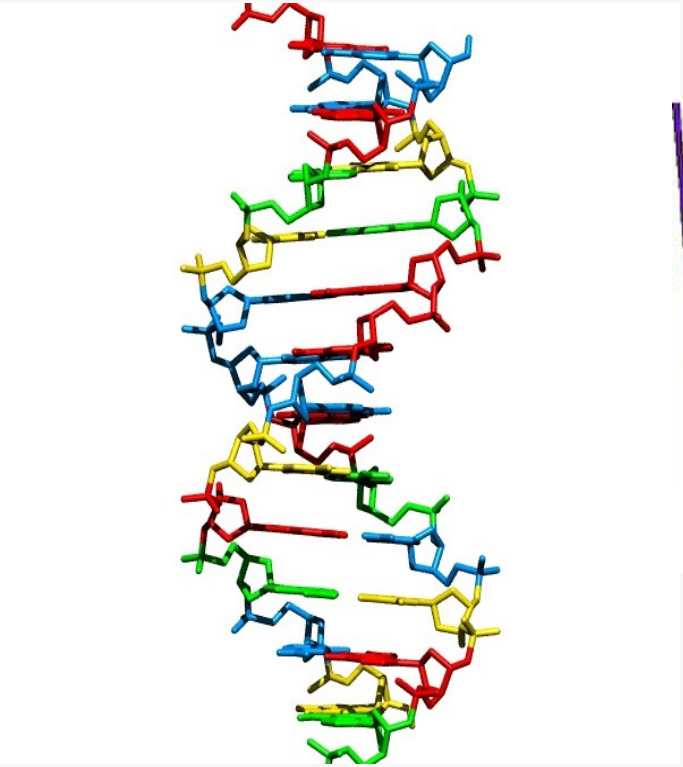
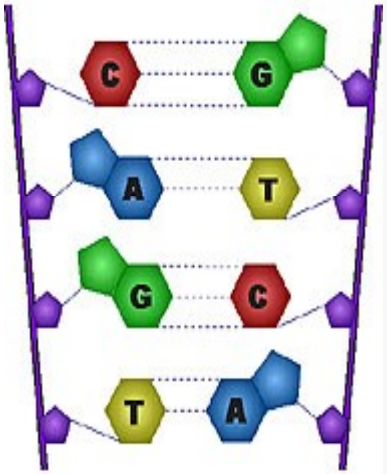
Information Theory for High Throughput Sequencing

Research supported by NSF Center for Science of Information.

