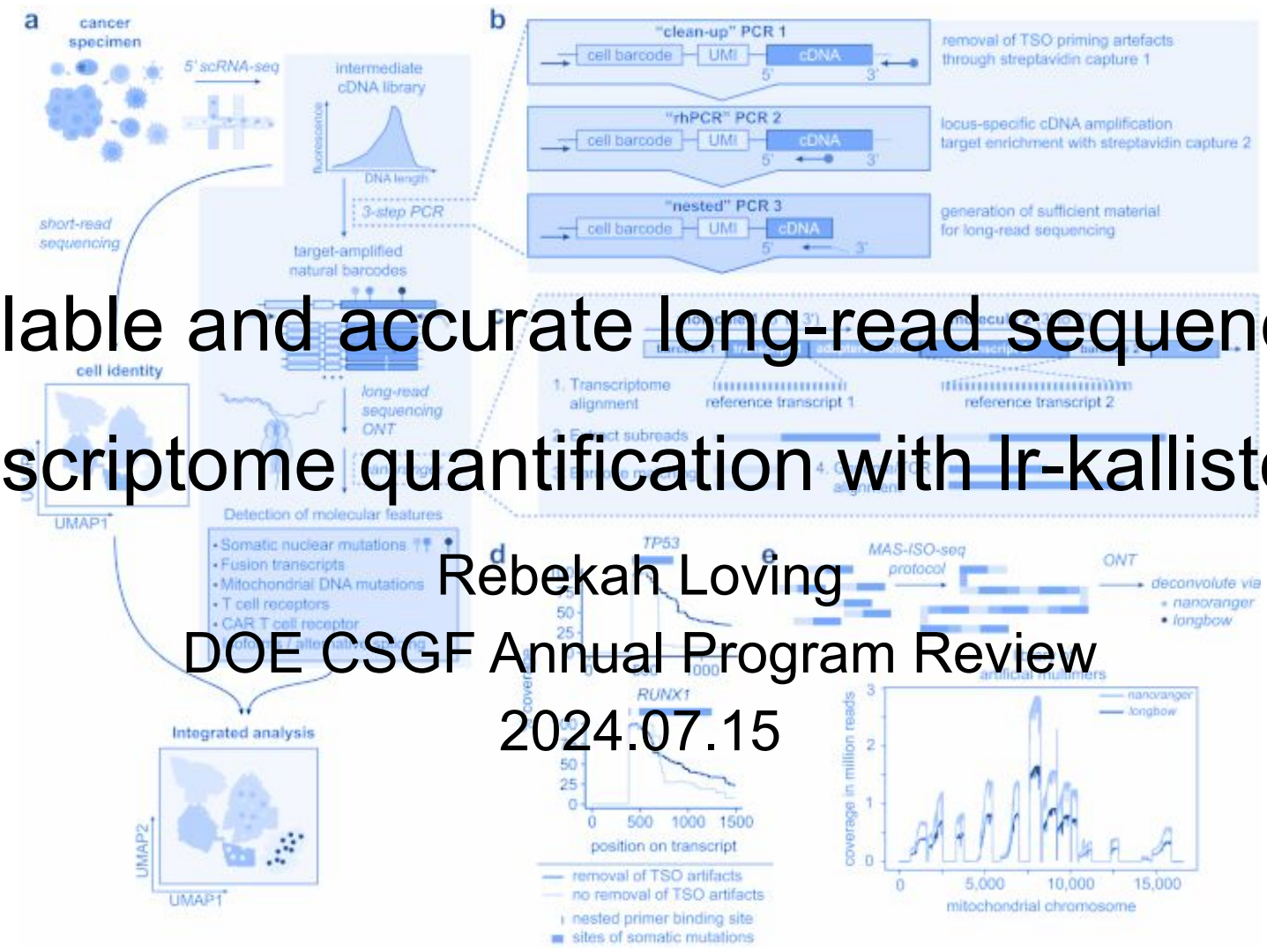


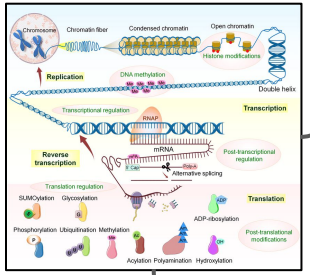
Scalable and accurate long-read sequencing transcriptome quantification with lr-kallisto

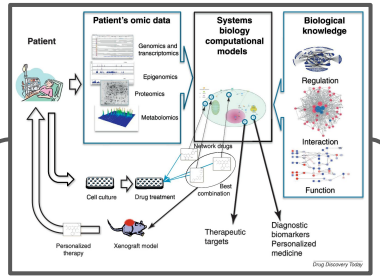
Rebekah Loving

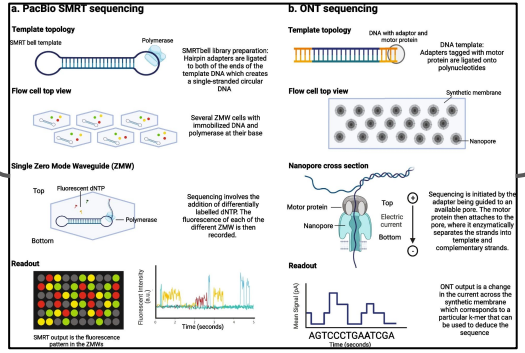
DOE CSGF Annual Program Review

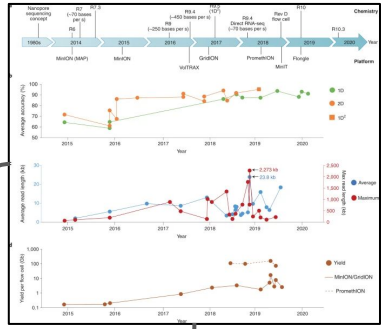
2024.07.15

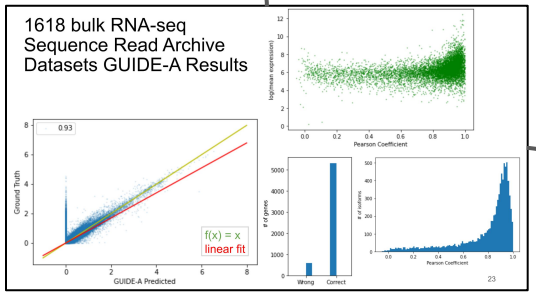
Roadmap

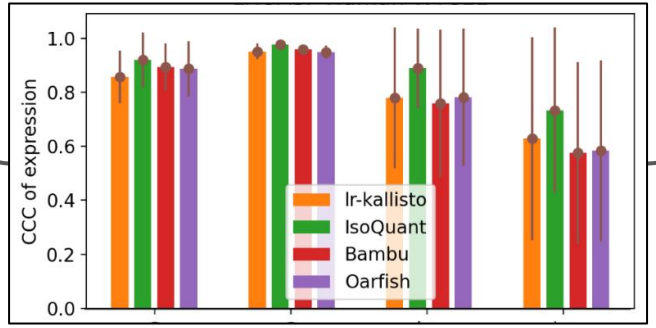

Chromosome
Chromatin fiber
Condensed chromatin
Open chromatin
Histone modifications
Replication
DNA methylation
Double helix
Transcription
Transcriptional regulation
RNAP
mRNA
Reverse transcription
Post-transcriptional regulation
Alternative splicing
Translation regulation
SUMOylation
Glycosylation
Translation
ADP-ribosylation
Phosphorylation
Ubiquitination
Methylation
Post-translational modifications
Acylation
Polyamination
Hydroxylation

