Working with STARR-Seq Data in Plants

Sacha Davis | Aug 19th 2022



Canada NRC·CNRC amii

Table of contents

01 **Introduction to Genomics + GEP**
Background on the biology of genes in aggregate

02 **The STARR-seq Method (+others)**
Measuring gene expression through self-transcription

03 **STARR-Seq in Plants**
Unique challenges posed by STARR-seq in Plantae

04 **Experiments + Directions**

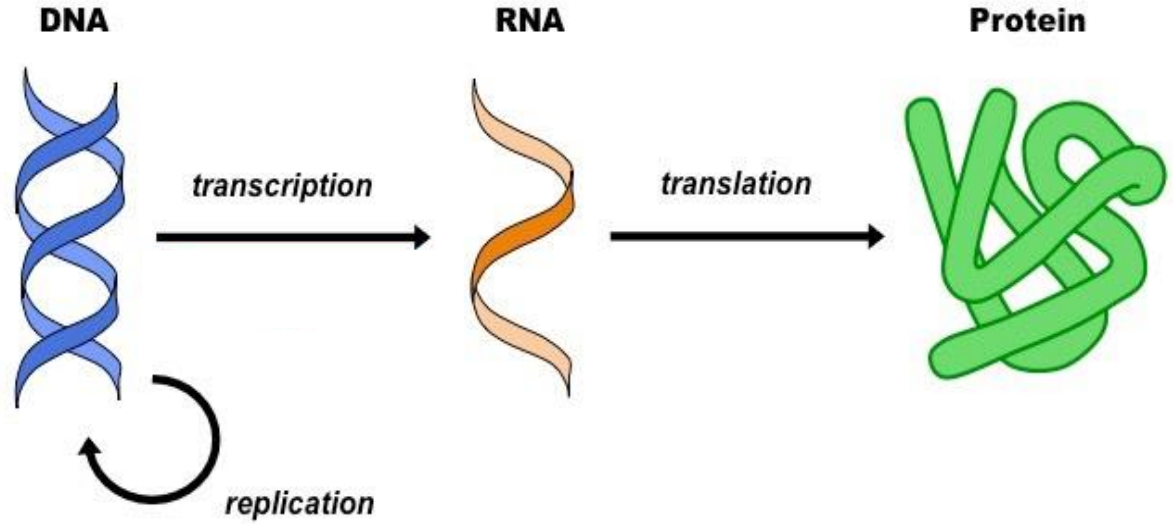