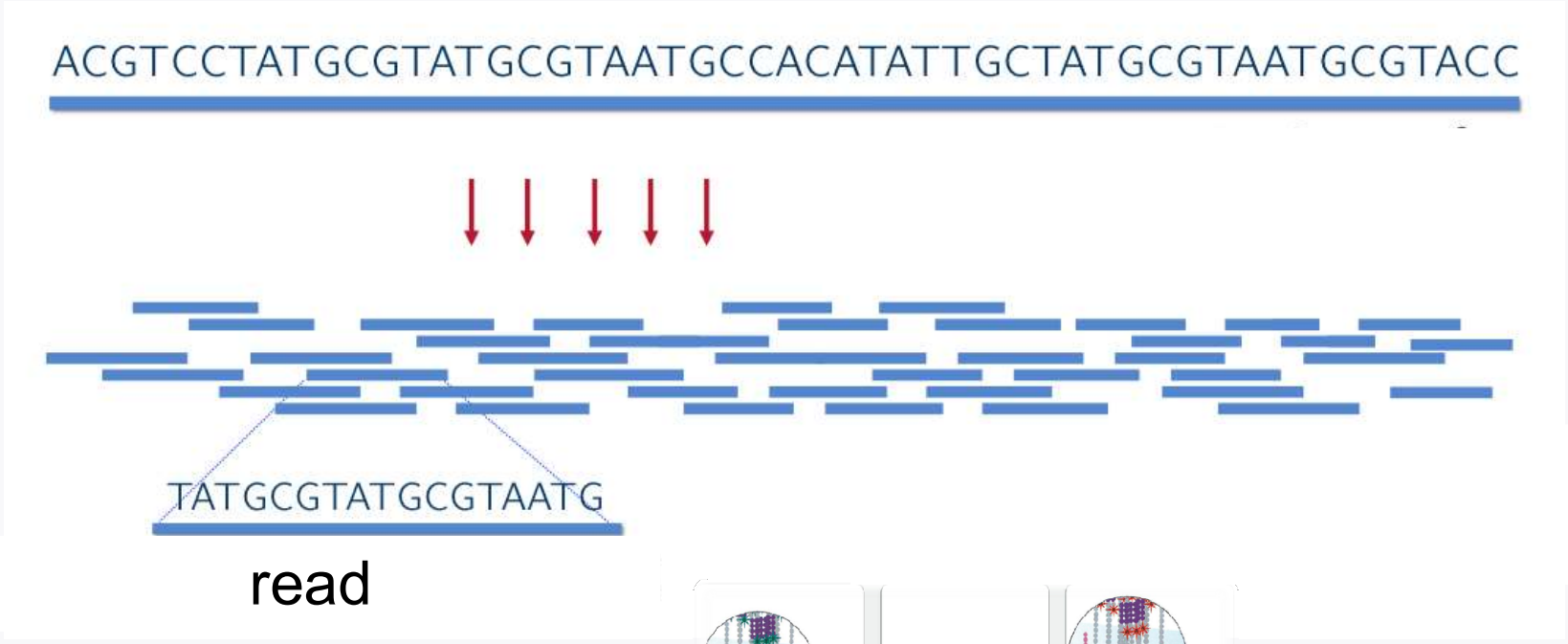
DNA sequencing

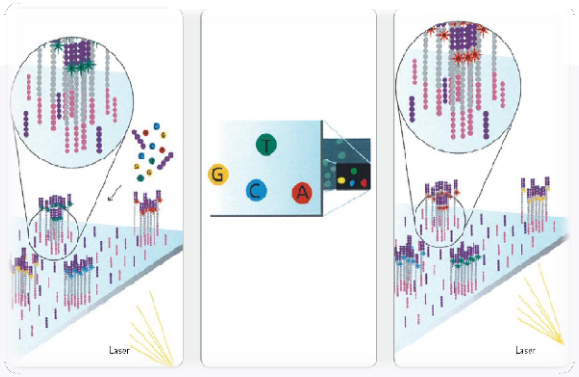




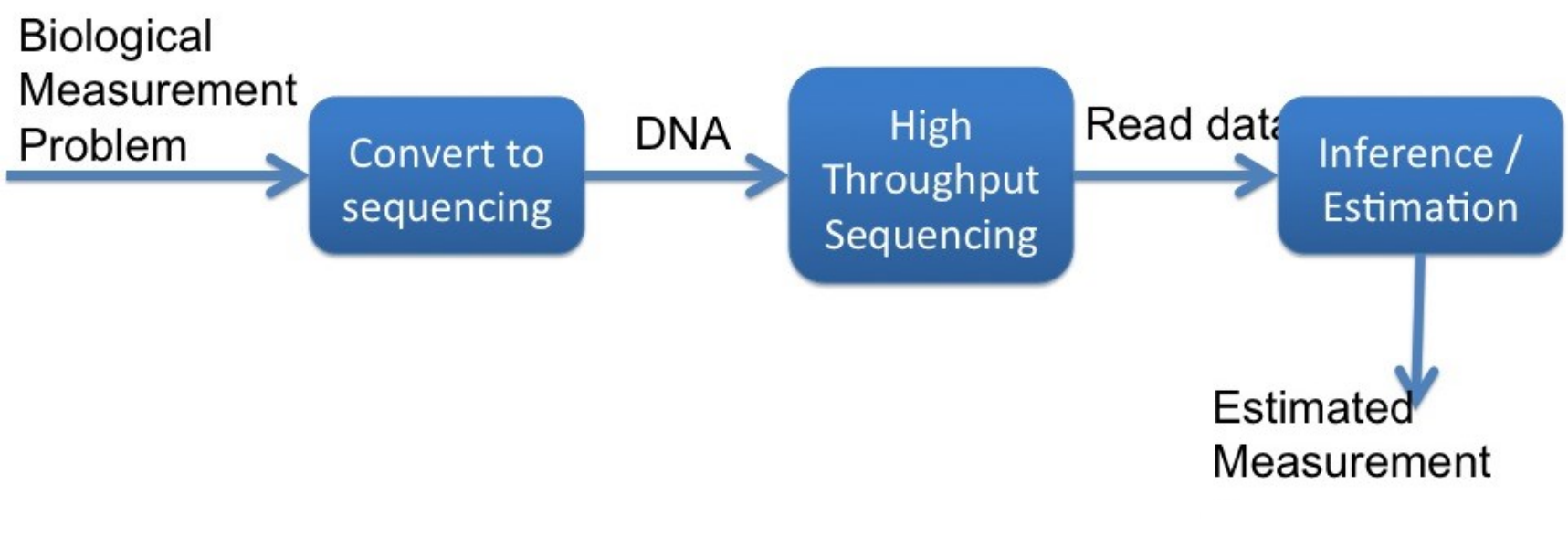
...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT...

Shotgun sequencing





High throughput sequencing: Microscope in the big data era



(Pachter diagram)

Genomic variations, 3-D structures, transcription, translation, protein interaction, etc.

Some computational problems

- *De novo* assembly
- Read mapping , SNP calling, quantification.
- Downstream association studies

Assembly as a software engineering problem

- A single sequencing experiment can generate 100's of millions of reads, 10's to 100's gigabytes of data.
- Primary concerns are to minimize time and memory requirements.
- No guarantee on optimality of assembly quality and in fact no optimality criterion at all.

Computational complexity view

- Formulate the assembly problem as a combinatorial optimization problem:
  - Shortest common superstring (Kececioglu-Myers 95)
  - Maximum likelihood (Medvedev-Brudno 09)
  - Hamiltonian path on overlap graph (Nagarajan-Pop 09)
- Typically NP-hard and even hard to approximate.
- Does not address the question of when the solution reconstructs the ground truth.

Information theoretic view

Basic question:

What is the quality and quantity of read data needed to reliably reconstruct?

Information theoretic approach to assembly design

I. DNA assembly

ShannonDNA:

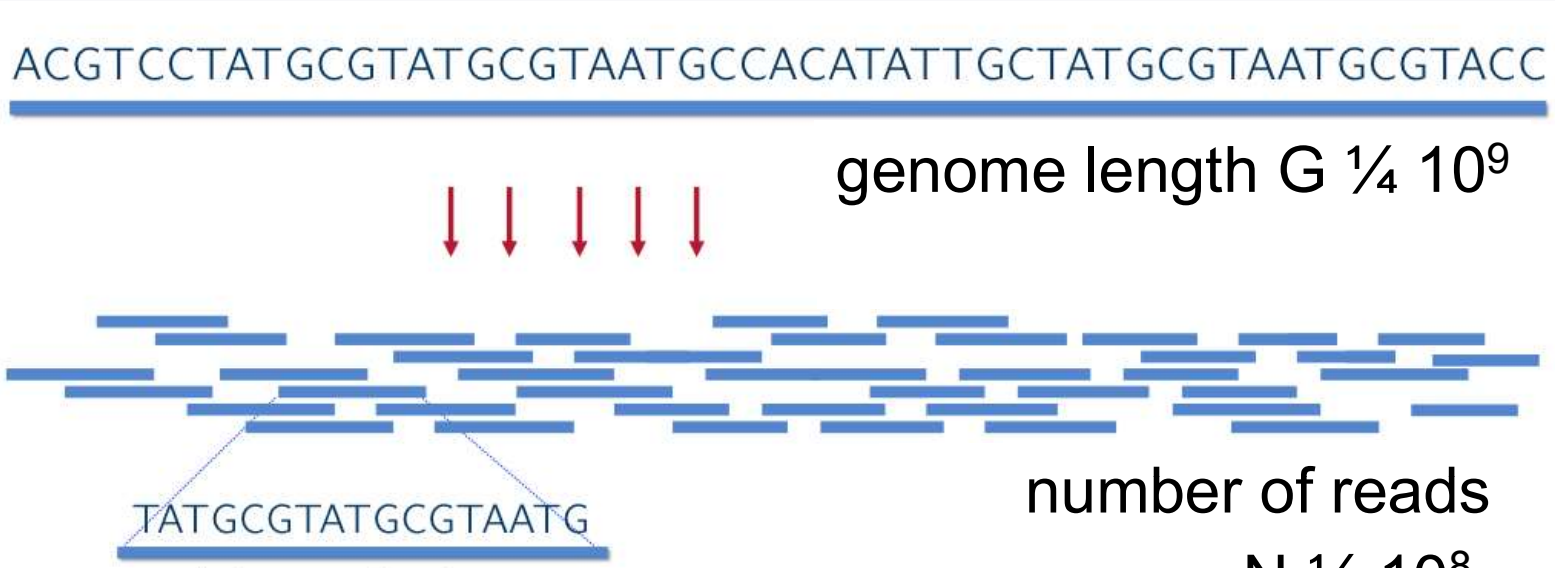
a *de novo* DNA assembler from long, noisy reads