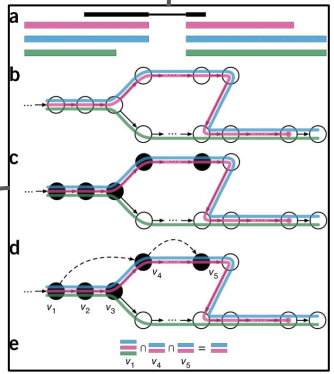

Patient
Patient's omic data
Genomics and transcriptomics
Epigenomics
Proteomics
Metabolomics
Systems biology computational models
Network design
Cell culture
Drug treatment
Best combination
Biological knowledge
Regulation
Interaction
Function
Personalized therapy
Xenograft model
Therapeutic targets
Diagnostic biomarkers
Personalized medicine
Drug Discovery Today


a. PacBio SMRT sequencing
Template topology
SMRT bell template
Polymerase
SMRTbell library preparation: Hairpin adapters are ligated to both of the ends of the template DNA which creates a single-stranded circular DNA
Flow cell top view
Several ZMW cells with immobilized DNA and polymerase at their base
Single Zero Mode Waveguide (ZMW)
Top
Fluorescent dNTP
Polymerase
Bottom
Sequencing involves the addition of differentially labelled dNTP. The fluorescence of each of the different ZMW is then recorded.
Readout
Fluorescent Intensity (au)
Time (seconds)
SMRT output is the fluorescence pattern in the ZMWs
b. ONT sequencing
Template topology
DNA with adaptor and motor protein
DNA template: Adapters tagged with motor protein are ligated onto polynucleotides
Flow cell top view
Synthetic membrane
Nanopore
Nanopore cross section
Motor protein
Nanopore
Top
Electric current
Bottom
Sequencing is initiated by the adapter being guided to an available pore. The motor protein then attaches to the pore, where it enzymatically separates the strands into template and complementary strands.
Readout
Mean Signal (pA)
AGTCCCTGAATCGA
Time (seconds)
ONT output is a change in the current across the synthetic membrane which corresponds to a particular k-mer that can be used to deduce the sequence

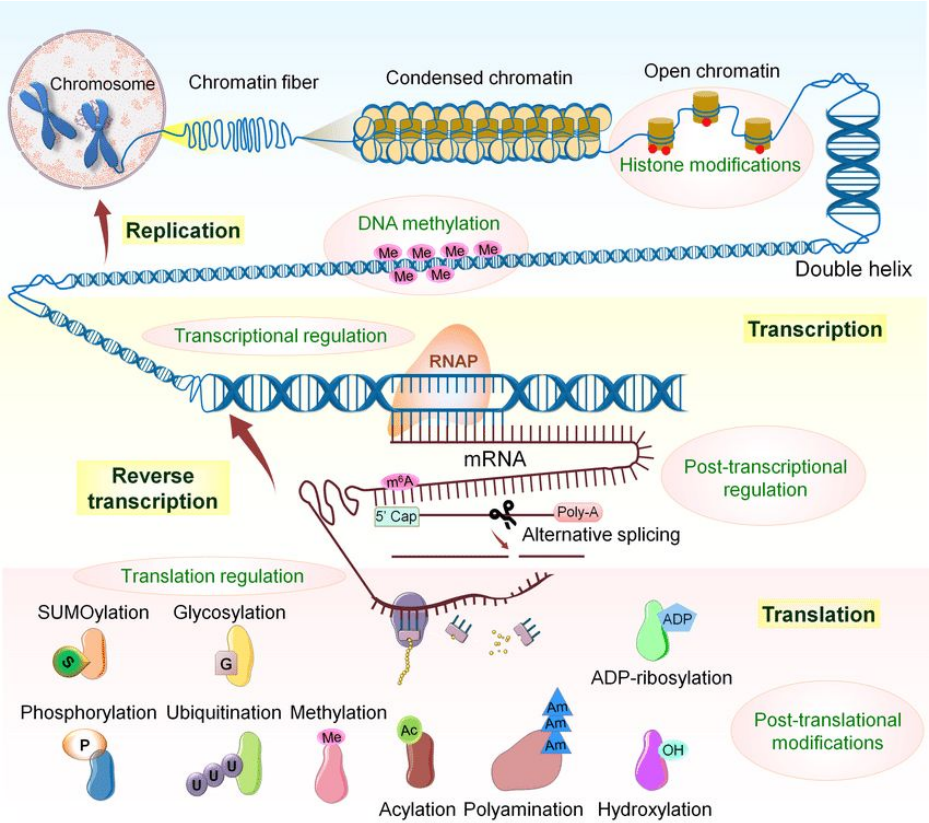

Nanopore sequencing concept
Chemistry
Platform
1980s
2014
2015
2016
2017
2018
2019
2020
Year
MinION (MAP)
MinION
GridION
PromethION
MinIT
VolTRAX
Flongle
Average accuracy (%)
Average read length (kb)
Yield per flow cell (Gb)
Yield
MinION/GridION
PromethION


1618 bulk RNA-seq Sequence Read Archive Datasets GUIDE-A Results
0.93
f(x) = x
linear fit
Ground Truth
GUIDE-A Predicted
highest expression
Pearson Coefficient
# of genes
Wrong
Correct
# of isoforms
23


