# High Throughput Sequencing Assembly: Information and Computation

David Tse

Stanford University and U.C. Berkeley

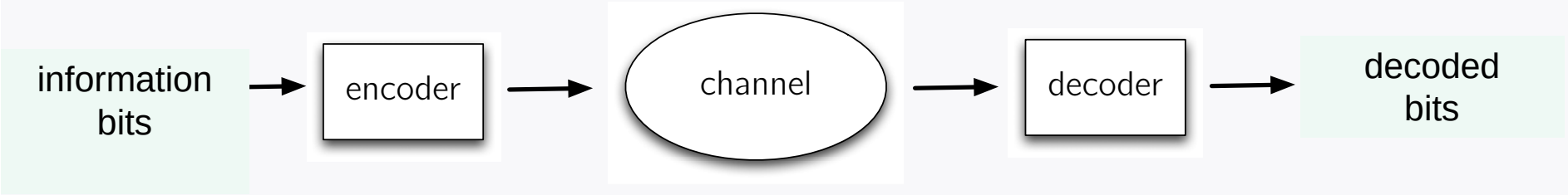
MBMC Workshop

USC

December 3, 2015

Joint work with Ilan Shomorony, Tom Courtade, Sreeram Kannan, Lior Pachter.
Research supported by NSF Center for Science of Information.

# A 60 year old quest in information theory




- In 1948, Claude Shannon showed that every channel has a capacity C.
- His method to achieve capacity uses random codes and very long block lengths.
- It took 60 years, but efficient capacity-achieving codes were finally found.

# Information and computation

C.E. Shannon

A.M. Turing

What is the information limit for communication?

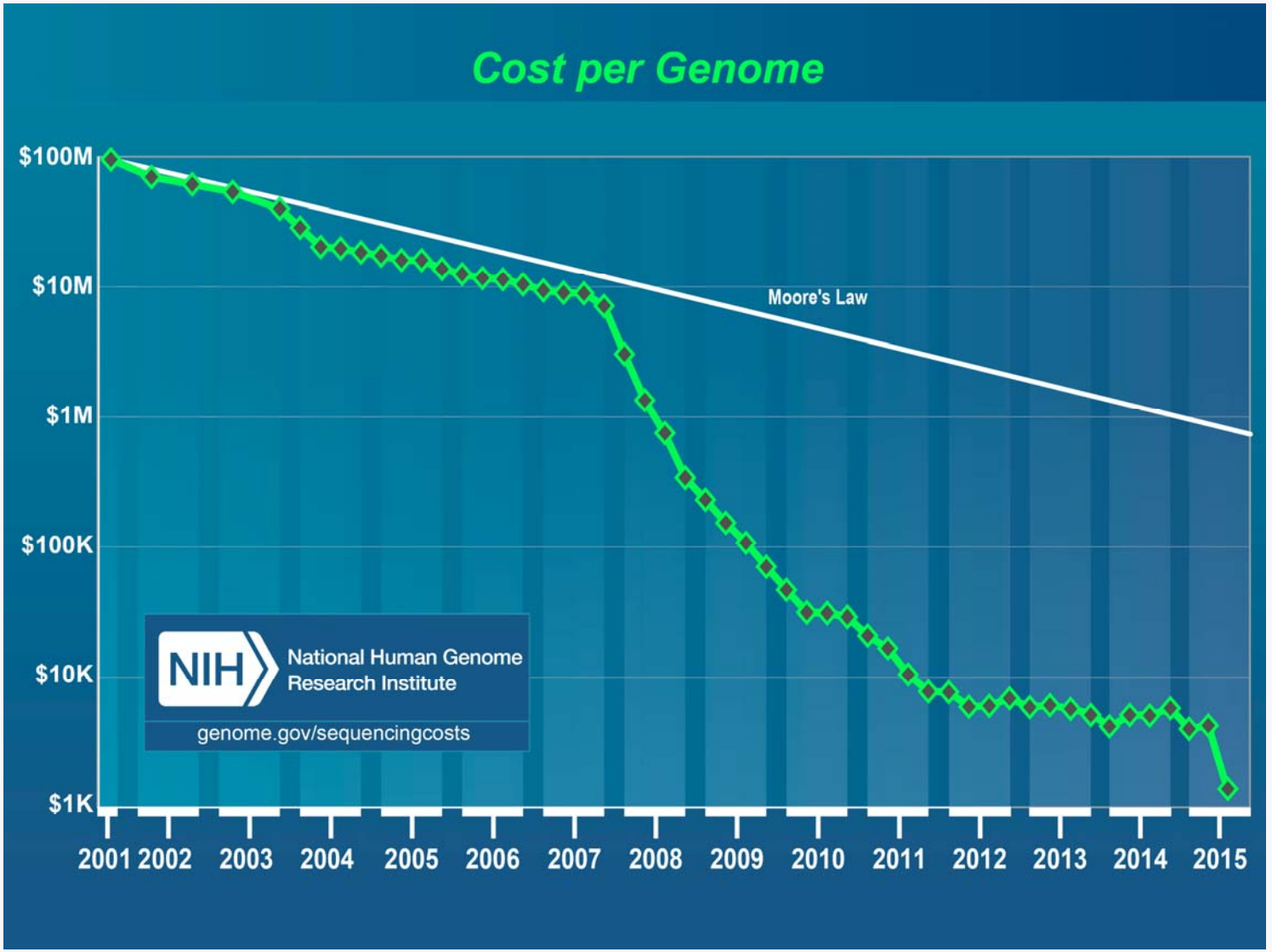
I only care about problem instances that matter.

