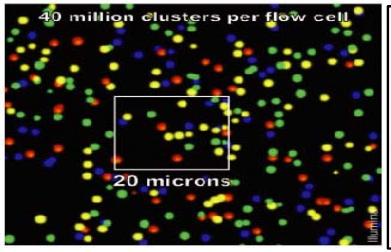
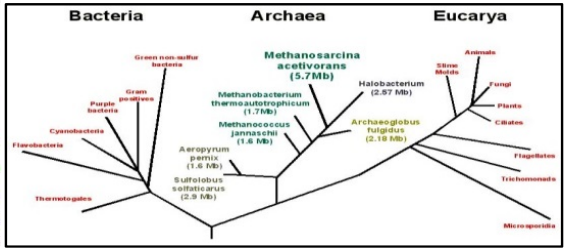
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



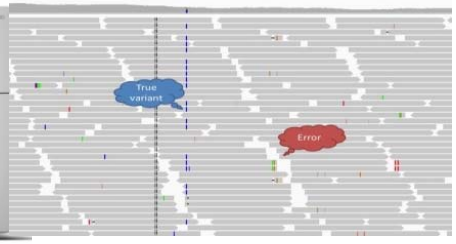


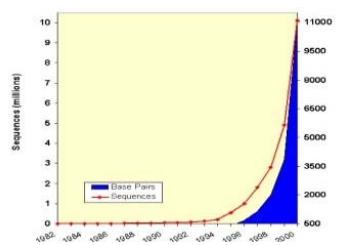
Explore Transcriptome using NGS

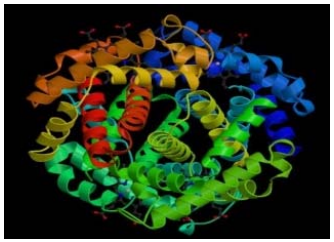
北京大学生物信息学中心 高歌

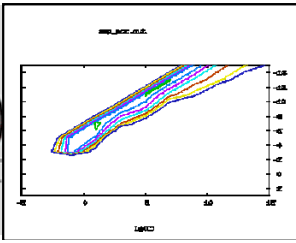
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

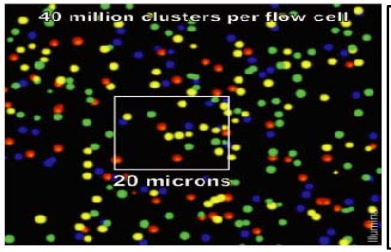


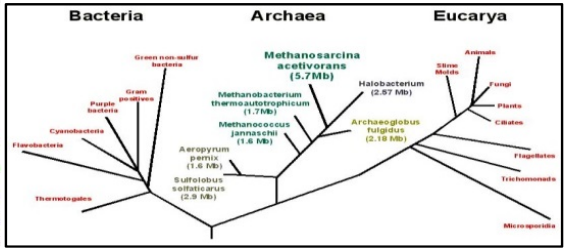






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



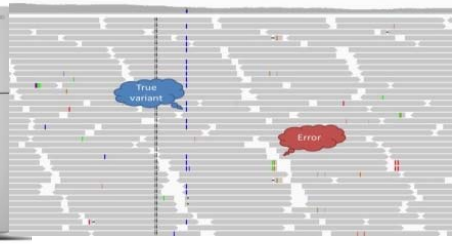


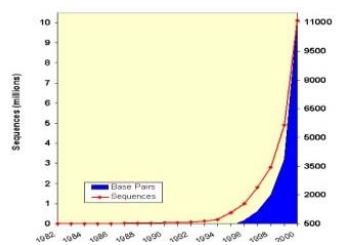
Unit 2:
RNA-Seq: Mapping & Assembling

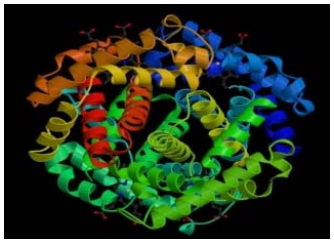
北京大学生物信息学中心 高歌

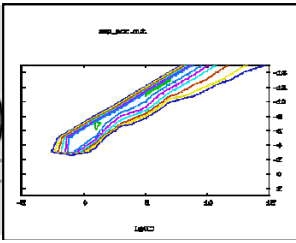
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University



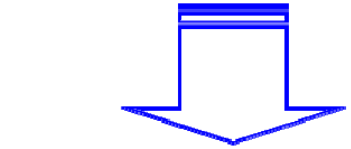






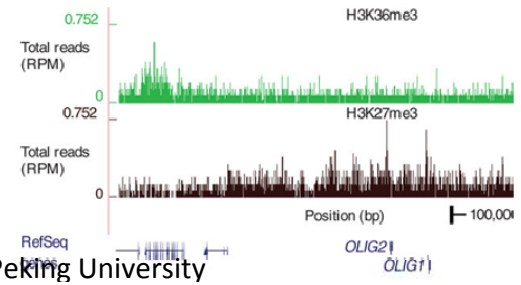
mapping

Reference Genome



Mapped alignment

Calling Genetic Variants



Measuring Abundance: RNA-Seq, ChIP-Seq, etc.

