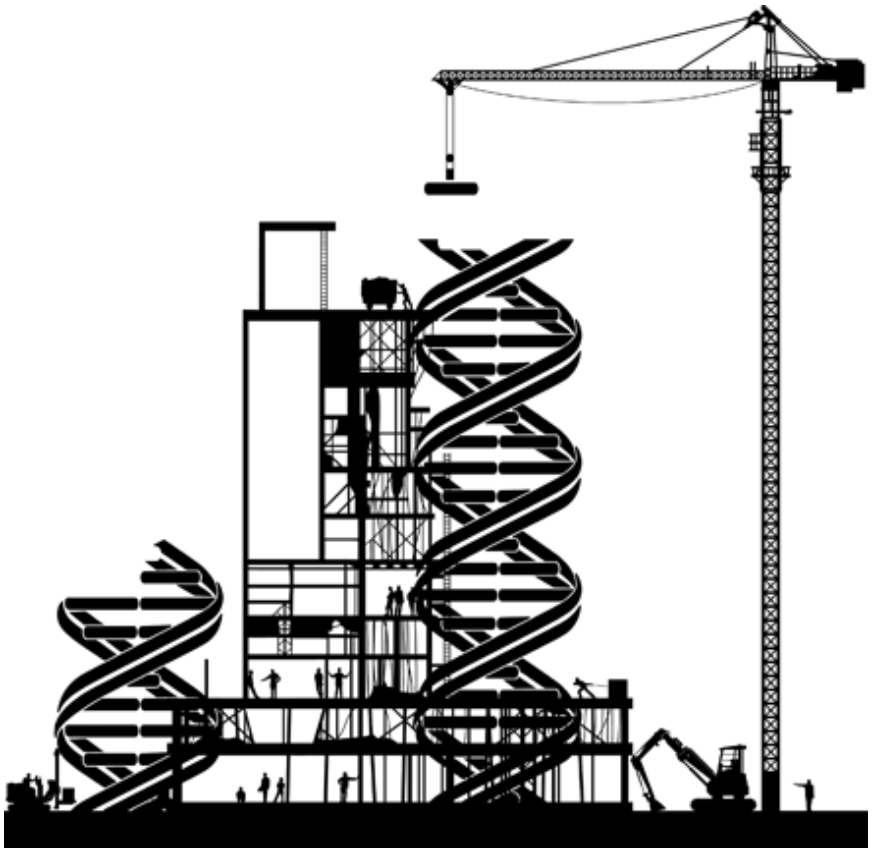
Measuring transcription in a SCRaMbLEd genome with Nanopore sequencing

Aaron Brooks / @scalefreegan

Steinmetz and Stegle Groups, EMBL Heidelberg

...in the not too distant future

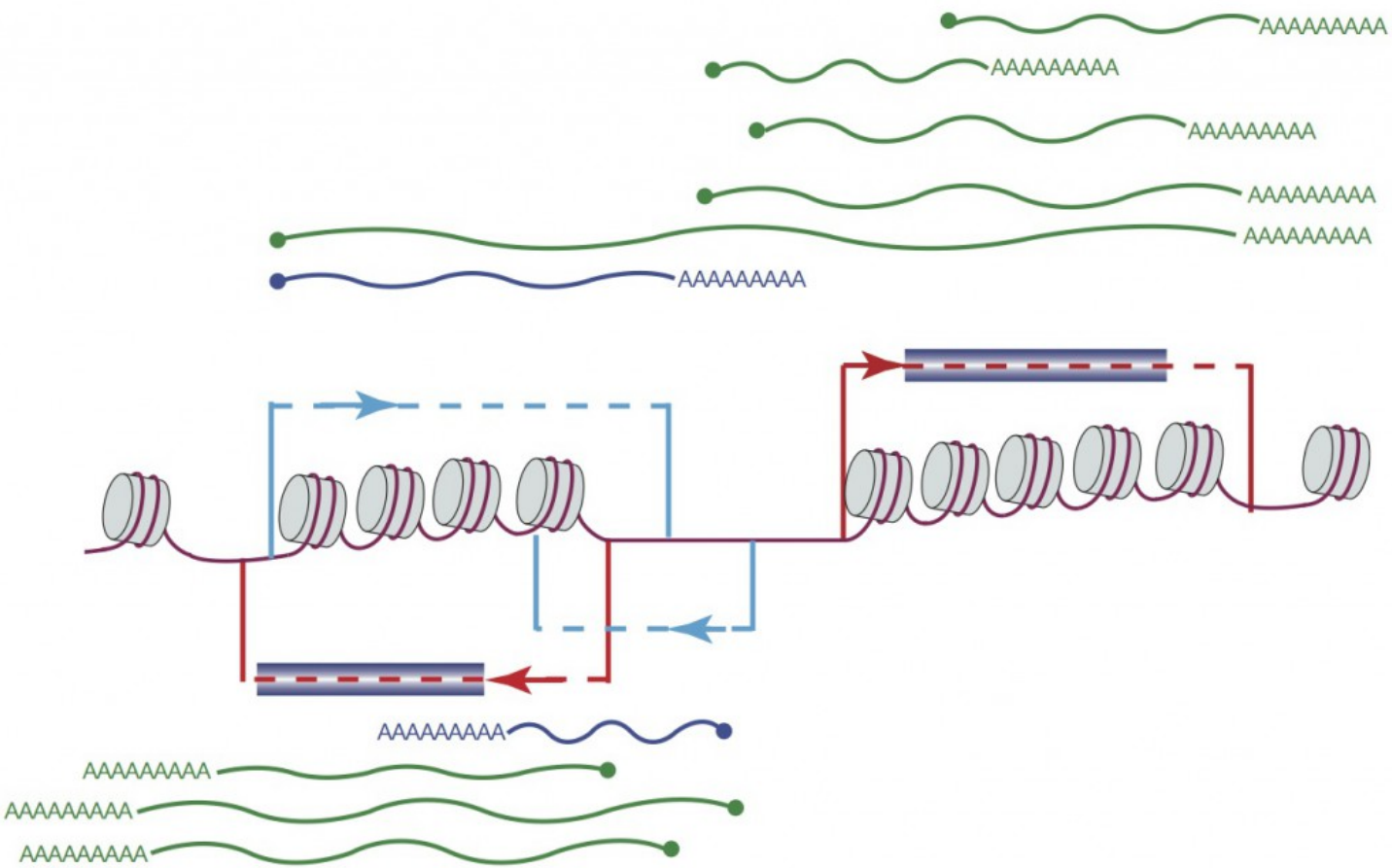


you are building a genome from scratch...

Where are you going to put the genes? In what order and which orientation? Does it matter?

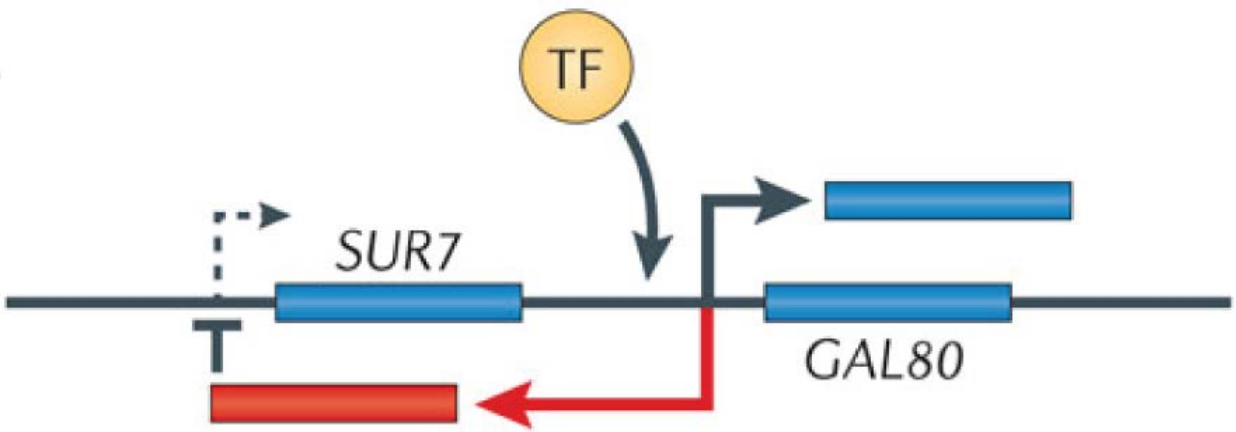
Science Magazine, Mar 28th 2014

Transcription is linked to genome architecture



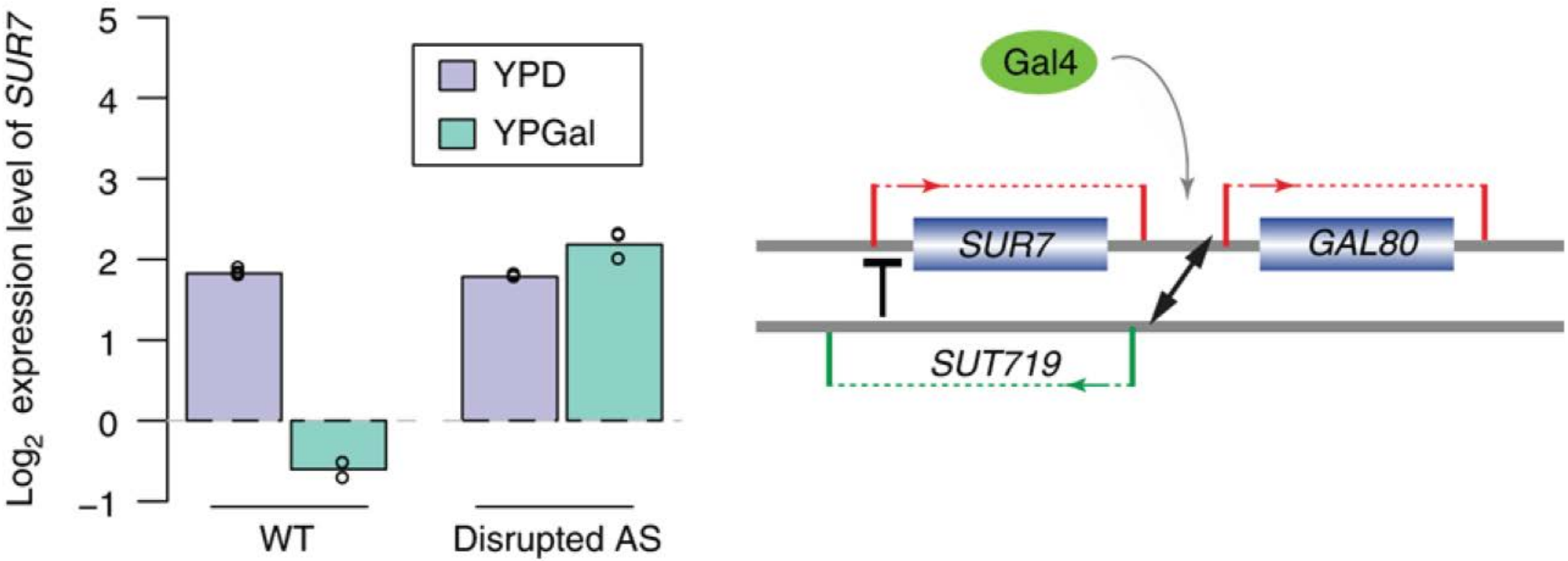
Pelechano Lab

Regulatory signals can propagate between neighbouring promoters



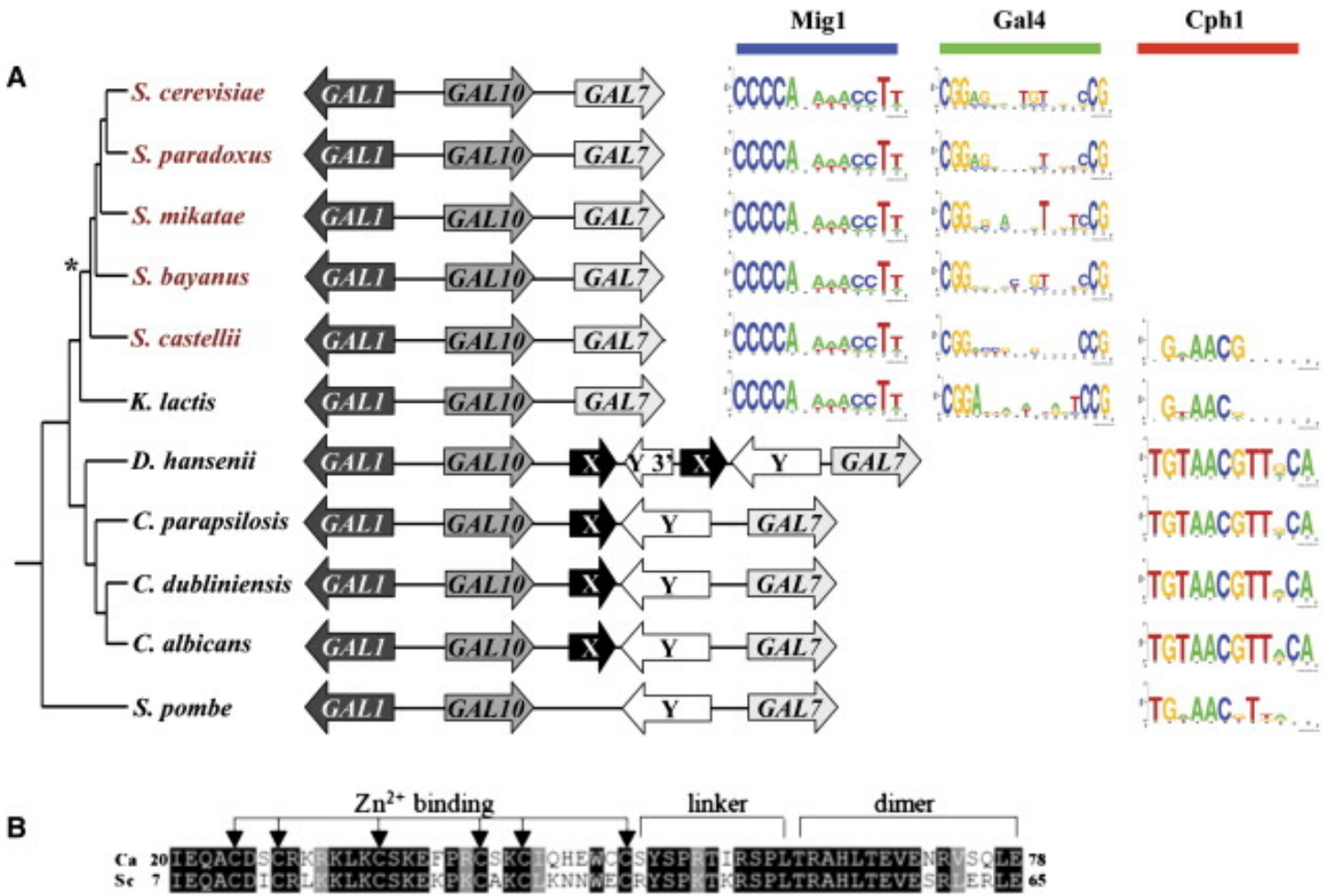
Xu et al 2011
Pelechano and Steinmetz 2013