01

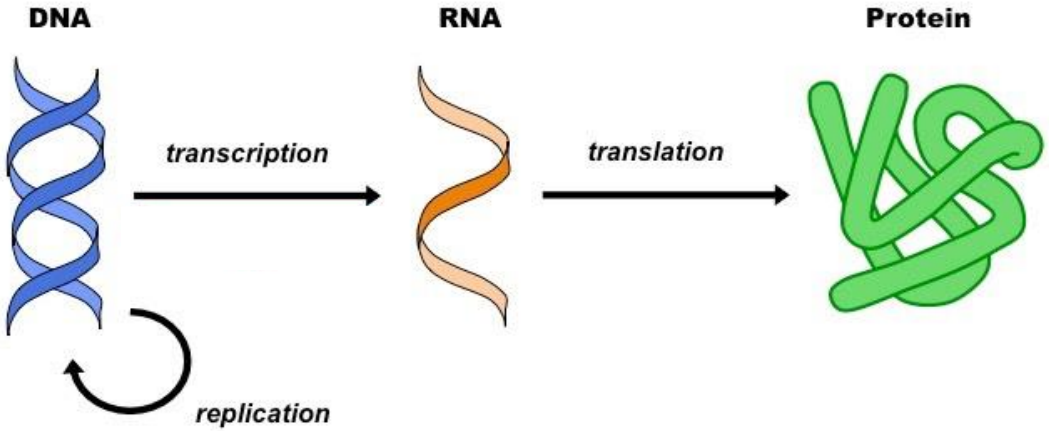
Introduction to Genomics + GEP

Background on the biology of genes in aggregate

Central Dogma of Biology



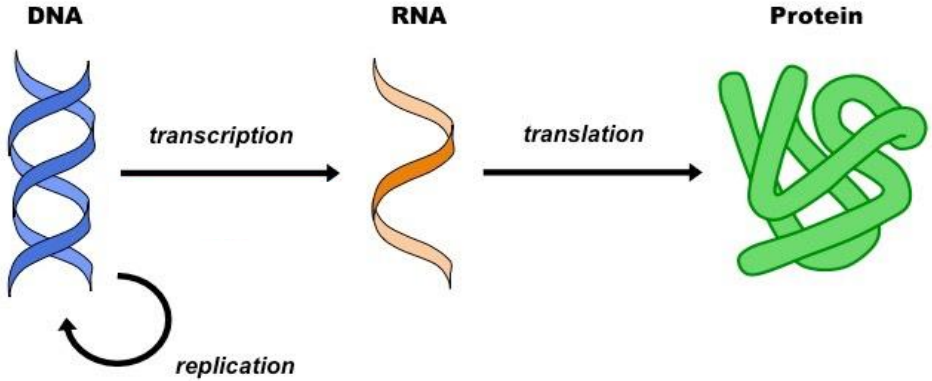
What is the problem?



Increasing Usefulness

Increasing Measurement Difficulty

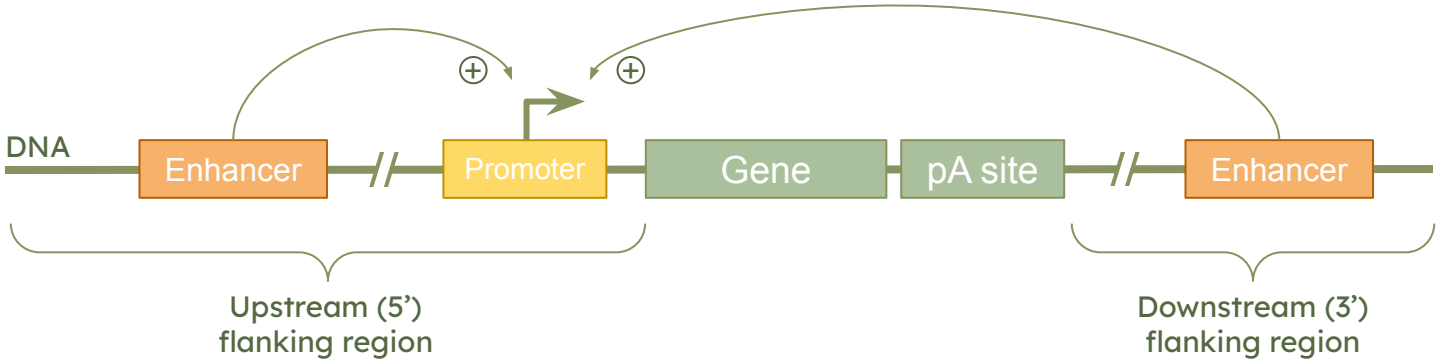
Approach



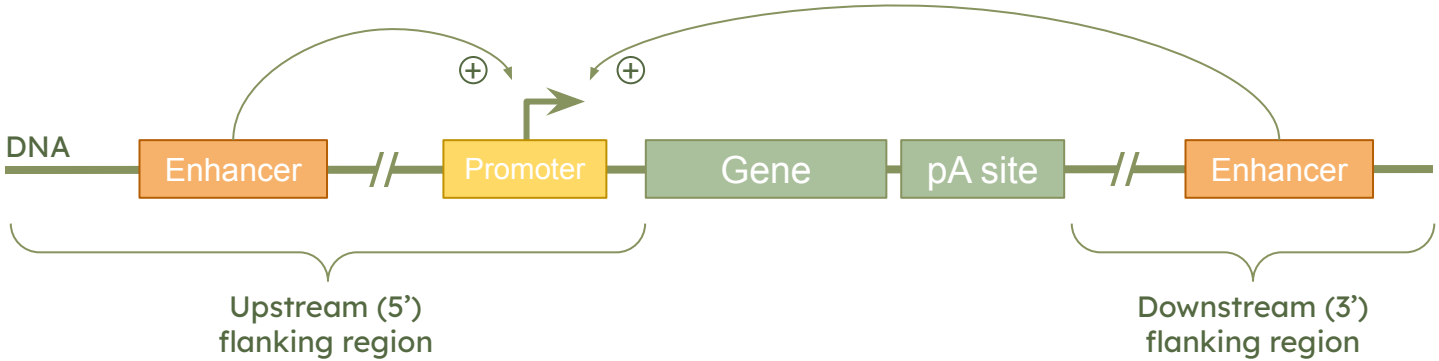
Use  to predict change in , which approximates  **