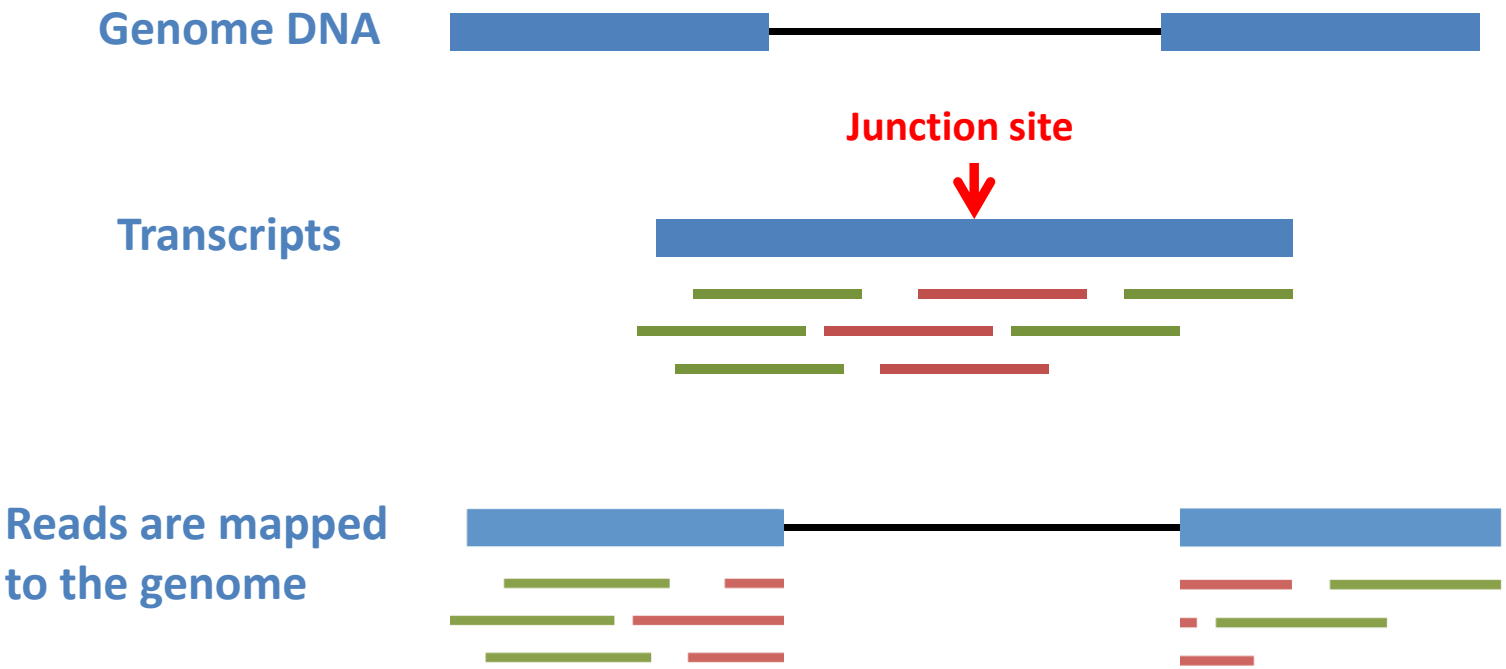
Mapping: Input Data


Reference Genome

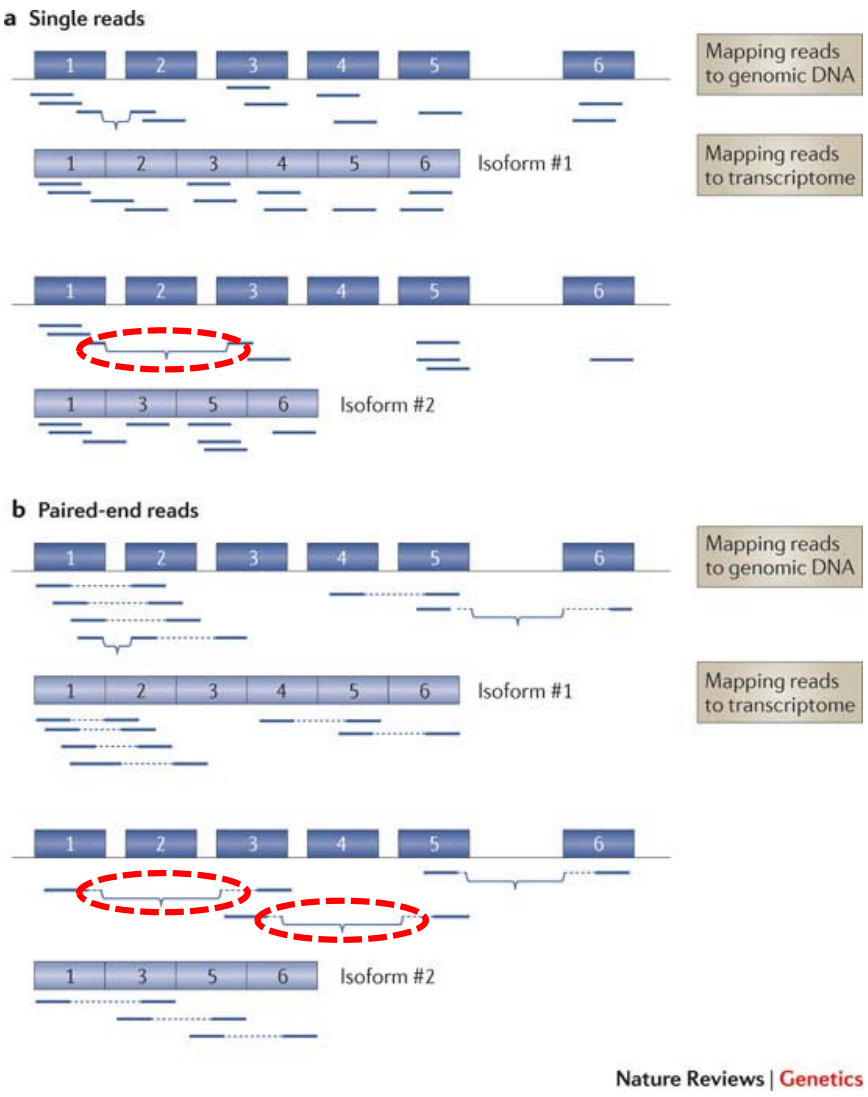
- Reference Genome
  - Nucleotide
  - **Length**: Hundreds of Mb *per* chromosome.
  - ~3 Gb in total (for human genome)