But optimal decoding of general codes is NP-hard!

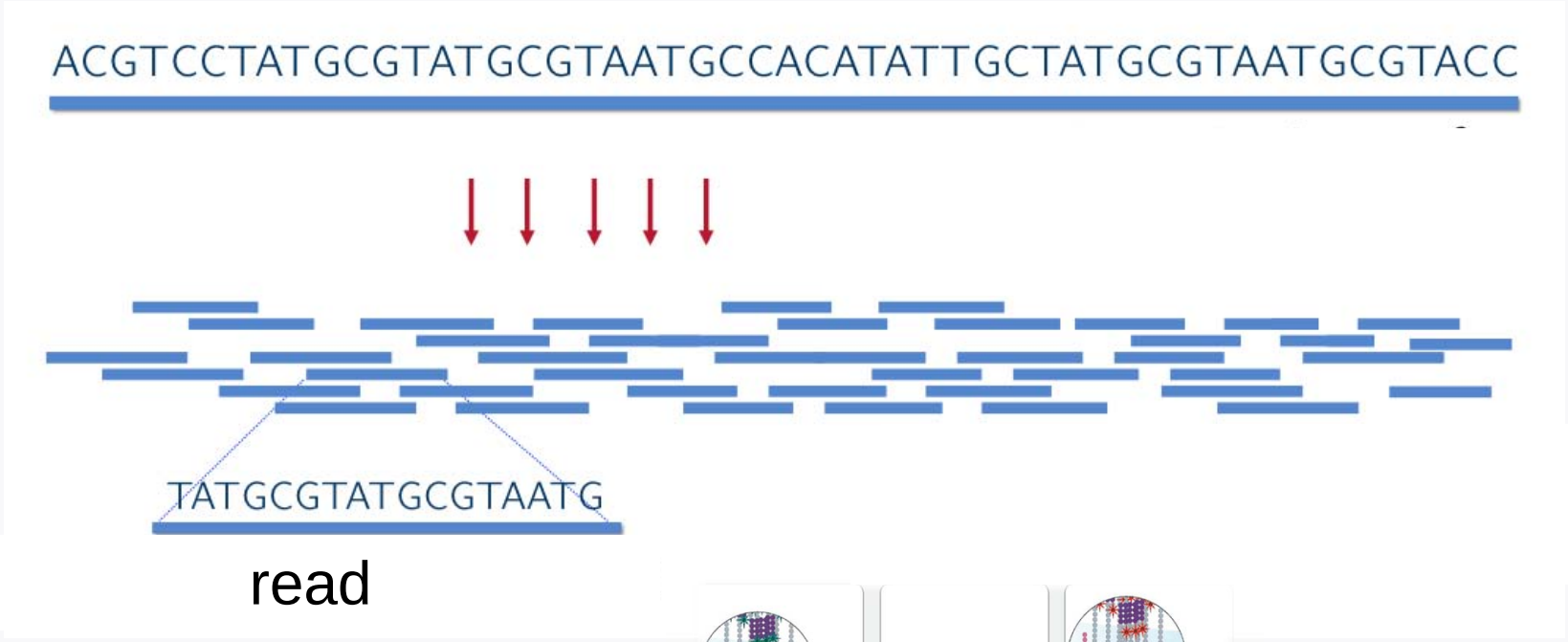
# Beyond communication

Can this way of thinking be broaden to other fields?

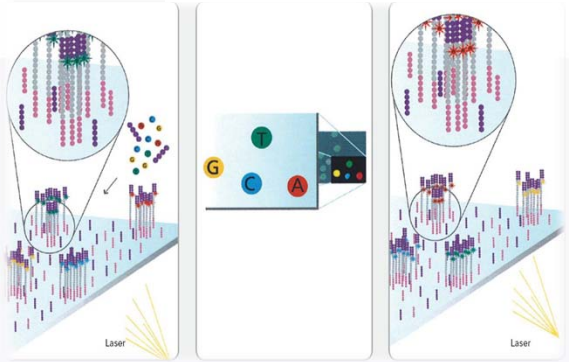
# High throughput sequencing revolution



# Shotgun sequencing




read



# The assembly problem: a gigantic jigsaw puzzle


A single sequencing experiment can generate 100's of millions of reads, 10's to 100's gigabytes of data.

# Computational complexity view

- Formulate the assembly problem as a combinatorial optimization problem.
- 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:**

How much read data is needed to reliably reconstruct?

# Key challenge: repeats

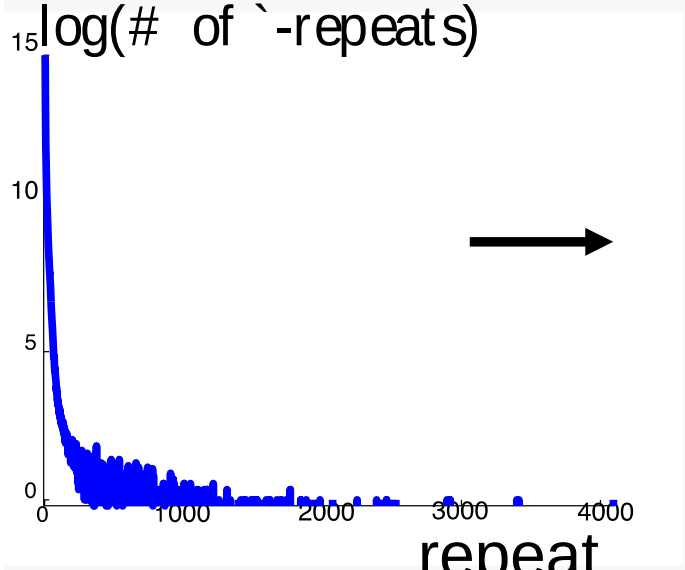
harder jigsaw puzzle

easier jigsaw puzzle

How exactly do the fundamental limits depend on repeat statistics?