CCC of expression
lr-kallisto
IsoQuant
Bambu
Oarfish

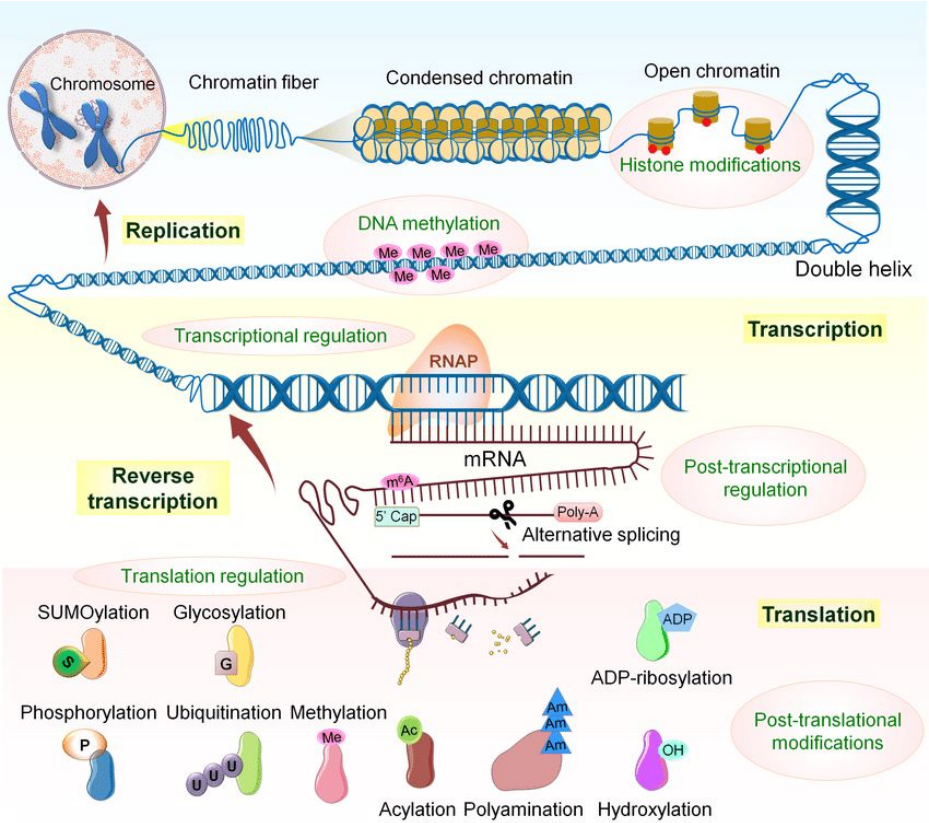

a
b
c
d
e
v_1
v_2
v_3
v_4
v_5

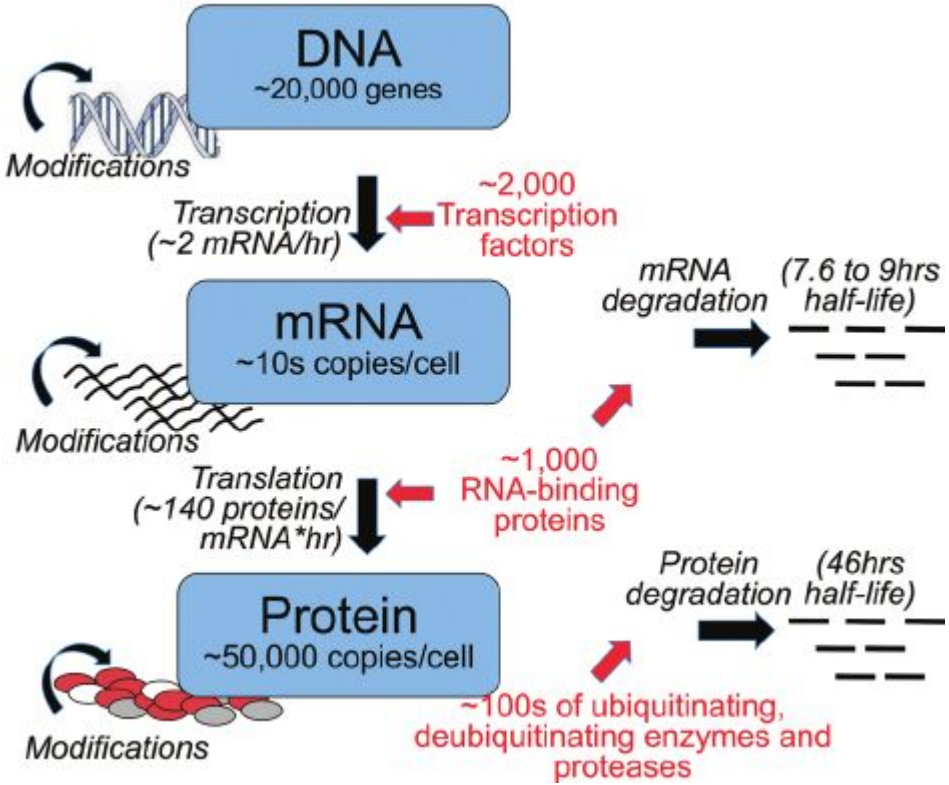
Central Dogma



CRISPR-Mediated Engineering across the Central Dogma in Plant Biology for Basic Research and Crop Improvement - Scientific Figure on ResearchGate.