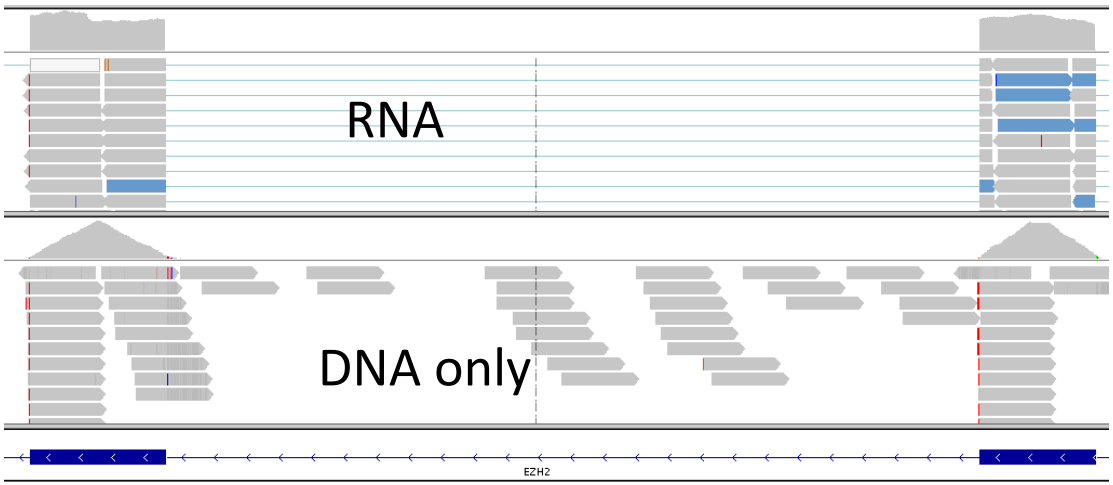
- Reads
  - Nucleotide, with **various qualities** (relatively **high error rate**: 1e-2 ~ 1e-5)
  - **Length**: 36~80 bp *per* read
  - Hundreds of Gbs *per* run