Antisense-mediated regulation of SUR7



Xu et al 2011

Conservation of organization despite regulatory divergence



Martchenko et al 2011

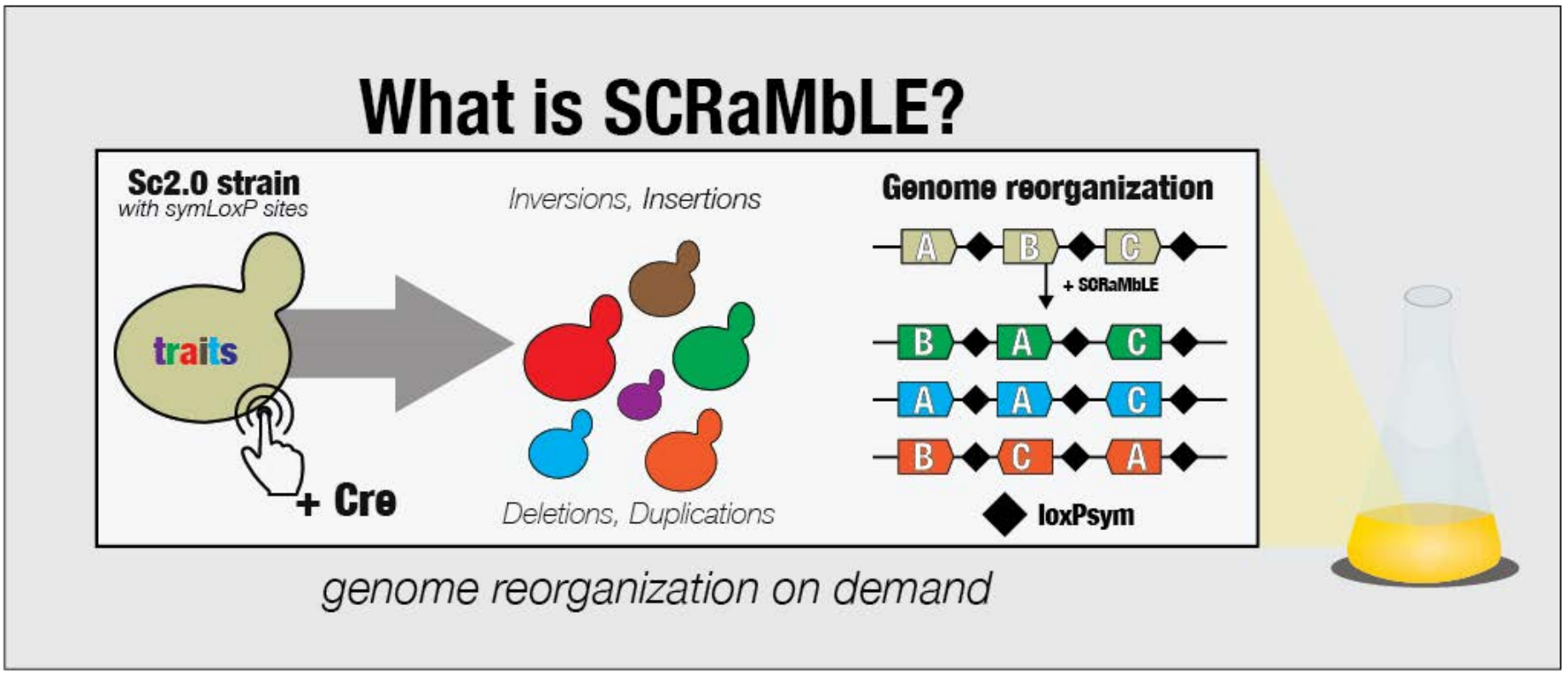
"...gene shuffling could be evolutionarily restricted in regions that support beneficial interactions between loci, whereas in other regions, ***new rearrangements could give rise to novel interaction networks that are mediated by ncRNAs and antisense transcripts.***"

Pelechano and Steinmetz 2013


Why are genomes organized as we observe them?

Gene shuffling with SCRaMbLE

Synthetic Chromosome Recombination and Modification by LoxP-mediated Evolution

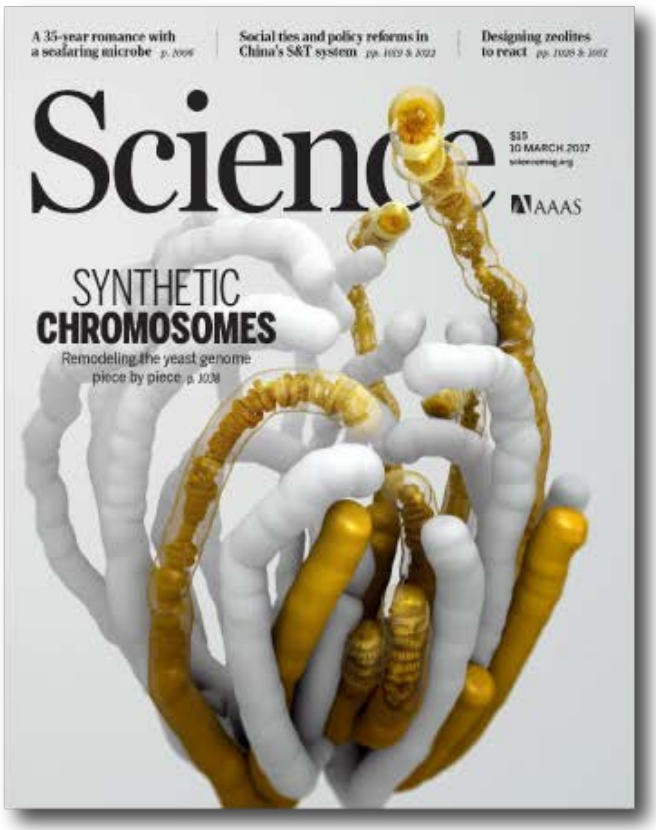


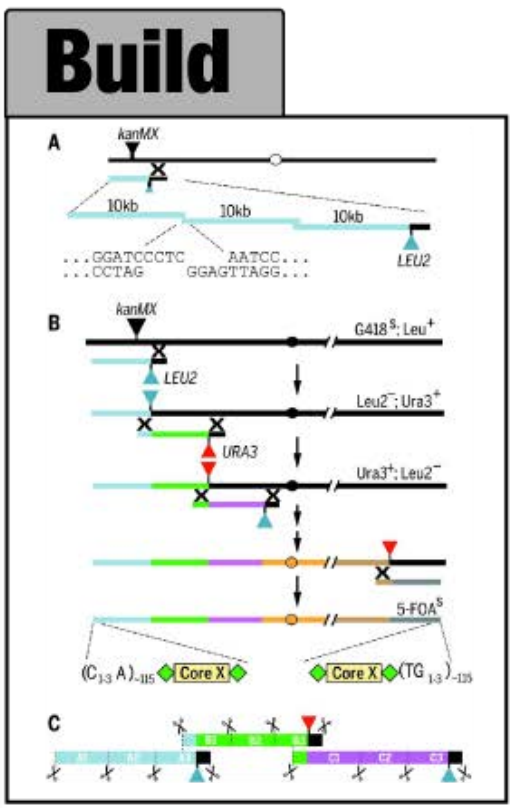
Dymond and Boeke 2012

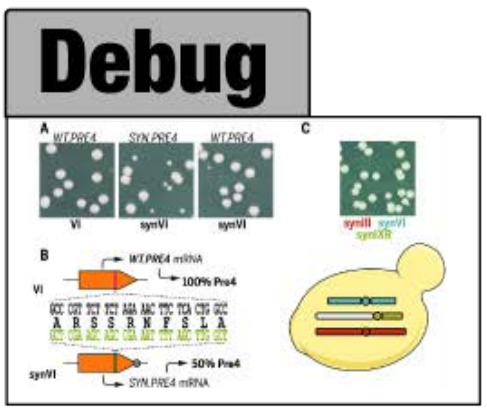
"A big step towards an artificial yeast genome"

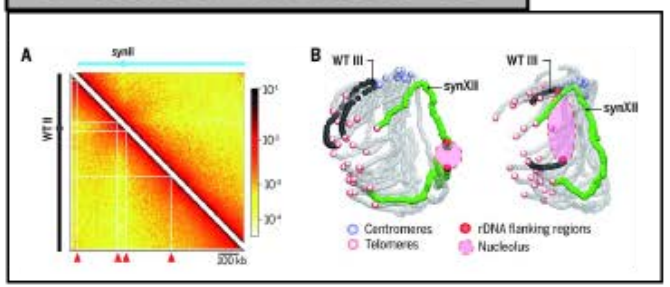
Economist, Mar 11th 2017

"Remodeling the yeast genome piece by piece"



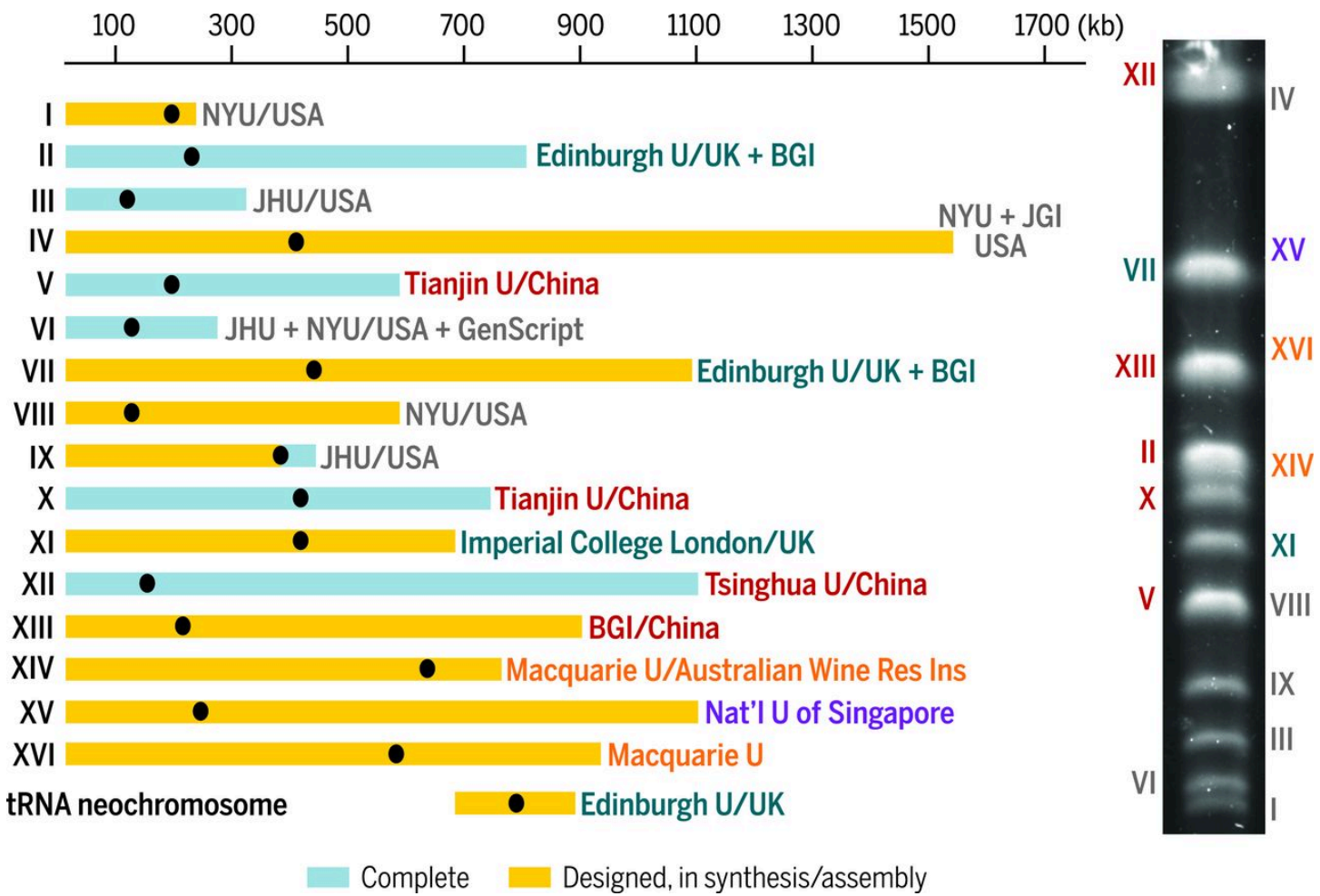






Synthetic Yeast Genome Project (Sc2.0) Science Special Issue

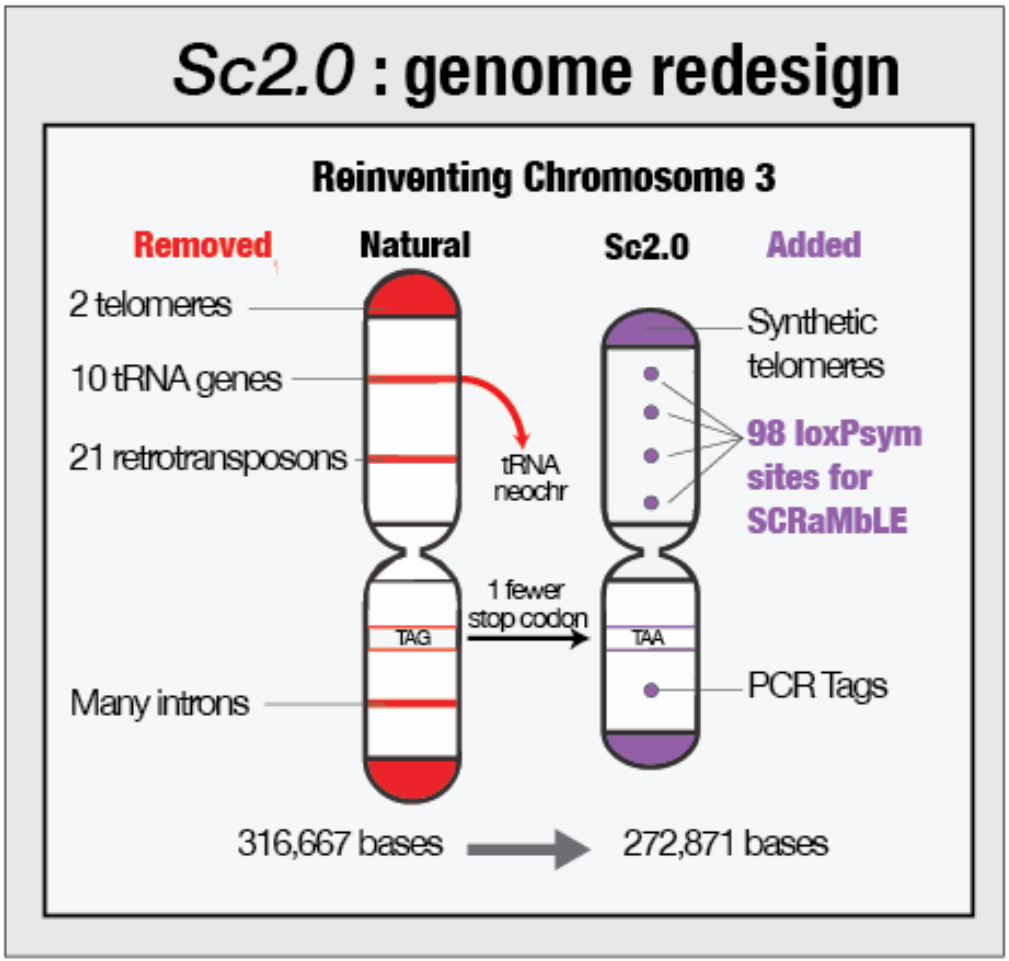
International consortium for design and synthesis



...synthesis is ongoing...

