a
matching M and a vertex
cover C in G of the same
cardinality.



- **M:** the five red edges
- **C:** the five blue vertices

König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

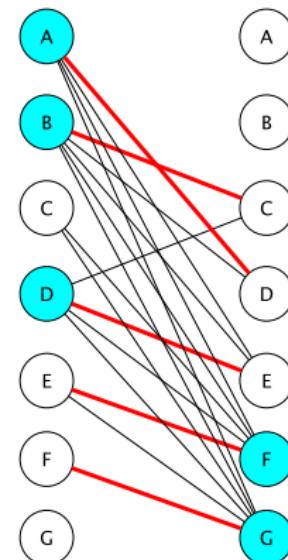
Bipartite

Proof (3): $K \rightarrow D$.

Let $T \subset S$ be the set of elements not contained in C .



- $C = \{A, B, D, F, G\}$
- $T = \{C, E\}$



Note that T is an **antichain** in the poset P (*why?*).

König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

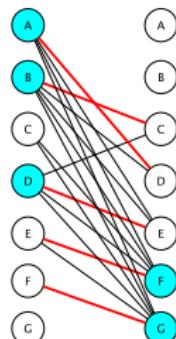
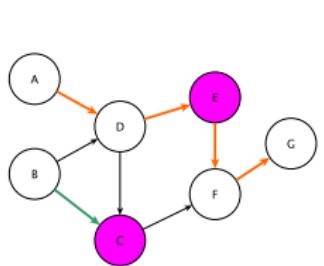
mapping

cufflinks

Bipartite

Proof (4): $K \rightarrow D$.

We now form a partition W of P into chains by declaring u and v to be in the same chain whenever there is an edge $(u; v) \in M$. Since C and M have the same size (by König's theorem), it follows that the antichain T (of the bipartite graph) and the partition W (of the DAG) have the same size.



König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

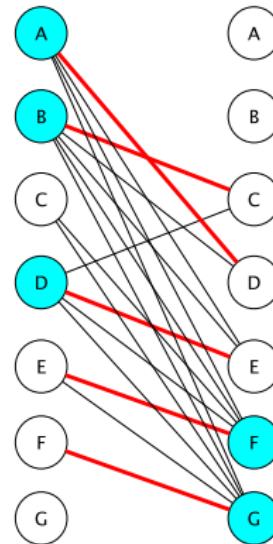
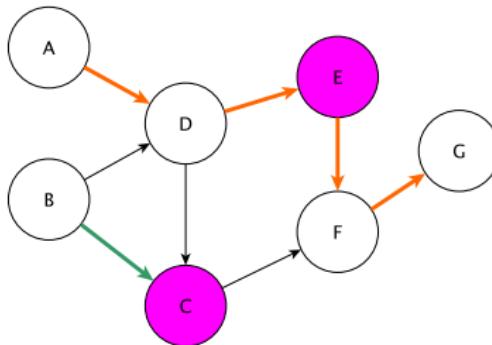
Microarrays

RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite



- Since C and M have the same size (König), it follows that T and W have the same size.
- Here: $T = \{C, E\}$ and $W = \{(A \rightarrow D \rightarrow E \rightarrow F \rightarrow G), (B \rightarrow C)\}$

König vs. Dilworth

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Proof: $K \rightarrow D$ (continued).

Therefore, we have shown that if the matching M and the vertex cover C have the same size (König), then the minimal number of chains (W) in our poset P has the same cardinality as the number of elements in an antichain of P (Dilworth), and the proof is finished. □

- A similar proof shows that Dilworth's theorem implies König's theorem (left as an exercise)
- Thus, we have shown that the two theorems are equivalent!

Reachability graph

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

The final ingredient we are missing is a way of finding a maximum cover in the bipartite graph, which will be termed the **reachability graph**.

- We will present a simple algorithm for finding a maximum cover in a reachability graph, using a simple bipartite graph to illustrate the algorithm.

Terminology (1)

RNA-seq (1)

Peter N.
Robinson

Microarrays

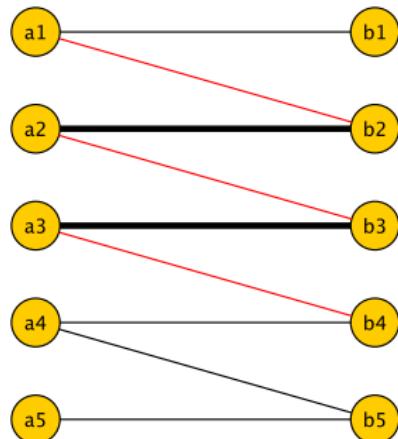
RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite

- The edges of a **matching** M are marked bold
- $v \in V$ is a **free vertex**, if no edge from M is incident to v (i.e., if v is not matched).
- Here, a_1, b_1, a_4, b_4, a_5 , and b_5 are free.