# Information limit of assembly

(Bresler, Bresler & T. BMC Bioinformatics 13)




$\frac{N}{N_{LW}}$

information limit

$L_{critical}$

Lander-Waterman coverage

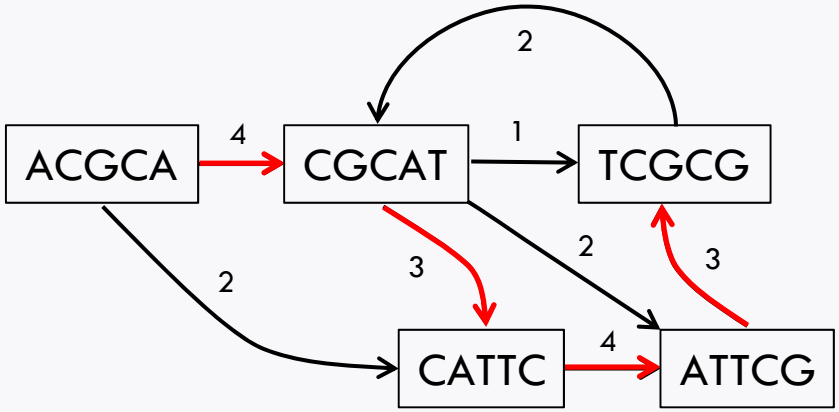
read length $L$

Human Chr 19
Build 37

# The assembly problem

(Shomorony, Courtade & T. 15)

- Formulated on a read-overlap graph

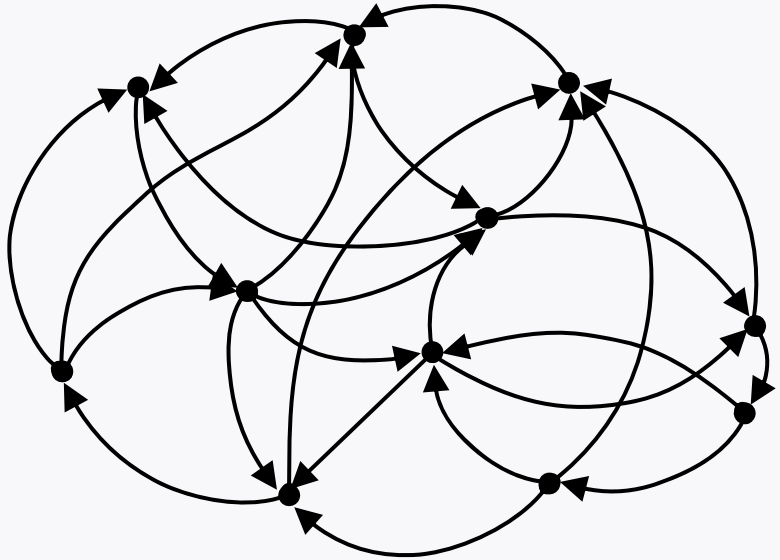



ACGCATTCGCG

- Sequence is a path that visits every node.
  - (Generalized) Hamiltonian Path
  - Finding shortest GHP is NP-hard