** in Canadian protein crops!

Structure of a Gene

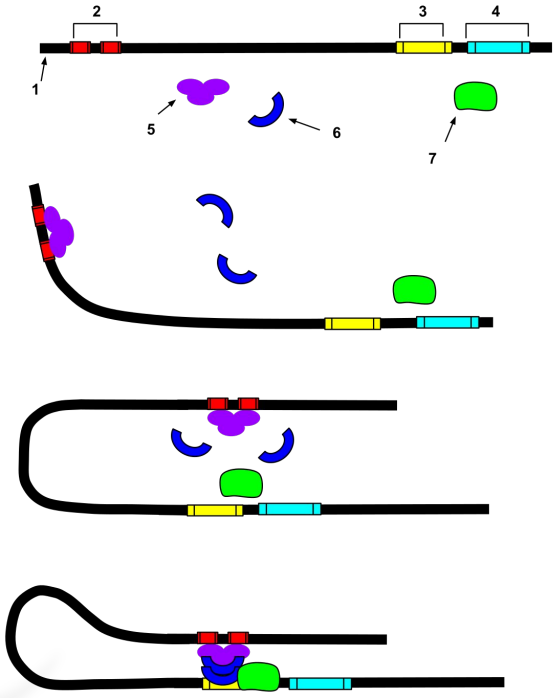


Structure of a Gene



Enhancers are sequences of DNA that, in proximity* to a gene, increase the likelihood of its transcription.

Enhancer Activity

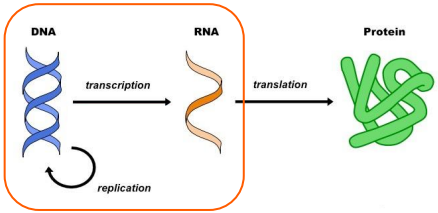


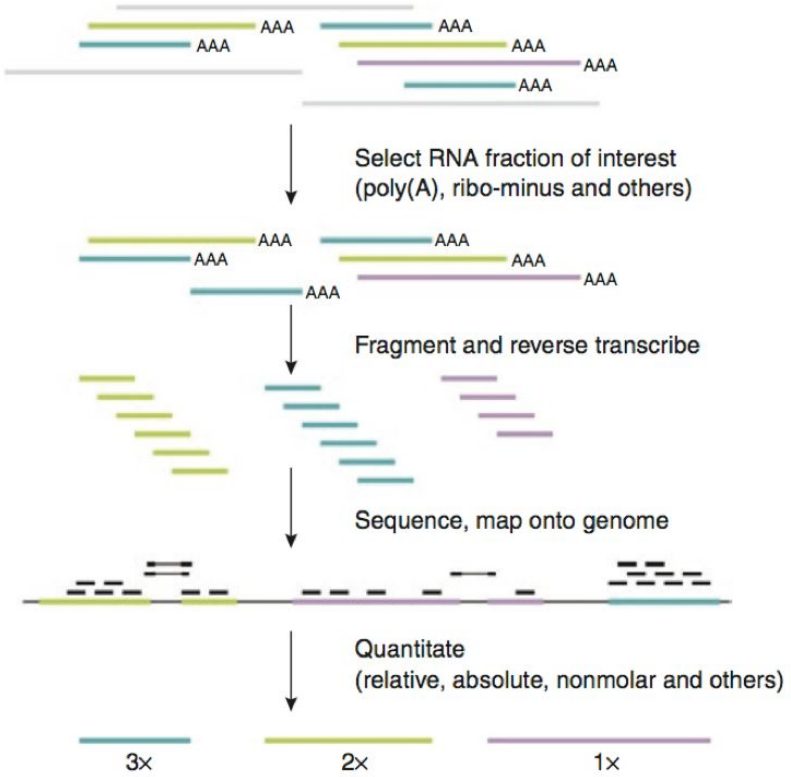
1. DNA
2. Enhancer
3. Promoter
4. Gene
5. Transcription Activator Protein
6. Mediator Protein
7. RNA Polymerase

Gene vs. Genome?

“A **gene** consists of enough DNA to code for one protein, and a **genome** is simply the sum total of an organism's DNA.”

Measuring Genome Activity: RNA-Seq





Problem Motivation

- With global climate change, political unrest, and an increasing population, we need to be smart about how we make food

World hunger rising as UN agencies warn of 'looming catastrophe'

Up to 828 million people, or nearly 10 percent of the world's population, were affected by hunger last year: UN report.

6 Jul 2022

WORLD • CLIMATE CHANGE

Climate, Not Conflict. Madagascar's Famine is the First in Modern History to be Solely Caused by Global Warming

August 15, 2022 – Record high food prices have triggered a global crisis that will drive millions more into extreme poverty, magnifying hunger and malnutrition, while threatening to erase hard-won gains in development. The

49 million people face famine as Ukraine war, climate disasters intensify

By Sarah Kaplan

June 13, 2022 at 6:00 a.m. EDT



- Luckily, there are a number of interesting avenues to explore for tackling these issues!