The next few slides on maximum mapping were adapted from lectures notes by C. Stein.

Terminology (2)

RNA-seq (1)

Peter N.
Robinson

Microarrays

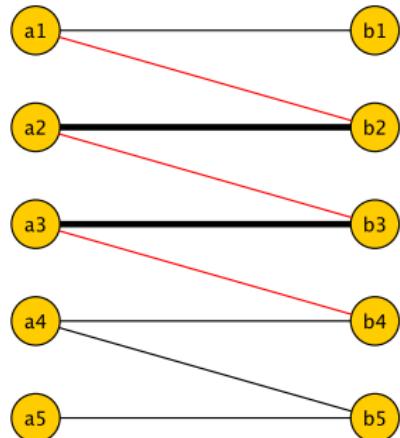
RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite

- P is an **alternating path** if P is a path in G , and for every pair of subsequent edges on P it is true that one of them is in M and another one is not.
- $\{a_1, b_1\}$ and $\{b_2, a_2, b_3\}$ are two examples of alternating paths,



Terminology (3)

RNA-seq (1)

Peter N.
Robinson

Microarrays

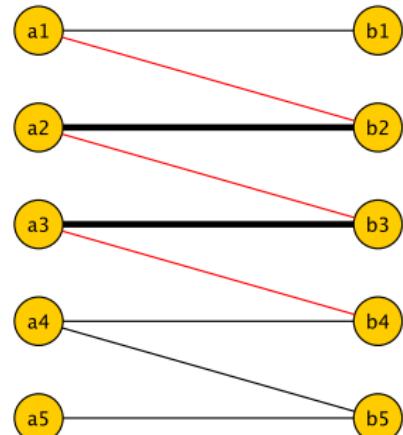
RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite

- P is an **augmenting path**, if P is an alternating path with a special property that its start and end vertex are free.
- $\{a_1, b_2, a_2, b_3, a_3, b_4\}$ is an augmenting path



a_1 and b_4 are free vertices because no edge from M (bold edges) is incident to them.

Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

The main idea for a simple algorithm to find a maximum matching on bipartite graphs exploits a fact about augmenting paths

- Given a matching M and an augmenting path P ,
 $M' = M \oplus P$ is a matching such that $|M'| = |M| + 1$.

Here, \oplus denotes the symmetric difference set operation: everything that belongs to both sets individually,

but doesn't belong to their intersection. Thus, $A \oplus B = (A \cup B) \setminus (A \cap B)$

Note that \setminus denotes set subtraction.

Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

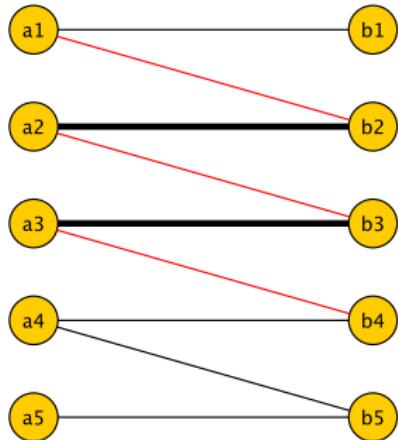
RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite

- **Proof:** every augmenting path P is alternating and starts and ends with a free vertex. Therefore, it must be odd length and must have one edge more in its subset of unmatched edges ($P \setminus M$) than in its subset of matched edges ($P \cap M$).
- Consider the augmenting path $P = \{(a1, b2), (b2, a2), (a2, b3), (b3, a3), (a3, b4)\}$ and the matching $M = \{(a2, b2), (a3, b3)\}$
- Then $M' = M \oplus P = \{(a1, b2), (a2, b3), (a3, b4)\}$



Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

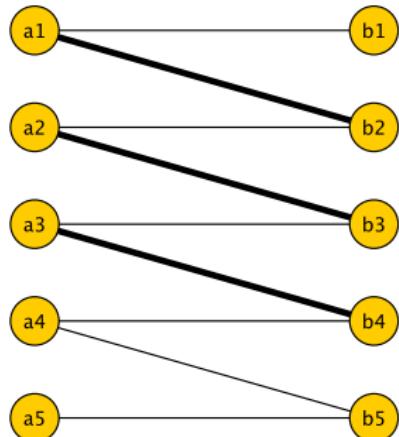
RNA-seq

Alternative
splicing
mapping

cufflinks

Bipartite

- The matching $M' = M \oplus P = \{(a1, b2), (a2, b3), (a3, b4)\}$
- Clearly, $|M'| = |M| + 1$.
- The operation of replacing the old matching M by a new one $M' = M \oplus P$ is called the augmentation over path P .



Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

The idea for an algorithm now becomes obvious. Starting with any matching in a bipartite graph G (e.g., an empty one), repeatedly find an augmenting path and augment over it, until there are no augmenting paths left.

Theorem

For a given bipartite graph G , a matching M is maximum if and only if G has no augmenting paths with respect to M .

Proof sketch.

If there is an augmenting path for a matching M of cardinality m , then by the above we can find a new matching with cardinality $m + 1$. 

Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

Algorithm 3 BIPARTITE-MATCHING(G)

- 1: $M = \emptyset$
 - 2: **repeat**
 - 3: $P = \text{AUGMENTING-PATH}(G, M)$
 - 4: $M = M \oplus P$
 - 5: **until** $P = \emptyset$
-

We now only need to show how to find an augmenting path in a bipartite graph.

Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

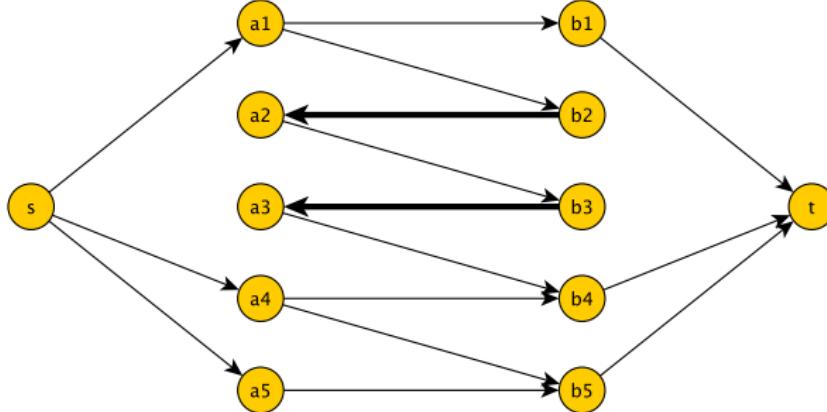
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



- Create a new graph by adding a source (s) and a sink (t) node
- Direct all matched edges from B to A , and all unmatched nodes from A to B . Add directed edges from the source to all unmatched nodes in A , and from all unmatched nodes in B to the sink

Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

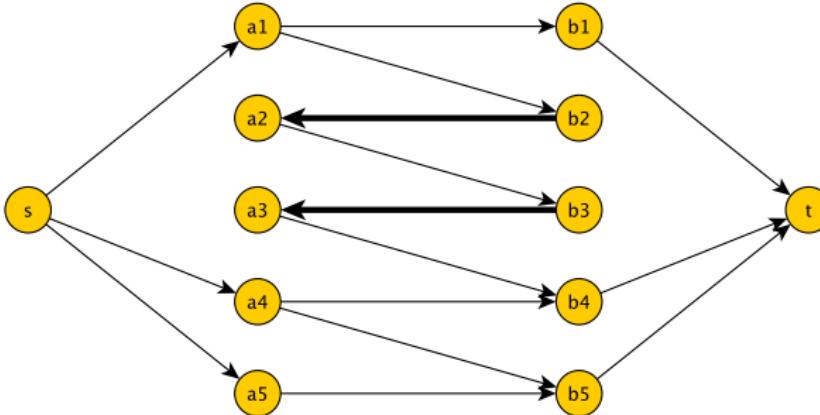
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



- Now all the directed paths in G are alternating
- A free vertex in B can be reached from a free vertex in A only via augmenting path.
- These paths can be found by performing a breadth-first-search (BFS) on the modified graph

Find augmenting path

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

The algorithm for finding an augmenting path can now be given as:

Algorithm 4 AUGMENTING-PATH(G, M)

- 1: Direct unmatched edges $A \rightarrow B$ and matched edges $B \rightarrow A$
 - 2: Attach source s and sink t to unmatched nodes
 - 3: Run BFS of G and identify a shortest path P from s to t
 - 4: Return $P \setminus \{s, t\}$
-

Maximum matching

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

- Let $m = |E|$ (number of edges) and $n = |V|$ (number of vertices)
- BFS is $\mathcal{O}(m)$
- A matching can be of size at most $\frac{n}{2} = \mathcal{O}(n)$, and each step of BIPARTITE-MATCHING adds one edge.
- Thus, BIPARTITE-MATCHING has an overall complexity of $\mathcal{O}(mn)$
- The Hopcroft-Karp Algorithm² improves on the simple algorithm and achieves a complexity of $\mathcal{O}(m\sqrt{n})$