II. RNA assembly

ShannonRNA:

a *de novo* RNA-Seq assembler from short reads

DNA Assembly



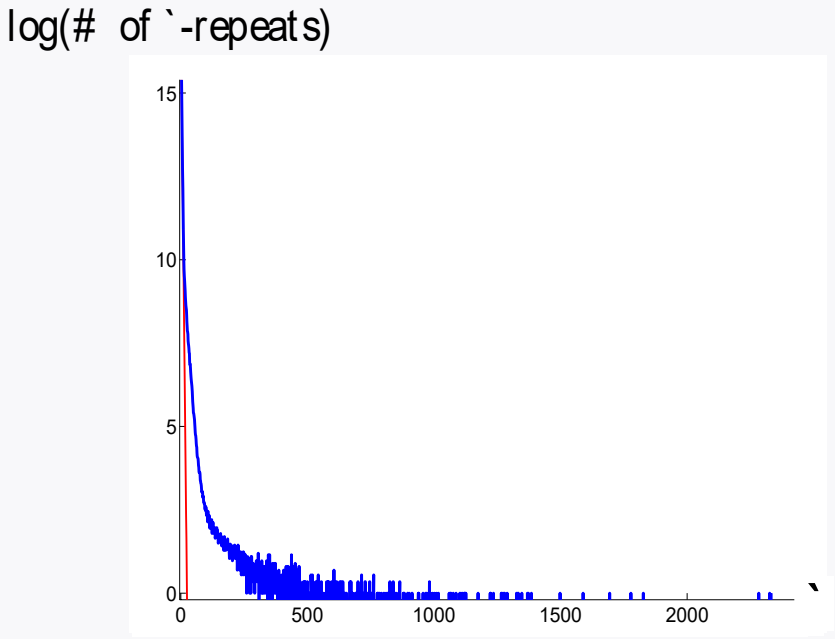
randomly sampled noisy reads

length L ¼ 100 - 1000

Reads are assembled to reconstruct the original DNA sequence.