Mapping Reads from RNA-Seq



Detection novel splicing isoforms through junction reads

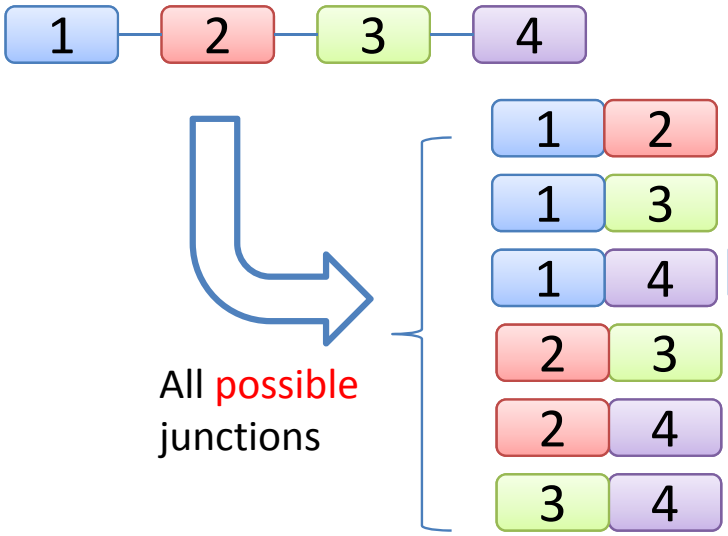




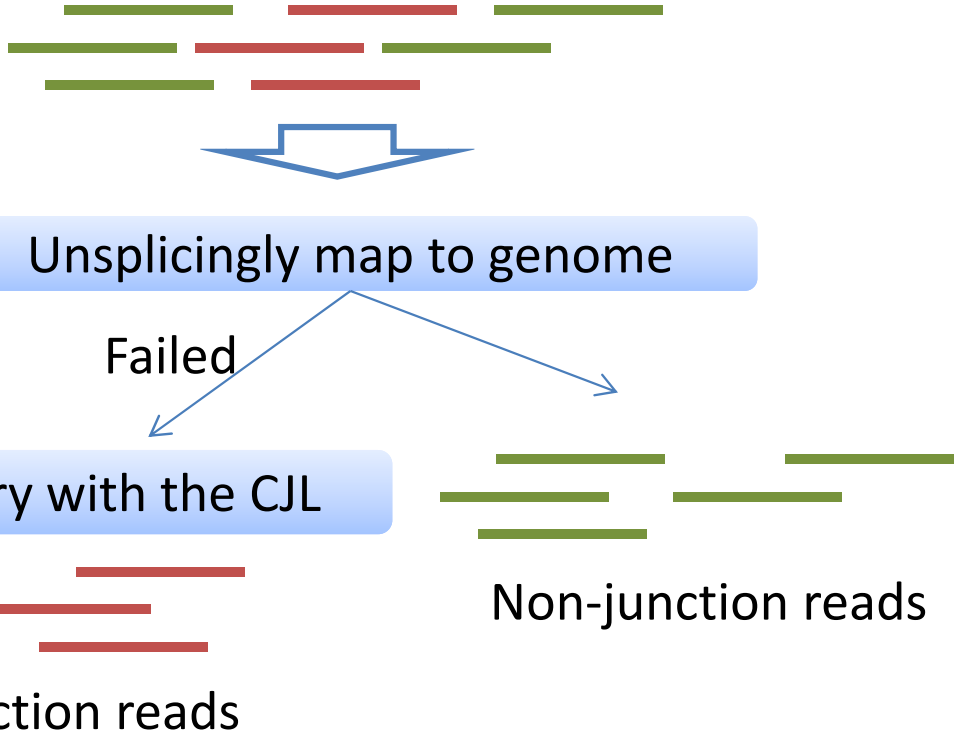
Mapping junction reads properly

Handle Junction Reads: “Join exon” strategy

1) **Build “conceptual junctions library” (CJL) for each known transcripts**

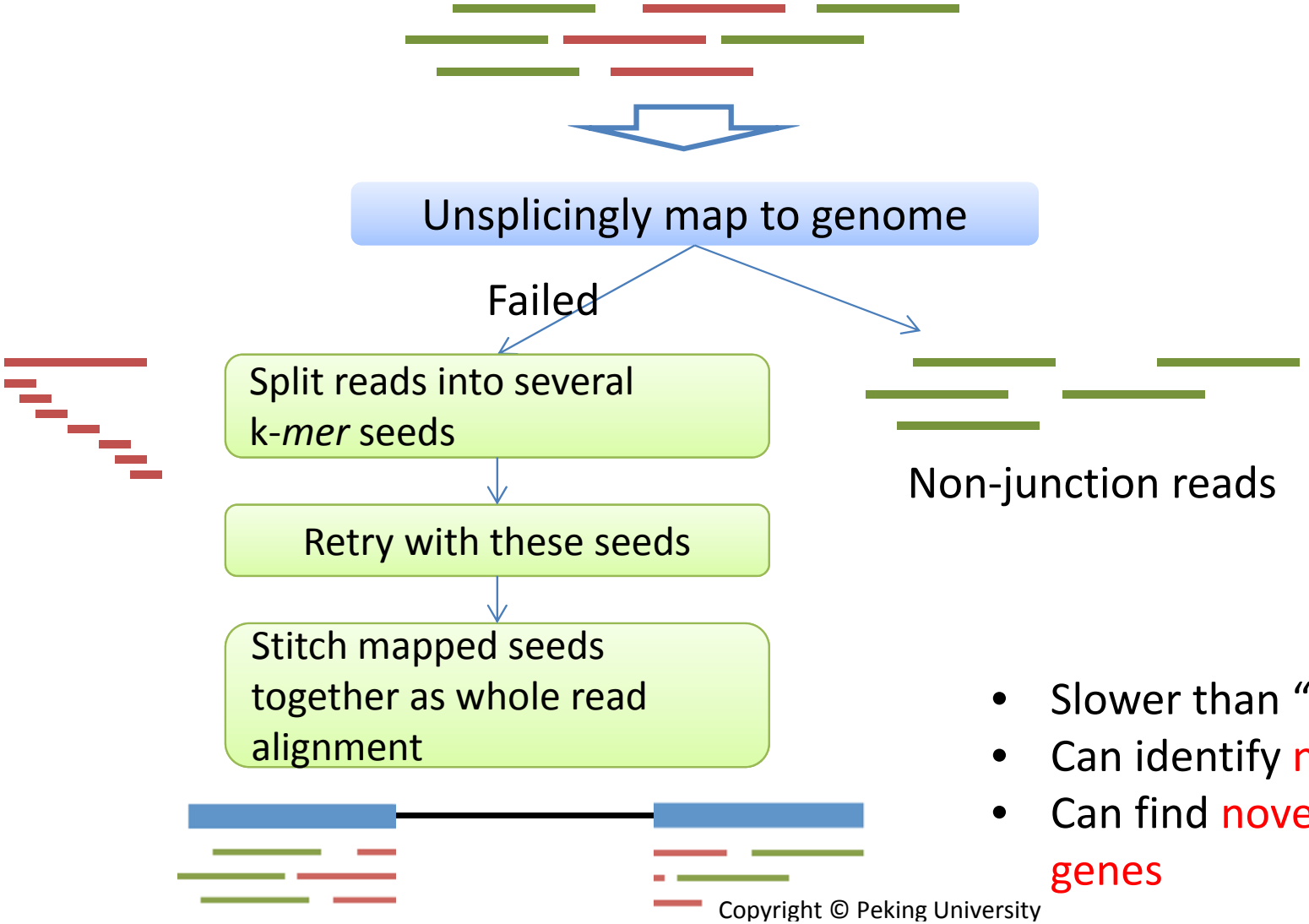


2) **Map RNA-Seq reads to the genome as well as the conceptual junction library**



- Fast
- Can identify novel splicing isoforms
- Can NOT find novel exons and novel genes