Richardson et al. 2017

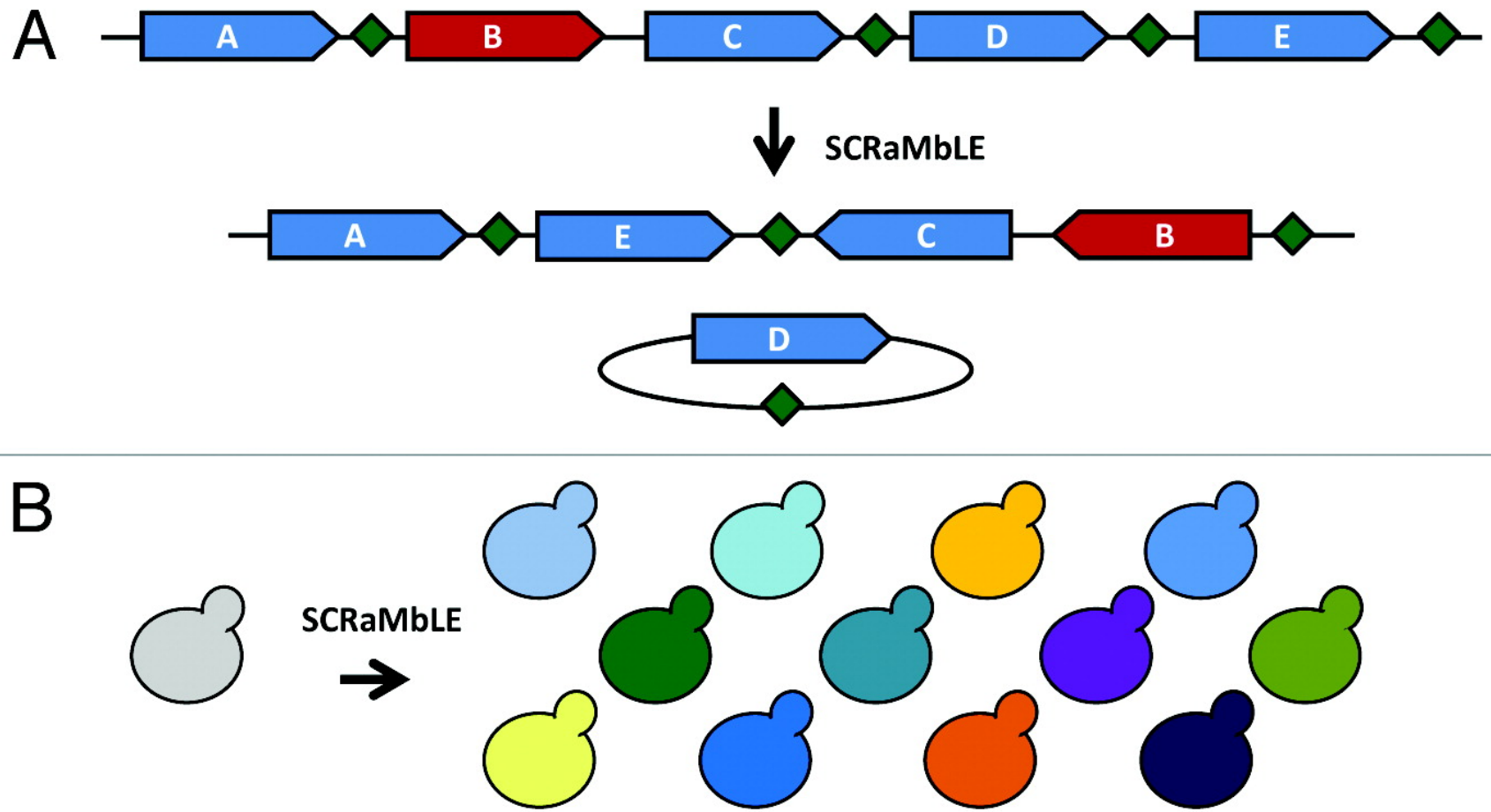
Encode sequence diversity by design: **Sc2.0**



Pennisi et al. 2014

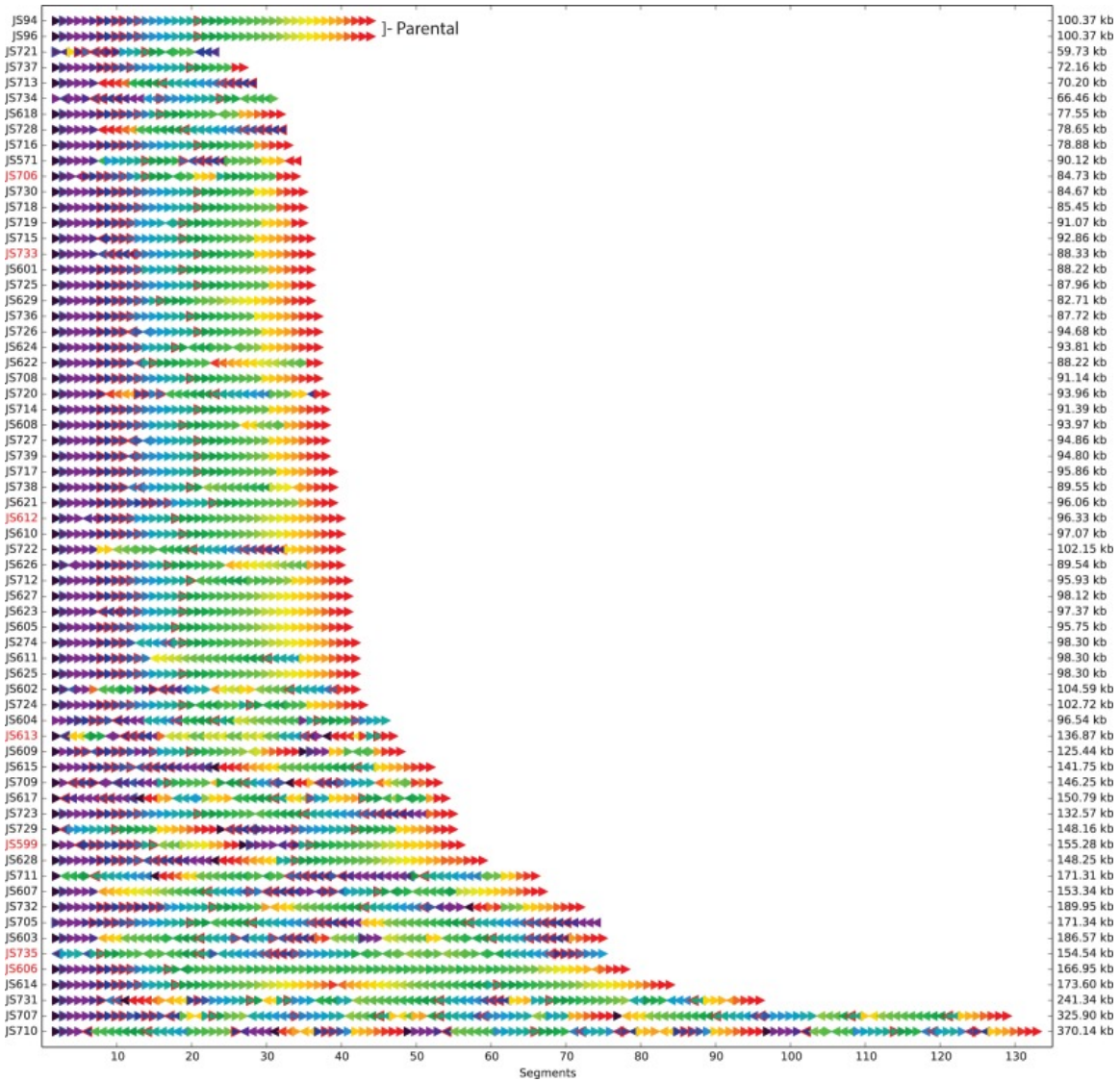
Gene shuffling with SCRaMbLE

Synthetic Chromosome Recombination and Modification by LoxP-mediated Evolution



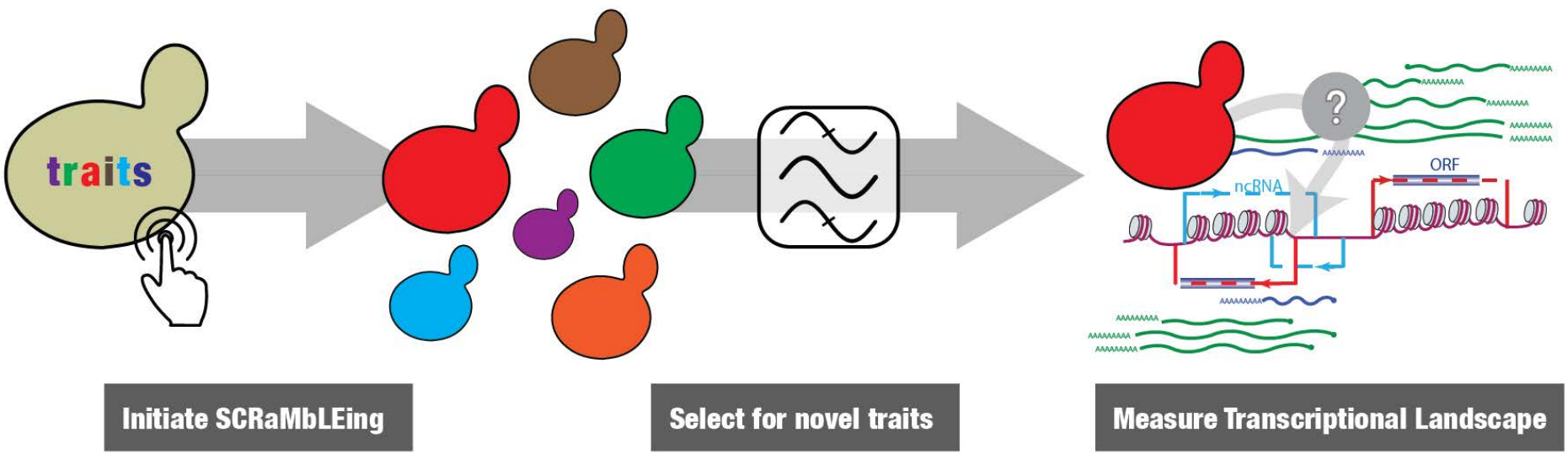
Dymond and Boeke 2012

Extensive heterogeneity among SCRaMbLE strains



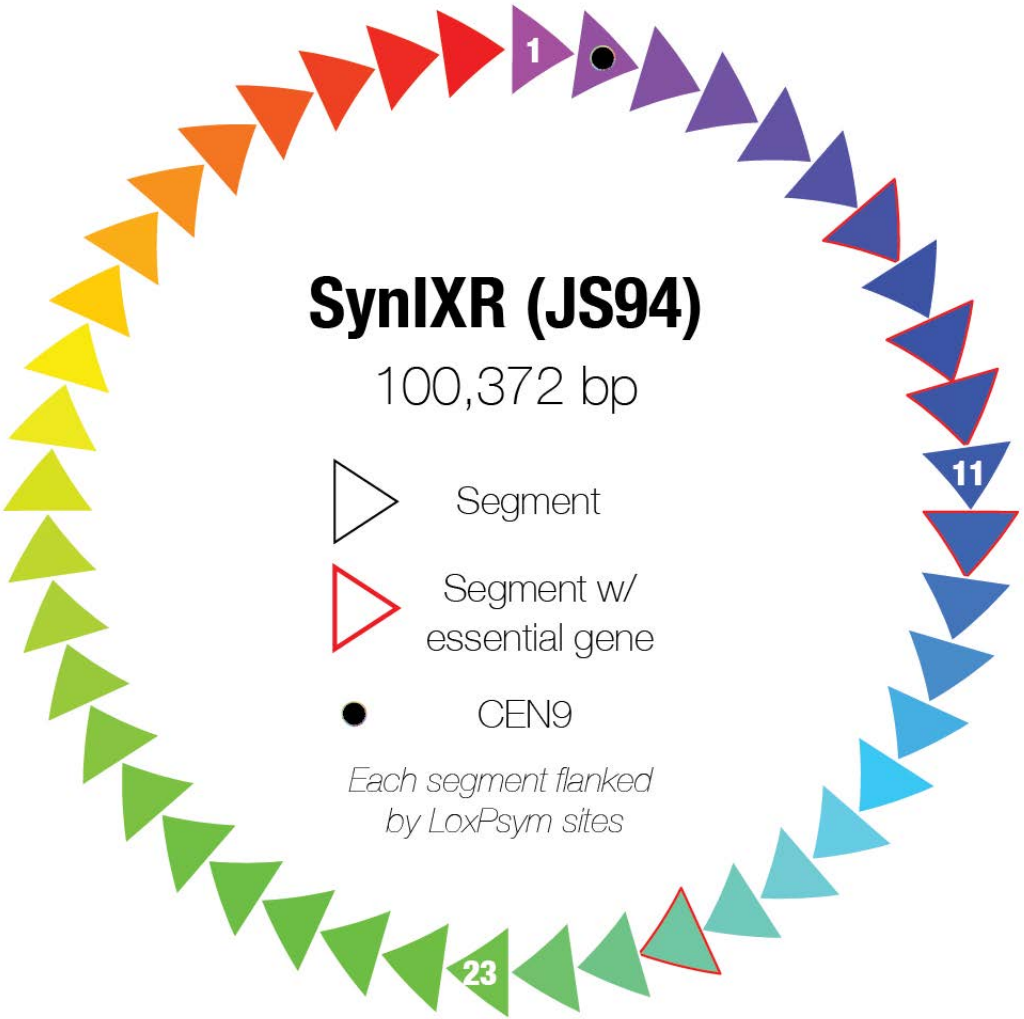
Shen et al. 2016

How is the transcriptional landscape modified by genome SCRaMbLEing?



What is the sequence of a SCRaMbLEd genome?

64 strains derived from SCRaMbLE of SynIXR



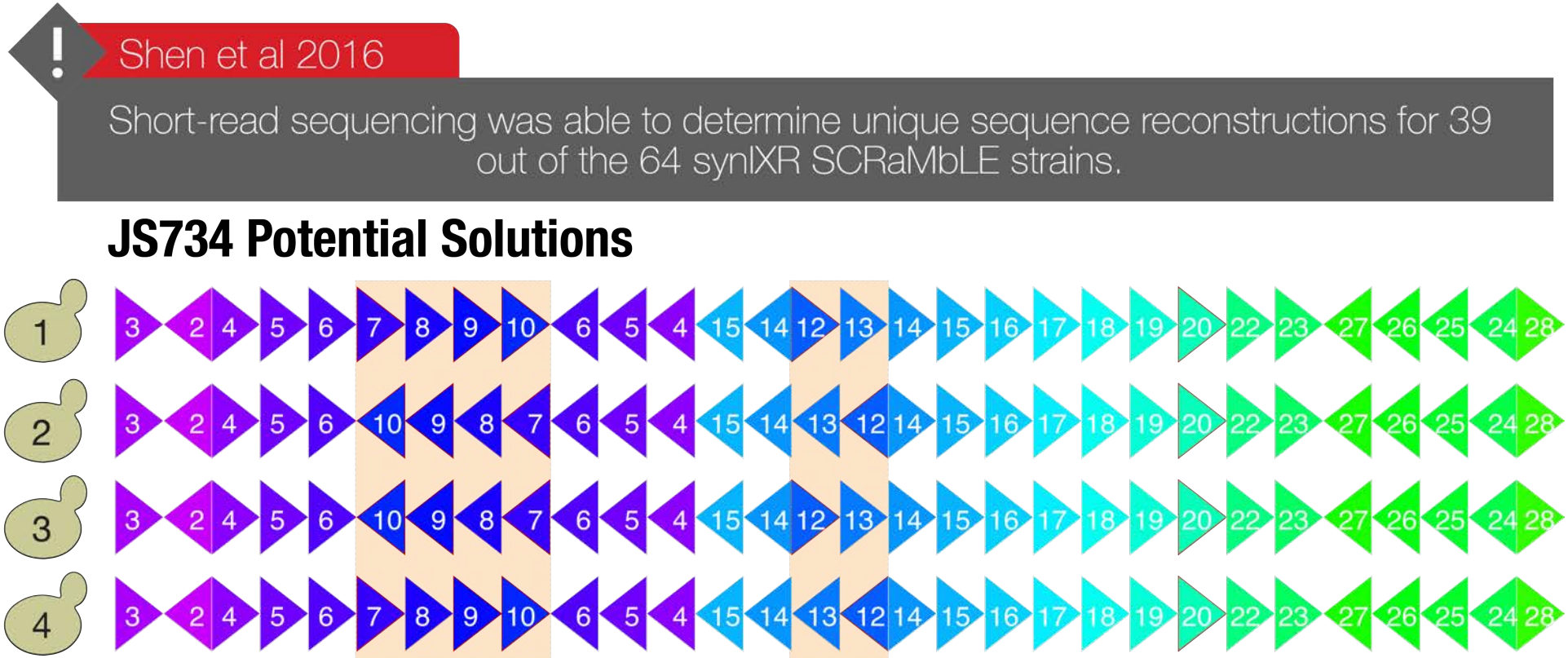
Each genome segment identified by color and number. ~1 gene per segment. SCRaMbLEgram