Model Schema



SQNC1: 1.19 log2-FC
SQNC2: 0.26 log2-FC



Potential *BNGRW* gene enhancers:

SQNC1: ...TAGAACCCTAGATCGCAT...
SQNC2: ...TAGAACGGTAACACGCAT...

Model





02

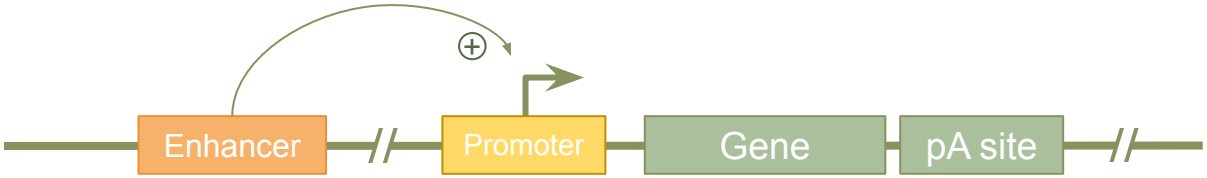
The STARR-seq Method (+others)

Measuring gene expression through self-transcription

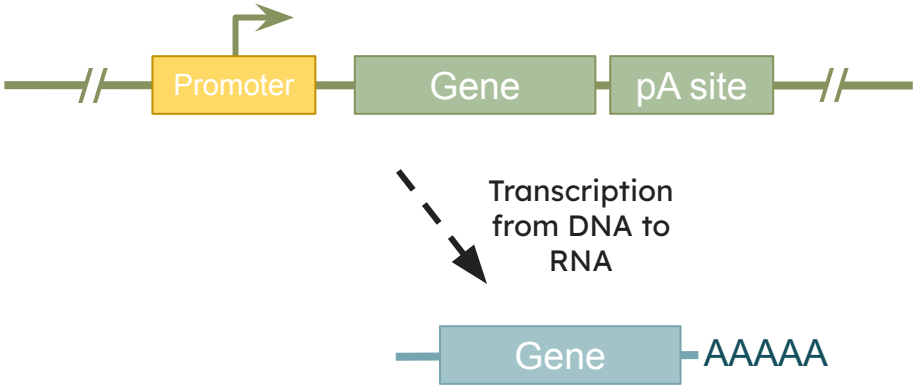
STARR-Seq Intuition

- We already touched on sequencing to measure gene activity
- How can we use this protocol to measure how much of an enhancement effect any given DNA sequence has?
  - STARR: Self-Transcribing Active Regulatory Region