Central Dogma

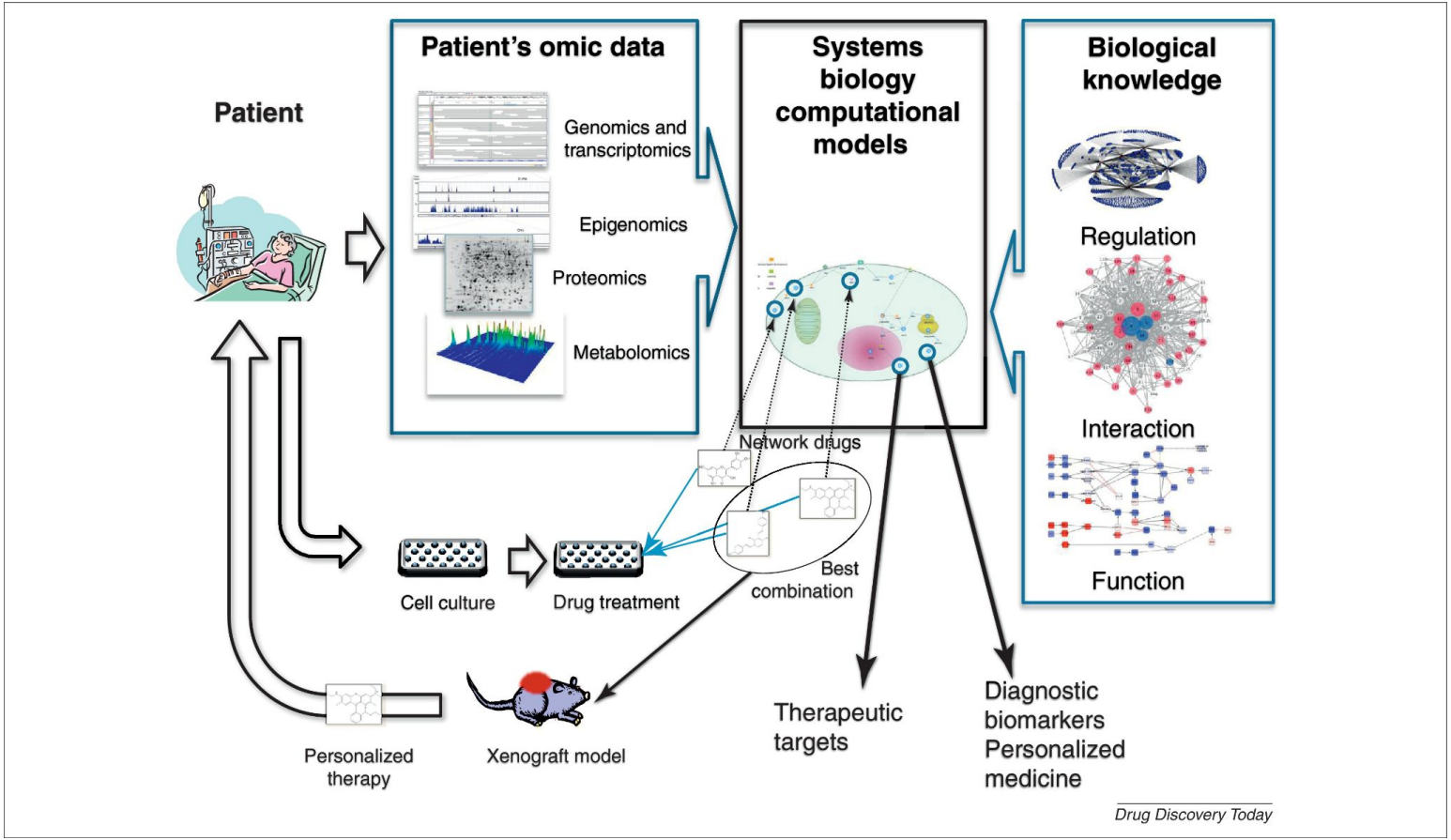




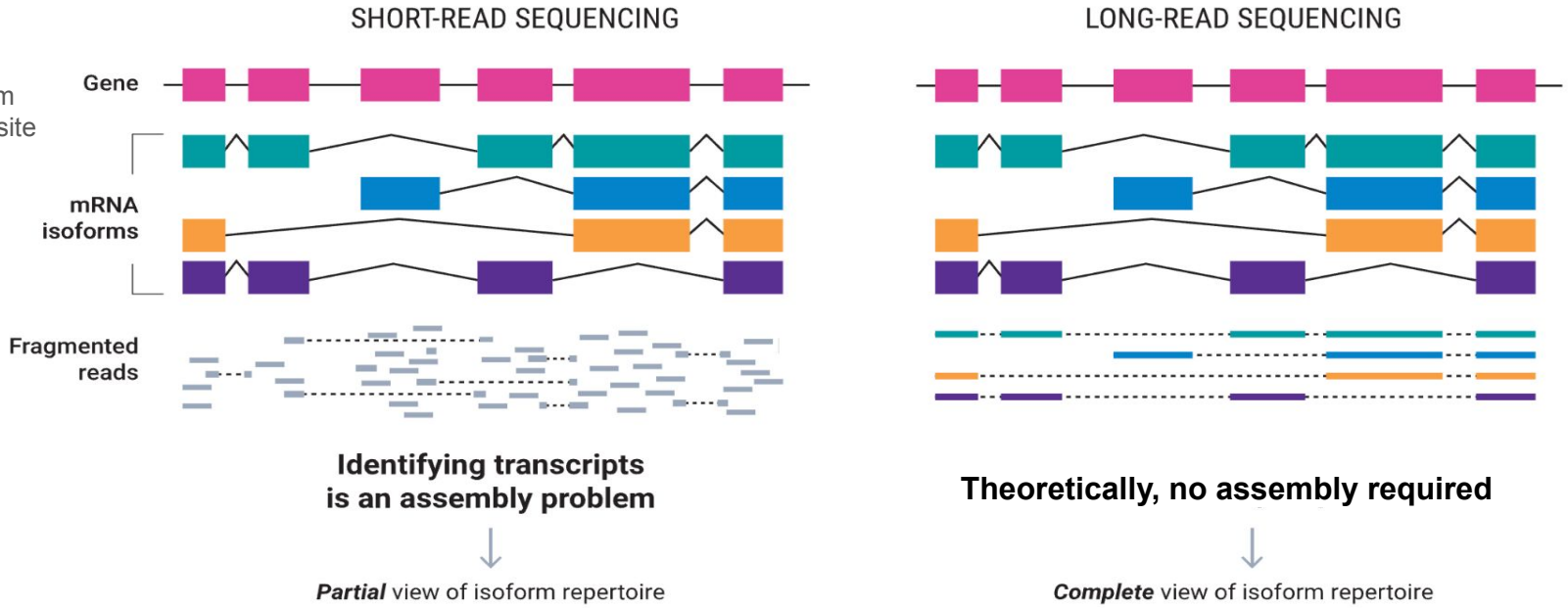
CRISPR-Mediated Engineering across the Central Dogma in Plant Biology for Basic Research and Crop Improvement - Scientific Figure on ResearchGate.

Next-generation analysis of gene expression regulation – comparing the roles of synthesis and degradation

Studying mRNA impact on drug discovery



Adapted from PacBio website

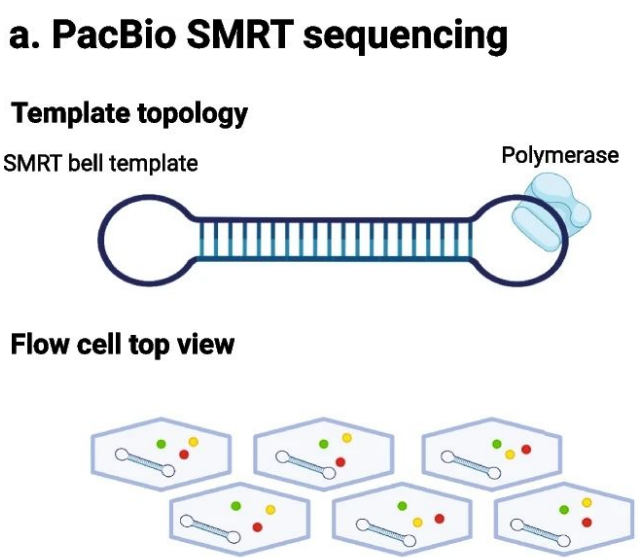


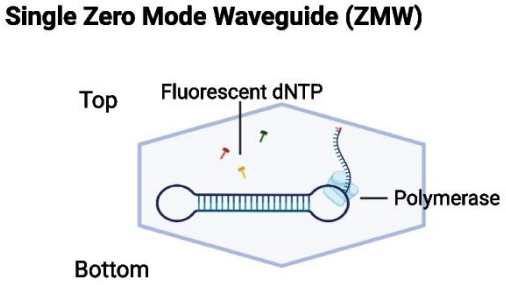
Long reads allow you to identify:

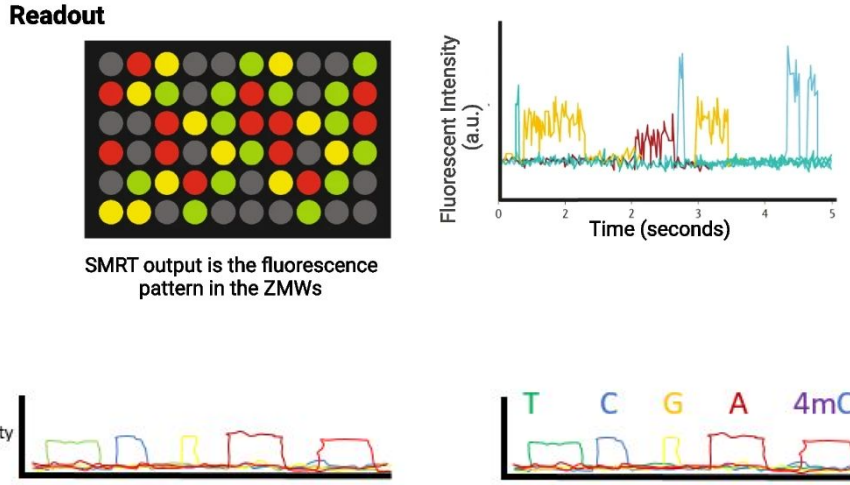
- Isoform resolution
- Improved annotations
- Gene fusions

New long read technology reduces error rate to .1-1.5%

PacBio's innovation of Circular Consensus Sequencing produced an error rate .1-.5%

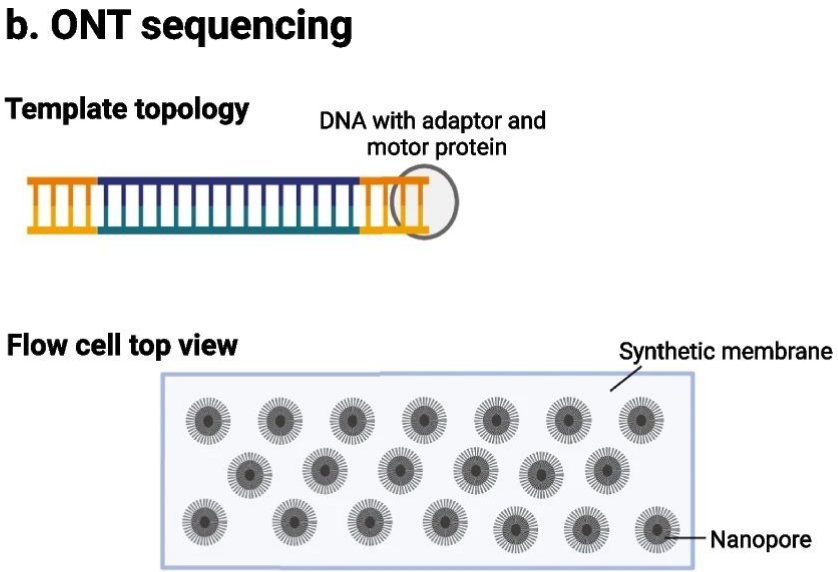


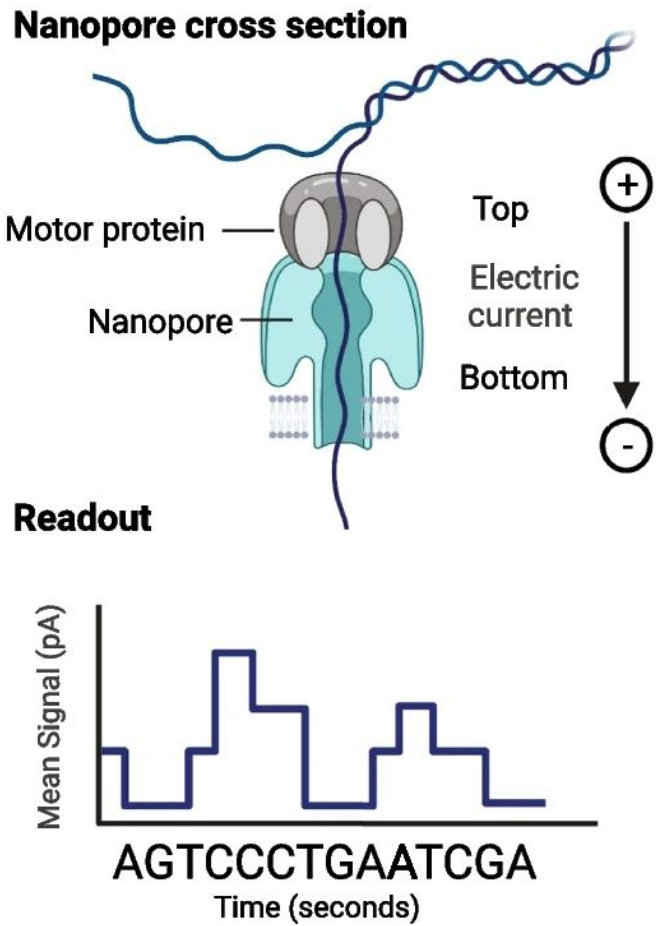




Adapted from PacBio website and Oehler, J.B., Wright, H., Stark, Z. *et al.* The application of long-read sequencing in clinical settings. *Hum Genomics* (2023).