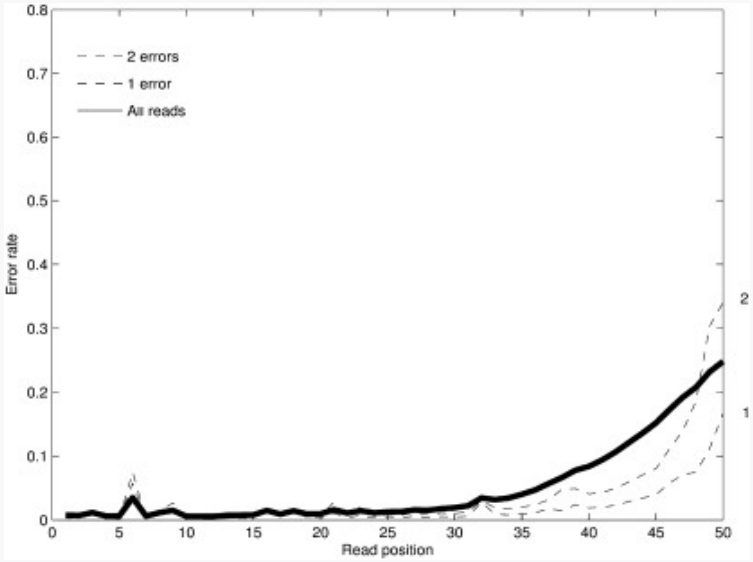
Challenges

Long repeats



Human Chr 22
repeat length histogram

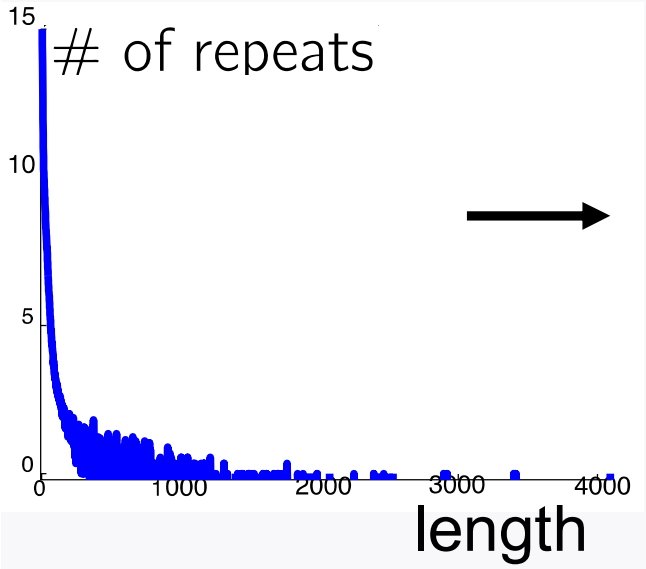
Noisy reads



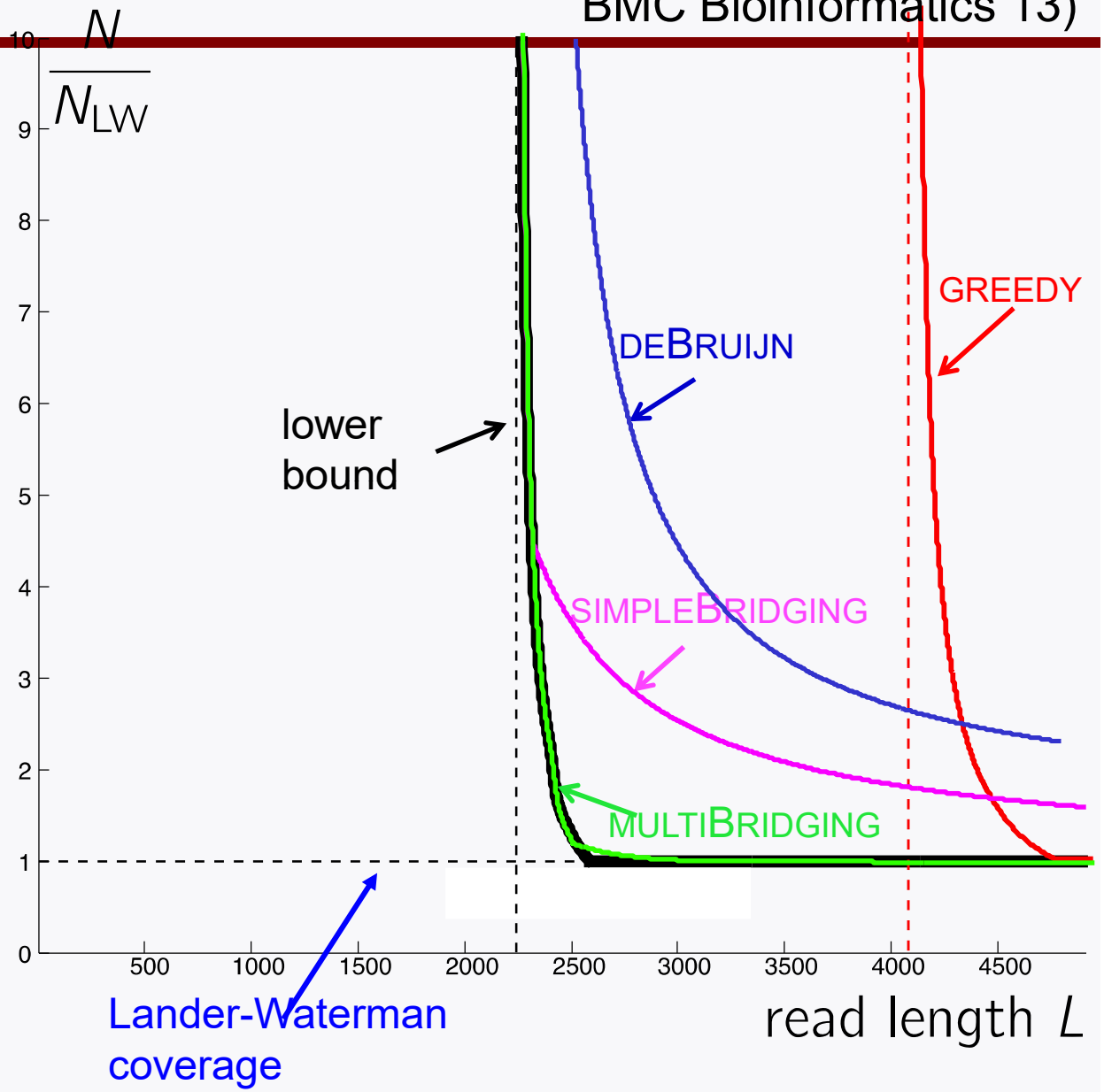
Illumina read error profile

Noiseless Reads

(Bresler, Bresler, T.
BMC Bioinformatics 13)

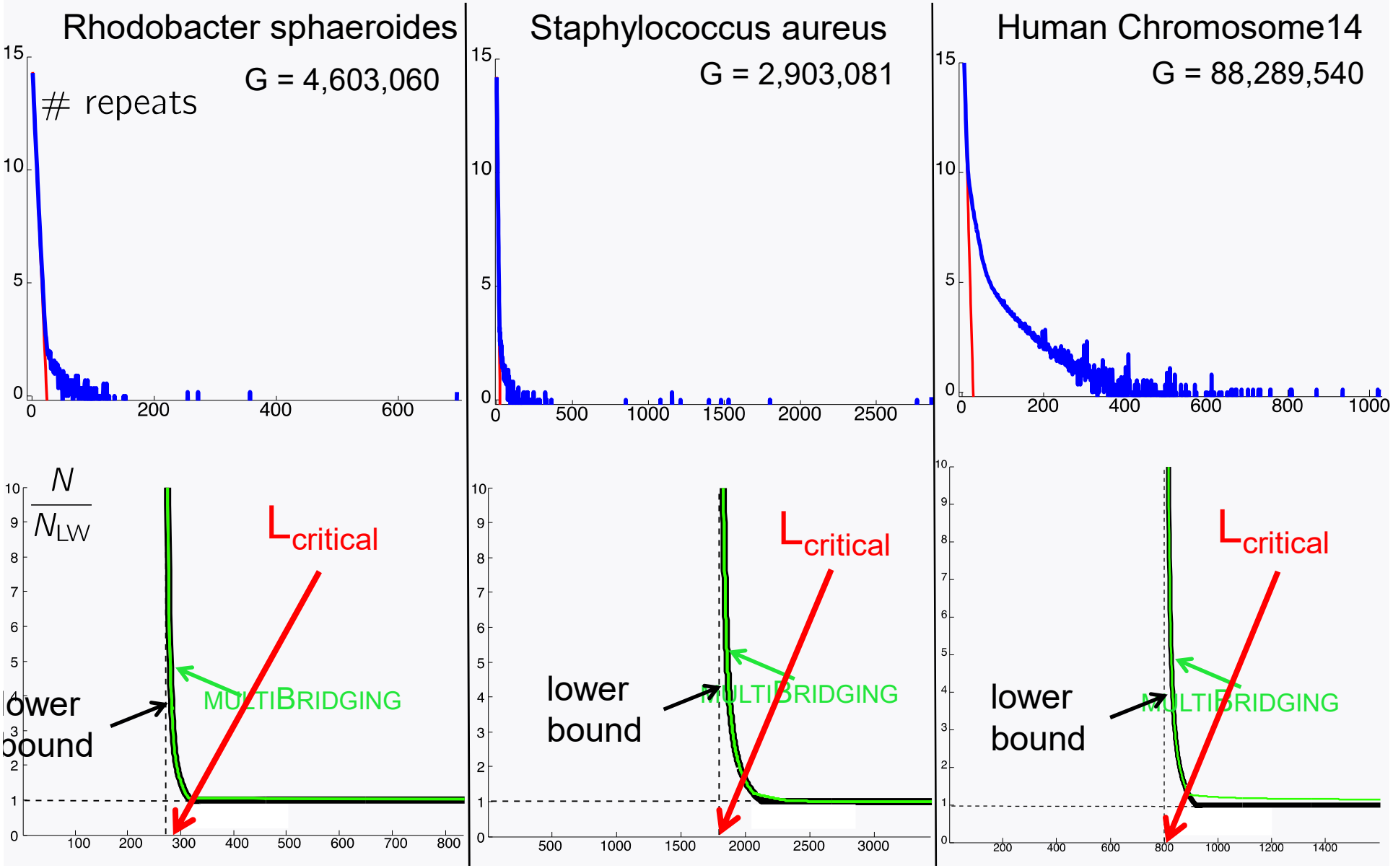


Human Chr 19
Build 37

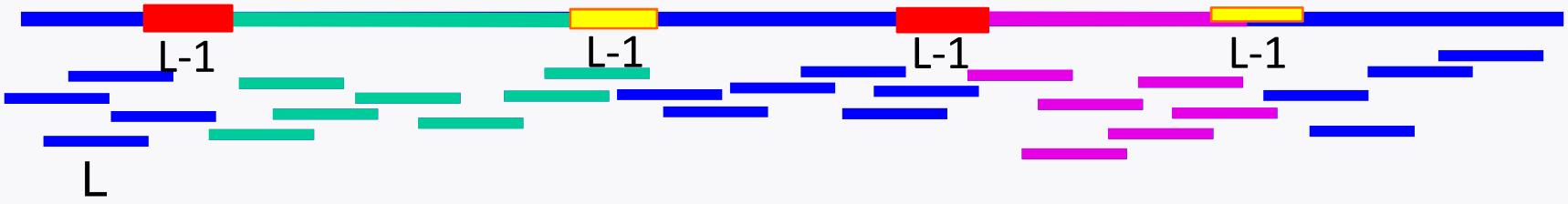


GAGE Benchmark Datasets

http://gage.cbcb.umd.edu/



Unresolvable repeat patterns: interleaved repeats



- If no such repeat pattern in the genome, then unique reconstruction with sufficient coverage depth.