# The assembly problem

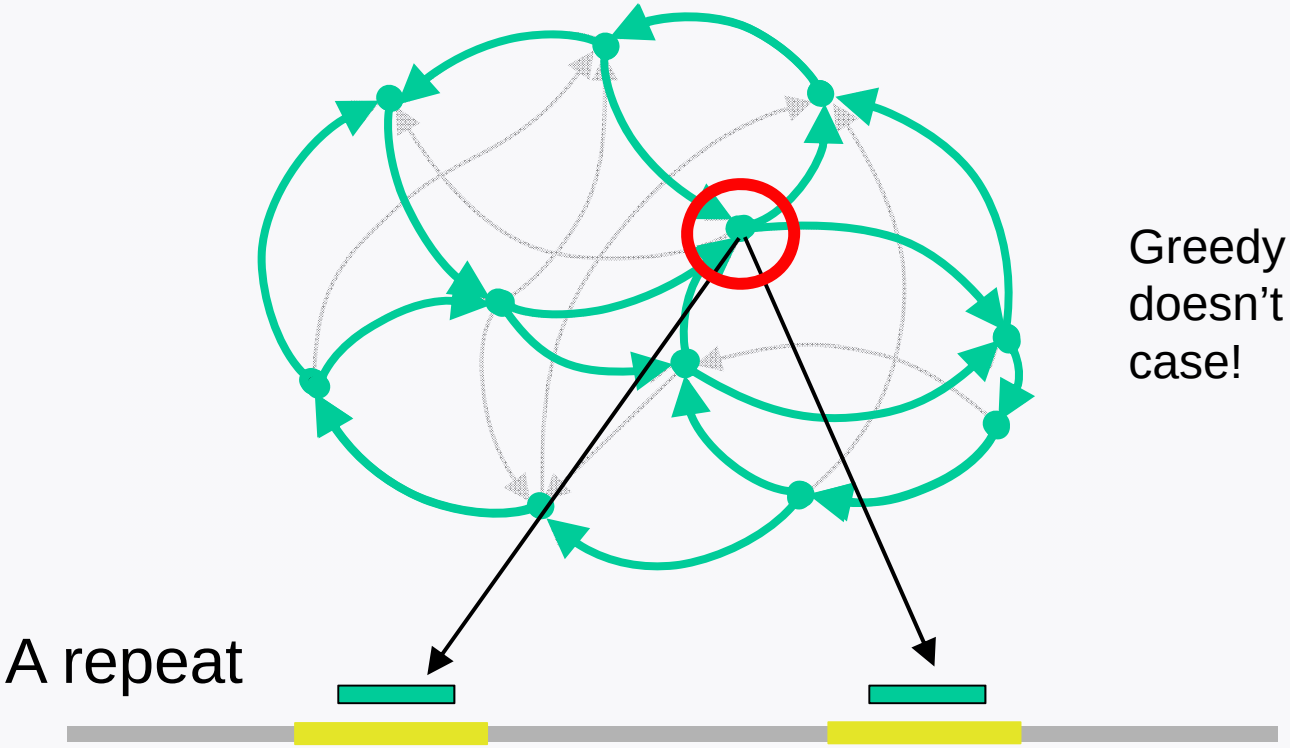
- Formulated on a read-overlap graph



- Sequence is a path that visits every node
  - (Generalized) Hamiltonian Path
  - Finding shortest GHP is NP-hard

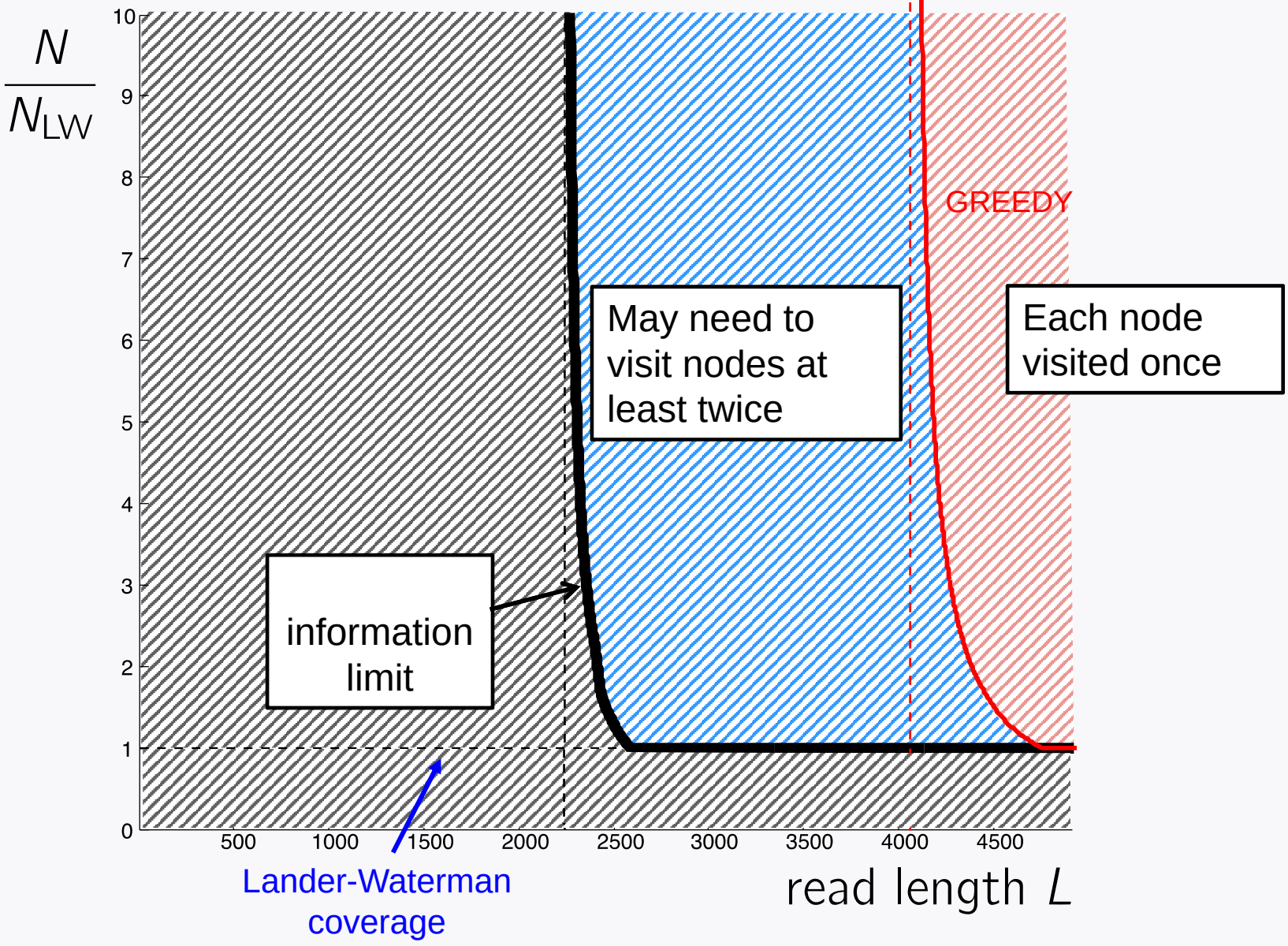
# How well does a greedy algorithm do?