New long read technology reduces error rate to .1-1.5%





Adapted from Oehler, J.B., Wright, H., Stark, Z. *et al.* The application of long-read sequencing in clinical settings. *Hum Genomics* (2023).

GUIDE-A Key Question

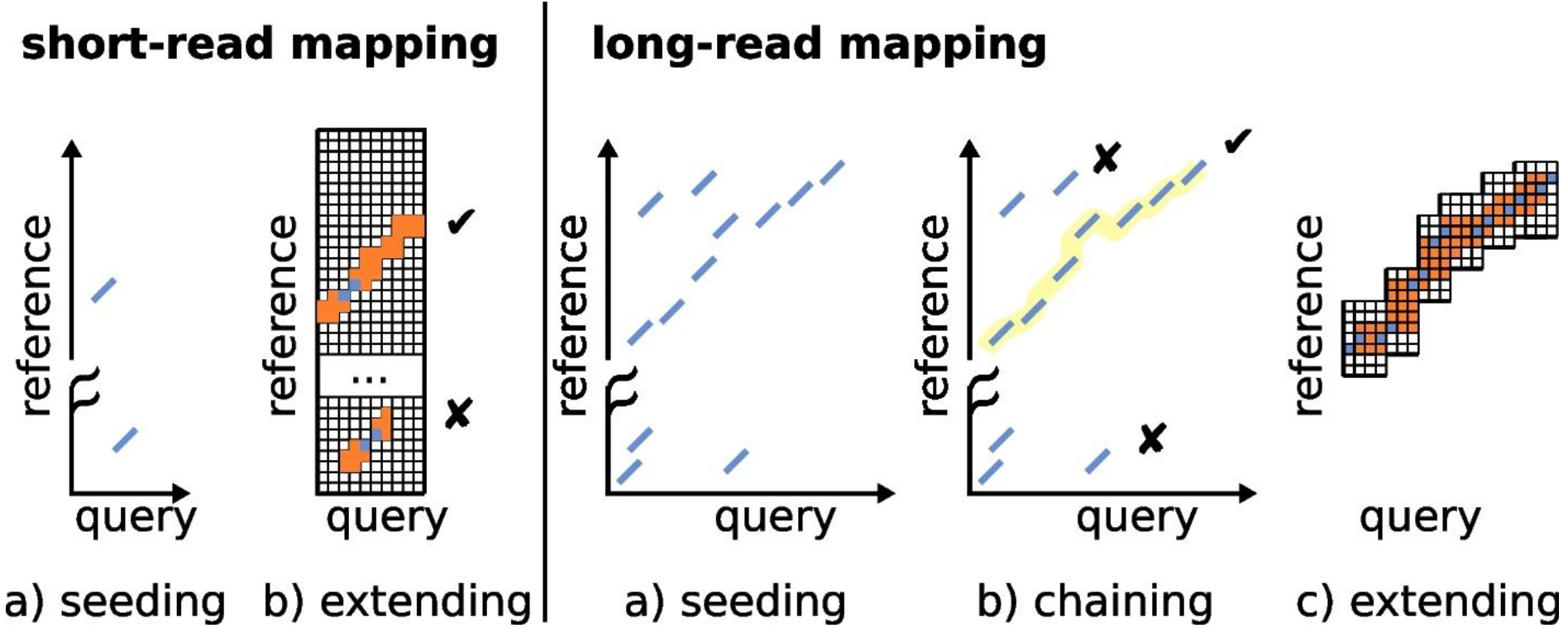
Does the 3’ end alone hold sufficient information to predict isoforms’ expression accurately, i.e. can the 3’ end reveal key alternative splicing patterns?

Challenges of RNA-seq mapping and quantification

Short read tools vs long read tools

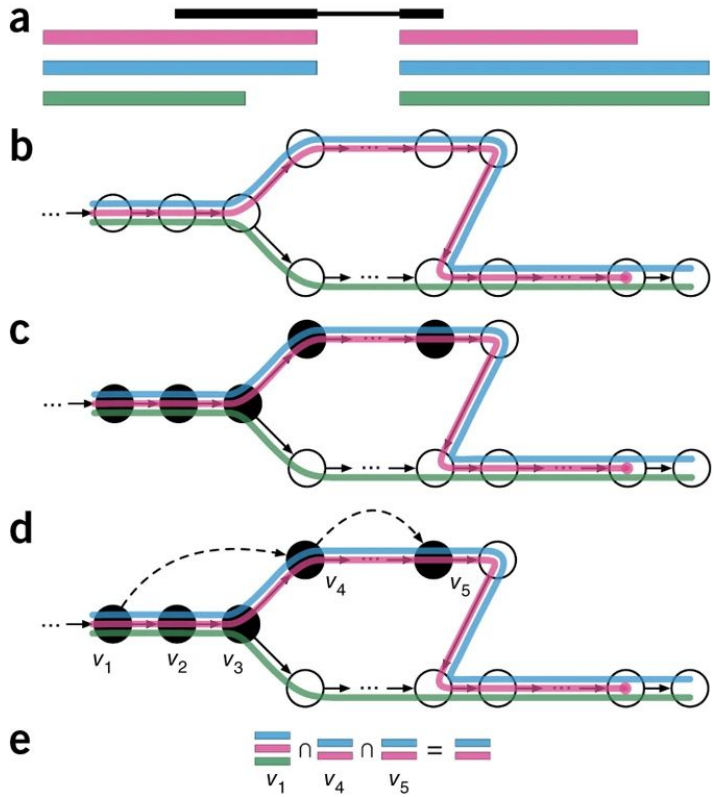
Challenges	**kallisto**	**salmon**	**Oarfish**	**bambu**	**IsoQuant**
Scalability and efficiency					
Accuracy					
Robustness (to error)					
Flexibility (to different assays)					
Alignment			minimap2	minimap2	minimap2