- $L_{critical}$ is the longest of such interleaved repeats.

Read Noise

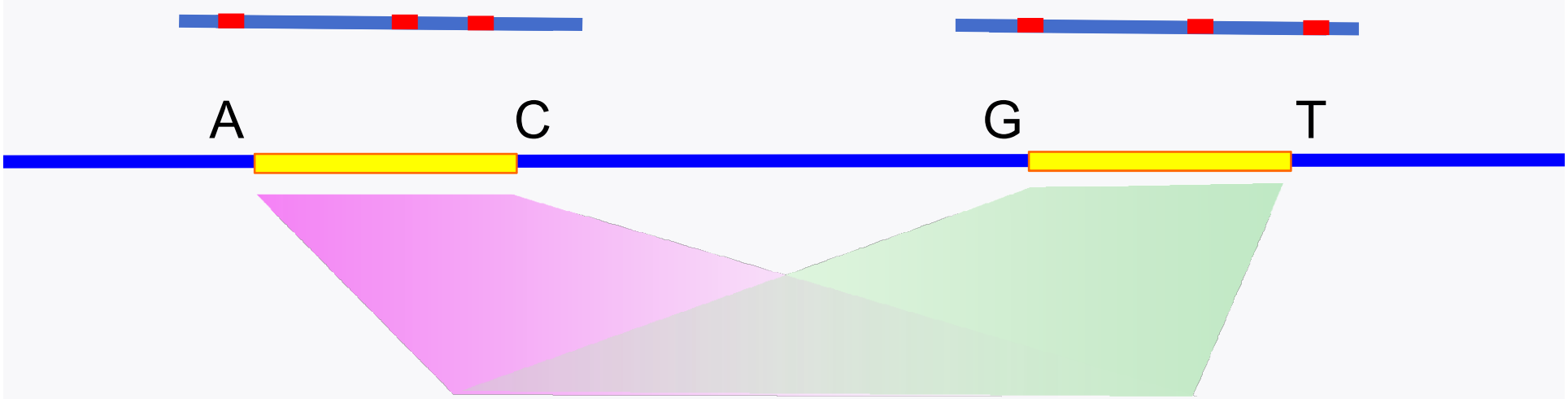
Each symbol corrupted by a noisy channel.

Assembly with noisy reads

(Lam,Khalak, T. Recomb-Seq 14)

Sensitive to noise ?

Information from random flanking region

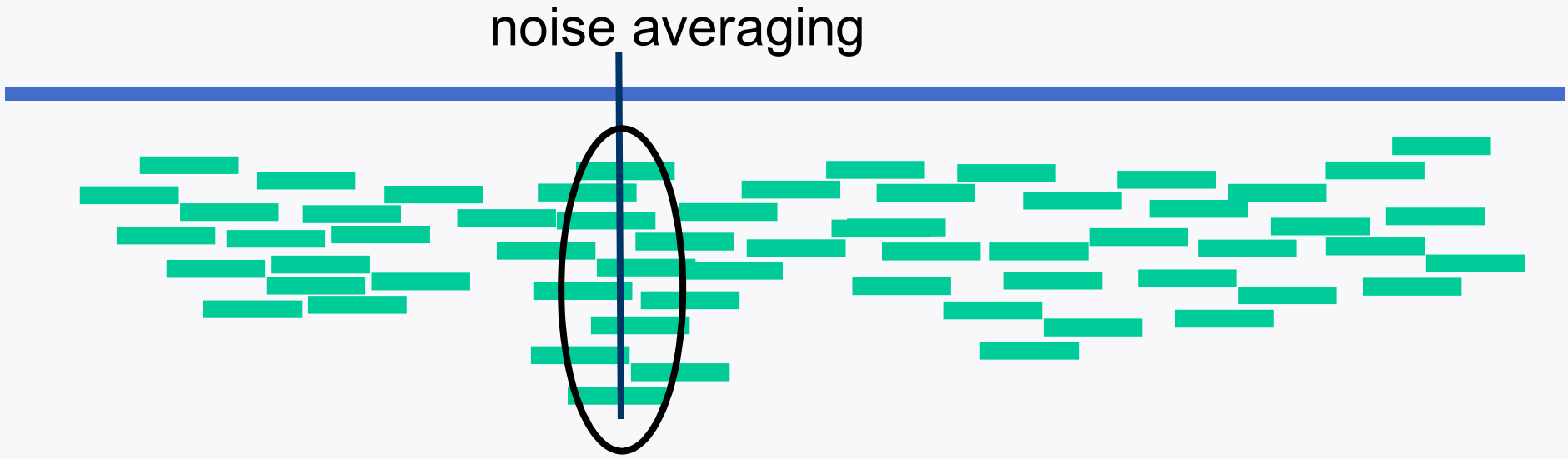


Copy 1 TAGCAGCAAATAGTT…CTGTTTGTT…TTGCC… GCCAGGATGT

Copy 2 TACGACGGAATAGTT…CTGTTTGTT…TTGCC… GTGACCACAG

Distance 0011011100000000…000000000…00000…0110110111

Information from coverage



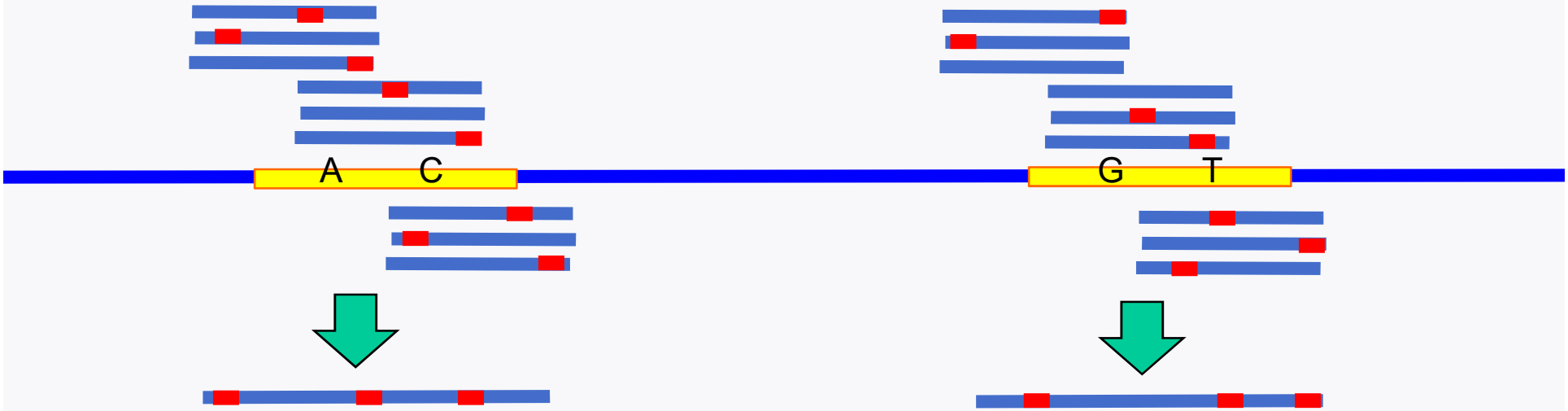
covering every position => most positions covered by many reads.