Many genomes cannot be assembled fully with short reads

JS734 has four potential solutions

Shen et al 2016

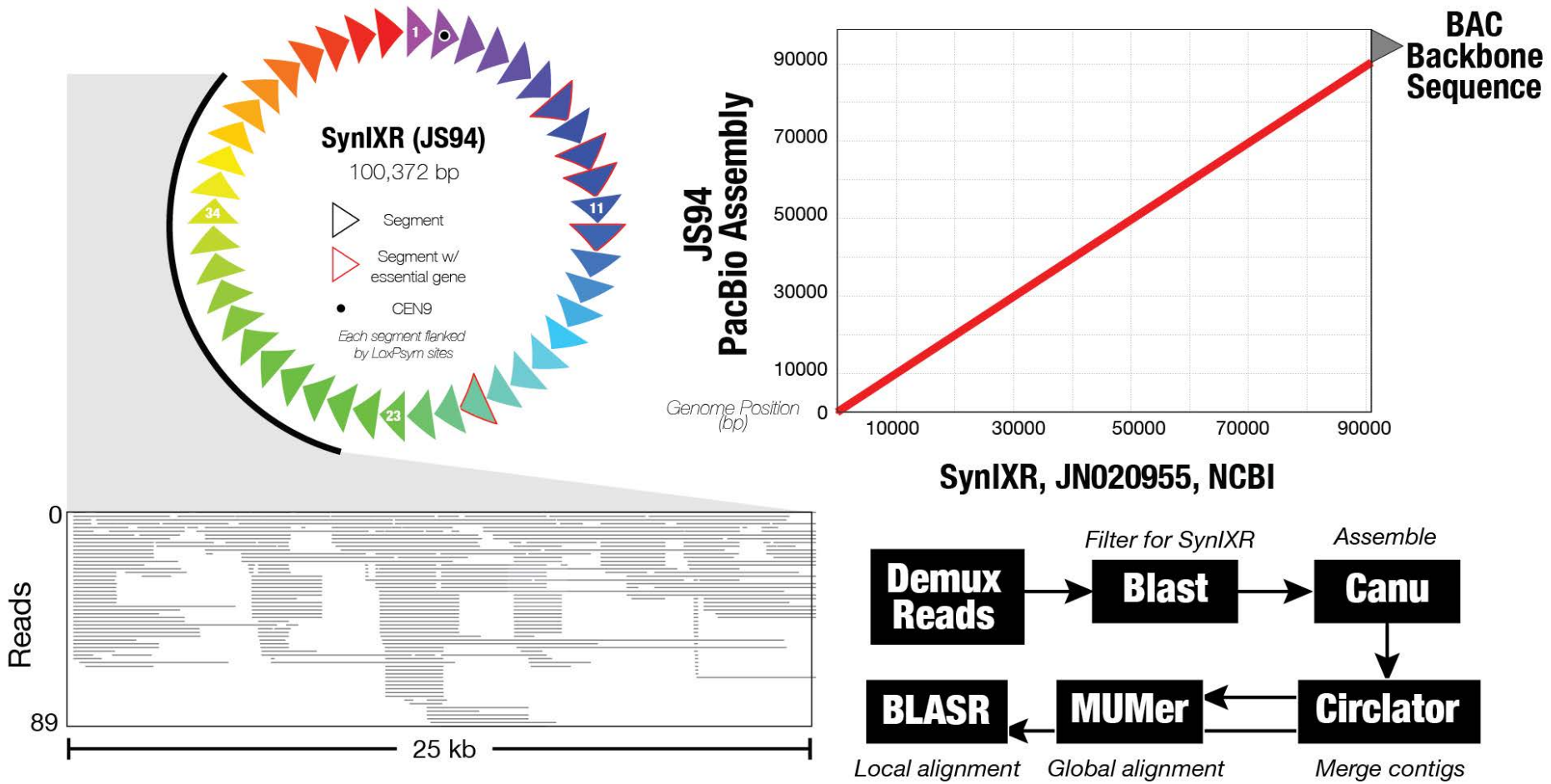
Short-read sequencing was able to determine unique sequence reconstructions for 39 out of the 64 synIXR SCRaMbLE strains.



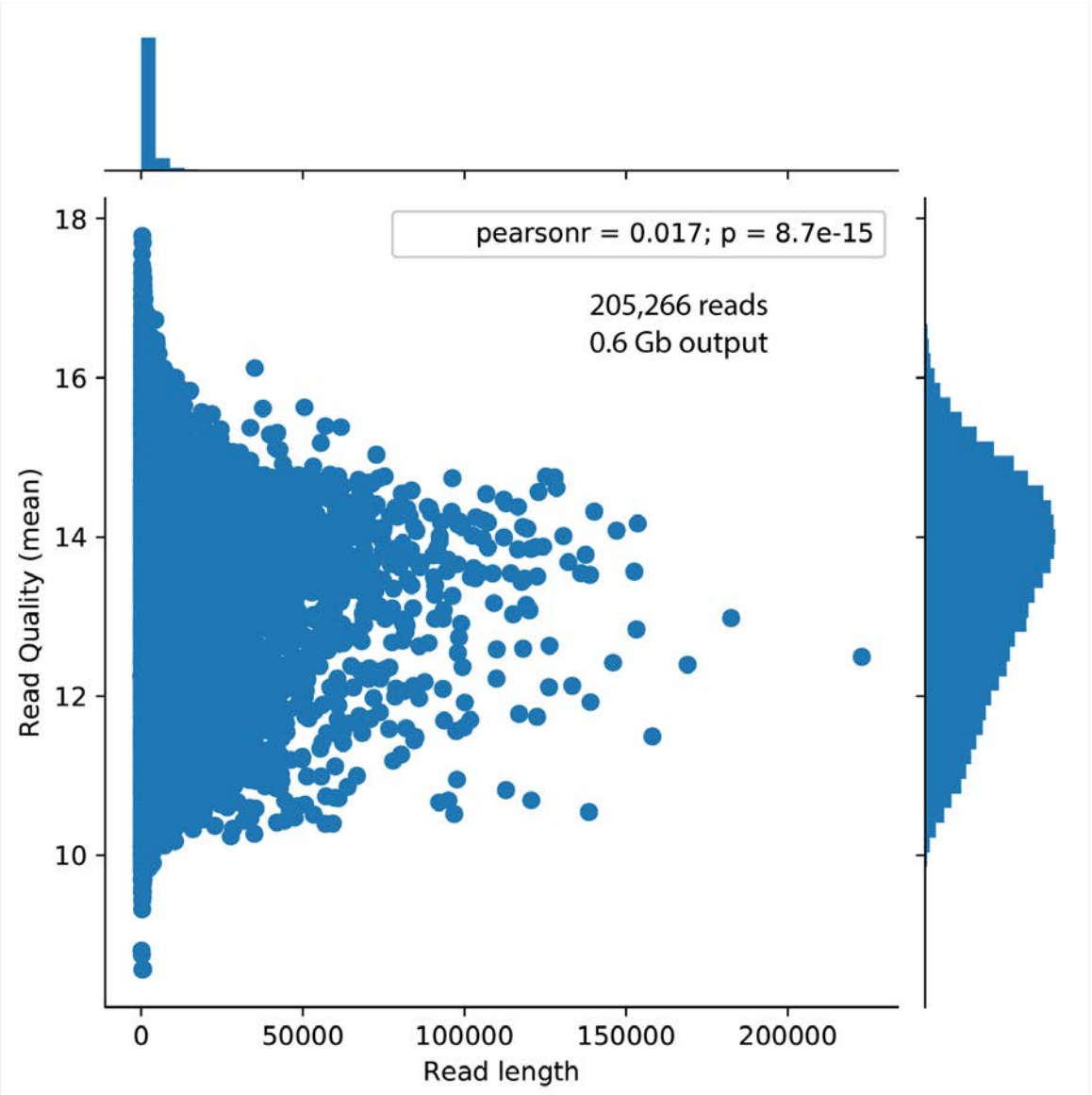
Shen et al. 2016

Workflow for *de novo* assembly with long-reads

Parental strain (JS94) can be solved with single SMRT cell

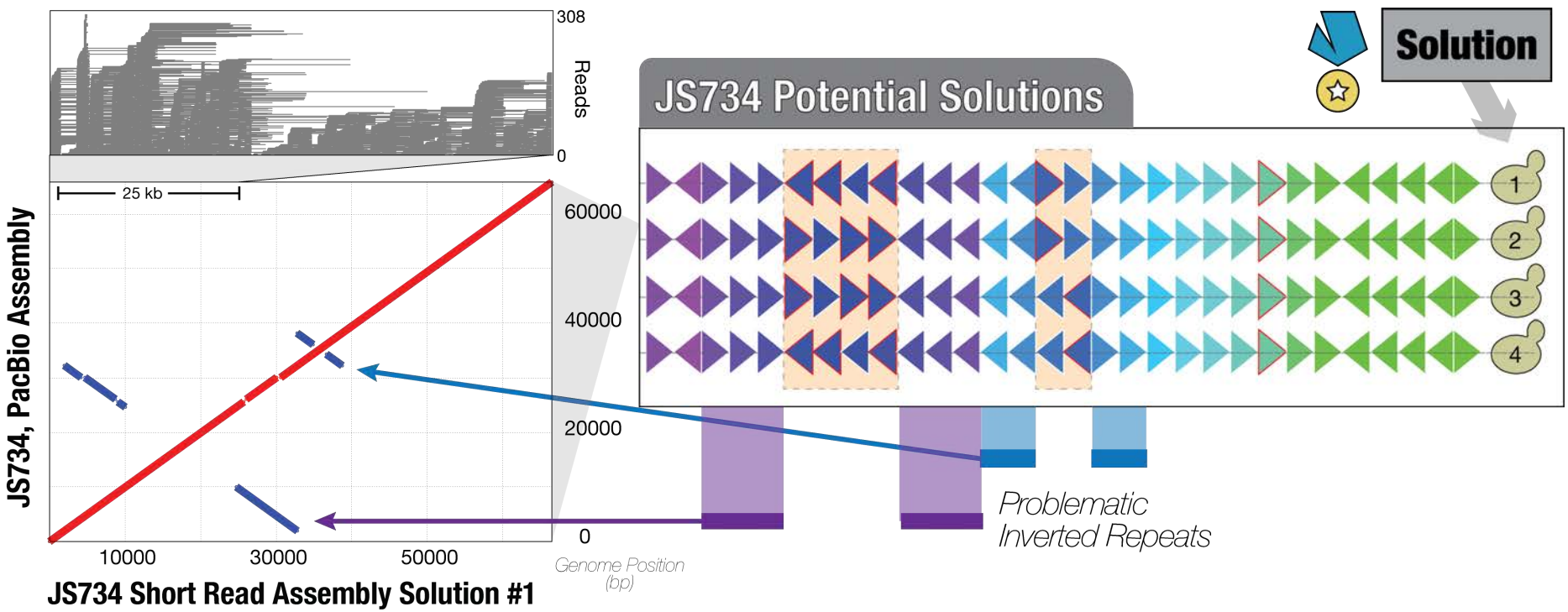


Our Nanopore sequencing output is below average

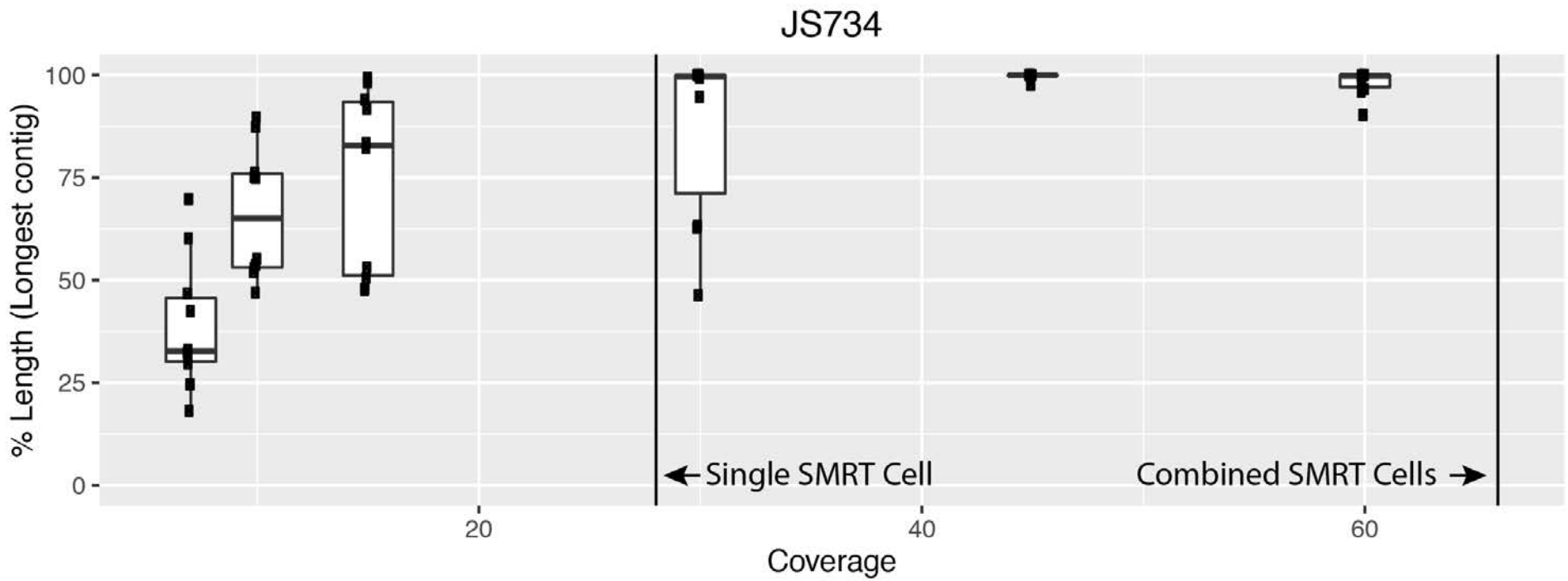


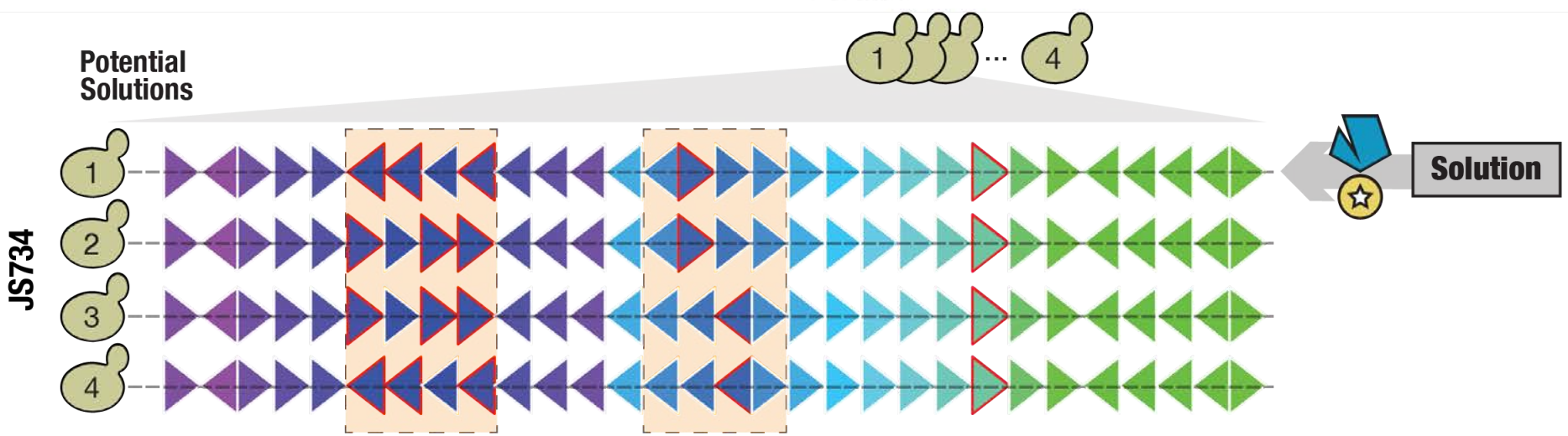
Near complete assembly of SCRaMbLE genomes with long-reads