Alignment challenges



Sahlin, K., Baudeau, T., Cazaux, B. *et al.* A survey of mapping algorithms in the long-reads era. *Genome Biol* 24, 133 (2023). https://doi.org/10.1186/s13059-023-02972-3

lr-kallisto Background

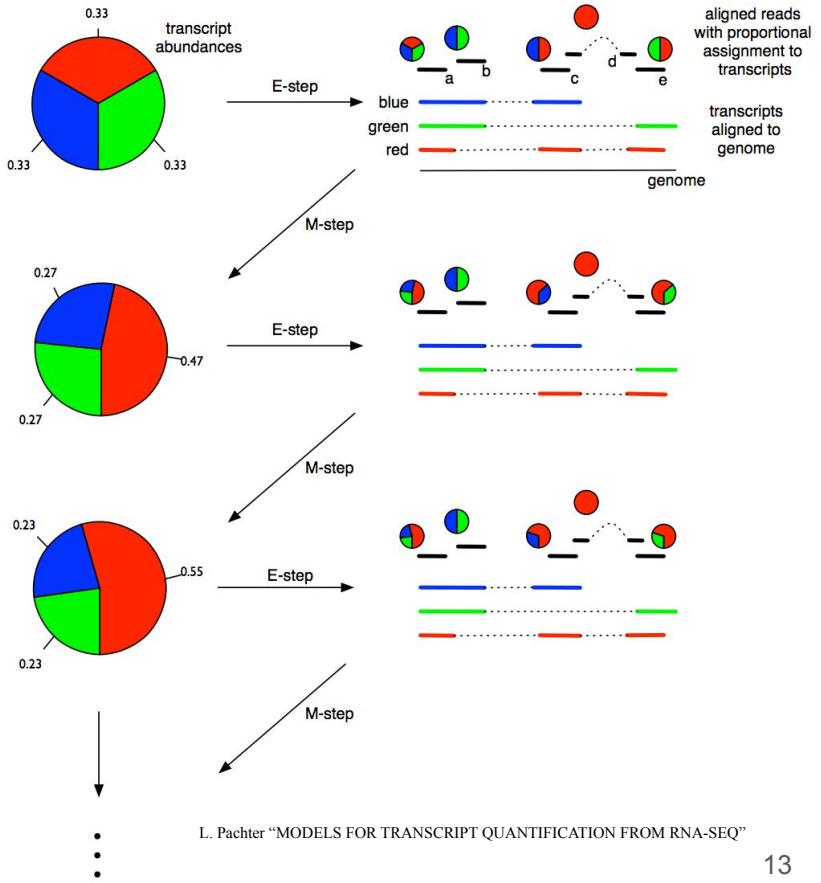


Bray, N., Pimentel, H., Melsted, P. *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* (2016)

lr-kallisto Background

$$L(\alpha) \propto \prod_{f \in F} \sum_{t \in T} \mathbf{y}_{f,t} \frac{\alpha_t}{l_t} = \prod_{e \in E} \left(\sum_{t \in e} \frac{\alpha_t}{l_t} \right)^{c_e}$$

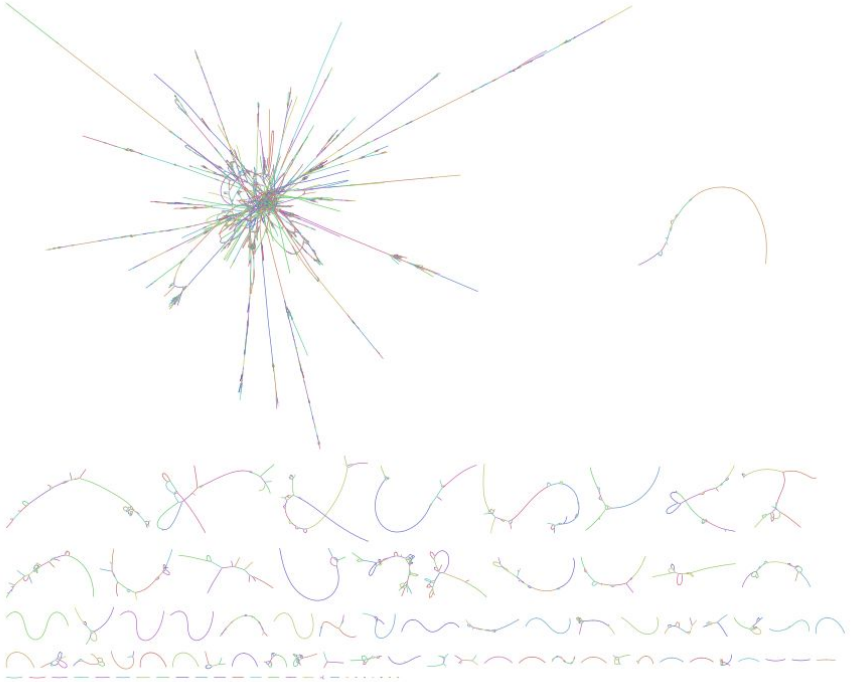
l_t is the (effective) length of transcript t



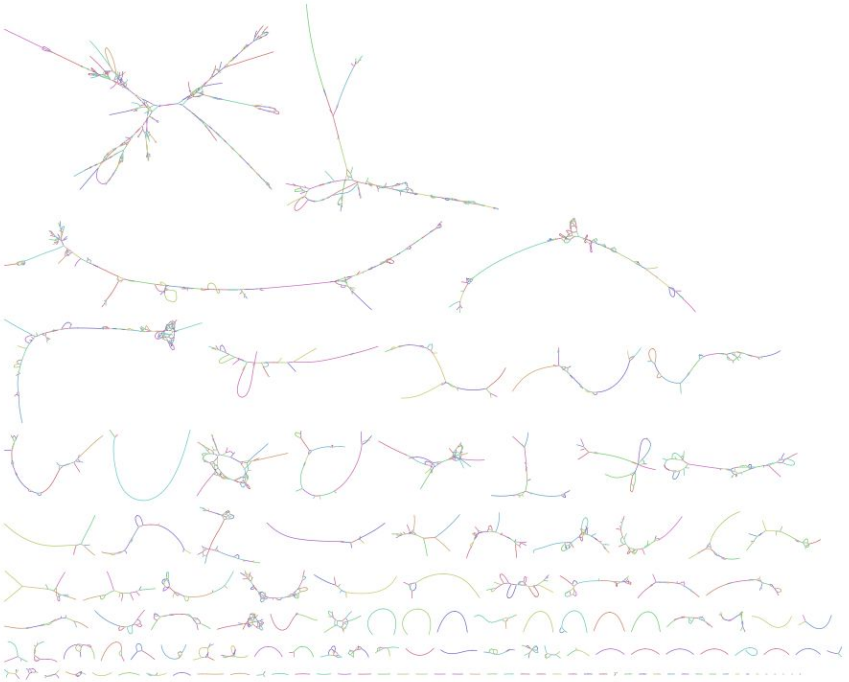
L. Pachter "MODELS FOR TRANSCRIPT QUANTIFICATION FROM RNA-SEQ"