Handle Junction Reads: “Split reads” strategy

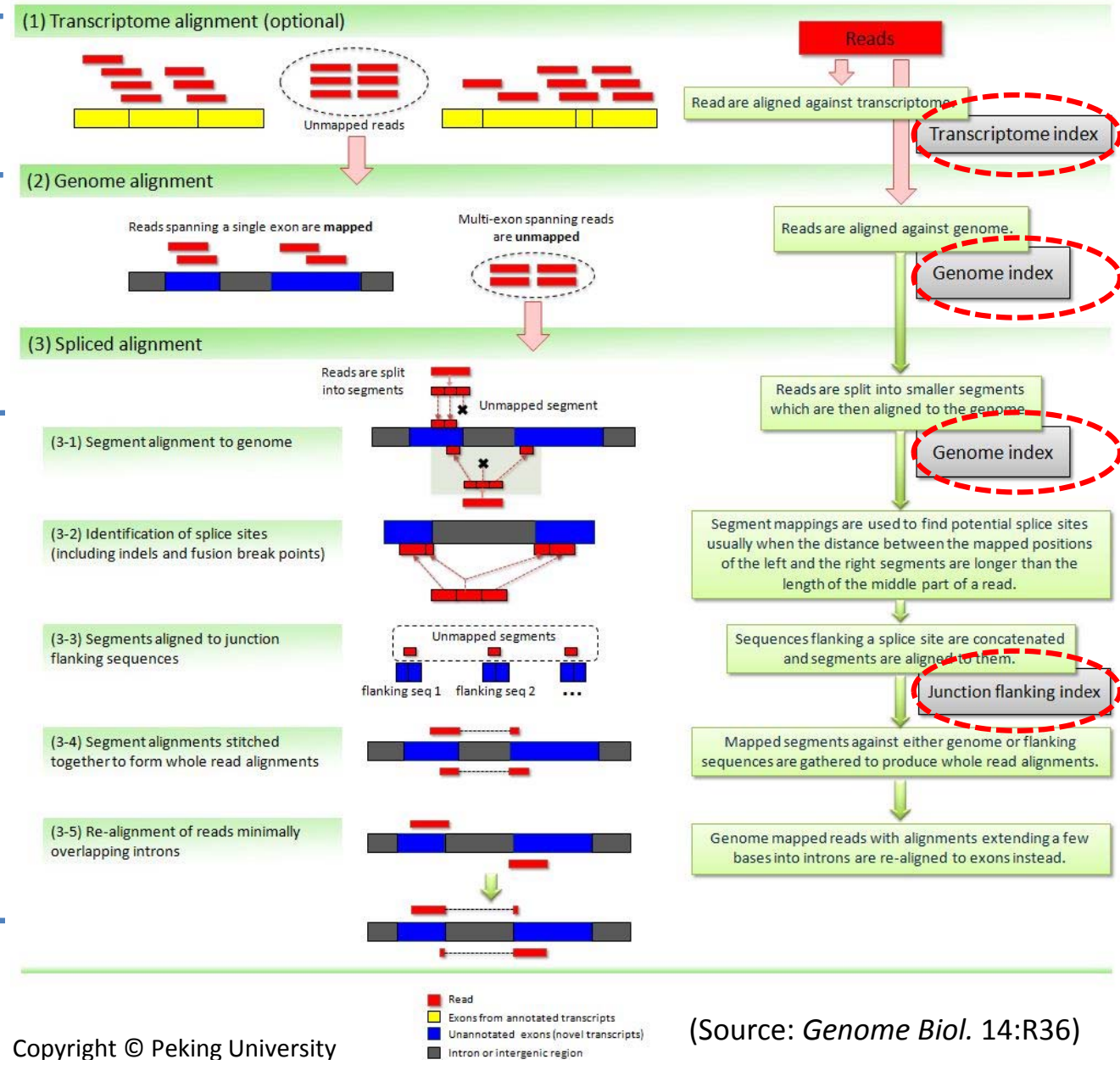


- Slower than “join exon”
- Can identify novel splicing isoforms
- Can find novel exons and novel genes

TopHat
A spliced read mapper for RNA-Seq

Join exon

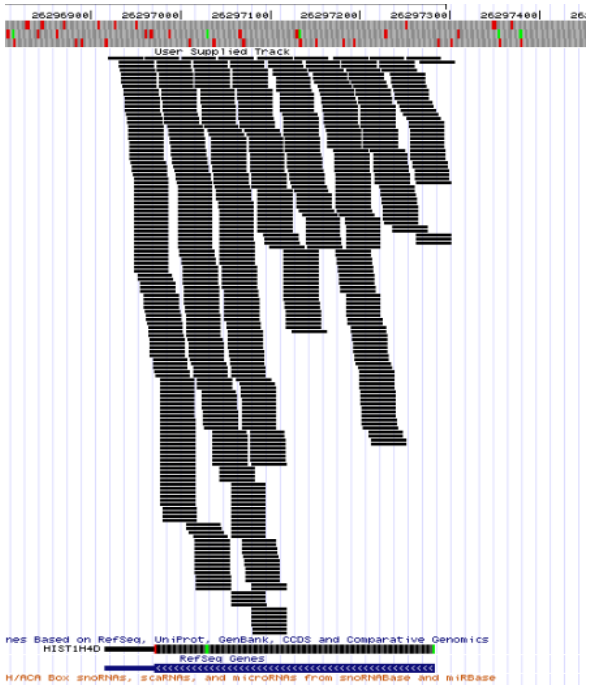
Split reads



(Source: *Genome Biol.* 14:R36)