- For each node, go to node with largest overlap.



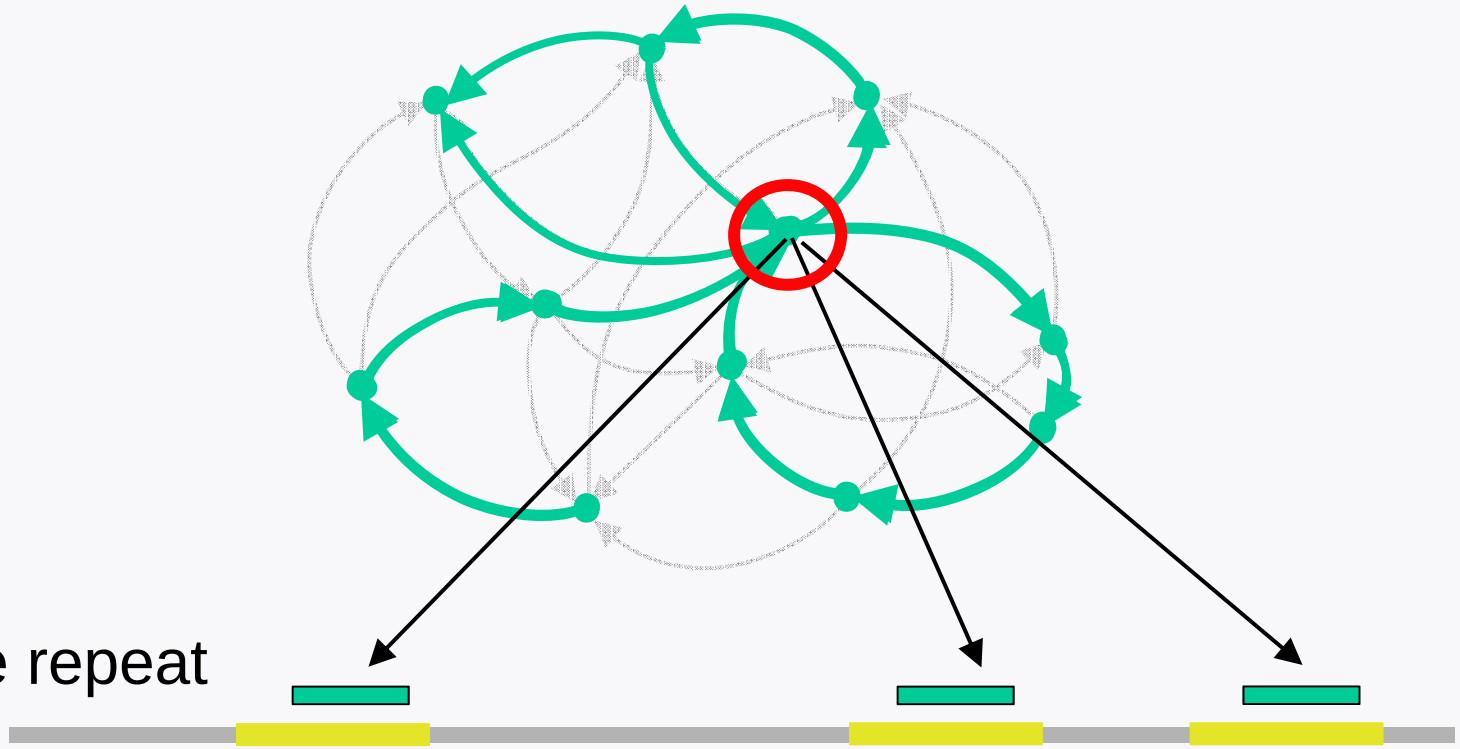

- What if true path visits a node twice?

# Information limit



# Do we need to visit nodes 3 or more times?

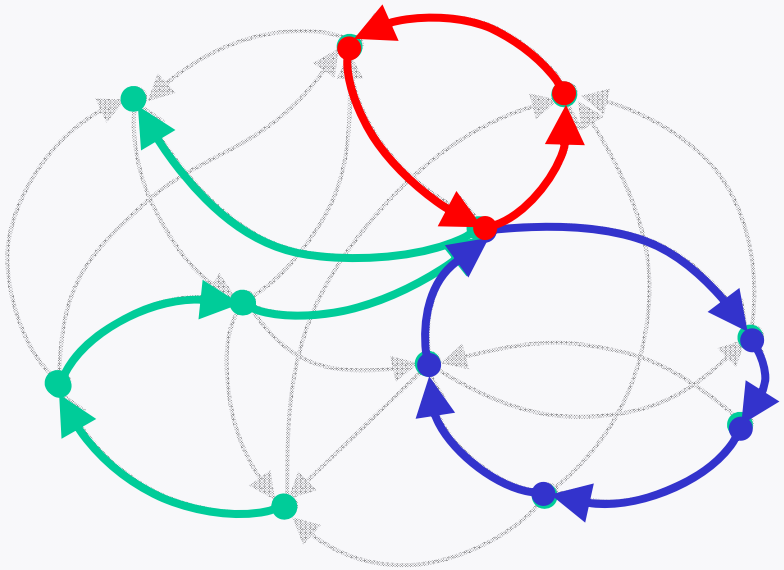
- What if true path visits a node three times?