STARR-Seq Intuition

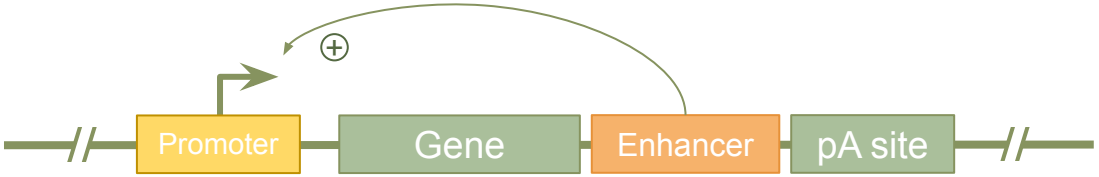


STARR-Seq Intuition

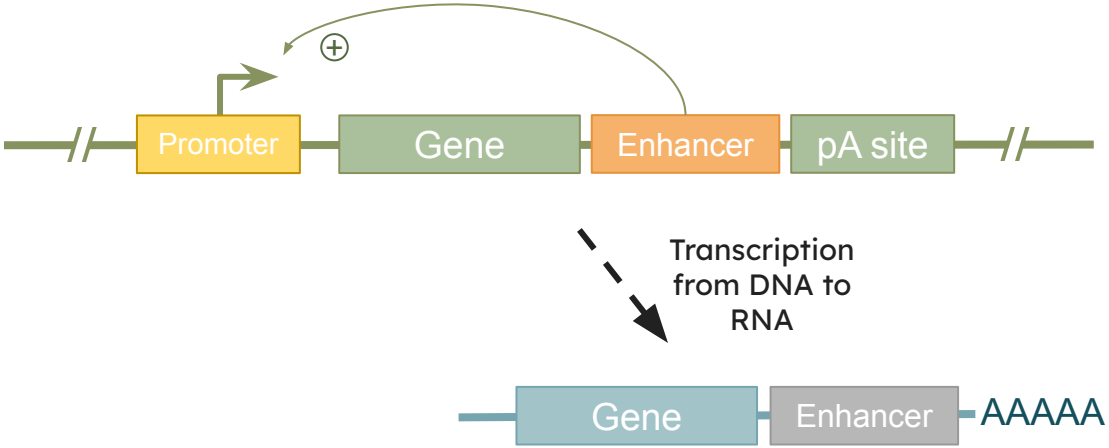


STARR-Seq Intuition

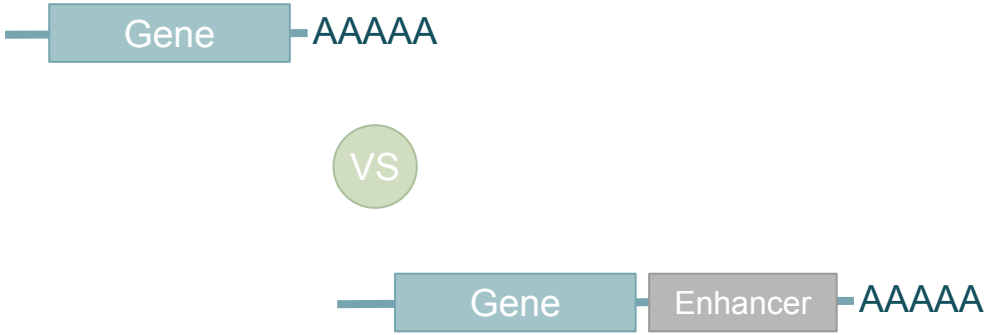




STARR-Seq Intuition



STARR-Seq Intuition

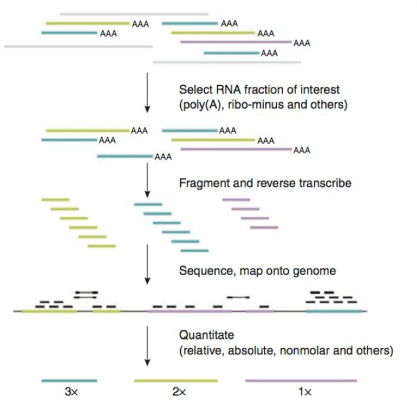


STARR-Seq Intuition



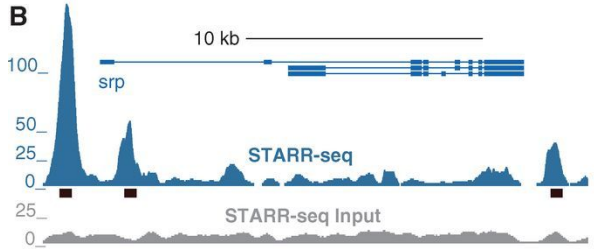
VS





STARR-Seq Intuition



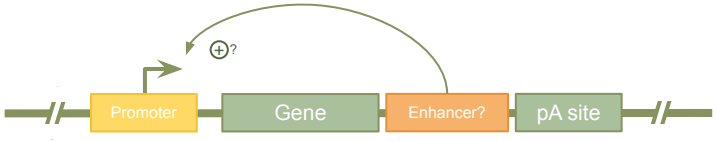


VS



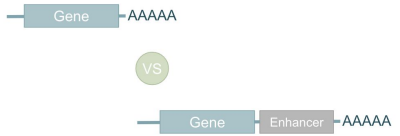
To create dataset:

1. Create candidate library:
   - Shred genome into fragments
   - Use prior knowledge
2. Clone candidates downstream of a minimal promoter
3. Measure expression of gene sequences with and without each enhancer
   - Can be done in a highly parallel manner
4. Calculate enhancement effect using:



a = norm baseline expression
b = norm treatment expression

target = $\log_2(a/b)$

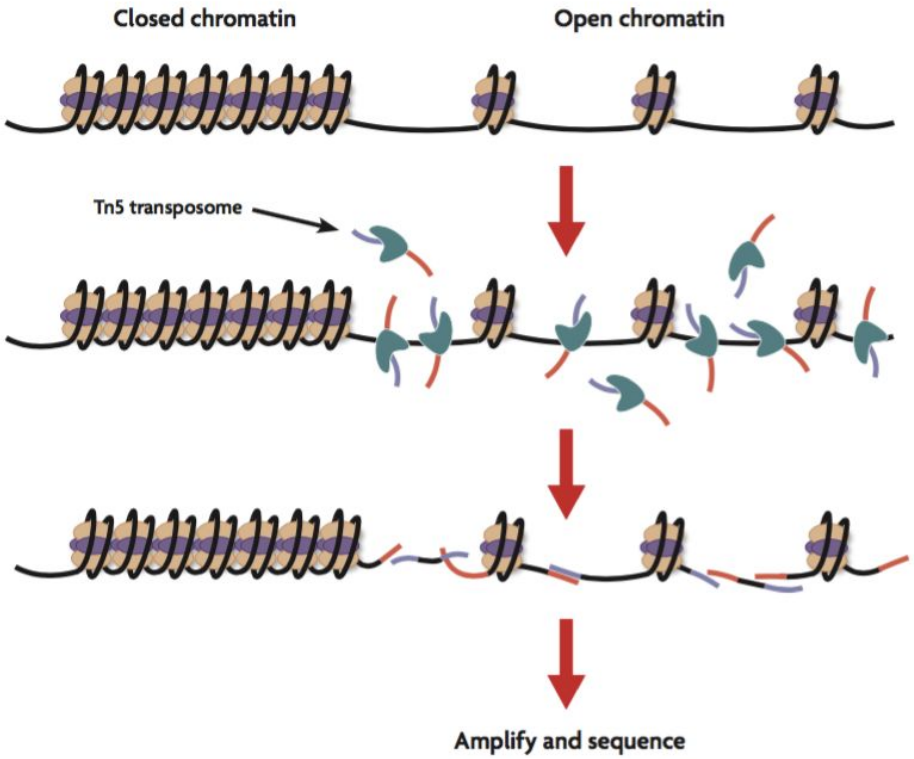


Stark Lab: Data



hk_target	set	sequence
1.711032	train	ATTCAGATTGCCTCTCATTGTCTCACCCATATTATGGGAACCAAAT...
1.974276	train	AAATGGCCGCTCAAGAAAAGGCTCGAATATATATTGCCTGCCTCTC...
0.606119	train	ATAAGGATCAAAAAGTCCTGATTTCCGAAATGGCGGTTCTCCTTCA...
3.780414	train	TTTCCATGACTGACTGGAATGGGTGGAGAACATCGCTTTGGGAGTG...
0.714676	train	TCTATCGACCCATAGCCGTAGTCGCTAGACCCGCCCTTCGGAGCAT...
...	...	...
-0.517860	val	GCAAATCATGAAGGTGGAATATATATATAGCAGATCTAAACCCTTC...
-1.155432	val	AAGTATTTTCCCAATTTCCCATACCGCACCCAGCTGCATTTCCCAT...
-0.150081	val	TTAATTTAAAGTACTTCCCATCAGCTTTTTGTGCATTTTTCCCATG...
0.882345	val	GTCGTCTTACACTCGCAGCAAAACTTGTAAATTAAAAATGCGTCCA...
1.361124	val	GCCAGCGGCAAAGTTGAAAAGGCAGAAATATGCTTTGTAGTTTAGT...