Key is to be able to align reads that belong together.

(Earlier work: Abolfazl et al, ISIT 13)

Multiple sequence alignment

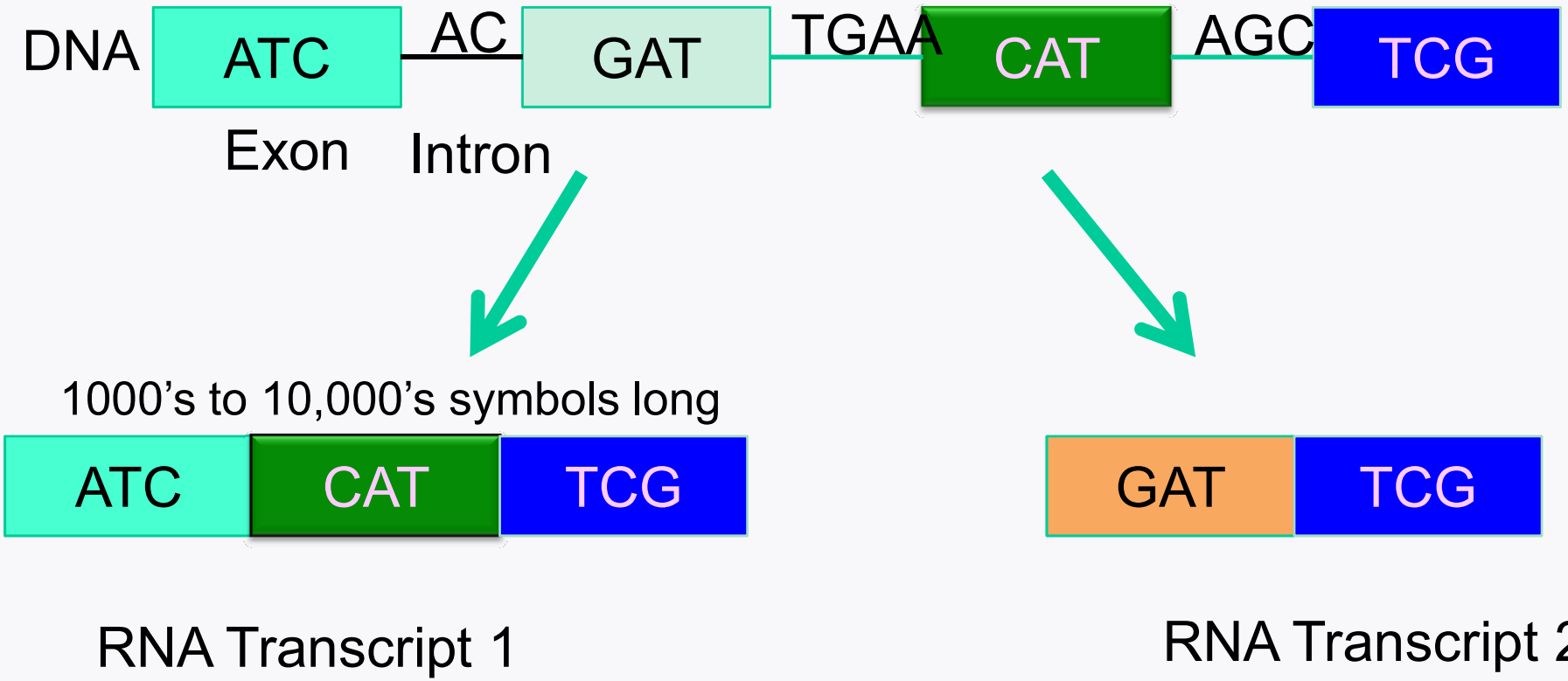
- Use flanking region as anchor to align reads close to boundary of approximate repeats
- Average across reads to correct errors
- Bootstrap to extend further into the interior of repeat.



Central dogma of molecular biology

RNA transcripts and their abundances capture the state of a cell at a given time.

Alternative splicing



Alternative splicing yields different isoforms.