Effect of *k*-mer length on DBG in kallisto

ii. Ex.: first 1000 transcripts, *k*-mer=31

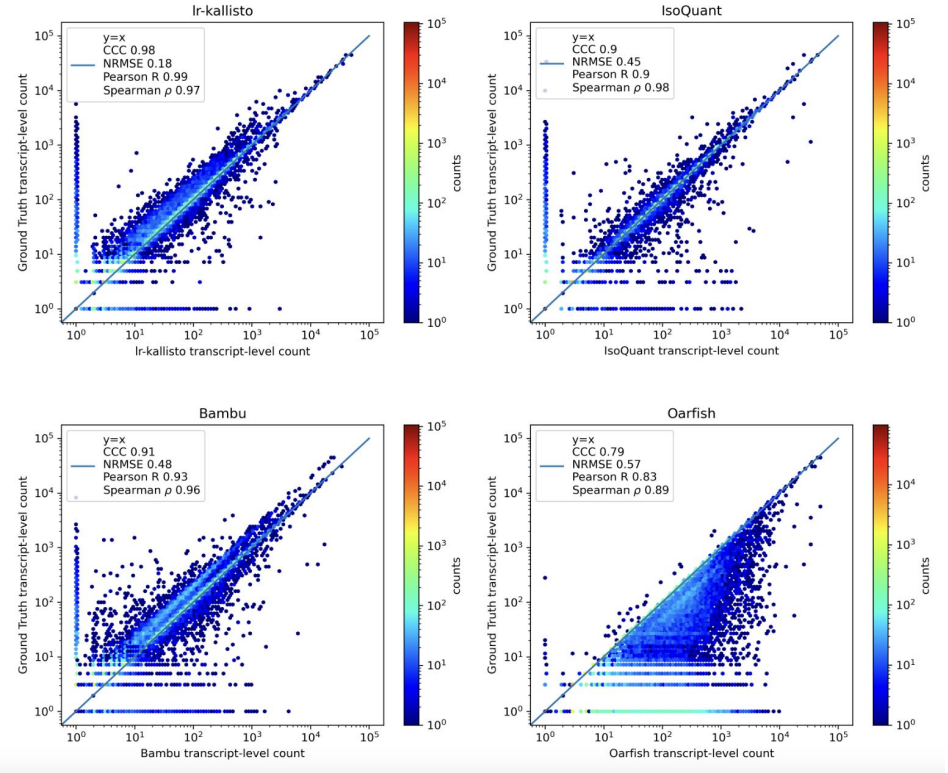


k-mer=63

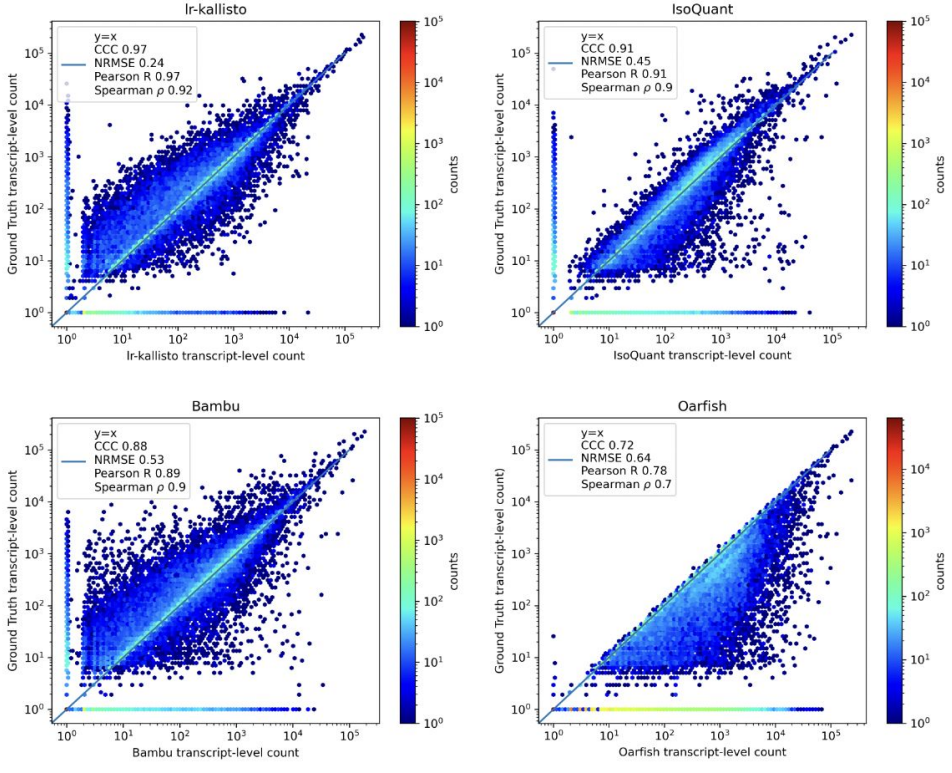


Simulation Benchmark

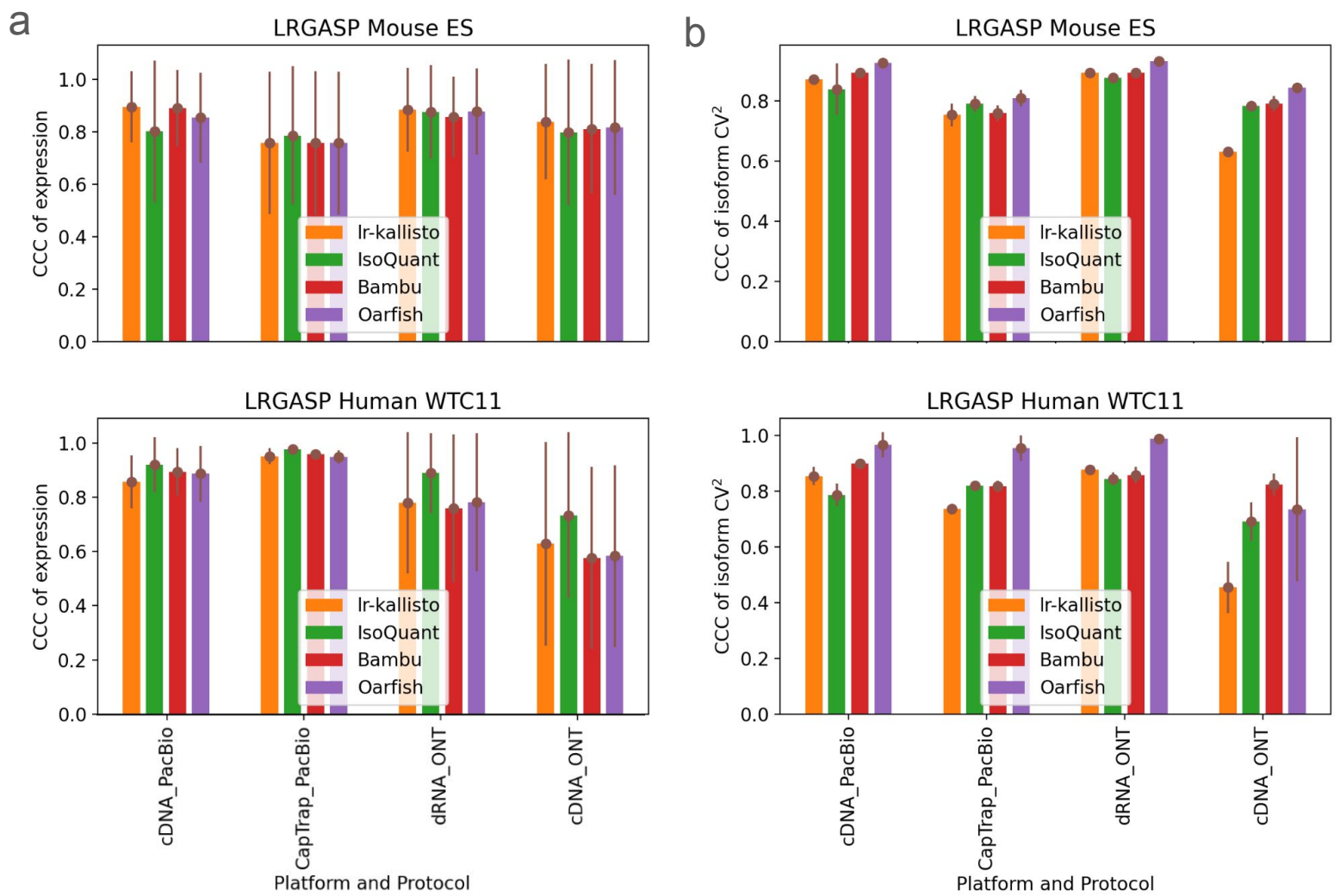
a PacBio IsoSeqSim Simulation



b ONT NanoSim R10.4 Simulation



LRGASP Reproducibility Benchmark



GUIDE-A: Gene Unambiguous Isoform Deduction Extraction - Algorithm

A machine learning approach for isoform prediction and gene/isoform network detection from 3’ RNA-seq data

GUIDE-A

Problem Setup

TCC = transcript compatibility counts
TPM = transcripts per million

Inputs: 3’ end Transcript Compatibility Counts, **TCC**

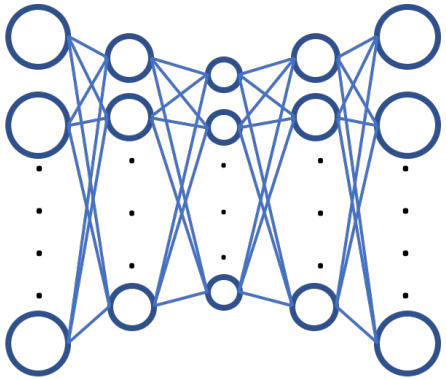
Outputs: full length Transcript Abundances, **TPM**

Find Mapping between **3’ end** TCC and **full length** TPM

Mapping is complex; can we use a modified autoencoder to uncover the **correlations** and discover the mapping?