A triple repeat

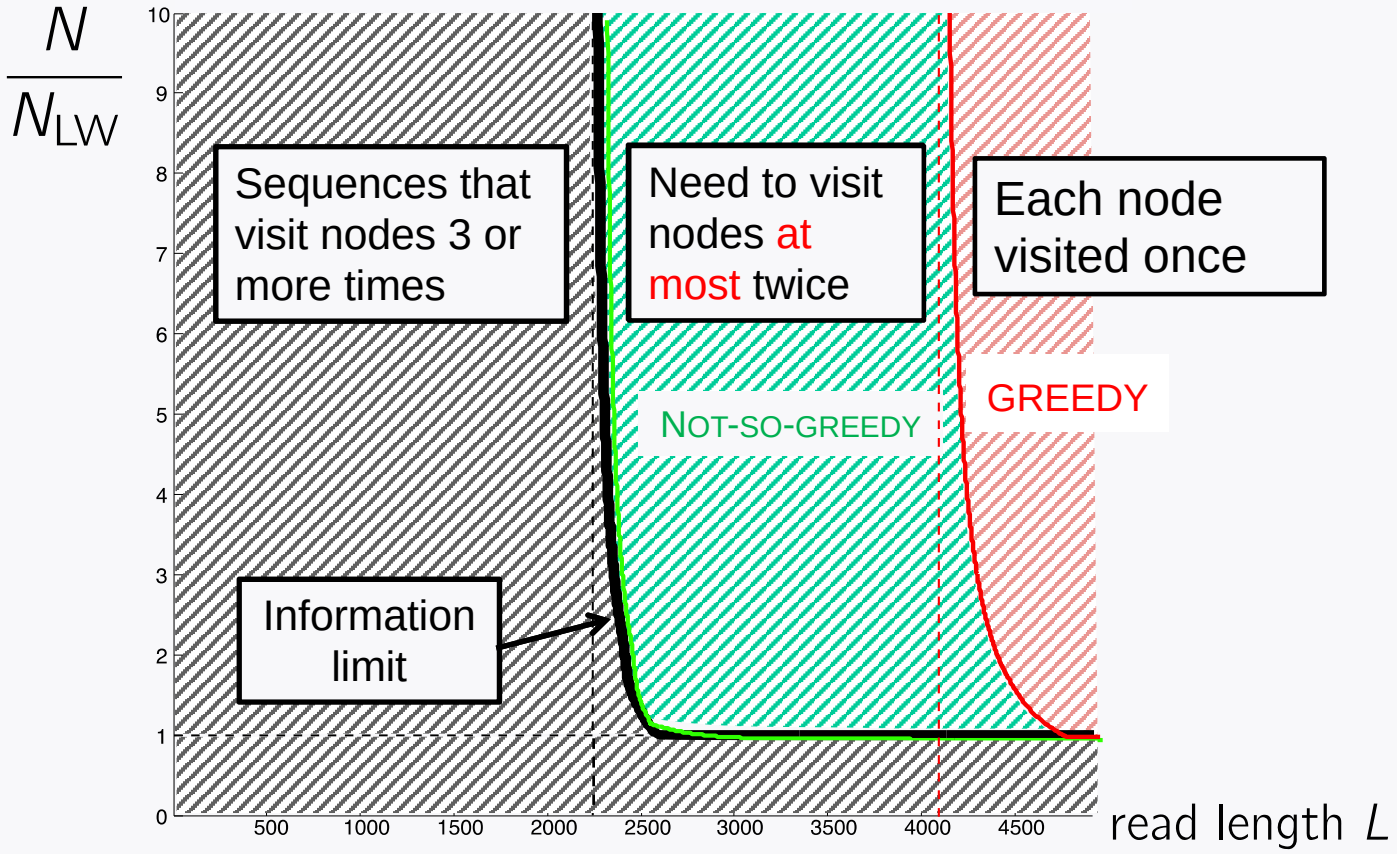
# Do we need to visit nodes > 2 times?

- What if true path visits a node three times?



- Both paths have the same likelihood!