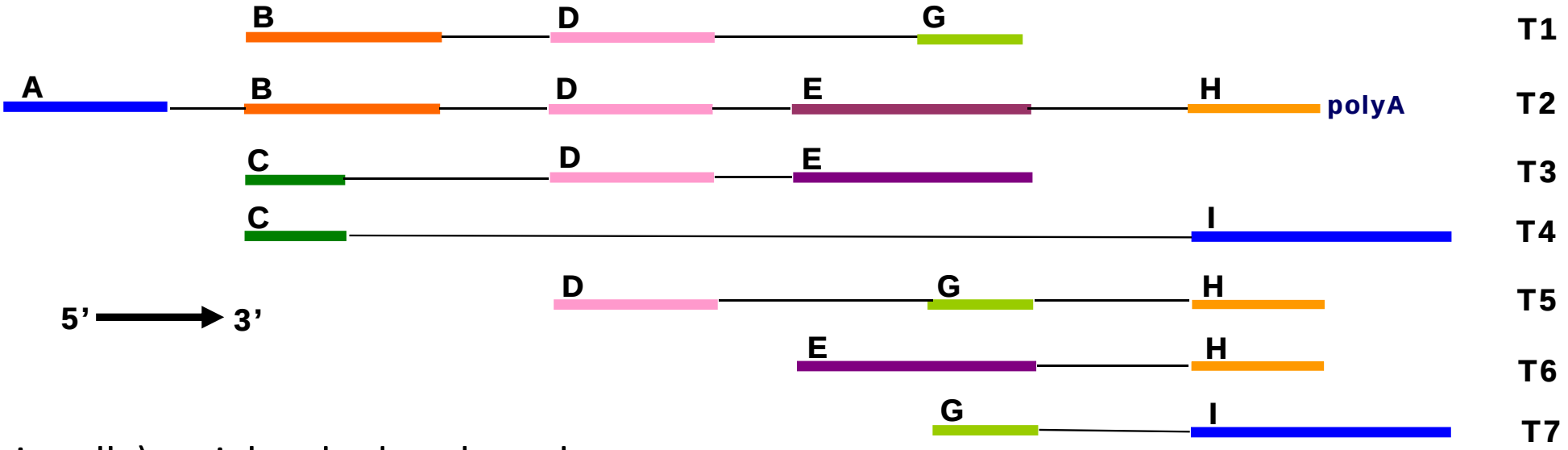
1) **<u>Assembly</u>**: reconstruct full-length transcript sequences from the (mapped) reads.

2) **<u>Quantification</u>**: estimate the expression abundance for each transcripts



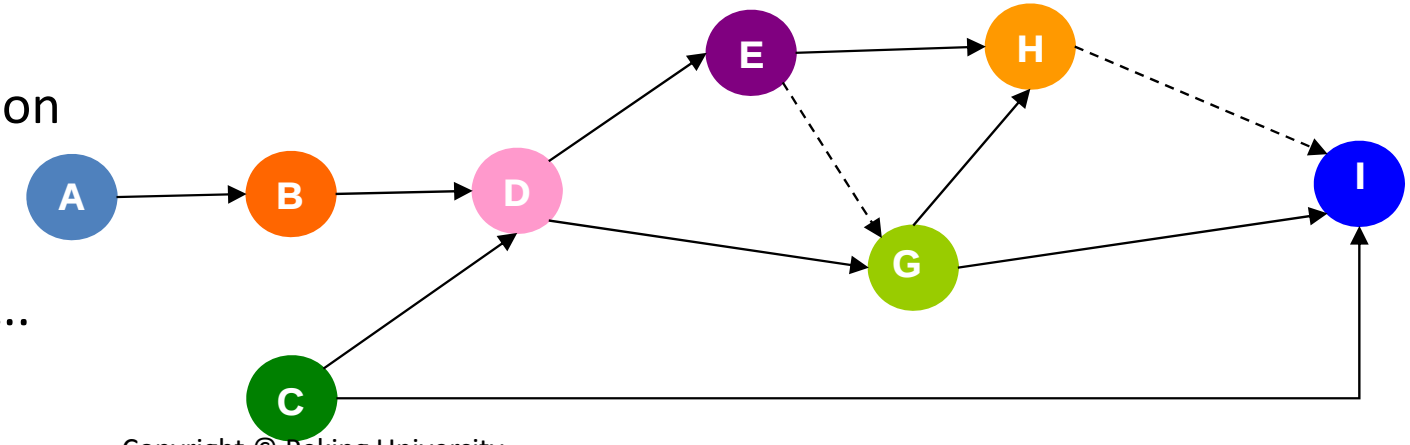
	B	C	D	E	F
1	gene	nsc1	nsc1 SE	nsc2	nsc2 SE
2	brain protein	18.9574	3.79952	21.5848	3.02241
3	Cluster Incl AW1	110.513	7.84625	114.894	7.95669
4	Cluster Incl AI8	235.873	35.6748	210.349	27.612
5	Cluster Incl AV3	47.4605	3.94976	29.6941	3.6586
6	Cluster Incl AV1	28.4527	3.74512	15.2986	3.62097
7	Cluster Incl AV1	80.302	6.45368	107.23	8.09591
8	Cluster Incl AV3	40.8113	5.13418	54.0835	3.18591
9	Cluster Incl AI1	53.1437	3.63392	58.635	5.50994