JS734 could only be solved by aggregating data from several SMRT cells



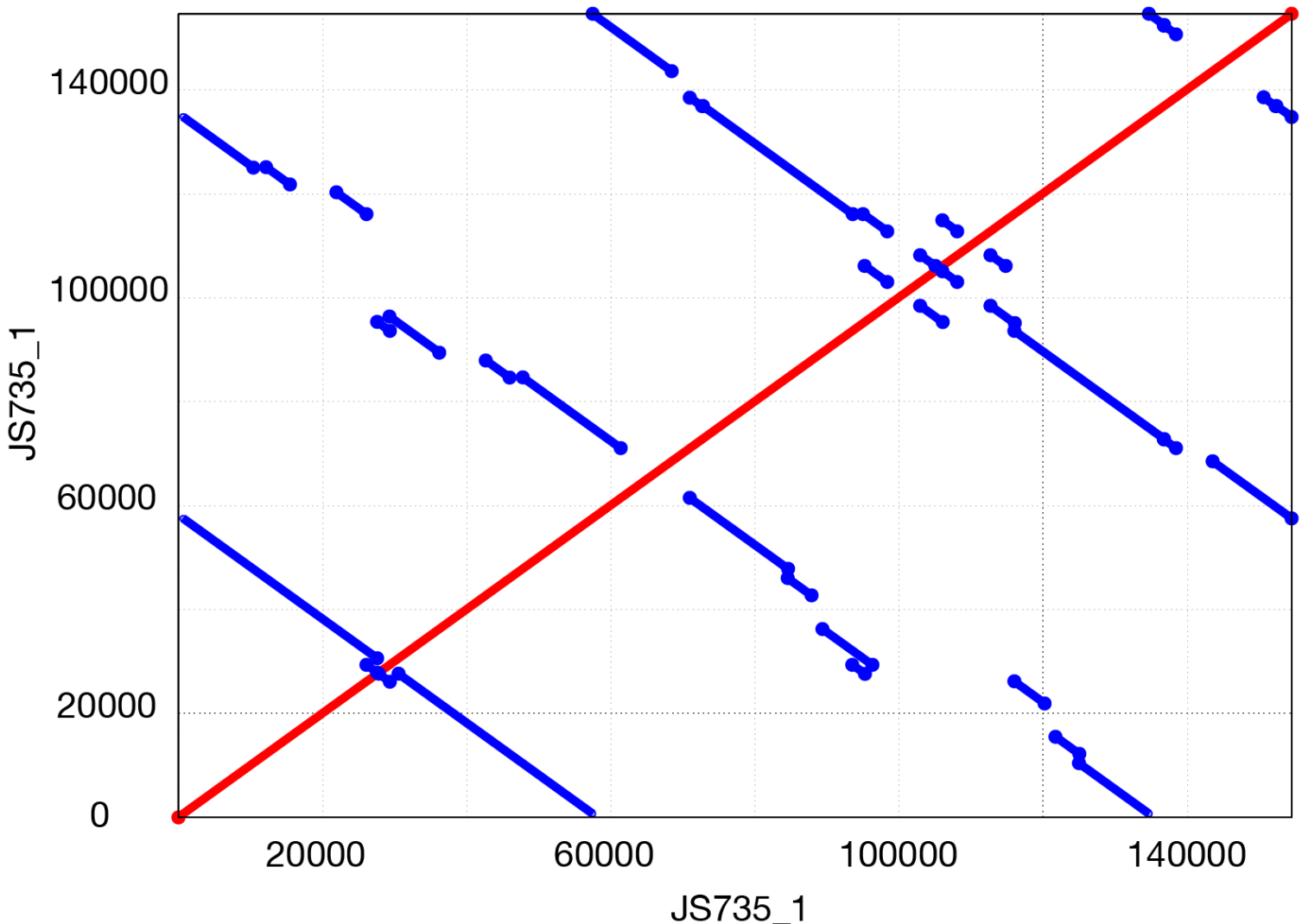
Complete assembly achieved frequently at high coverage for JS734



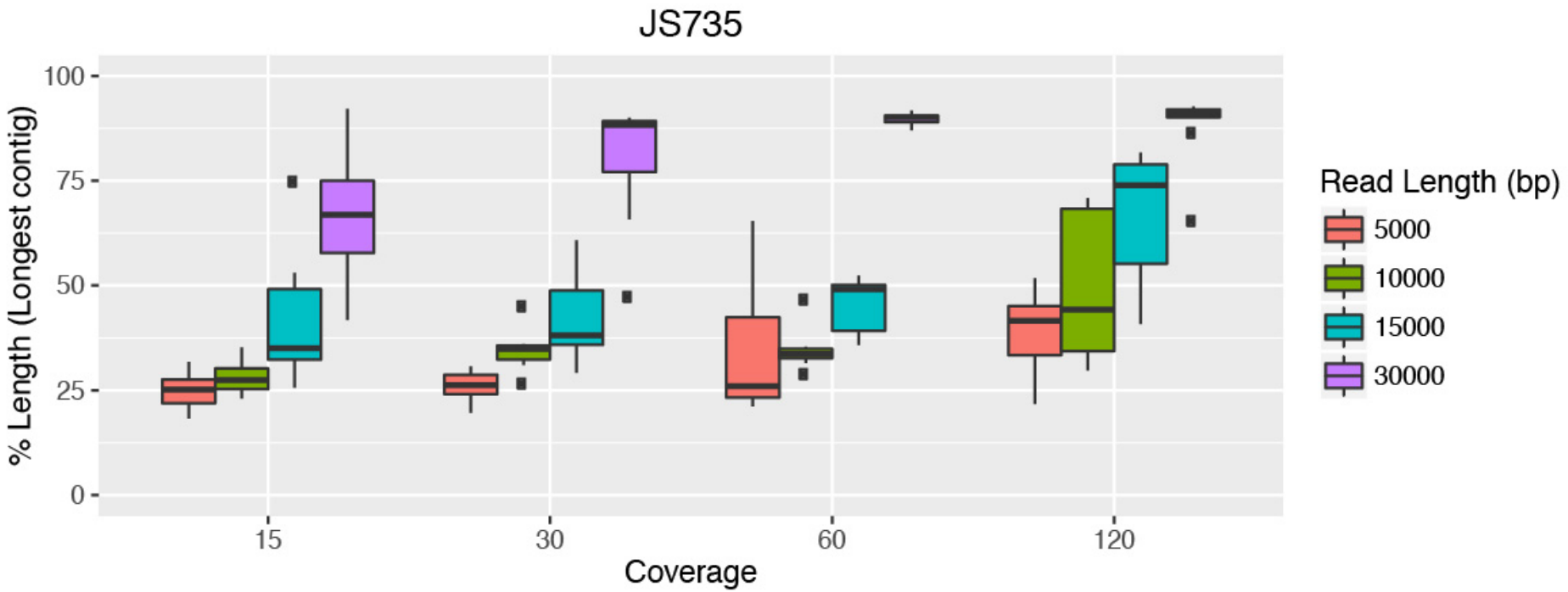


Some SCRaMbLE genomes are full of inverted repeats

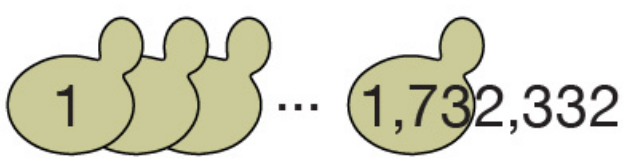
Will increased coverage still lead to complete assembly?



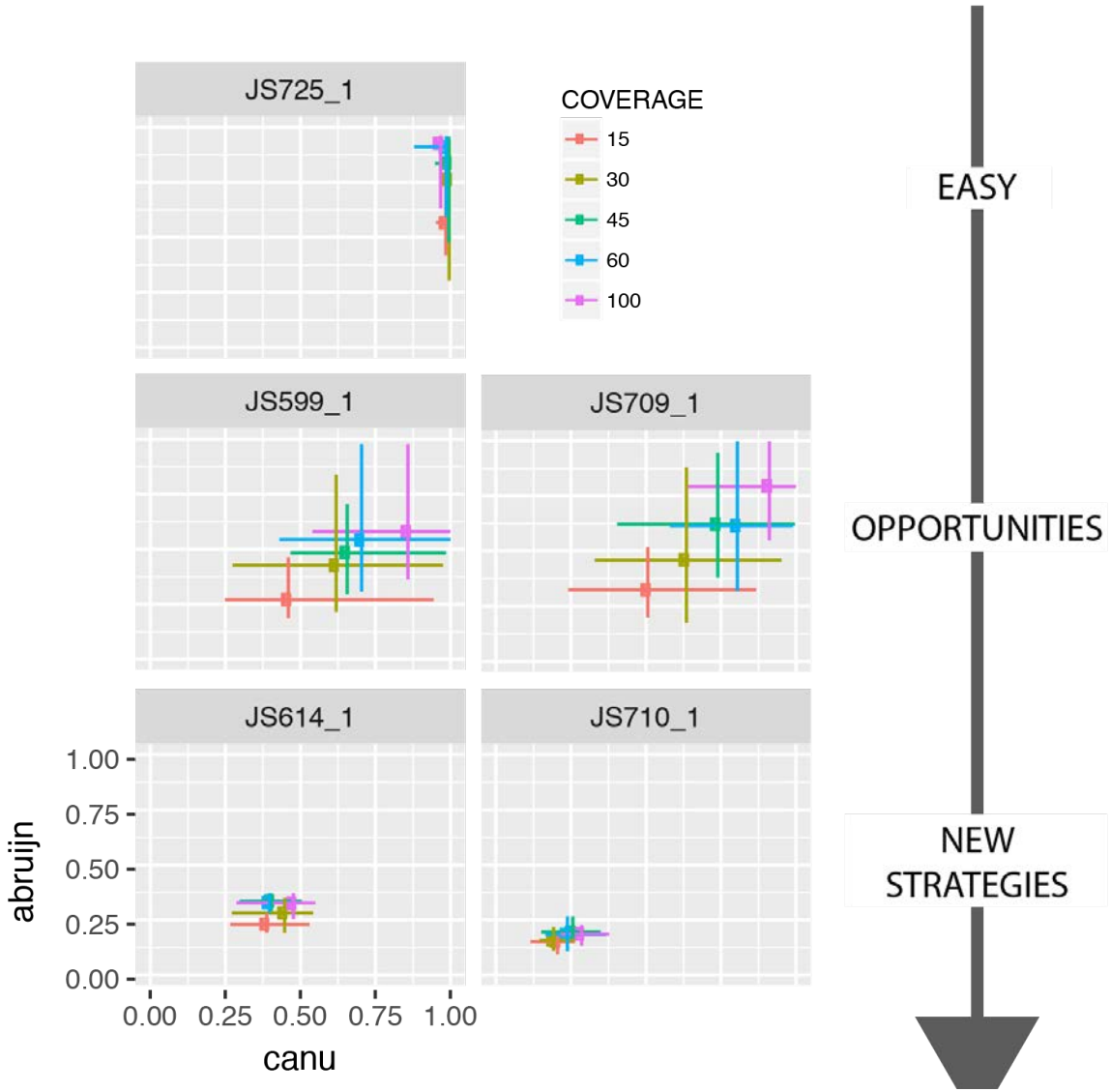
Some strains require a combination of depth and increased read length



Potential Solutions



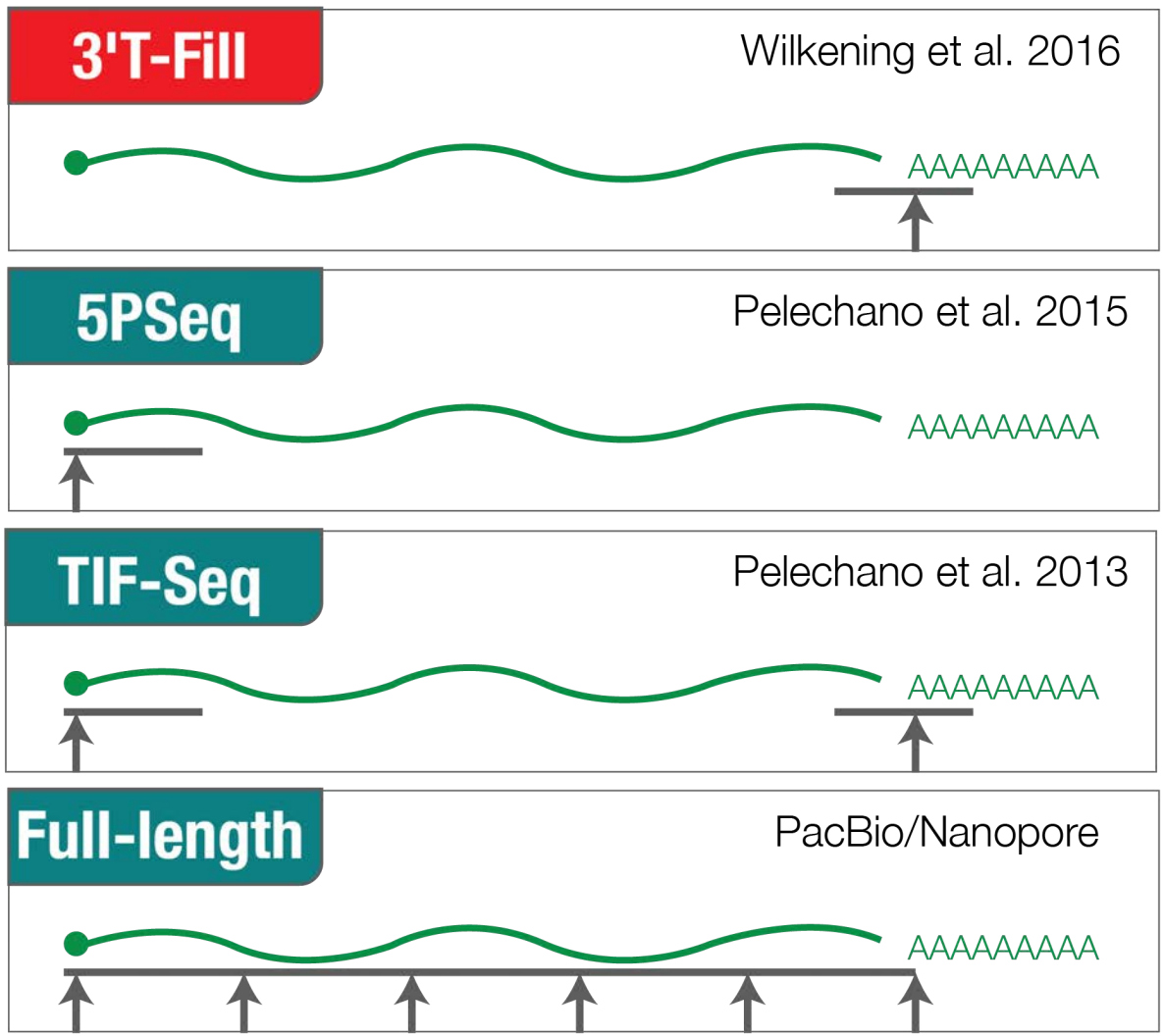
Variable difficulty of *de novo* assembly for SynIXR SCRaMbLEs