Transcriptome

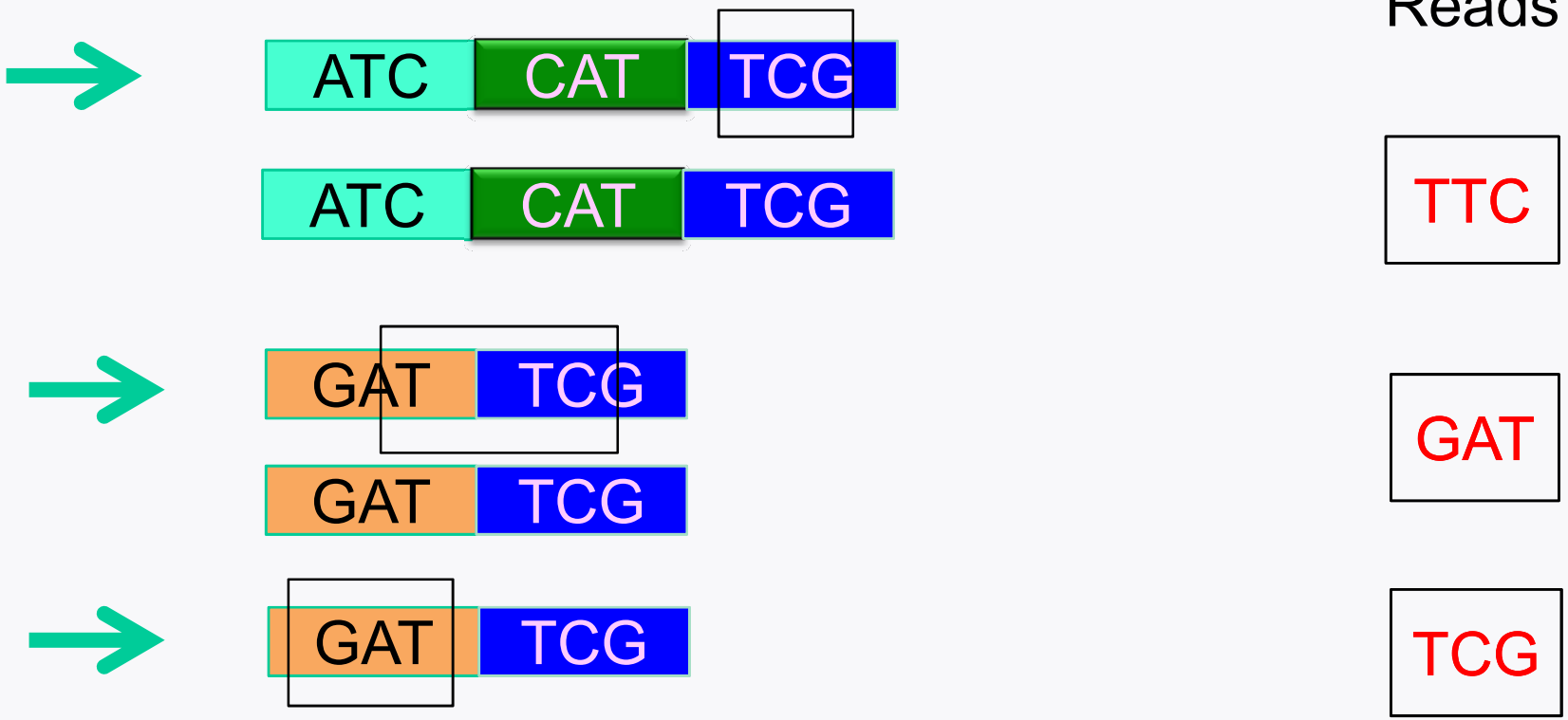
20 copies in cell

30 copies in cell

- Different transcripts are present at different abundances.
- Transcriptome is the mixture of transcripts from all the genes.
- Human transcriptome has 10,000's of transcripts from 20,000 genes.

RNA-Seq

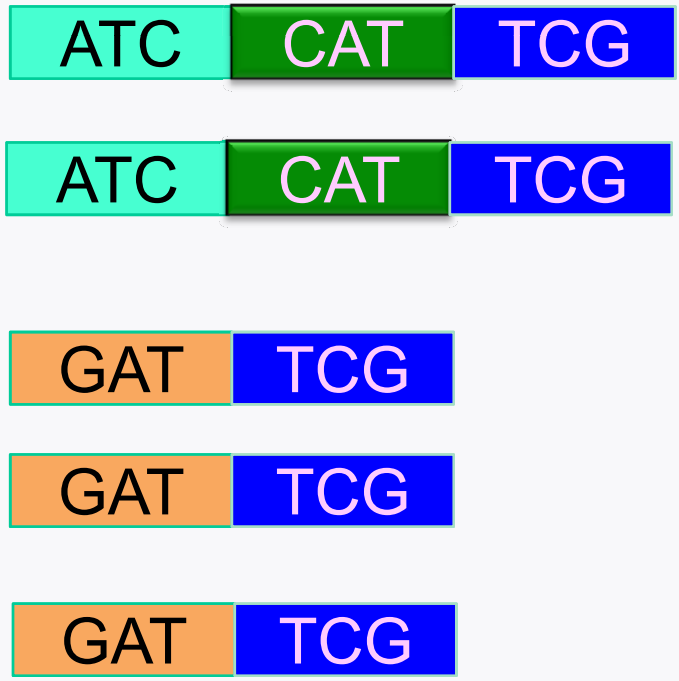
(Mortazavi et al, Nature Methods 08)



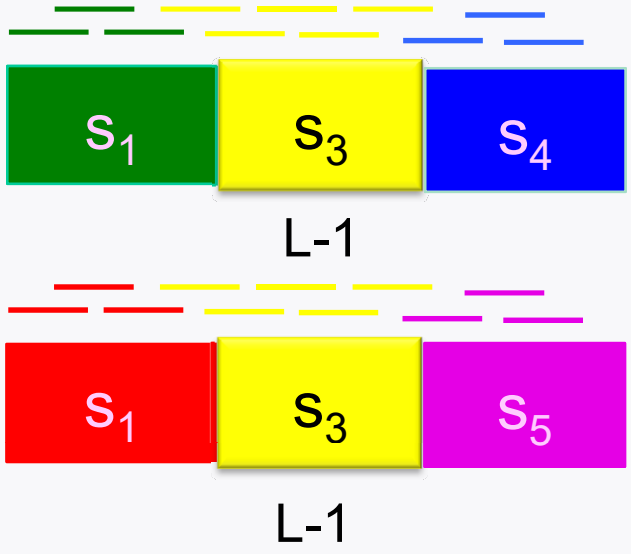
What is $L_{critical}$ for a transcriptome?

- $L_{critical}$ is lower bounded by the length of the longest interleaved repeat in any trancript

- It can potentially be much larger due to inter-transcript repeats of exons across isoforms.



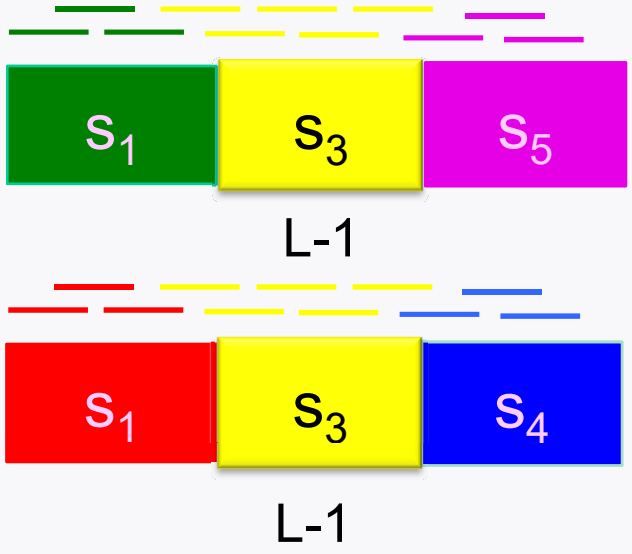
Ambiguity due to inter-transcript repeats