# Information limit

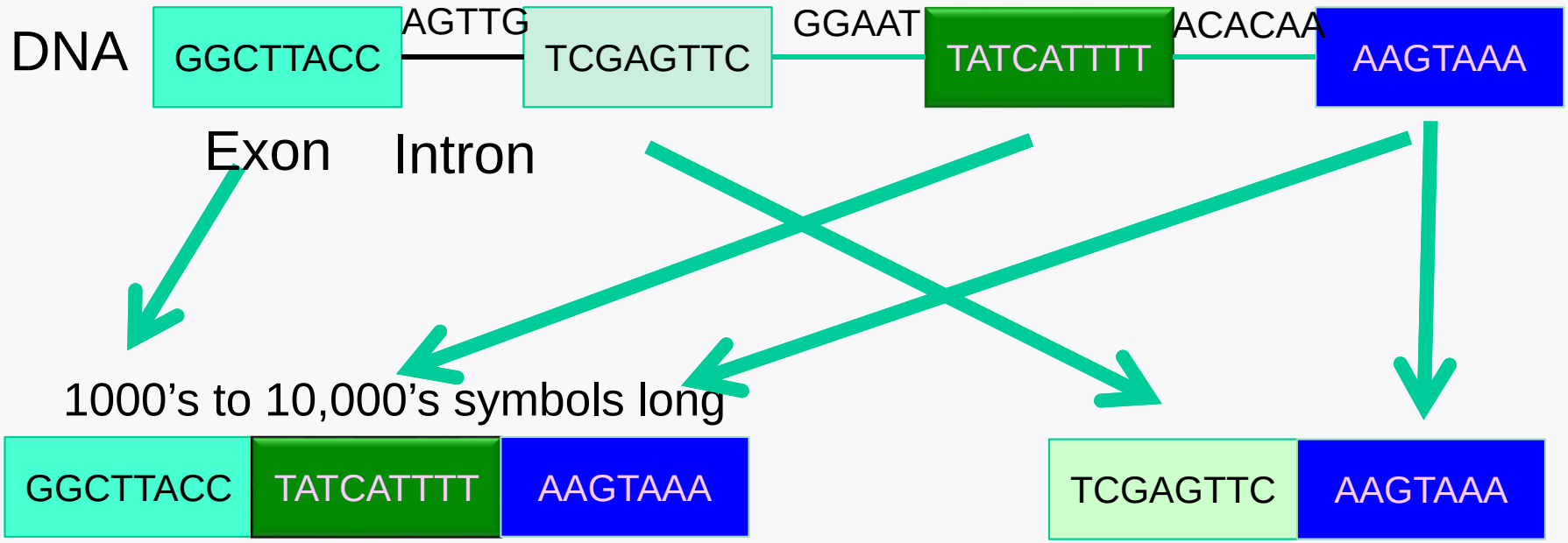



- "Not-so-greedy" algorithm (Shomorony, Courtade & T. 2015)
  - Only uses two best extensions.
  - With further pruning, complexity linear in number of reads.

# Information then computation

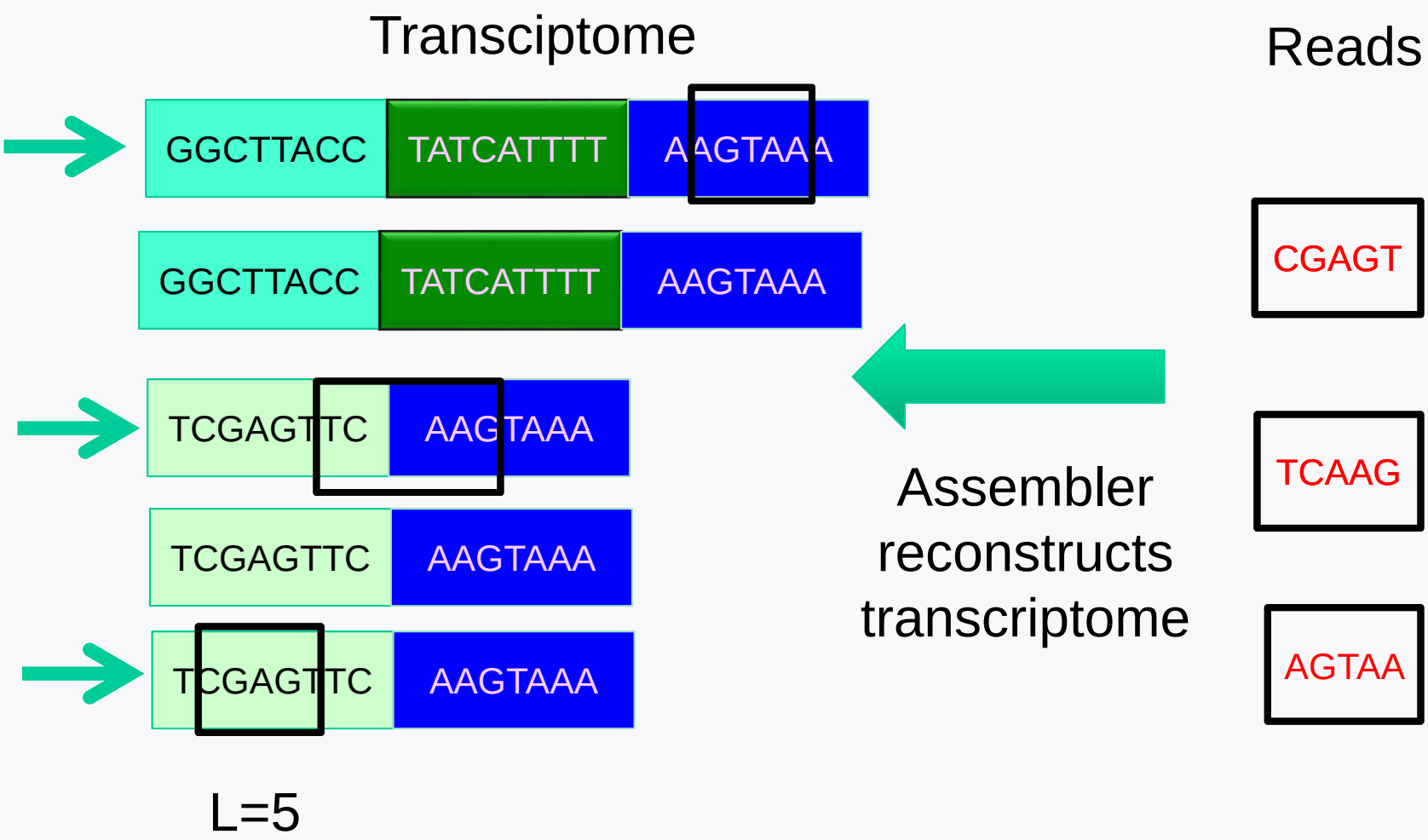
- NP-hardness is a worst-case measure of complexity over all problem instances.
- But we only care about the instances where there is enough data to reconstruct.
- It happens that those instances are also efficiently solvable.