Assembling as a graph traveler



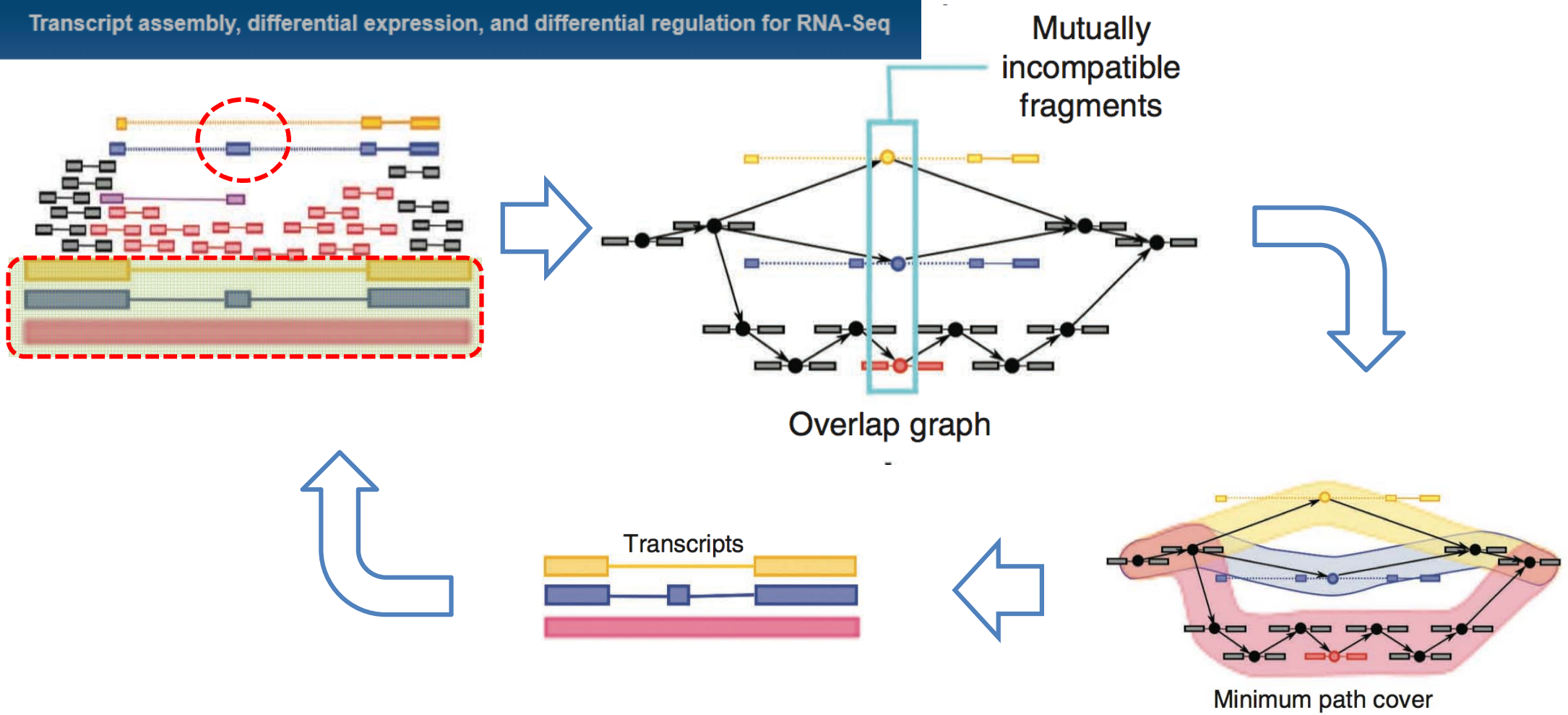
(Rationally) weighted edges based on various supporting evidences

- Existing transcript data: junction reads
- Sequence context: splicing junction sites, polyA signals, ….
- Existing annotation



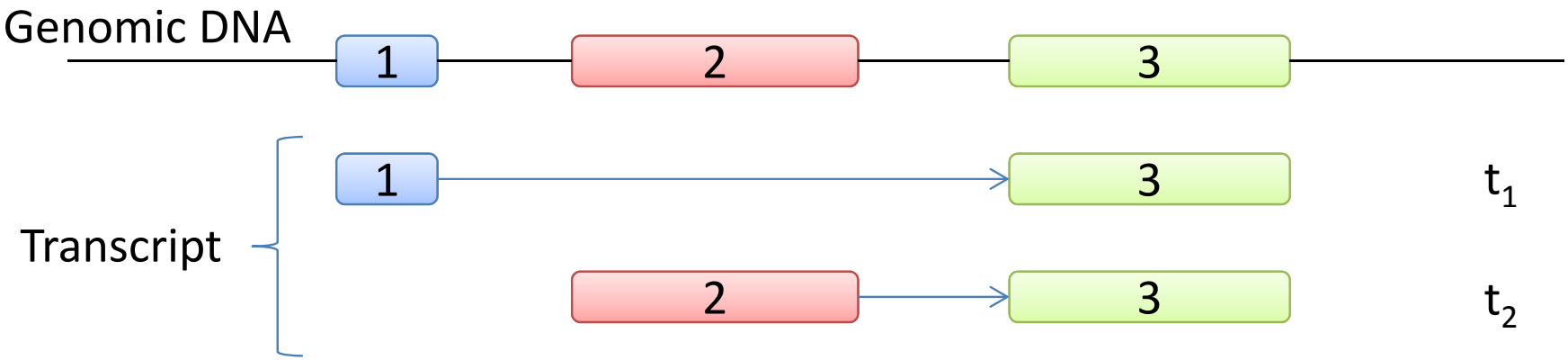
Cufflinks

Transcript assembly, differential expression, and differential regulation for RNA-Seq