²Which we will not cover here

cufflinks

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

- There are now a number of **additional steps** designed to extend and disambiguate the transcript models
- The fraction of mRNAs that contain an exon – the "Percent Spliced In" (PSI or Ψ) value – can be estimated as the ratio of the density of inclusion reads (i.e. reads per position in regions supporting the inclusion isoform) to the sum of the densities of inclusion and exclusion reads.
- The bipartite graph is weighted as to whether potentially adjacent fragments have similar Ψ values
- We will not discuss this further here

Alternative Transcript Events

Skipped exon (SE)



Retained Intron (RI)



Alternative 5' splice site (A5SS)



Alternative 3' splice site (A3SS)



Mutually exclusive exons (MXE)



Mutually exclusive 5' UTRs (5pMXE)



Mutually exclusive 3' UTRs (3pMXE)



Tandem UTRs



Total

Constitutive exon or region

Alternative exon or extension

cufflinks: Summary

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

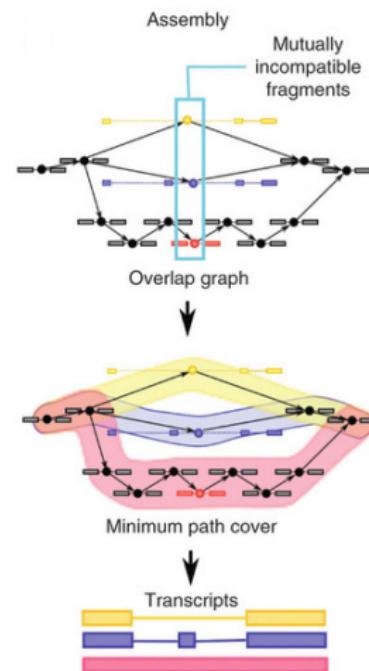
Alternative
splicing

mapping

cufflinks

Bipartite

- Identify incompatible fragments that must have originated from distinct mRNA splice forms
- Connect compatible fragments in an overlap graph
- Paths through the graph correspond to mutually compatible fragments
- Minimum path cover → transcripts
- Transcript abundance estimation



cufflinks

RNA-seq (1)

Peter N.
Robinson

Microarrays

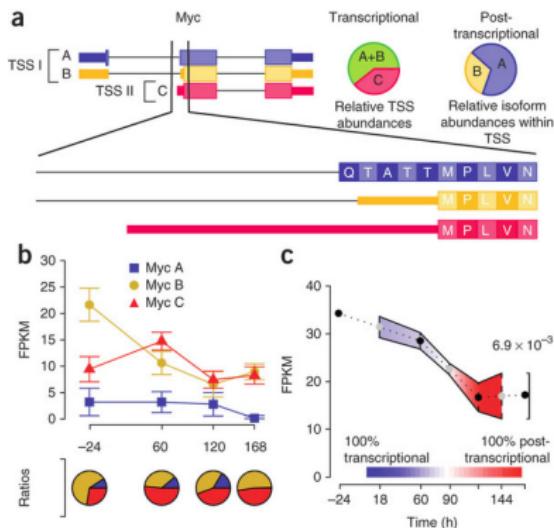
RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite



- As an example: Cufflinks identifies three isoforms of the *Myc* gene
- The three isoforms of Myc have distinct expression dynamics.

Summary

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

- In this lecture, we have looked at some algorithms used for mapping RNA-seq reads to the individual isoforms of a gene
- This is a key step towards analyzing alternative splicing
- Read mapping algorithms were adapted to take spliced reads into account (tophat)
- A graph algorithm was used to encode our biological knowledge about splicing (compatible and incompatible splice patterns) and identify isoforms (cufflinks)
- Next week: differential expression analysis with RNA-seq

Finally

RNA-seq (1)

Peter N.
Robinson

Microarrays

RNA-seq

Alternative
splicing

mapping

cufflinks

Bipartite

- Email: peter.robinson@charite.de
- Office hours by appointment

Further reading

- Trapnell C et al. (2009) TopHat: discovering splice junctions with RNA-Seq *Bioinformatics* **25**:1105-11.
- Trapnell C et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**:511-5.
- Trapnell C, et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks *Nat Protoc* **7**:562-78: An extremely useful “How To” (tutorial) that is highly recommended to get hands on experience with RNA-seq analysis