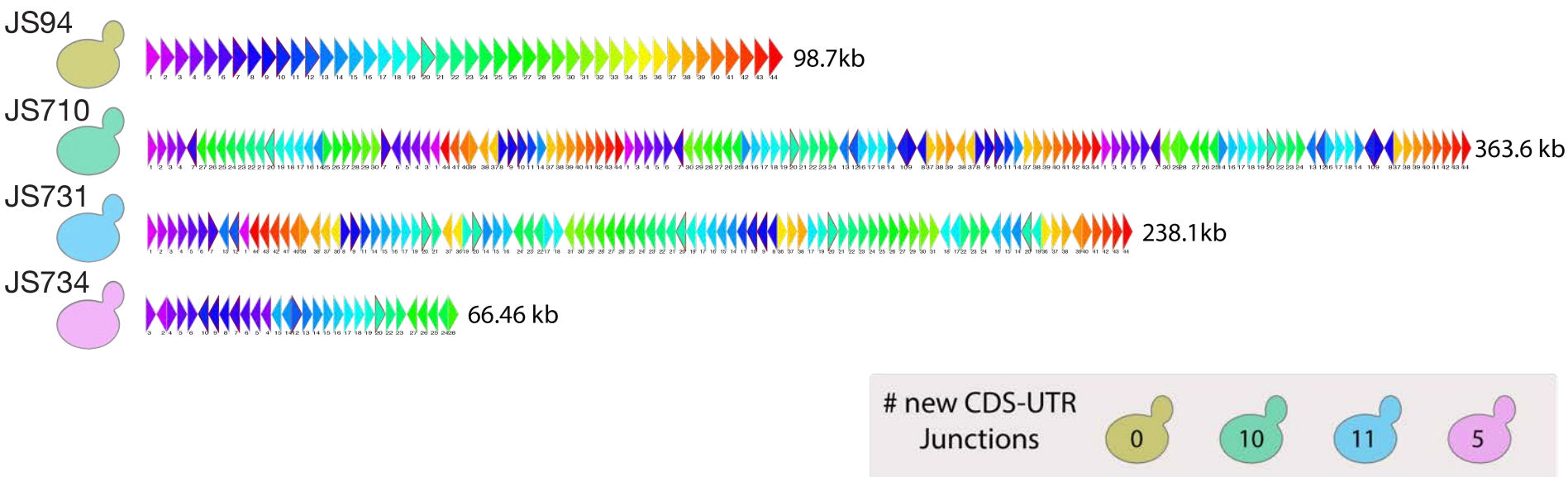
Fixed mean read length = 3.98 kb

How is a SCRaMbLEd genome transcribed?

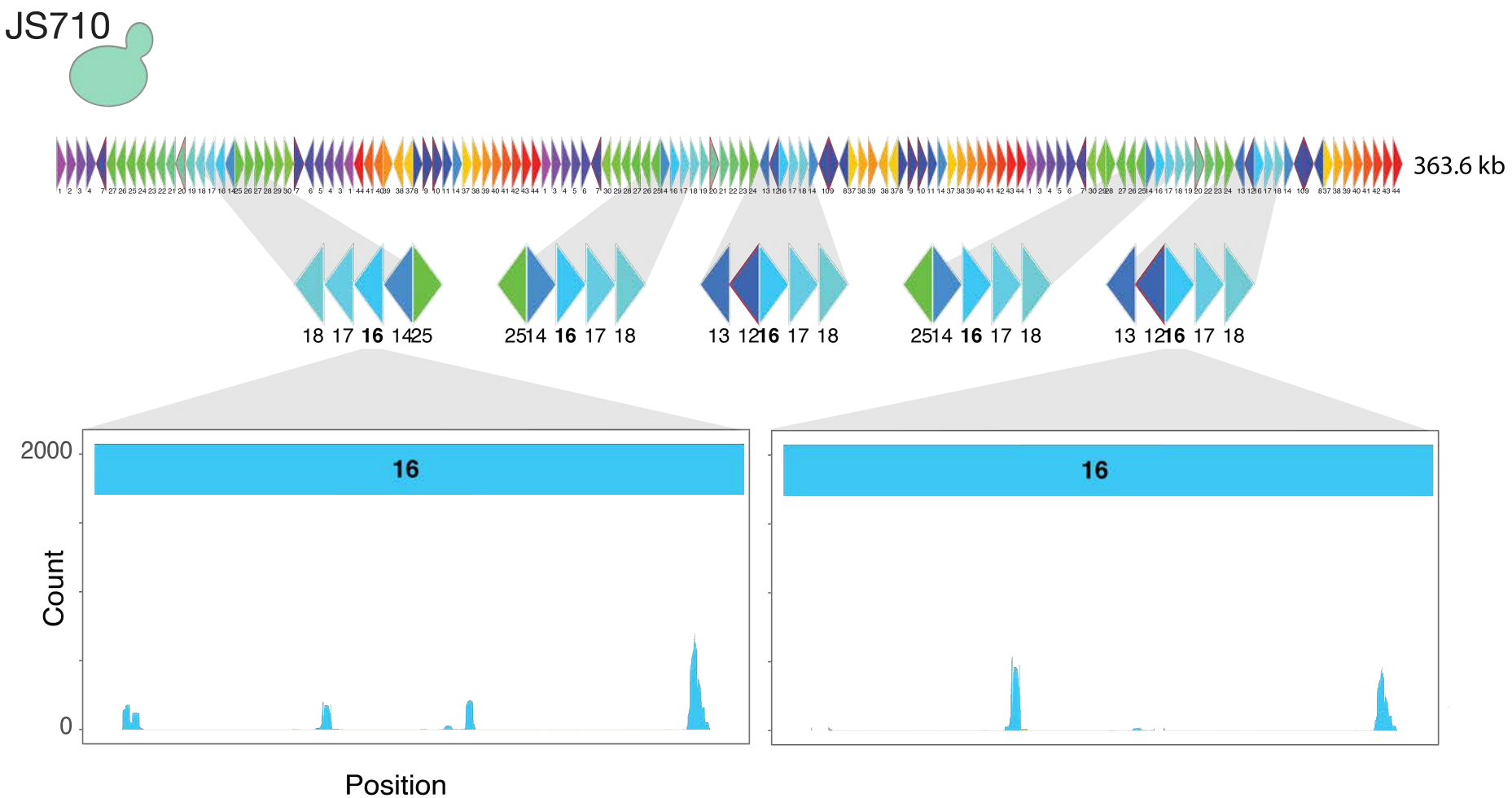
Sequencing methods to measure changes in the transcriptome



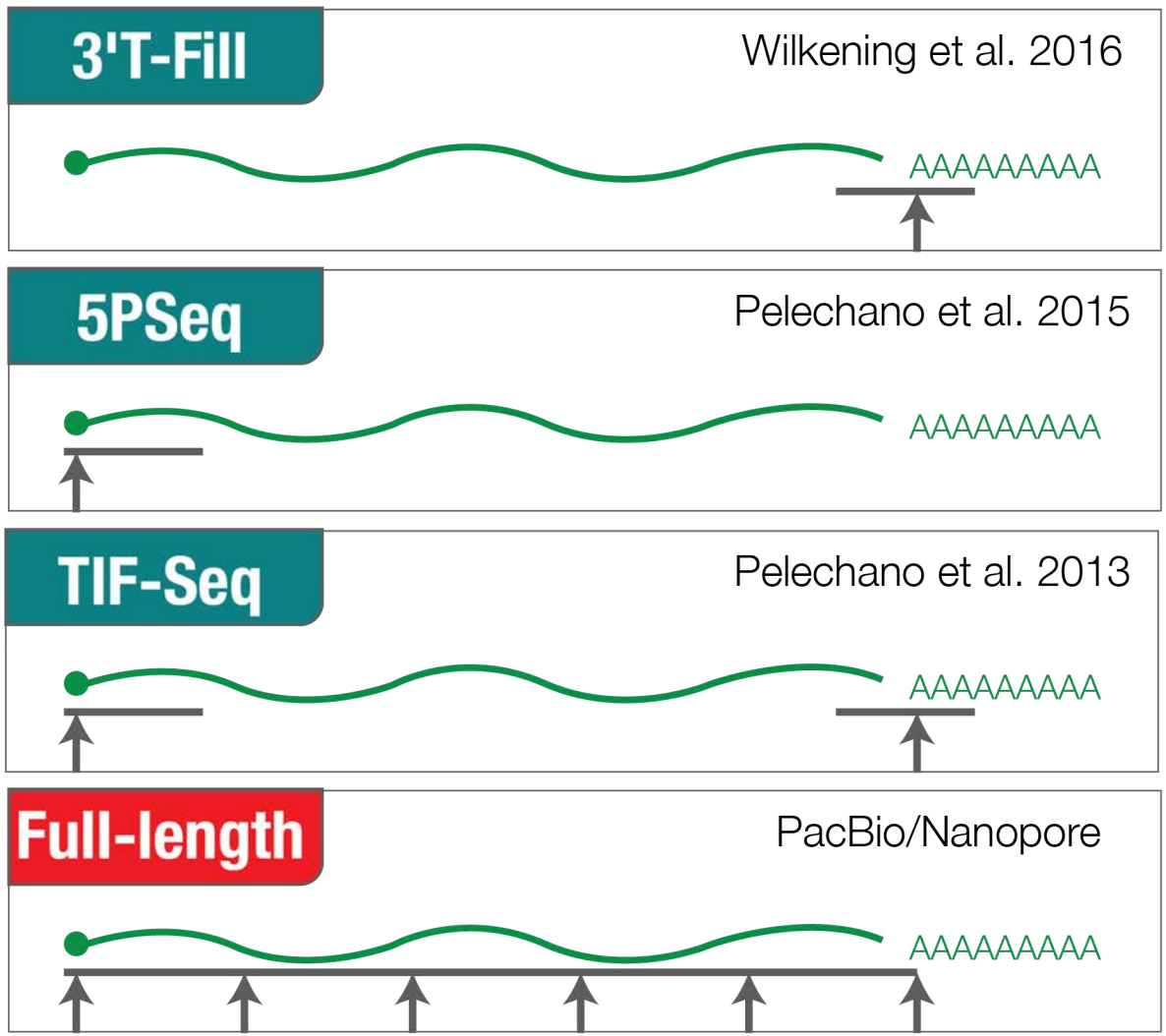
SynIXR SCRaMbLE Strains



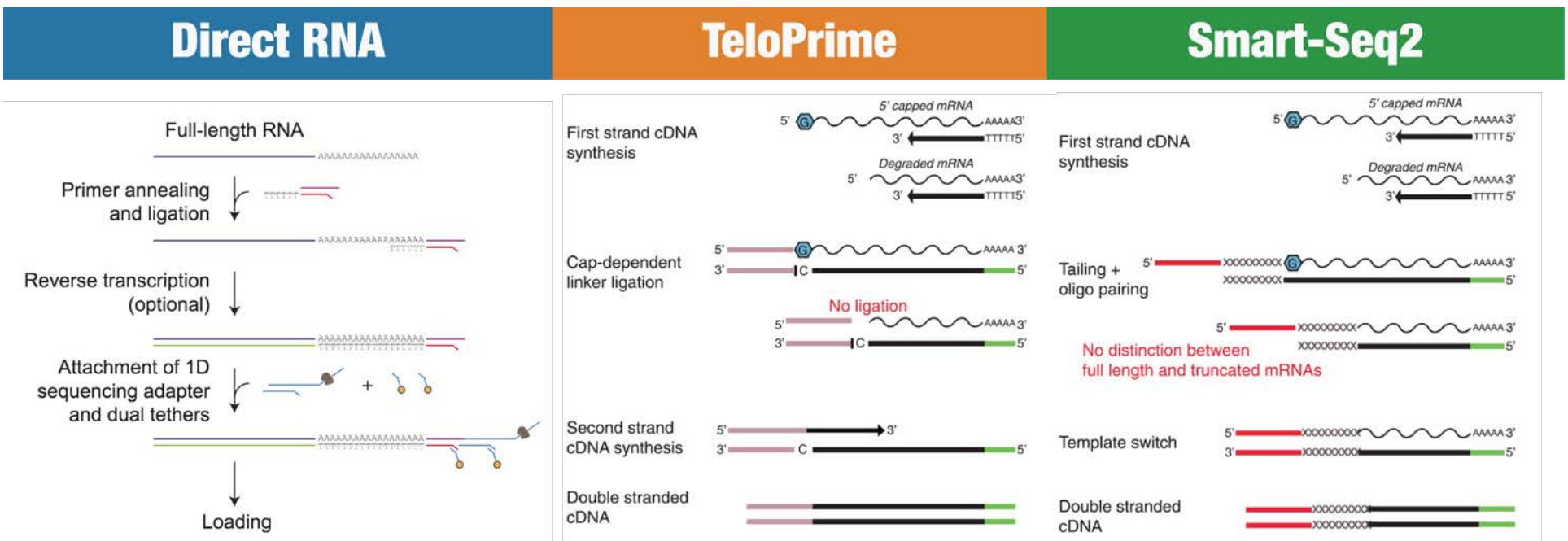
Mapping 3'-ends is problematic with short reads