Trapnell C. *et al*. *Nature Biotechnology* 2010

Quantifying as a maximum likelihood inference

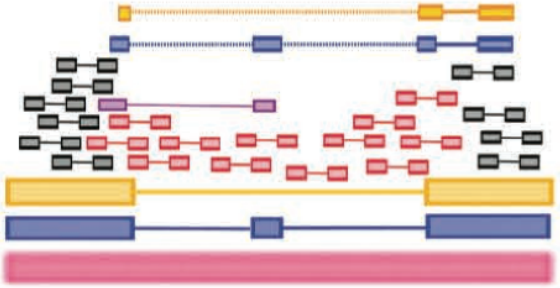


e_1=20 e_2=40 e_3=60

$$\begin{cases} e_1 = t_1 \\ e_2 = t_2 \\ e_3 = t_1 + t_2 \end{cases}$$

$$\begin{cases} t_1 = e_1 = 20 \\ t_2 = e_2 = 40 \end{cases}$$

Abundance estimation **Cufflinks**



Transcript coverage

Fragment length

$$L(\rho|R) = \prod_{r\in R} Pr(rd.\ aln. = r)$$

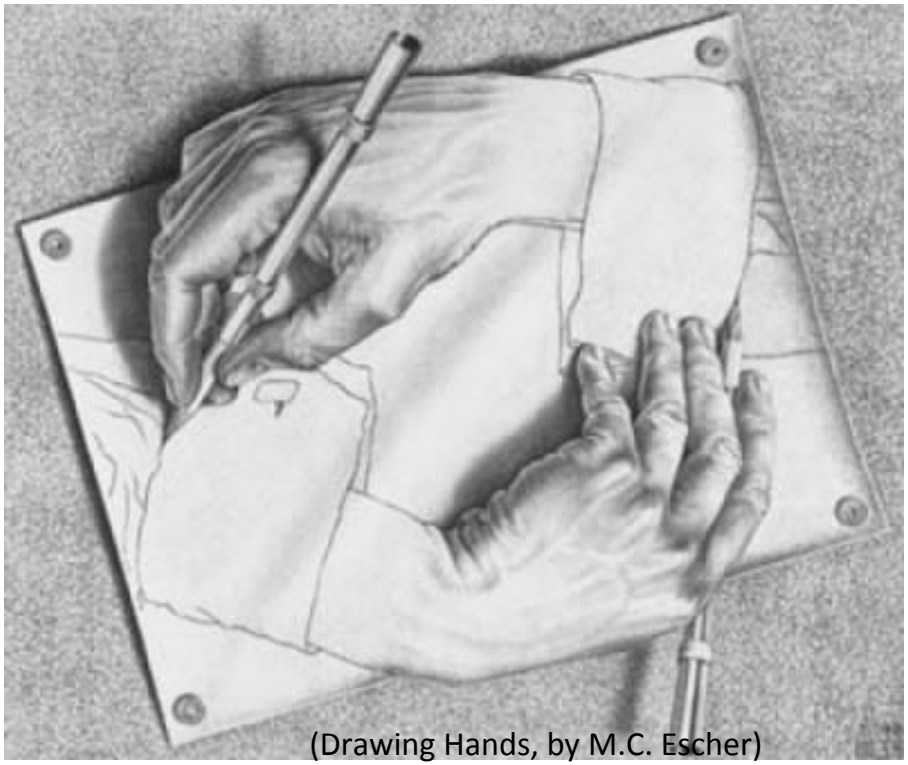
$$= \prod_{r\in R}\sum_{t\in T} Pr(rd.\ aln. = r|trans. = t)Pr(trans. = t)$$

$$= \prod_{r\in R}\sum_{t\in T} \frac{\rho_t \tilde{l}(t)}{\sum_{u\in T}\rho_u \tilde{l}(u)} Pr(rd.\ aln. = r|trans. = t)$$

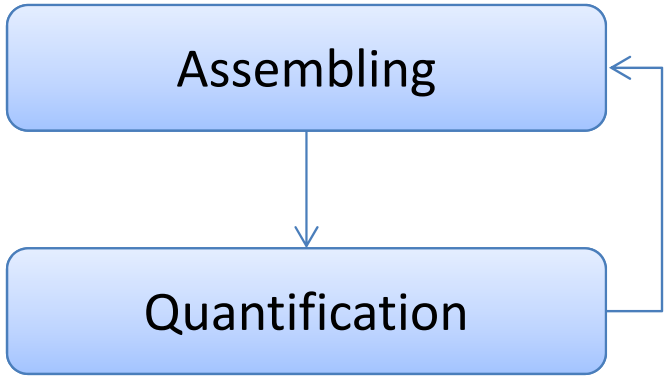
$$= \prod_{r\in R}\sum_{t\in T} \frac{\rho_t \tilde{l}(t)}{\sum_{u\in T}\rho_u \tilde{l}(u)} \left(\frac{F(I_t(r))}{l(t) - I_t(r) + 1}\right)$$

$$= \prod_{r\in R}\sum_{t\in T} \alpha_t \left(\frac{F(I_t(r))}{l(t) - I_t(r) + 1}\right),$$

(Nat Biotech 28:511)



(Drawing Hands, by M.C. Escher)



Summary Questions

- Could the split reads strategy also help for mapping DNA resequencing reads? Explain

- Could you show a case that the simple counting quantification in page 16 does NOT work? Explain

生物信息学：导论与方法
Bioinformatics: Introduction and Methods