Aside: ATAC Sequencing





03

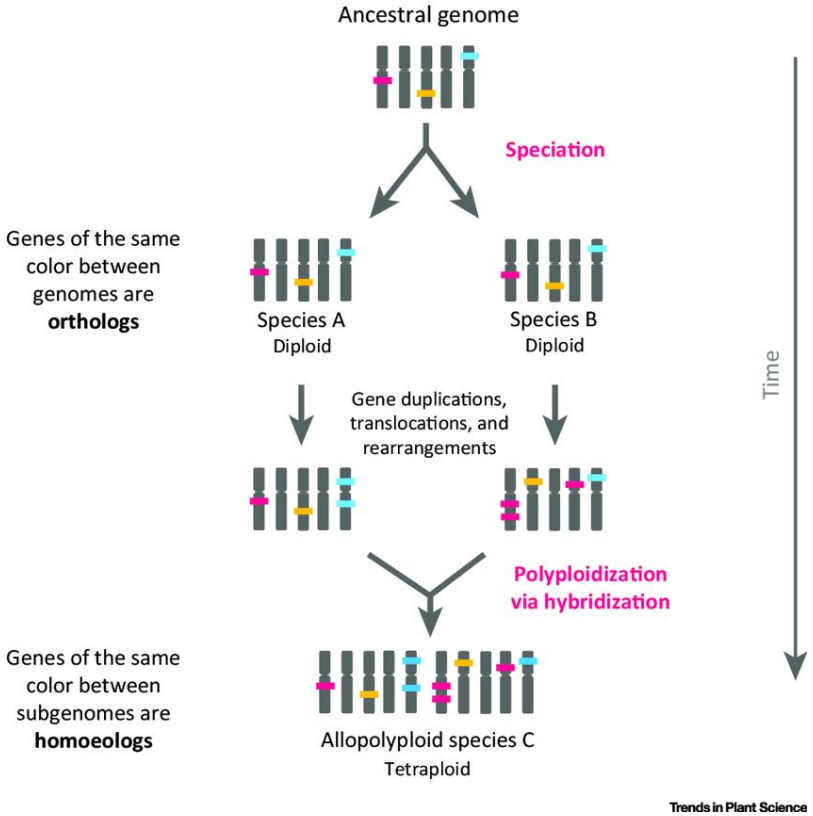
STARR-Seq in Plants

Unique challenges posed by STARR-seq in Plantae

Plants vs. Animals

- Genomes have a number of differences!
  - Gene frequency, size, space
  - GC content
  - Transposable element types and frequencies
  - Speed of evolution
  - Polyploidy**

Polyploidy


Ancestral genome
Speciation
Genes of the same color between genomes are **orthologs**
Species A
Diploid
Species B
Diploid
Gene duplications, translocations, and rearrangements
Time
Polyploidization via hybridization
Genes of the same color between subgenomes are **homoeologs**
Allopolyploid species C
Tetraploid
Trends in Plant Science

Datasets



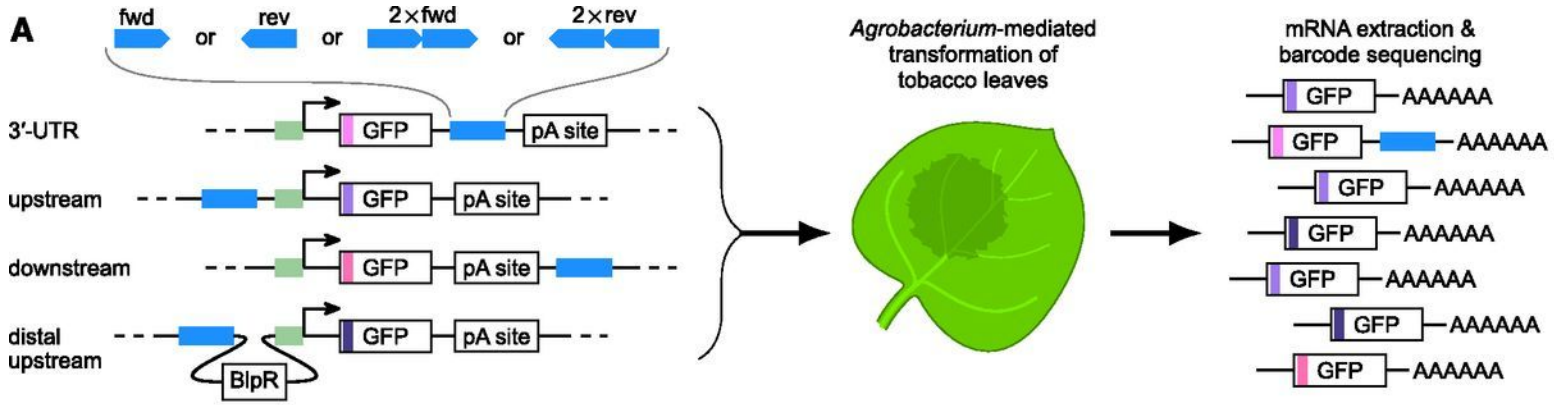
- Canola
  - Organellar STARR-seq (mitochondria + chloroplast)
  - ATAC-seq
- Arabidopsis
  - Genome STARR-seq
  - BAC STARR-seq
- Tobacco*
  - STARR-seq (ground-truth-adjacent!)
- Rice
  - STARR-seq (in progress)







Tobacco Dataset



(A) Scheme of the tobacco STARR-seq assay. All constructs are driven by the 35S minimal promoter (green). Enhancers (blue) are inserted in the indicated orientation and position. Barcodes (shades of purple) are inserted in the GFP open reading frame. BlpR, phosphinothricin resistance gene; pA site, poly-adenylation site.



04

Experiments + Directions

Recall: Model Schema



SQNC1: 0.98 log2-FC
SQNC2: 0.26 log2-FC



Potential *BNGRW* gene enhancers:

SQNC1: ...TAGAACCCTAGATCGCAT...
SQNC2: ...TAGAACGGTAACACGCAT...