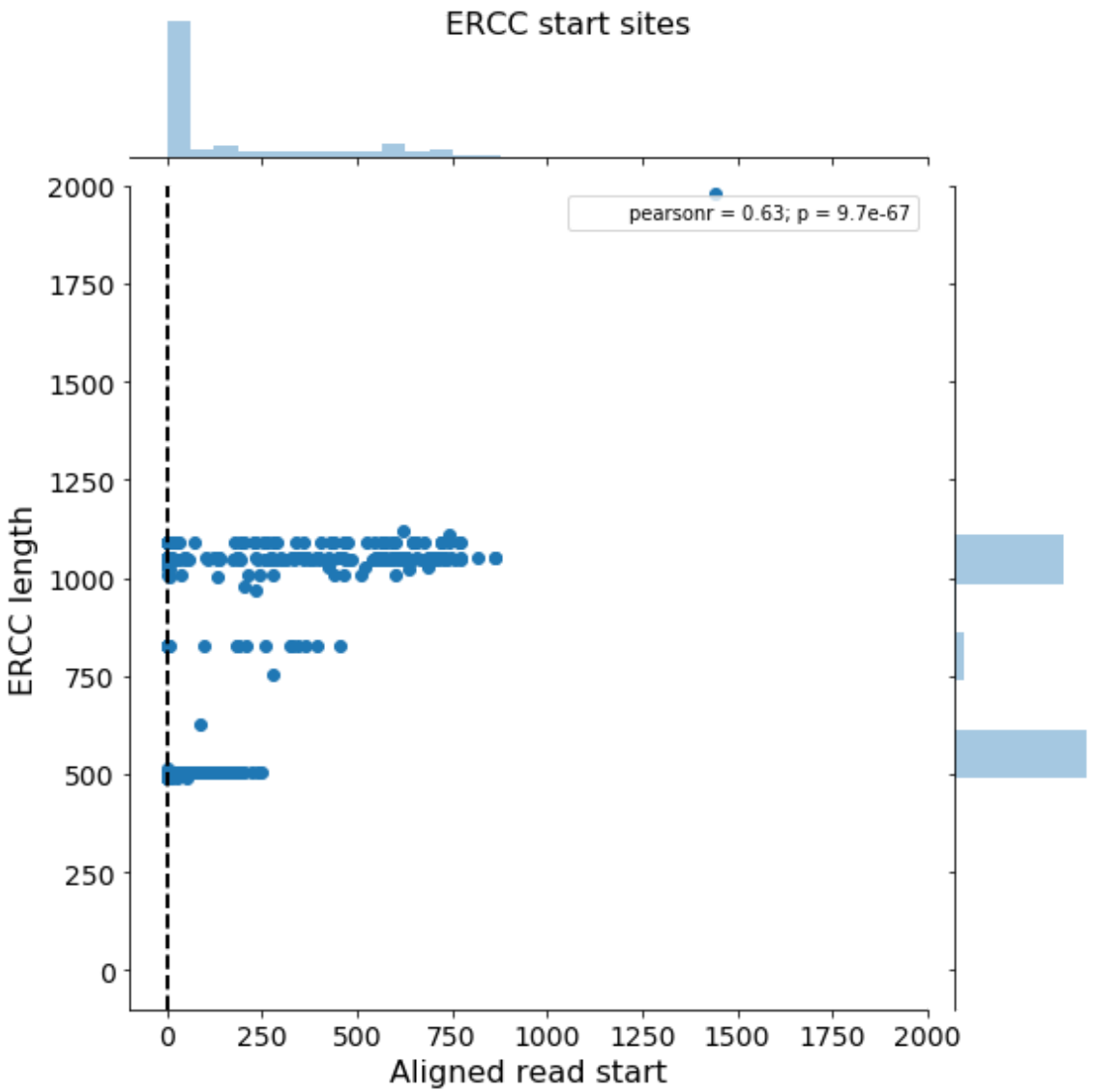
Sequencing methods to measure changes in the transcriptome



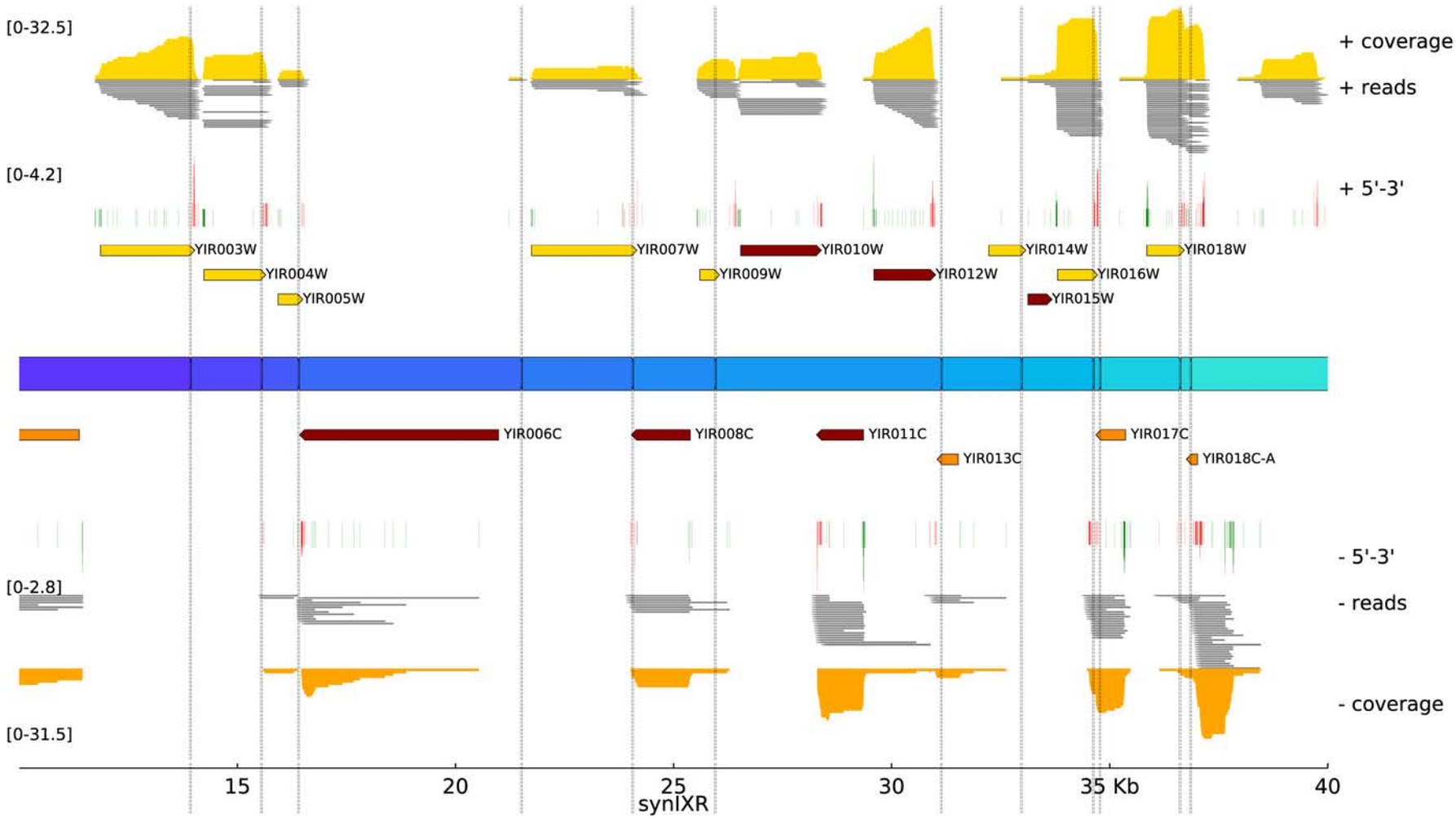
Multiple options for full-length RNA prep

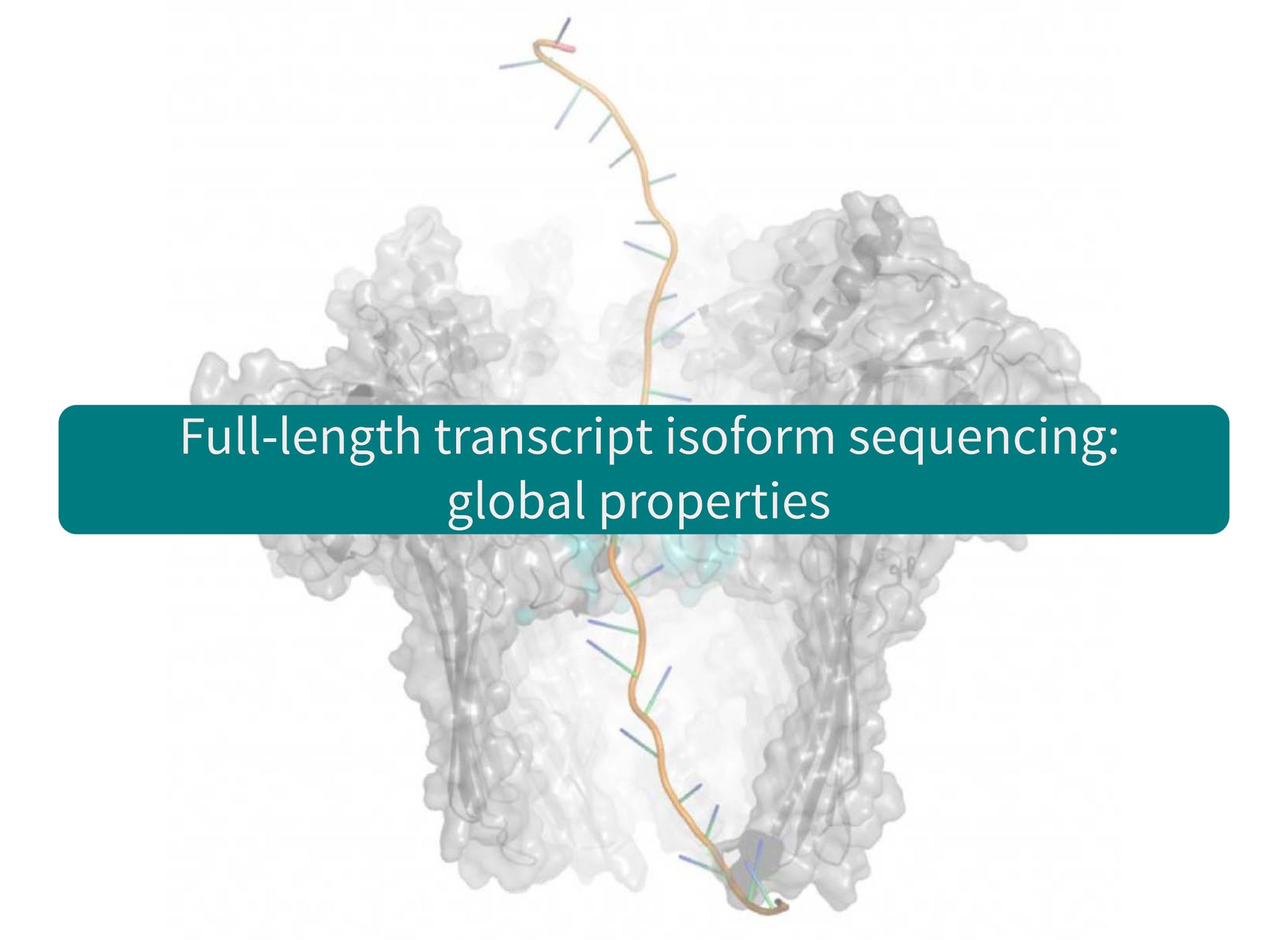


Are Smart-Seq2 reads not full-length?



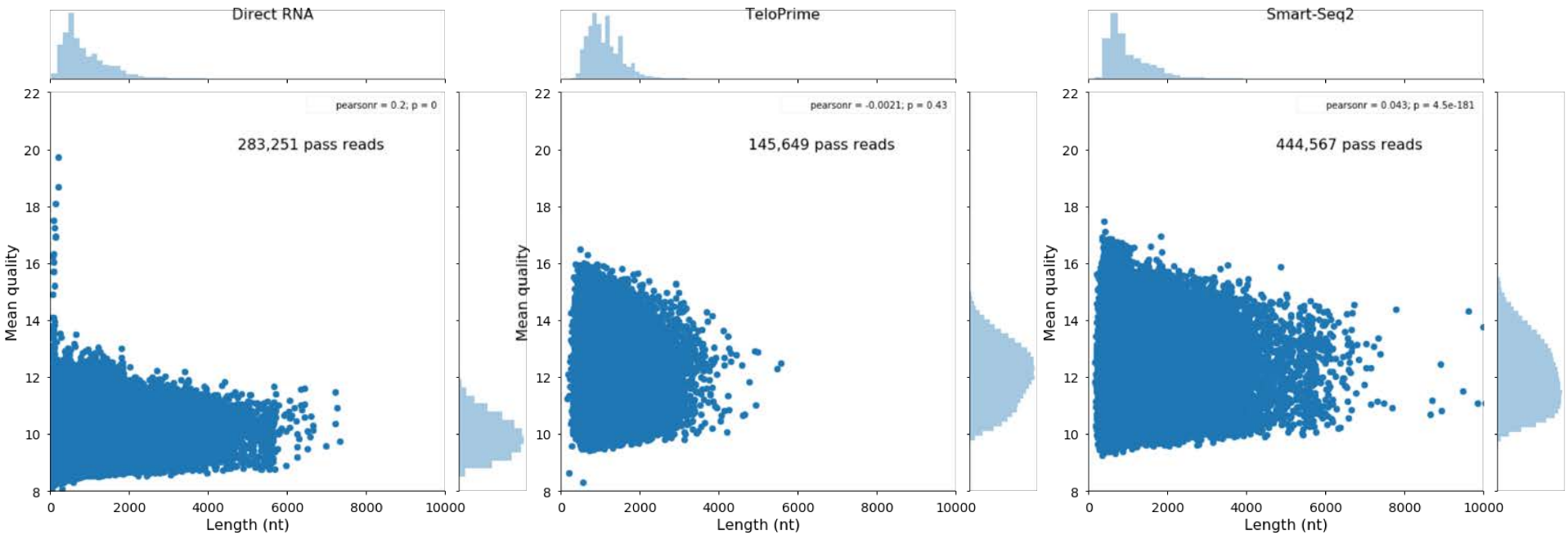
Strand-specificity with direct RNA sequencing