transcript 1

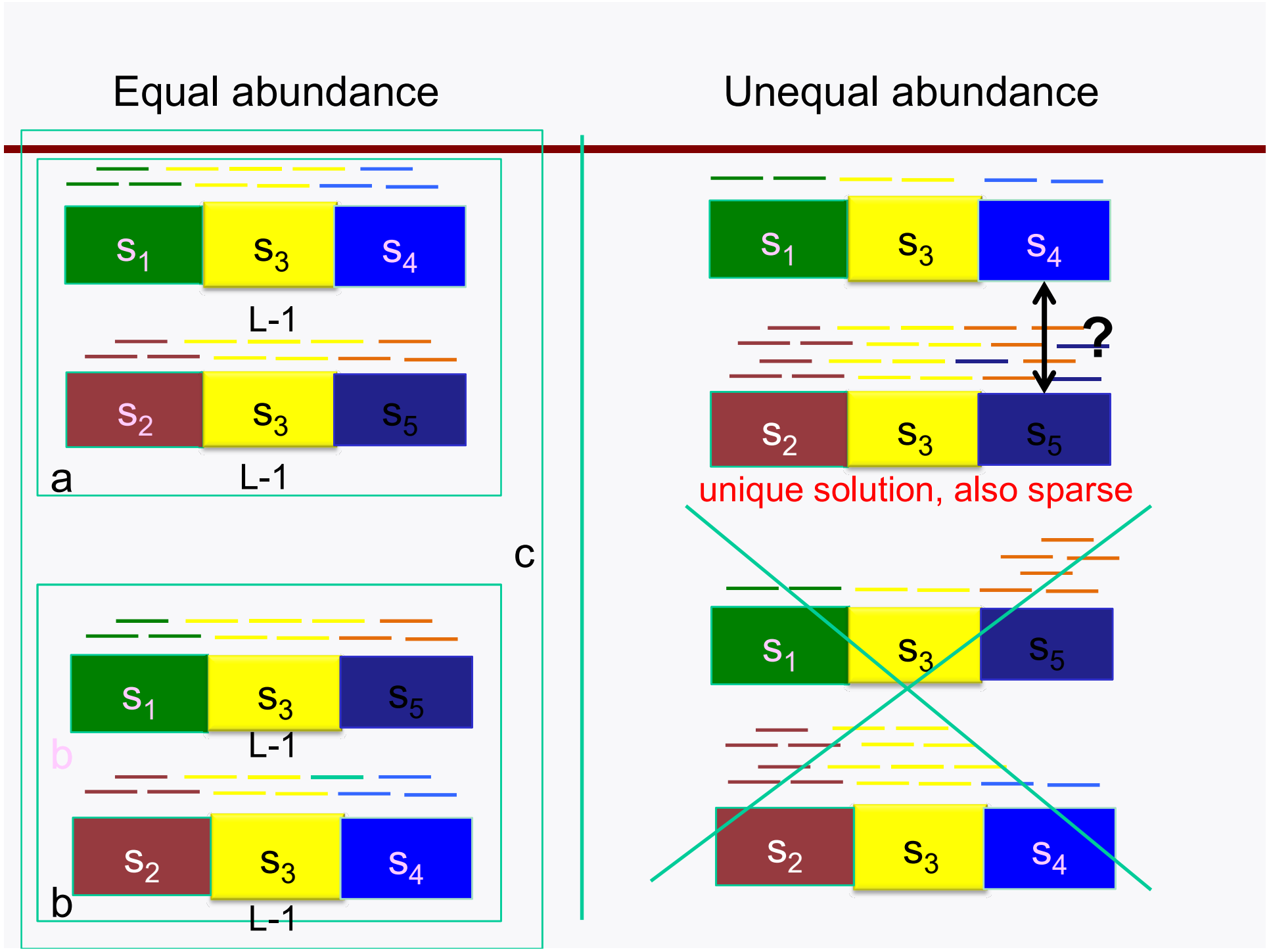
transcript 2

Ambiguity due to inter-transcript repeats

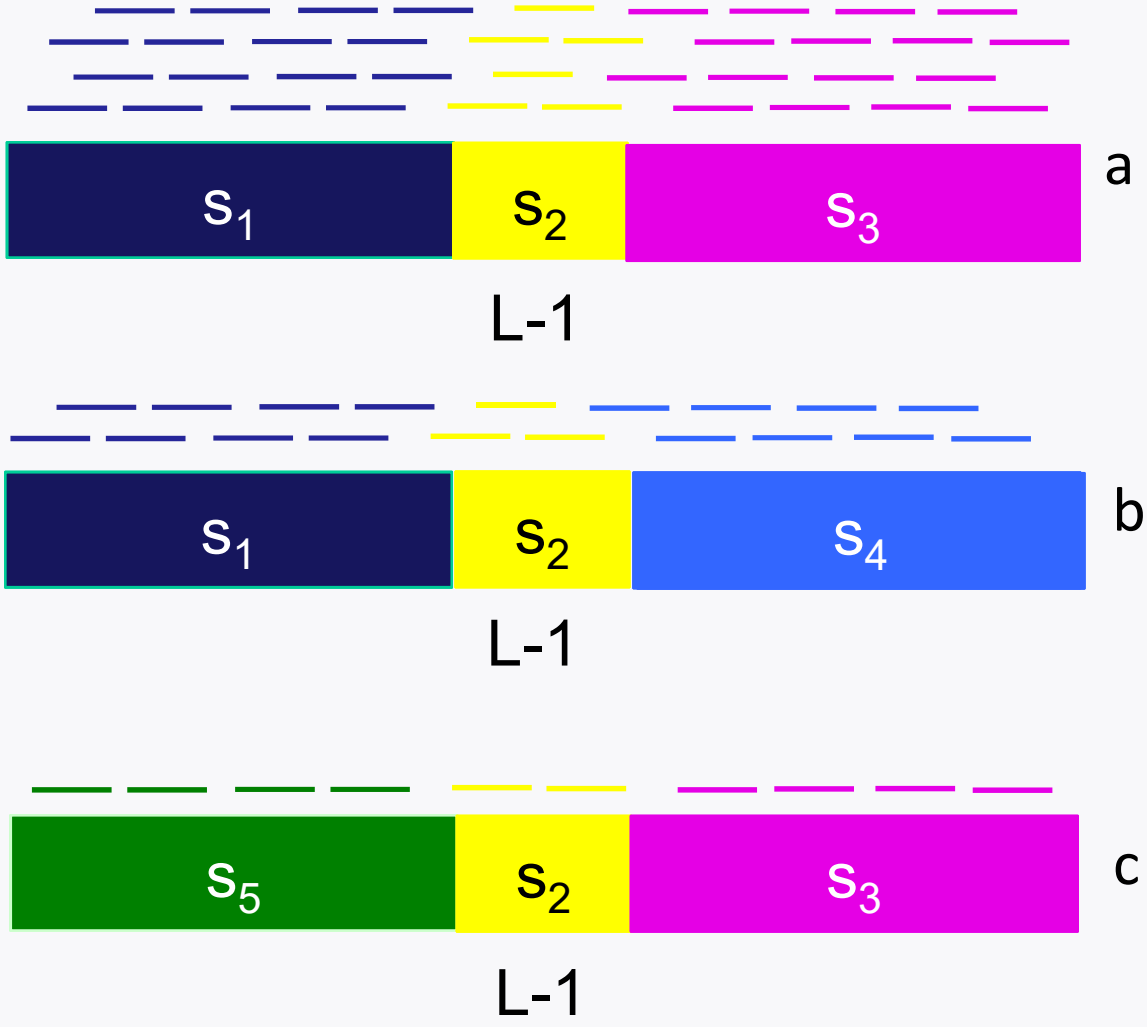


transcript 1

transcript 2


Equal abundance
Unequal abundance
s1
s3
s4
L-1
s2
s3
s5
a
L-1
c
s1
s3
s5
b
L-1
s2
s3
s4
b
L-1
s1
s3
s4
?
s2
s3
s5
unique solution, also sparse
s1
s3
s5
s2
s3
s4

Unresolvable inter-transcript repeats



Assembly algorithm architecture