GUIDE-A

Architecture and Setup



Each layer is a **Linear Unit**.

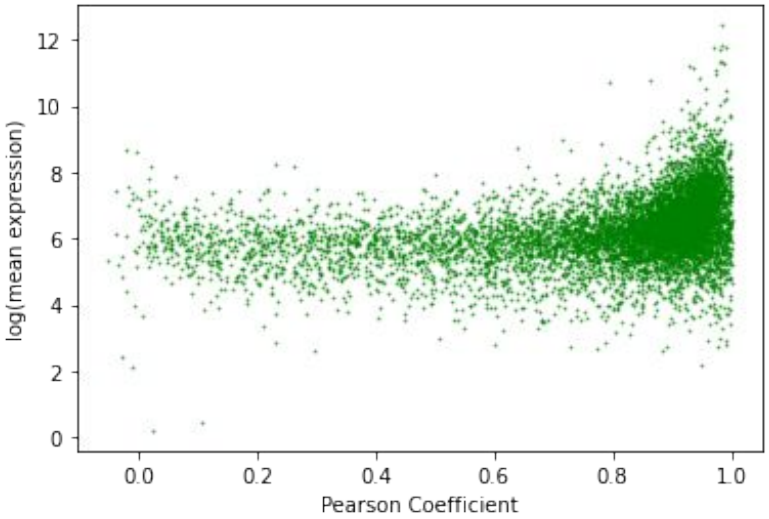
Rectified Linear Unit (ReLu) activation is used after each layer.

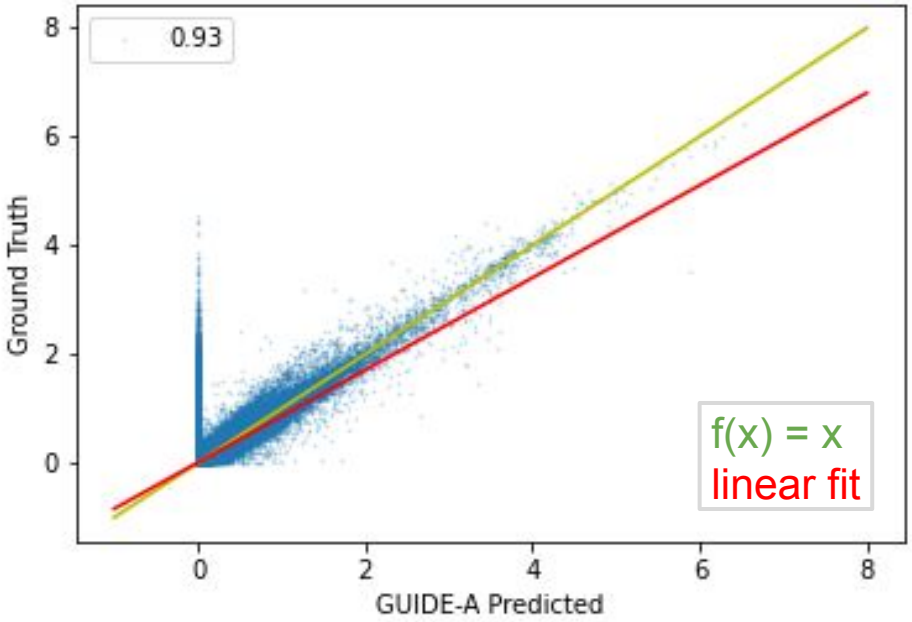
Batch normalization is used preceding the output layer.

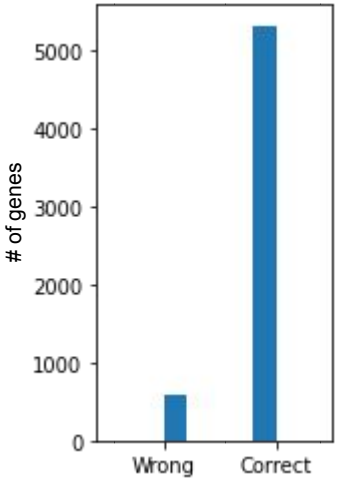
Loss is calculated via **Sum Squared Errors (SSE)**.

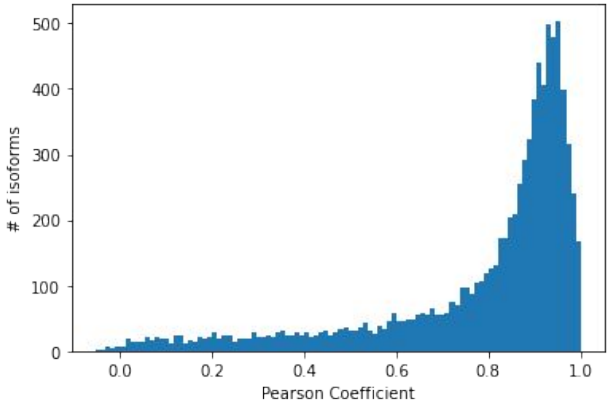
Training set is 80%, validation is 10%, and test is 10%, which was created by random sampling without replacement.

1618 bulk RNA-seq Sequence Read Archive Datasets GUIDE-A Results

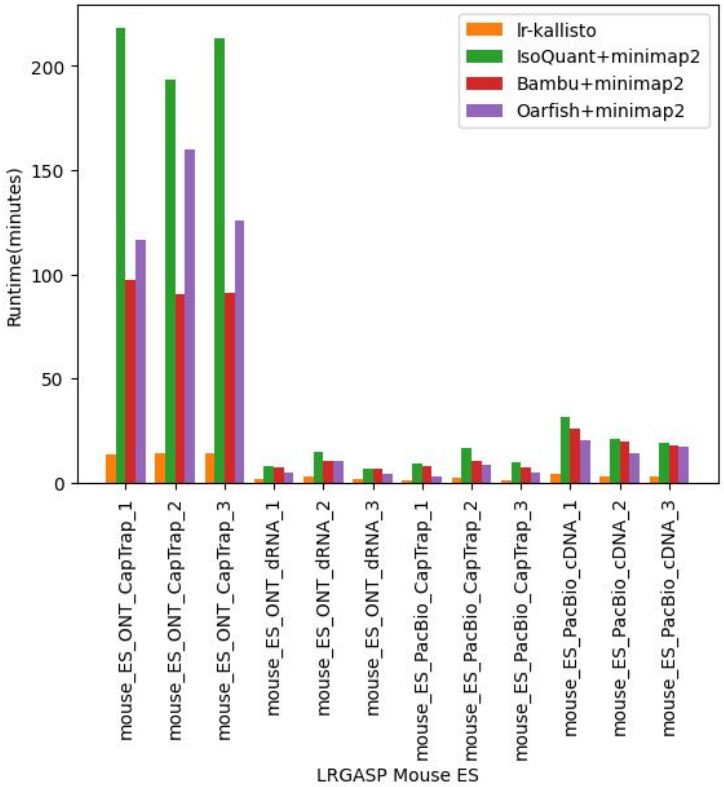








Runtime Efficiency



Acknowledgements

Pachter Lab
Lior Pachter
Delaney
Laura
Tara
Kayla
Maria
Nikki
Ángel
Kristjan
Taleen
Gennady
Cat
Meichen
Lambda
Joe Rich

Mortazavi Lab
Ali Mortazavi
Dana Wyman
Fairlie Reese
Jaz Sakr
Liz Rebboah

Wold Lab
Barbara Wold
Diane Trout
Brian Williams

Committee Members:
Matt Thomson
Pietro Perona

Funding:
DOE CSGF
IGVF Consortium