# From DNA to RNA




The human transcriptome has 10,000's of such transcripts from 20,000 genes.

# RNA-Seq assembly




Both intra and inter-transcript repeats.

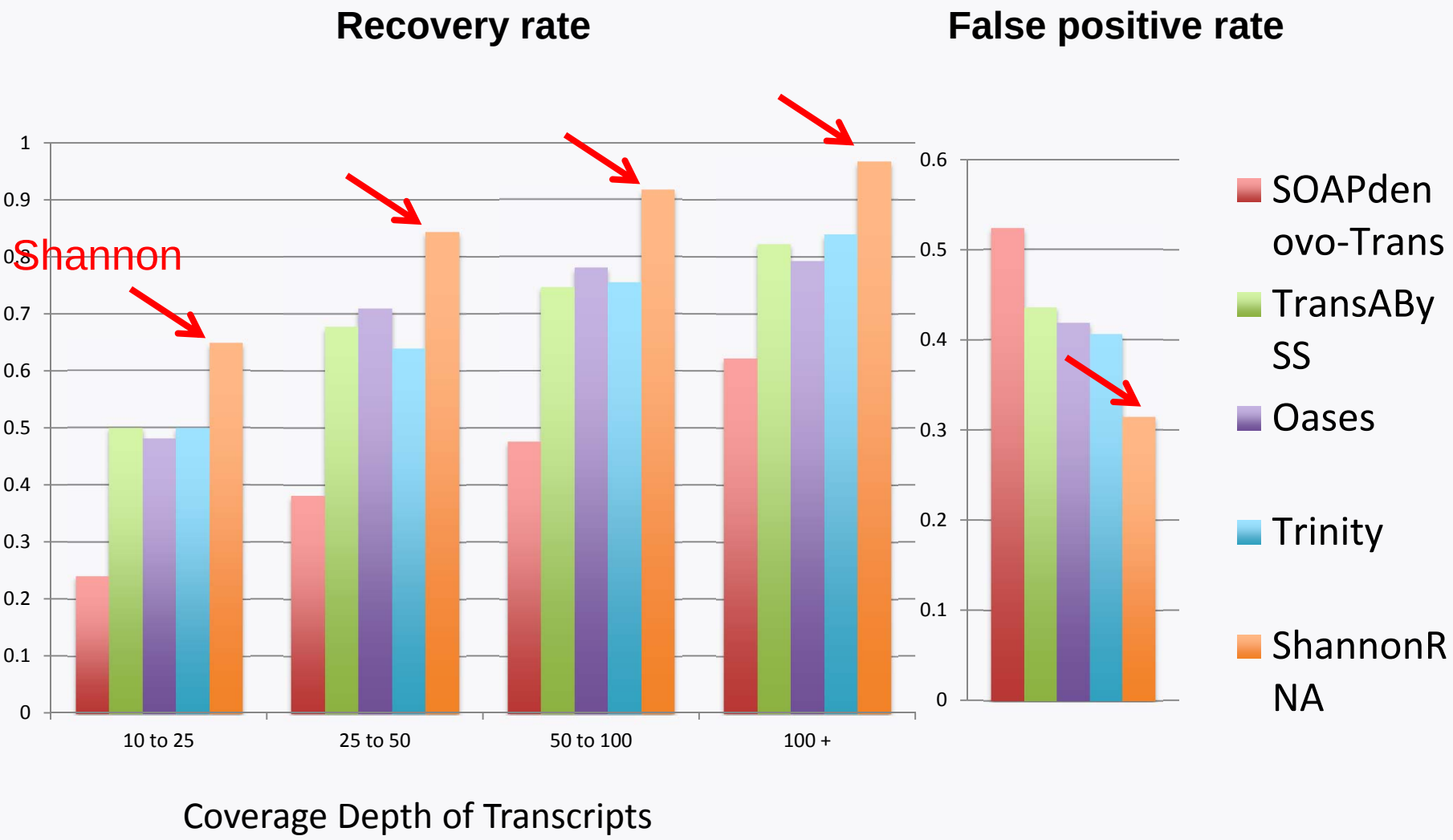
# RNA assembler: Shannon

Kannan, Pachter & T. 15

- Characterized information limit due to both intra and inter-transcript repeats.
- Intra-transcript repeats resolved using an algorithm similar to DNA assembly.
- Inter-transcript repeats resolved using an algorithm that finds sparsest flow in a graph.
- Sparsest flow is NP-hard, but we show that the reconstructible instances can be computed in linear time.

# Shannon: evaluation on real data

Read length=50, 135 Million reads, human embryonic stem cell (Au et al, PNAS 2013)



# Conclusion

- Information theory is about fundamental limits.
- It is also a constructive theory.
- It overcomes computationally intractable problems by focusing on tractable instances.