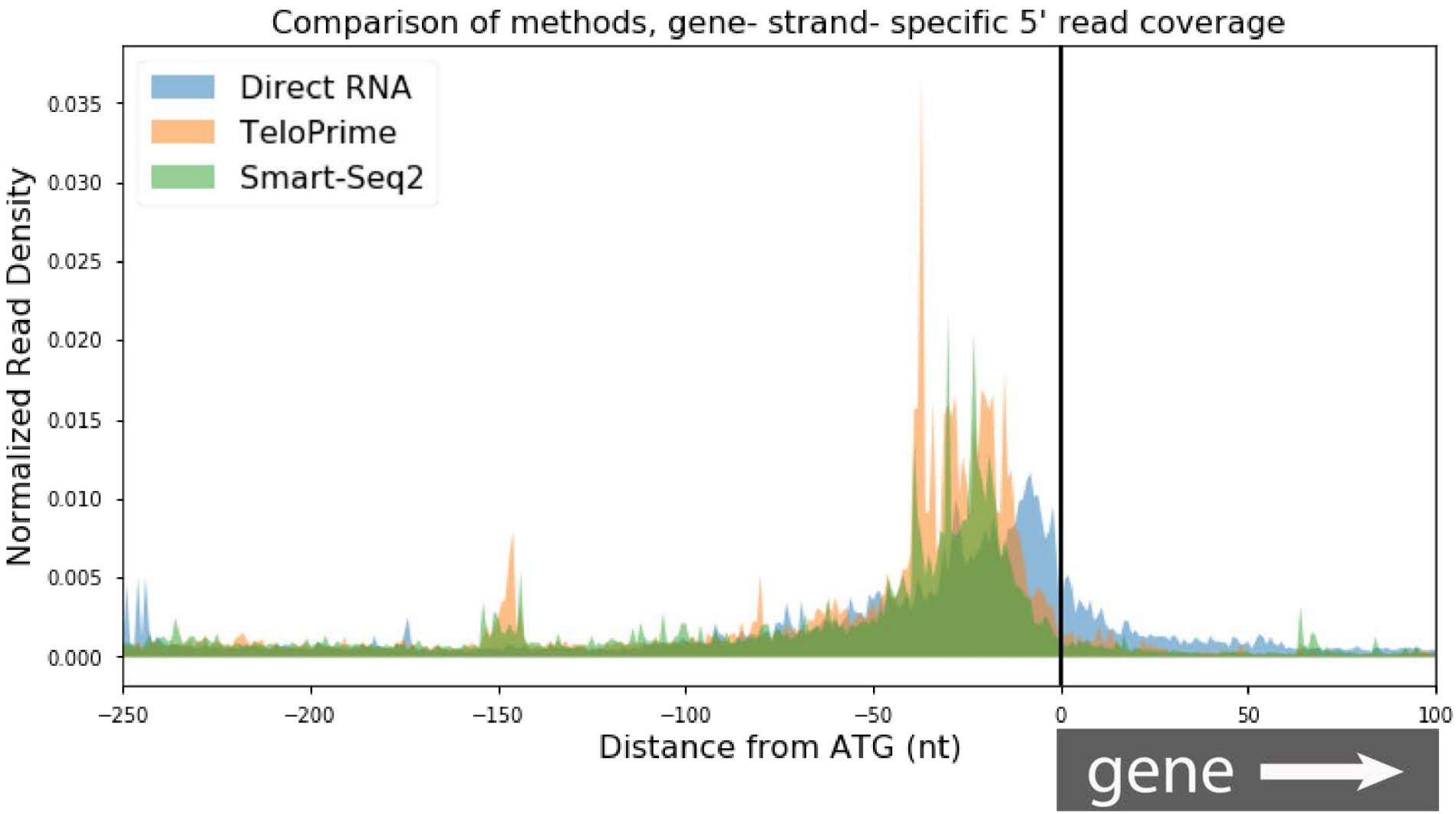


Full-length transcript isoform sequencing: global properties

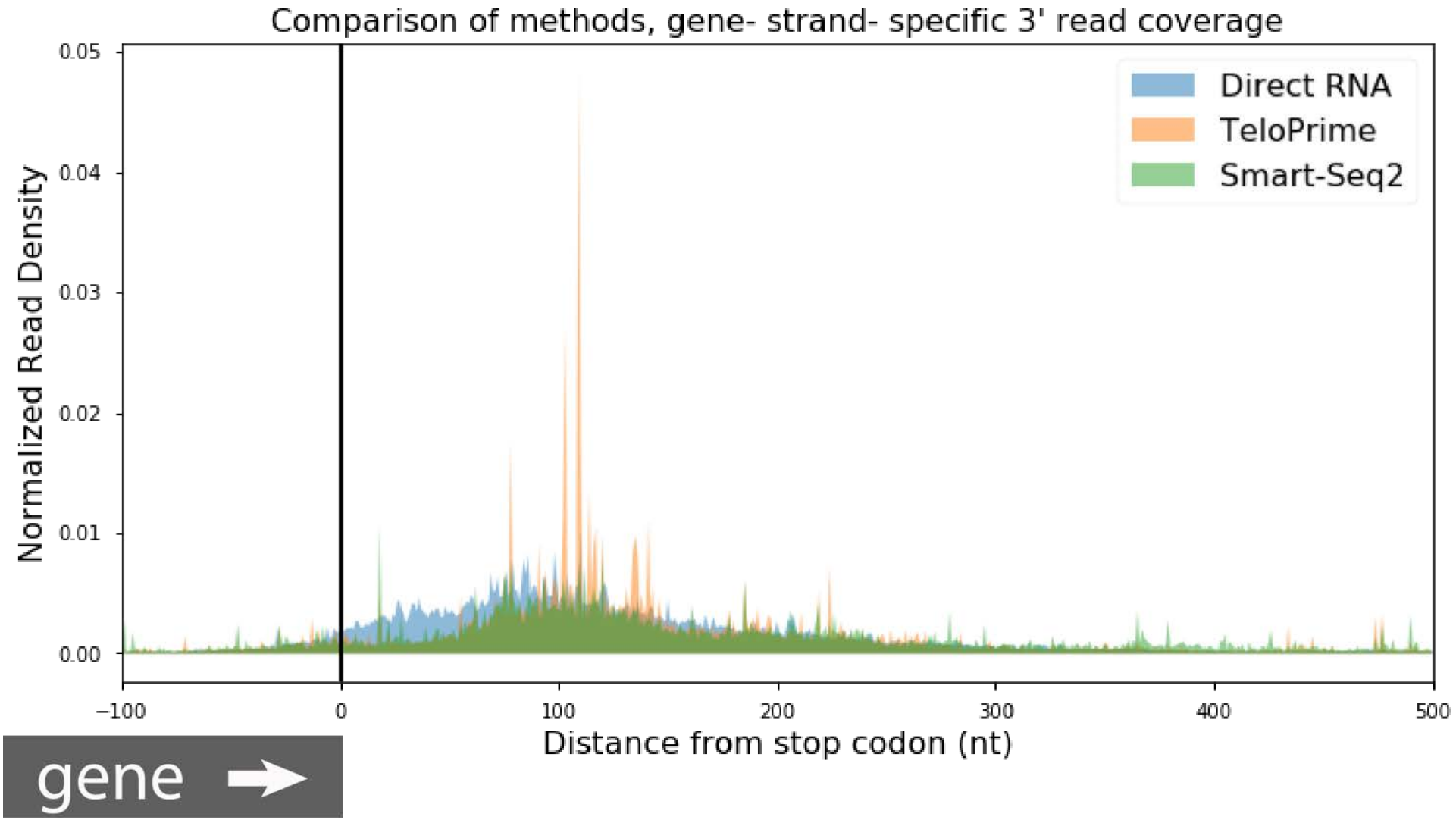
Small differences in read lengths and quality



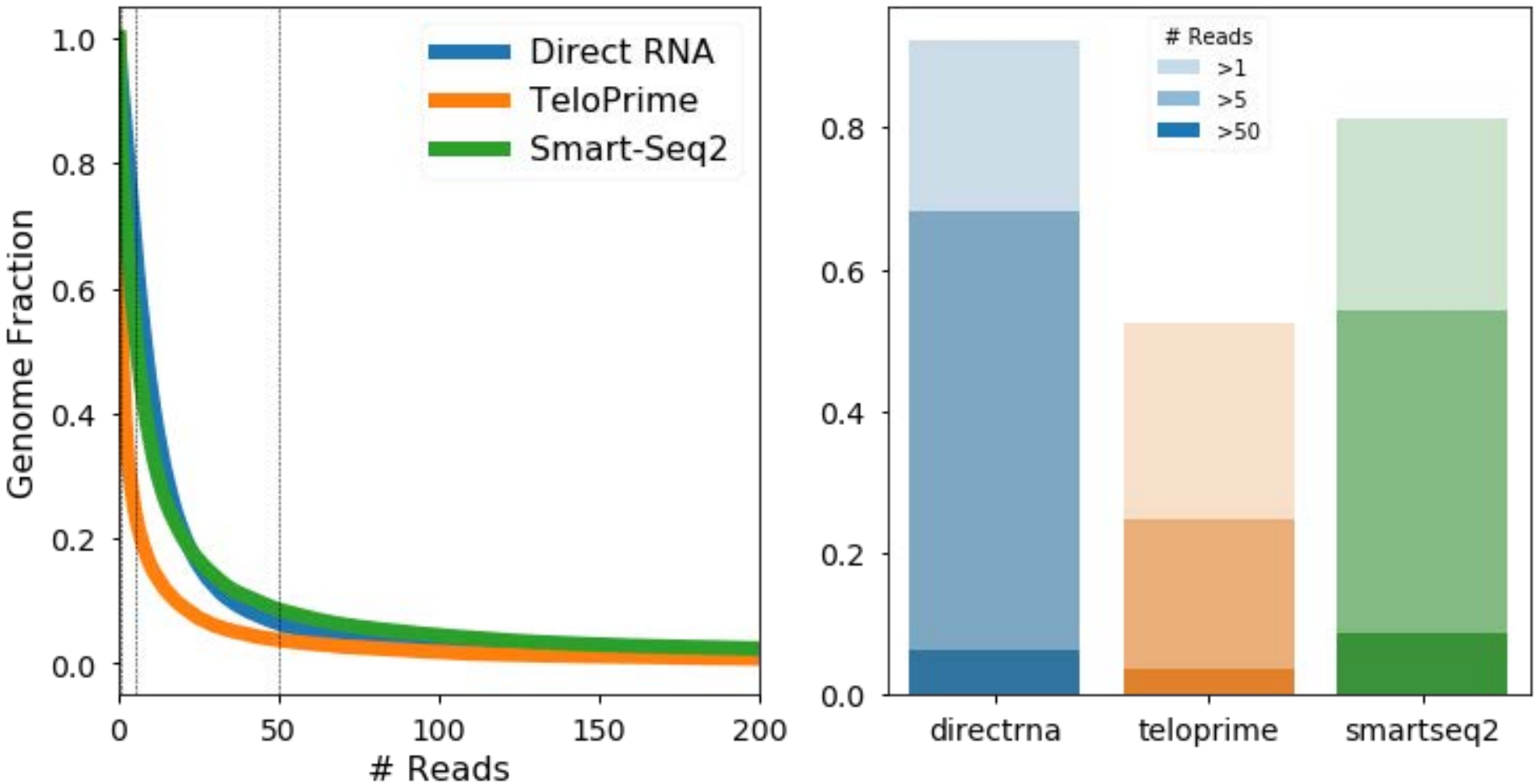
Global genic 5'-end profile



Global genic 3'-end profile



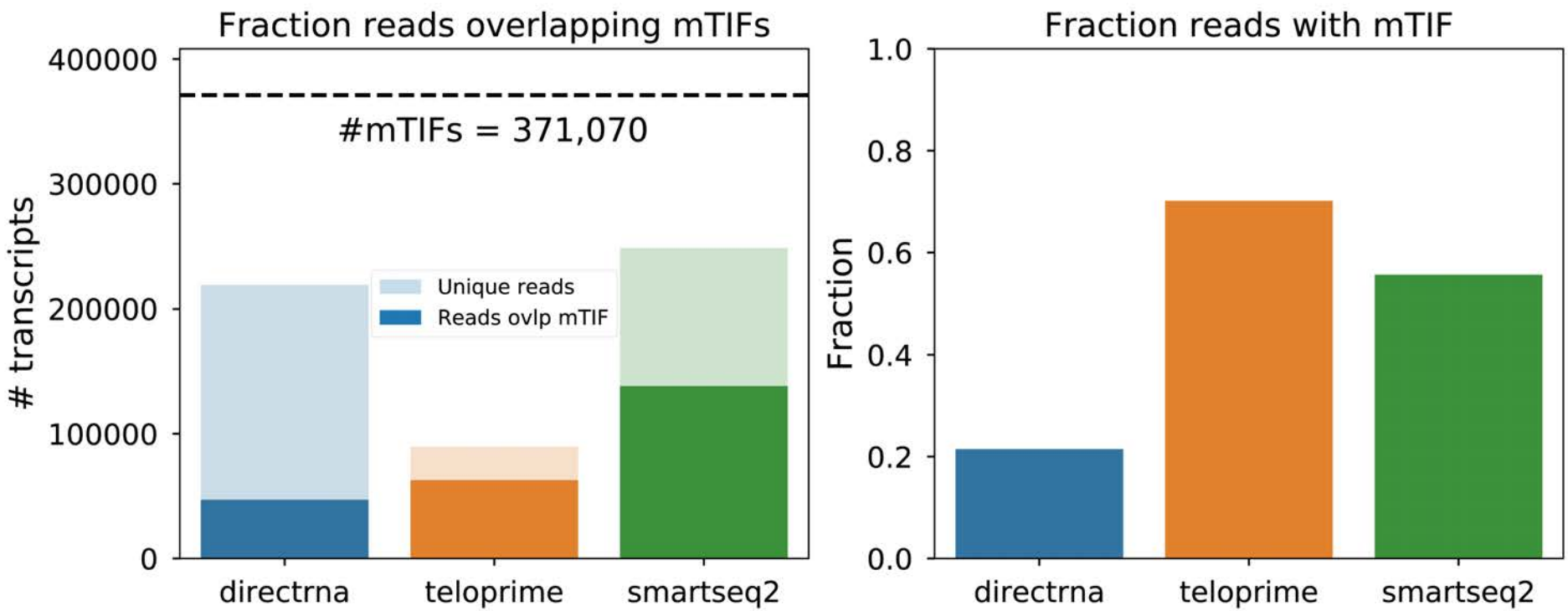
What fraction of the genome is expressed?



"A total of 85% of the genome is expressed in rich media."

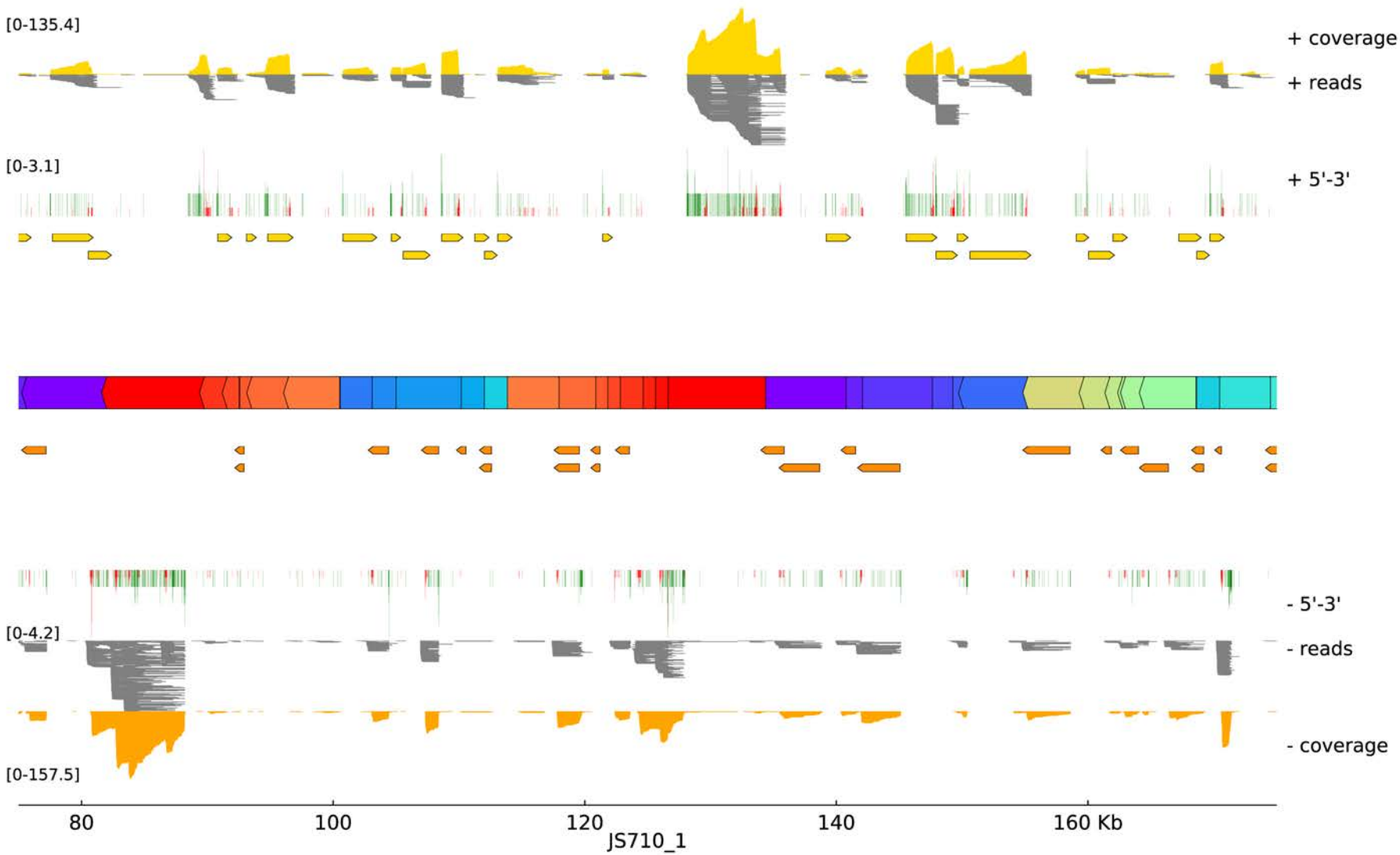
David et al. 2006

Rediscovery of previously described transcript isoforms



Pelechano, Wei and Steinmetz 2013

New transcriptional landscape in SCRaMbLEd cells



Summary

- **Direct RNA** sequences "everything", is inherently strand-specific, no PCR
- **TeloPrime** is specific for capped molecules, protocol has to be modified to achieve Nanopore input requirements
- **Smart-Seq2** can amplify degradation products and possibly other non-full-length transcripts, but...
- We can use dual UMIs with Smart-Seq2 to identify PCR duplicates