Model



Predictive Task

- Sequence-agnostic
  - Sequence representation: nucleotide frequencies
    - LR, FFNN, XGBoost

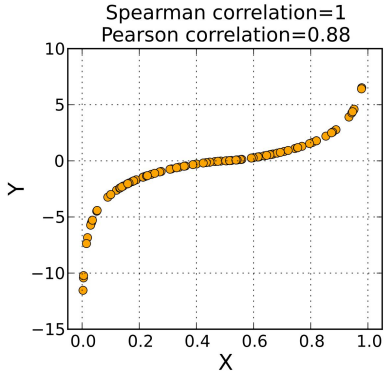
Target	Sequence
0.04	CGCCCCTACTCTACTATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTT
0.08	CTACTCTACTATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTT
0.06	CTACTATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTTTATATA
0.01	ATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTTTATATATAATA(
-0.07	TCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTTTATATATAATAGAAAA

Target	A	T	C	G
0.04	62	57	18	8
0.08	62	62	14	7
0.06	62	63	12	8
0.01	64	62	10	9
-0.07	65	59	11	10

- Not sequence-agnostic
  - CNN, RNN

Performance Measurement

- Two metrics:
  - Pearson correlation coefficient (r^2)
    - Deviation magnitude
  - Spearman correlation coefficient (ρ)
    - Rank



- Spearman coefficient tends to be slightly more lenient and more interesting

Already Explored

- Training many different models to scope expected performance on different datasets
- Replication studies
- Examining differences between genomic and organellar STARR-seq data in plants
- DNA sequence representation learning (using frequency and convolution-based features)
- Transfer learning experiments between plant and animal cells
- Auxiliary task learning

Goal #1

- We have two types of STARR-seq data for Arabidopsis
  - Large, noisy dataset (`nrows=2973418`)
  - Small, concentrated dataset (`nrows=6991`)
- Which of the two is better for model training? *Can we leverage both for a more informed model?*
- Hypothesis: By pre-training models on the small dataset, we will see some improvement when training & testing on the large dataset

Goal #1: Preliminary Results

		R2				Spearman			
Set-Up		Train Genome	Val Genome	Test Genome	Test BAC	Train Genome	Val Genome	Test Genome	Test BAC
Train on genome, test on BAC	Mean	0.2938	0.2692	0.2673	0.5107	0.5414	0.5375	0.5372	0.8156
	Std Dev	0.0040	0.0007	0.0025	0.0069	0.0020	0.0006	0.0020	0.0051

		R2				Spearman			
Set-Up		Train BAC	Val BAC	Test BAC	Test Genome	Train BAC	Val BAC	Test BAC	Test Genome
Train on BAC, test on genome	Mean	0.7323	0.6015	0.6668	-0.1019	0.8651	0.7935	0.8107	0.4453
	Std Dev	0.0252	0.0210	0.0245	0.0519	0.0129	0.0139	0.0126	0.0129

Goal #2

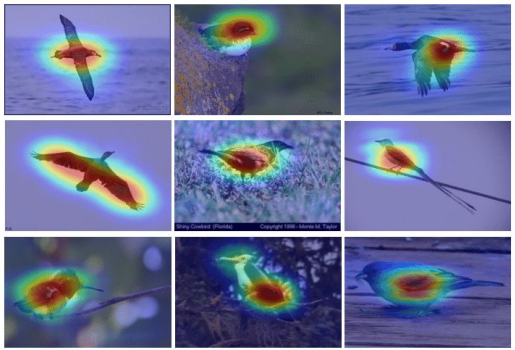
- Can we see value by combining STARR-seq datasets from different organisms?
- What can we do in terms of pre-processing that will allow the model to best capture interesting transferrable biological information?
- Hypothesis: by using ATAC accessibility data to allow us to focus on sequences that are “under the spotlight”, we can train better, more generalizable models

Goal #3

- Question the models we're training!
  - How can we do this? Come up with new methods of evaluation that make sense for biology?
- Characterize sequences that the models are getting right/wrong
- See what properties models that are good at capturing biological information have

Goal #4

- Biologically validate our models!
- NRC wet lab:
  - Test some *de novo* enhancer candidates from saliency maps
  - Evolution mechanism: can we iteratively improve upon known enhancer sequences?



Goal #5 (Moonshot)

- Propose method for sanity checking newly created STARR-seq datasets using transfer learning
  - Concordance implies interesting biological information is being captured
  - For a STARR-seq dataset to be considered trustworthy, a model trained on this dataset should achieve some level of performance when tested on a “ground truth” dataset
    - Tobacco dataset is the closest we have!
  - Can somehow be weighted based on genomic closeness?

Thanks!

Questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution