# Nascent Transcript Identification

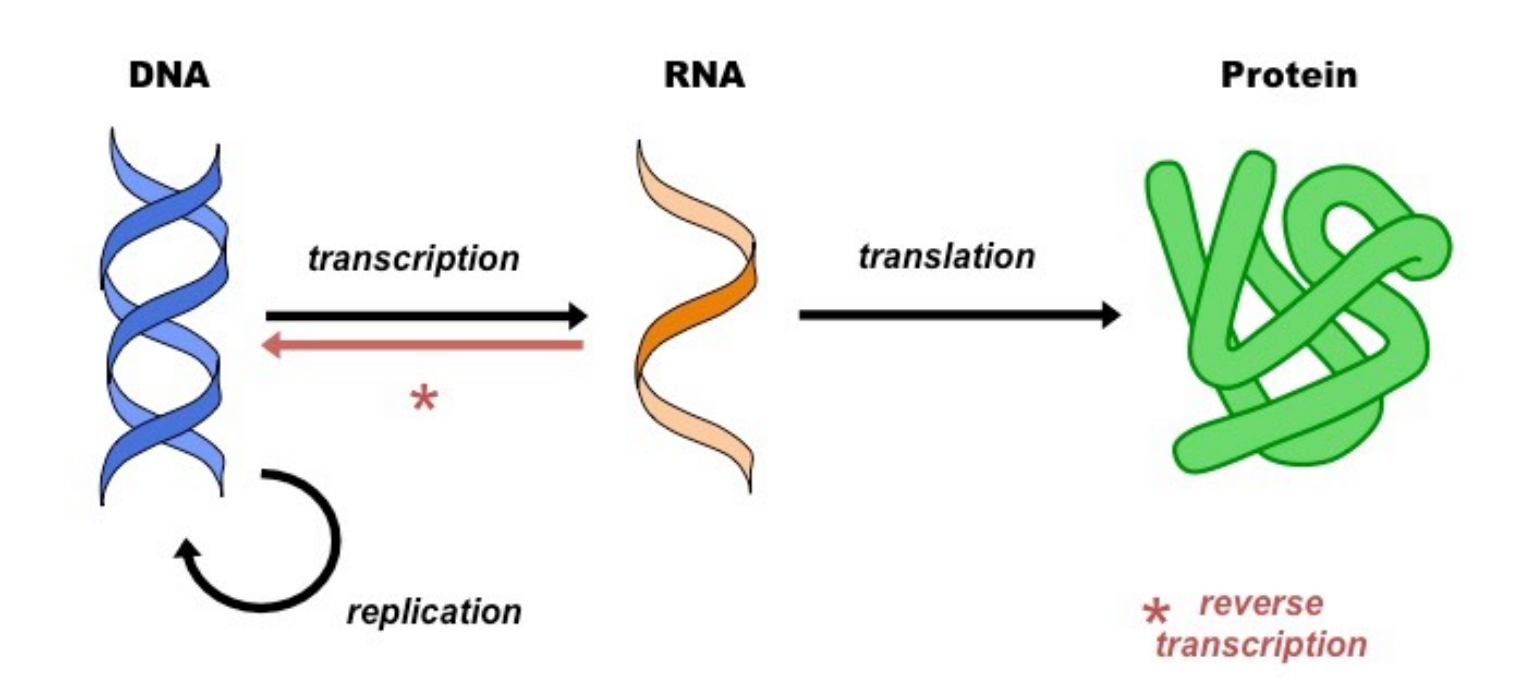
Edmund Miller

2022-04-11 Mon

# Background

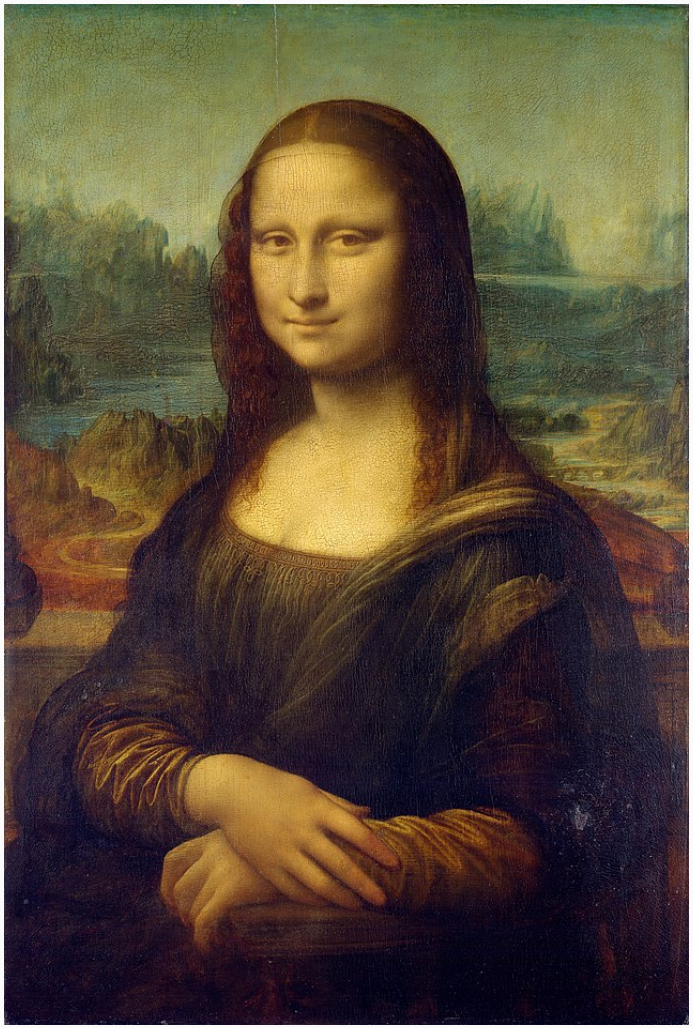
# Gene Regulation

# Central Dogma of Biology

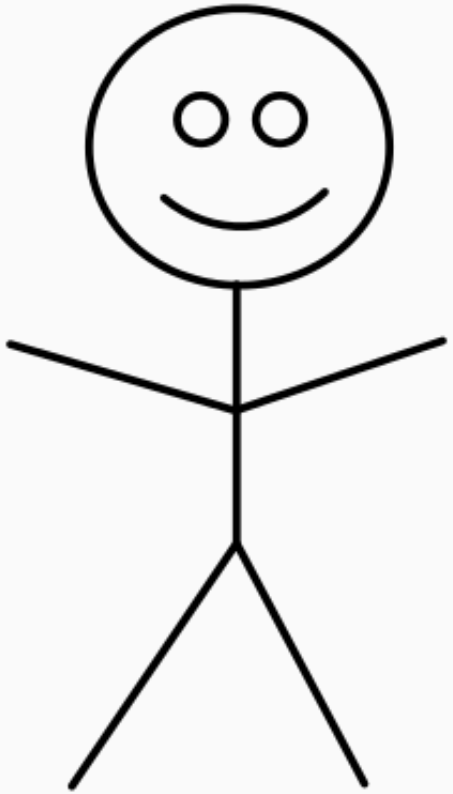


# A Simple View of Gene Expression

## Gene Expression



## Central Dogma



# Enhancers

- Cis-acting DNA sequences that can increase the transcription of genes [1]
- The human genome contains hundreds of thousands of enhancers
- Thought to work through DNA looping
- Evidence of Enhancer-Promoter interaction from cross-linking assays(3c)

[1] (Pennacchio et al. 2013)

# Topologically Associating Domain (TAD)

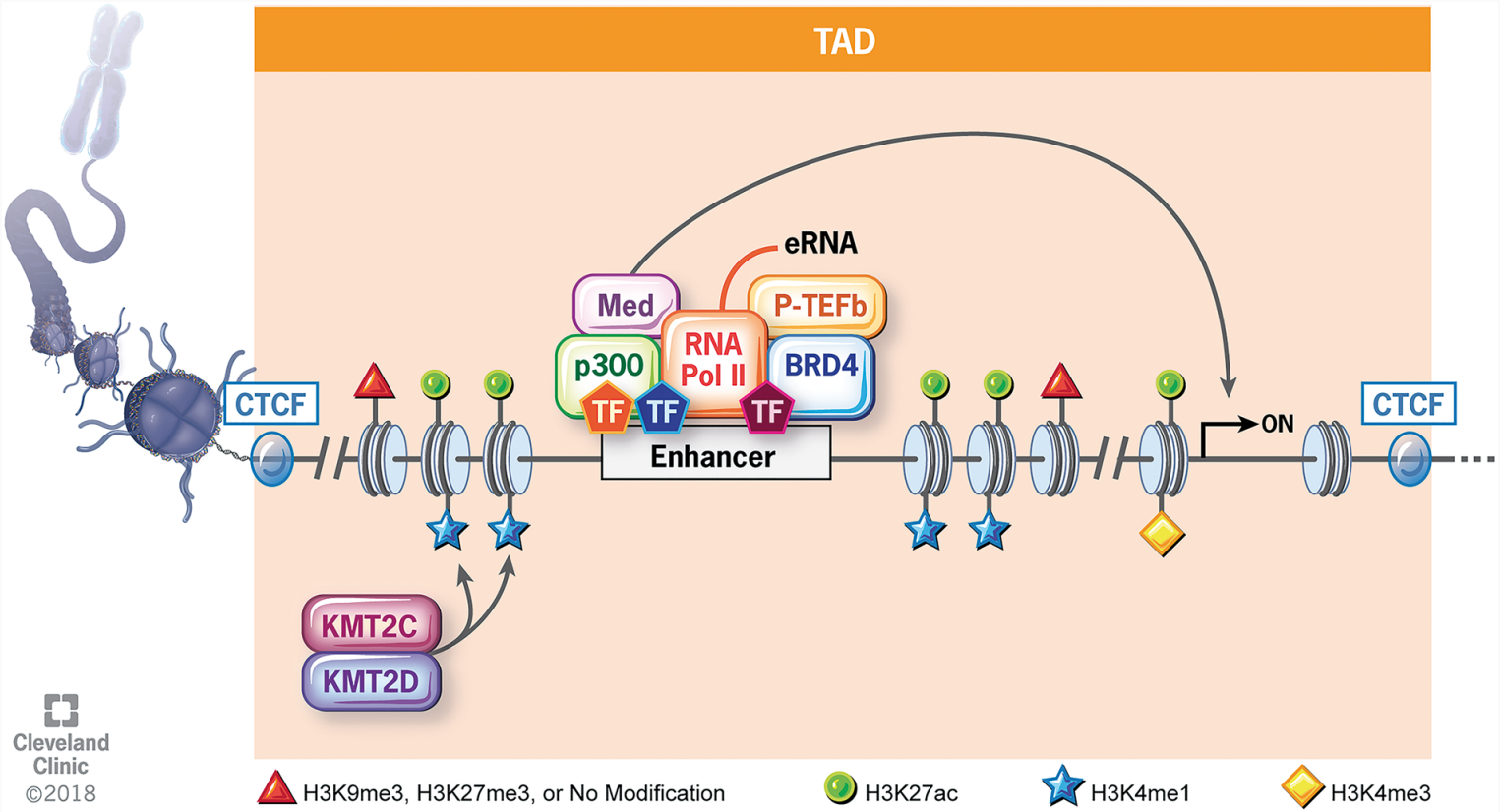



**Figure 1:** General topography of an active enhancer (Banerji, Rusconi, and Schaffner 1981).

# Enhancer-promoter Looping

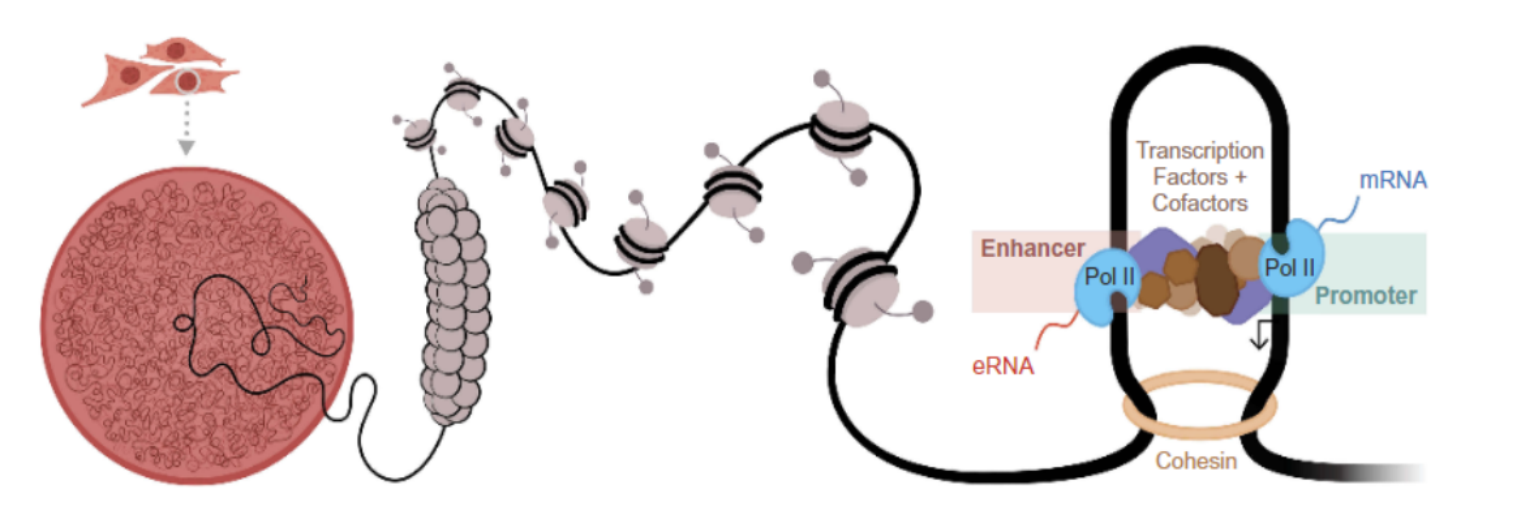



**Figure 2:** (Murakami, Nagari, and Kraus 2017)

# Why are Enhancers difficult to identify?

1. Scattered across the 98% of the human genome that does not encode proteins [1]
2. Enhancers location relative to their target gene (or genes) is highly variable. They can be upstream, downstream, or within introns. [1]
3. Enhancers do not necessarily act on the respective closest promoter but can bypass neighbouring genes to regulate genes located more distantly along a chromosome [1]

# Why are Enhancers difficult to identify?

1. One Enhancer can regulate multiple genes [2] and One gene can be regulated by multiple enhancers [3]
2. The general sequence code of enhancers, if one exists at all, is poorly understood. [1]
3. The activity of enhancers can be restricted to a particular tissue or cell type

[2](Mohrs et al. 2001)
[3](Kim et al. 2018)

# Multiple Enhancers can regulate one gene

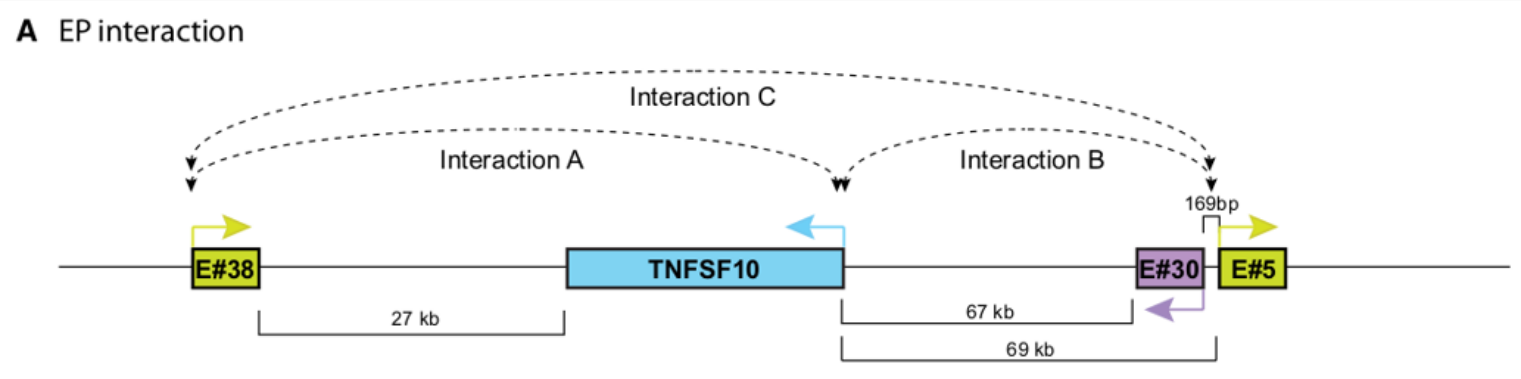



**Figure 3:** (Kim et al. 2018)

# eRNAs Introduction

- Sense and antisense transcripts from enhancers
- Play a role in the activation because the knockdown of a subset of the eRNAs resulted in decreased gene transcription [4]
- Used to identify active enhancers

[4](Ulf Andersson Ørom 2010)

# Global Transcriptional Activity Dynamics Reveal Functional Enhancer RNAs

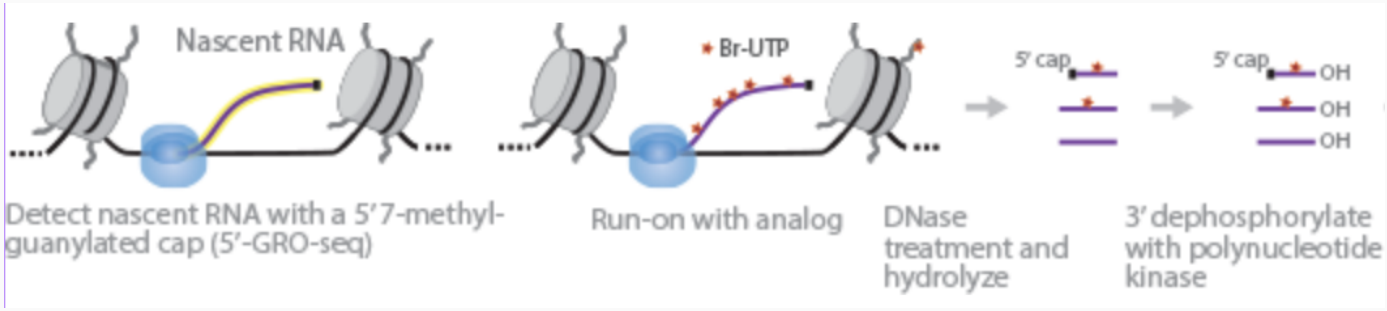
# Global Transcriptional Activity Dynamics Reveal Functional Enhancer RNAs

## Global transcriptional activity dynamics reveal functional enhancer RNAs

Yoon Jung Kim,[1,2] Peng Xie,[1,2] Lian Cao,[1] Michael Q. Zhang,[1] and Tae Hoon Kim[1]

[1]*Department of Biological Sciences and Center for Systems Biology, University of Texas at Dallas, Richardson, Texas 75080, USA*

# GRO-Seq Overview



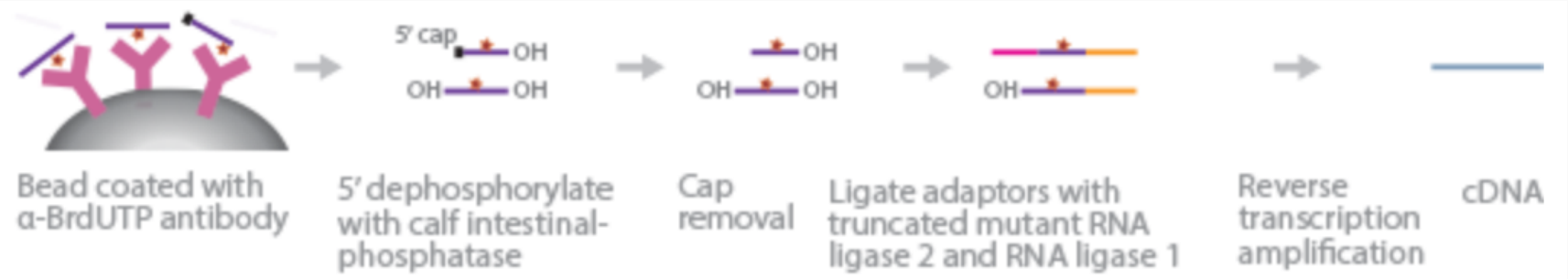




**Figure 4:** Illumina Sequencing Methods

# GRO-Seq pros and cons

**Pros:**

- Maps nascent capped 5' RNA sequence at any given time
- Determines activity of transcription sites
- No prior knowledge of transcription sites needed

**Cons:**

- Limited to cell cultures and other artificial systems due to incubation with labelled nucleotides

# Kim et. al 2018 Summary

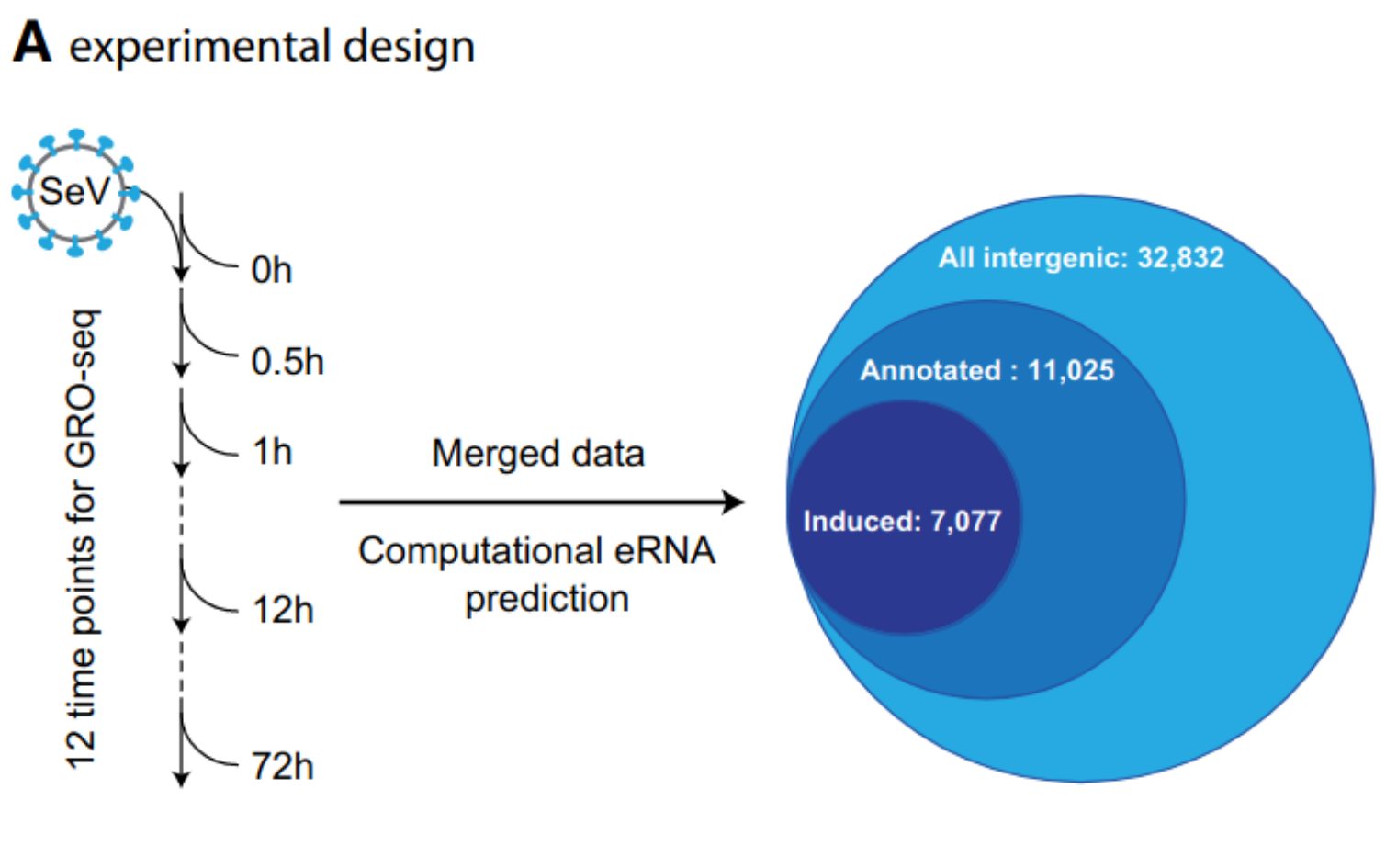
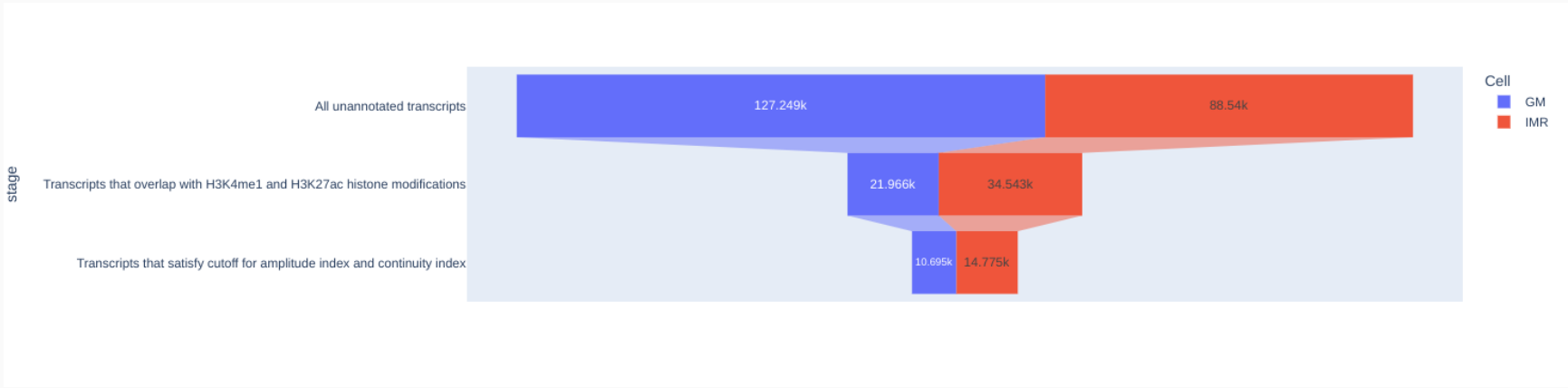



**Figure 5:** (Kim et al. 2018)

# Kim et. al 2018 Summary



# Kim et. al 2018 Summary

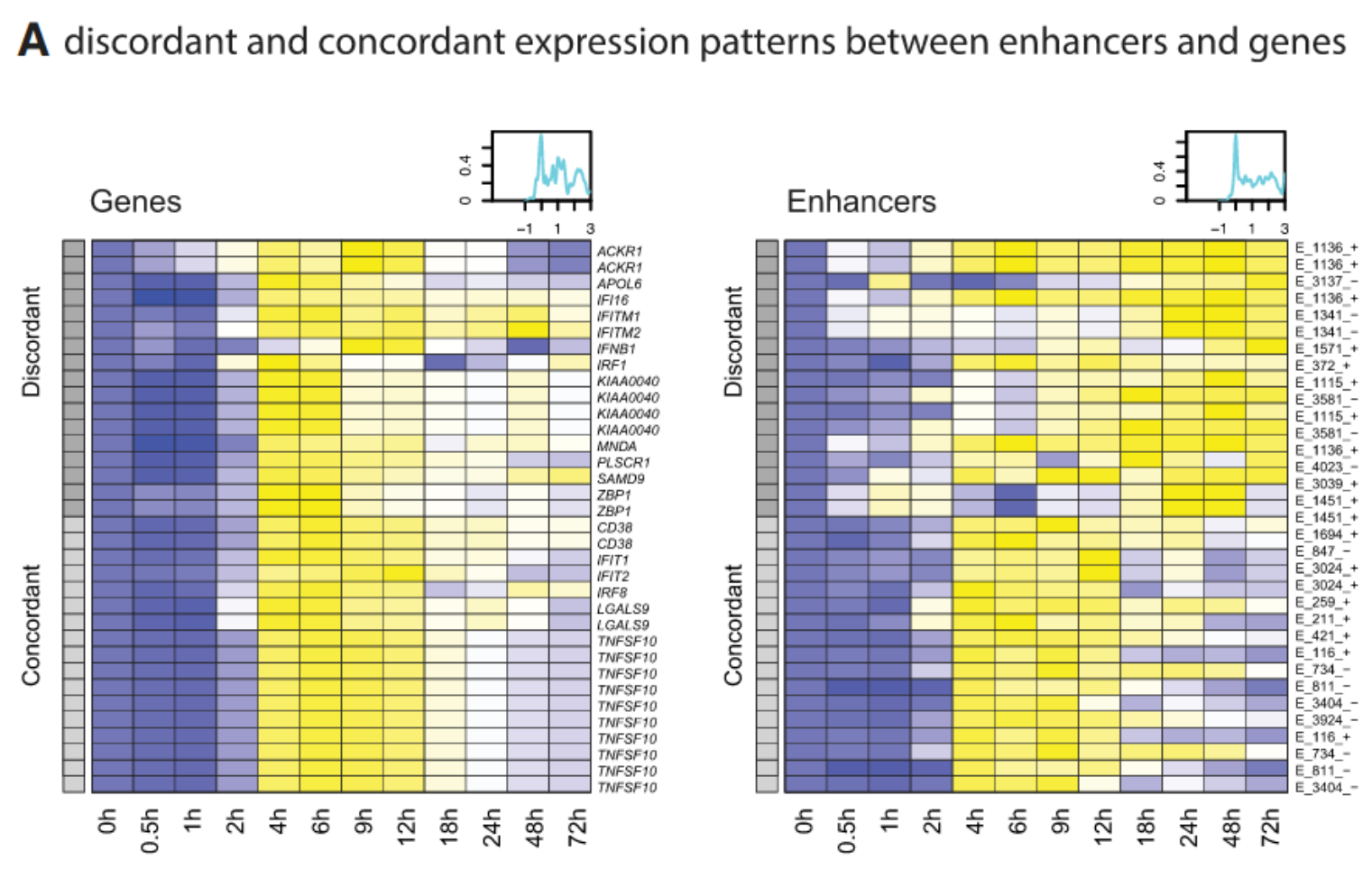



**Figure 6:** (Kim et al. 2018)

# Kim et. al 2018 Summary

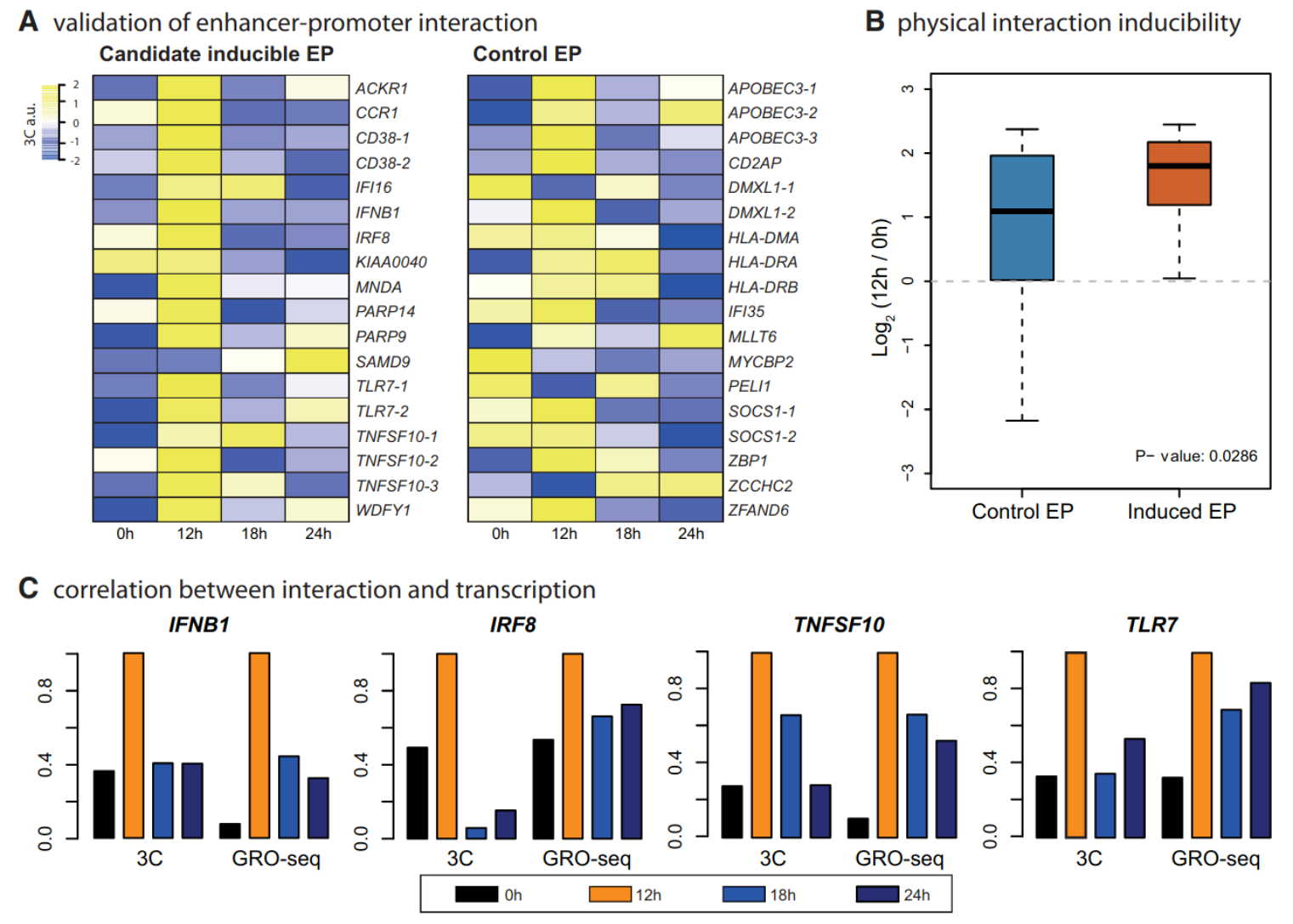



**Figure 7:** (Kim et al. 2018)

# Reproduction with Parallel IMR Dataset

- Wrote workflow using snakemake
- Goal was to reproduce GM results
- Achieved 80% of predicted eRNAs due to difficulty with nascent transcript identification

# Hypothesis

- Does the standardization of secondary analysis and use of transcription start sites for calling enhancer RNAs improve the accuracy full transcript identification?
- Using the streamlined process of transcript identification can new dynamics and classes of eRNAs can be indentified from massively parrallel processing of publicly accessible nascent trancript assay across cell lines?

# Aims

# Aim 1: Create a best practice secondary analysis pipeline for nascent transcripts

# Standardizing Snakemake Workflow

- January 2020
- Template
- Universal Commands
- Testing
- CI/CD
- Wrappers

# nf-core Paper

Correspondence | Published: 13 February 2020

## The nf-core framework for community-curated bioinformatics pipelines

Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen

*Nature Biotechnology* **38**, 276–278 (2020) | Cite this article

**10k** Accesses | **174** Citations | **173** Altmetric | Metrics

# Main concepts of nf-core

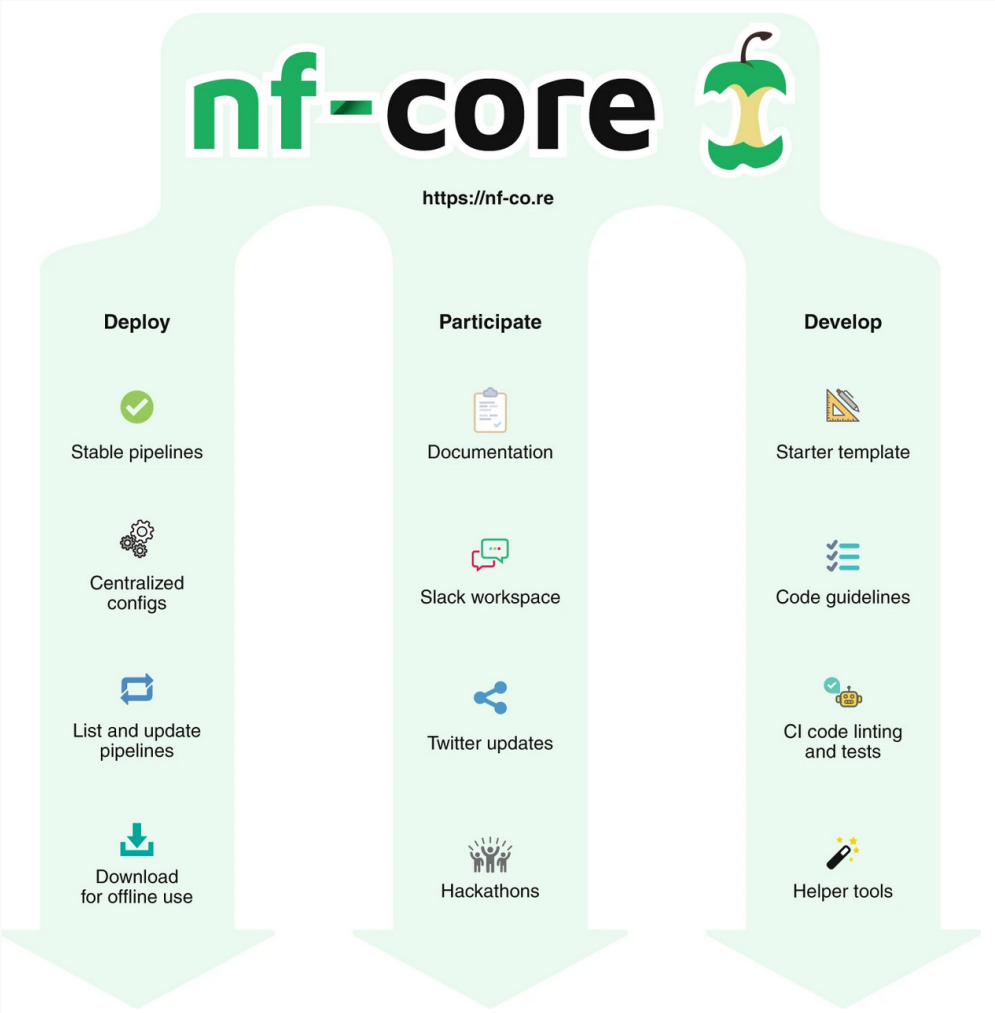



**Figure 8:** (Ewels et al. 2020)

# nf-core Getting started

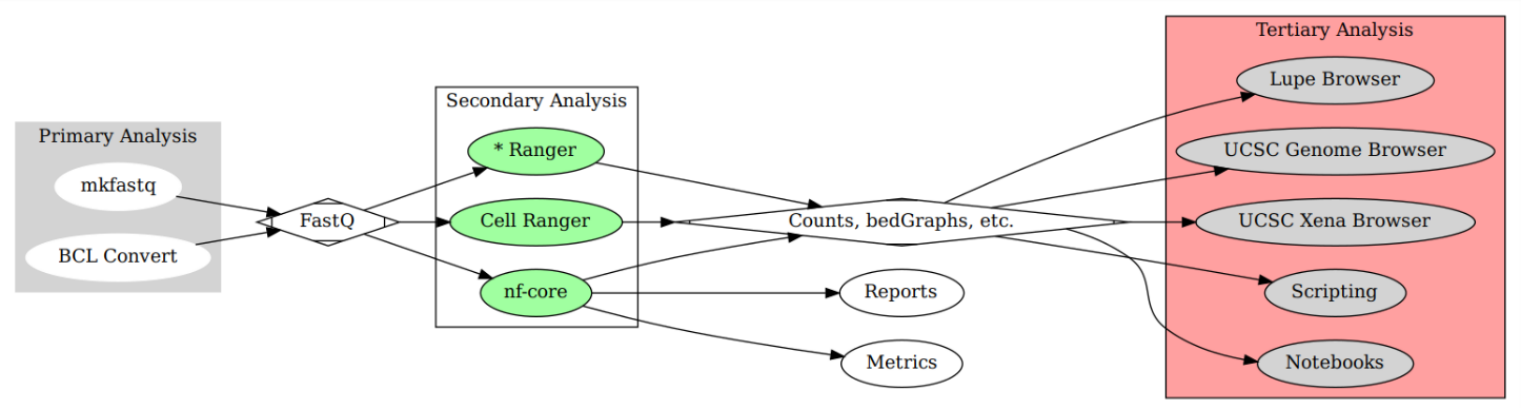
```
# Install nextflow
curl -s https://get.nextflow.io | bash
mv nextflow ~/bin/

# Launch the Nascent pipeline
nextflow run nf-core/nascent \
    --input samplesheet.csv \
    --genome GRCh38 \
    -profile docker
```

# Inheriting nf-core Nascent

- Breaking our analysis up into smaller pieces
- nf-core portion includes Quality Checks, alignment, graph generation, transcript identification, and transcript quantification
- Downstream analysis is then a seperate nextflow workflow
- Data engineering/Data Science split

# Primary-Secondary-Tertiary Analysis



# Nascent Goals

- Benchmark aligners to find best practices(Lots of opinions, no hard numbers)
- Handle alignment, QC, Genome graph generation, and naive transcript identification

# Aim 2: Take advantage of New Developments to improve eRNA annotation

# New developments

- CHM13 Released
- PINTS and transcriptional regulatory elements (TREs) matrix

# CHM13

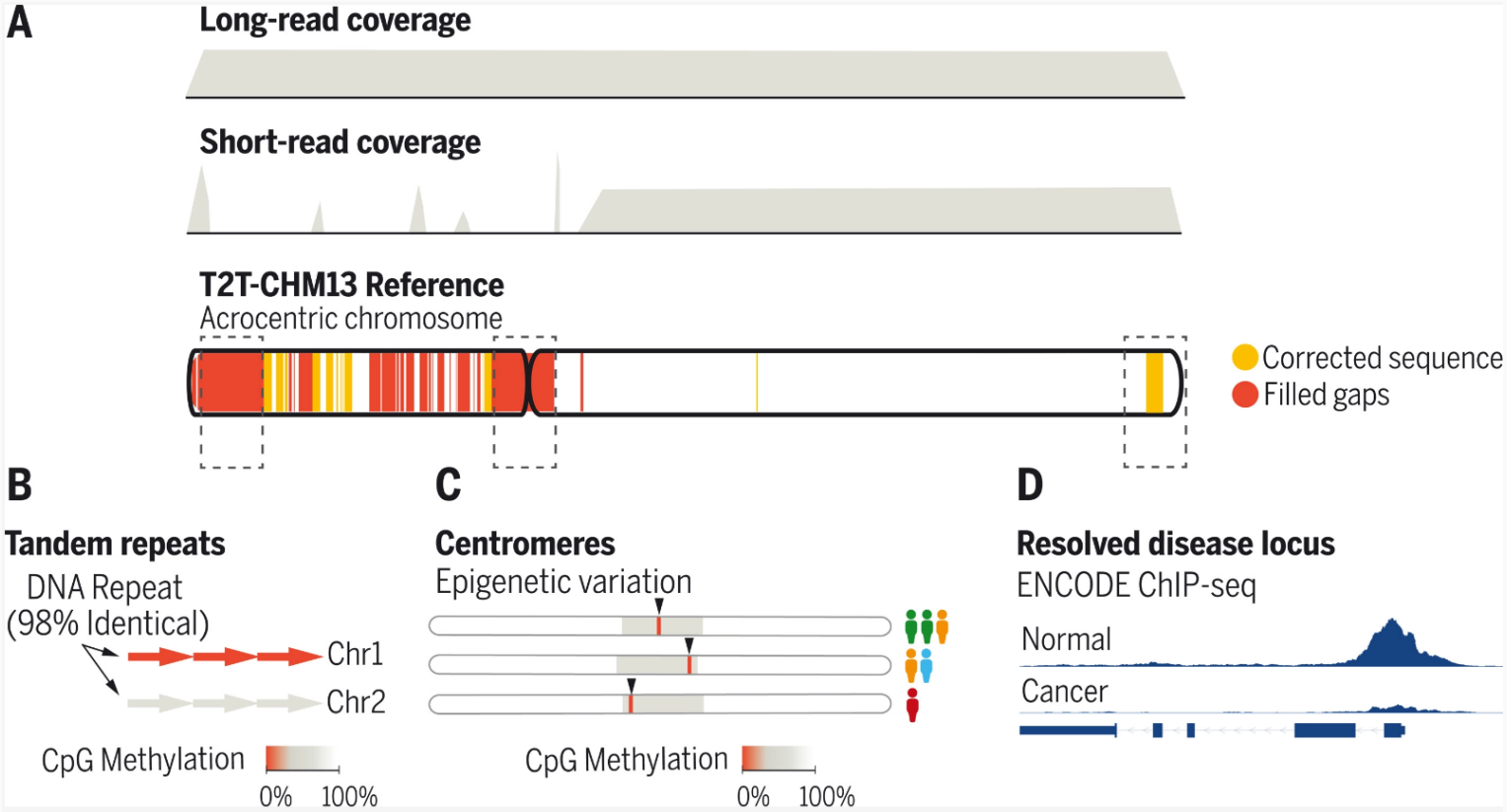



**Figure 9:** (Gershman et al. 2022)

# CHM13

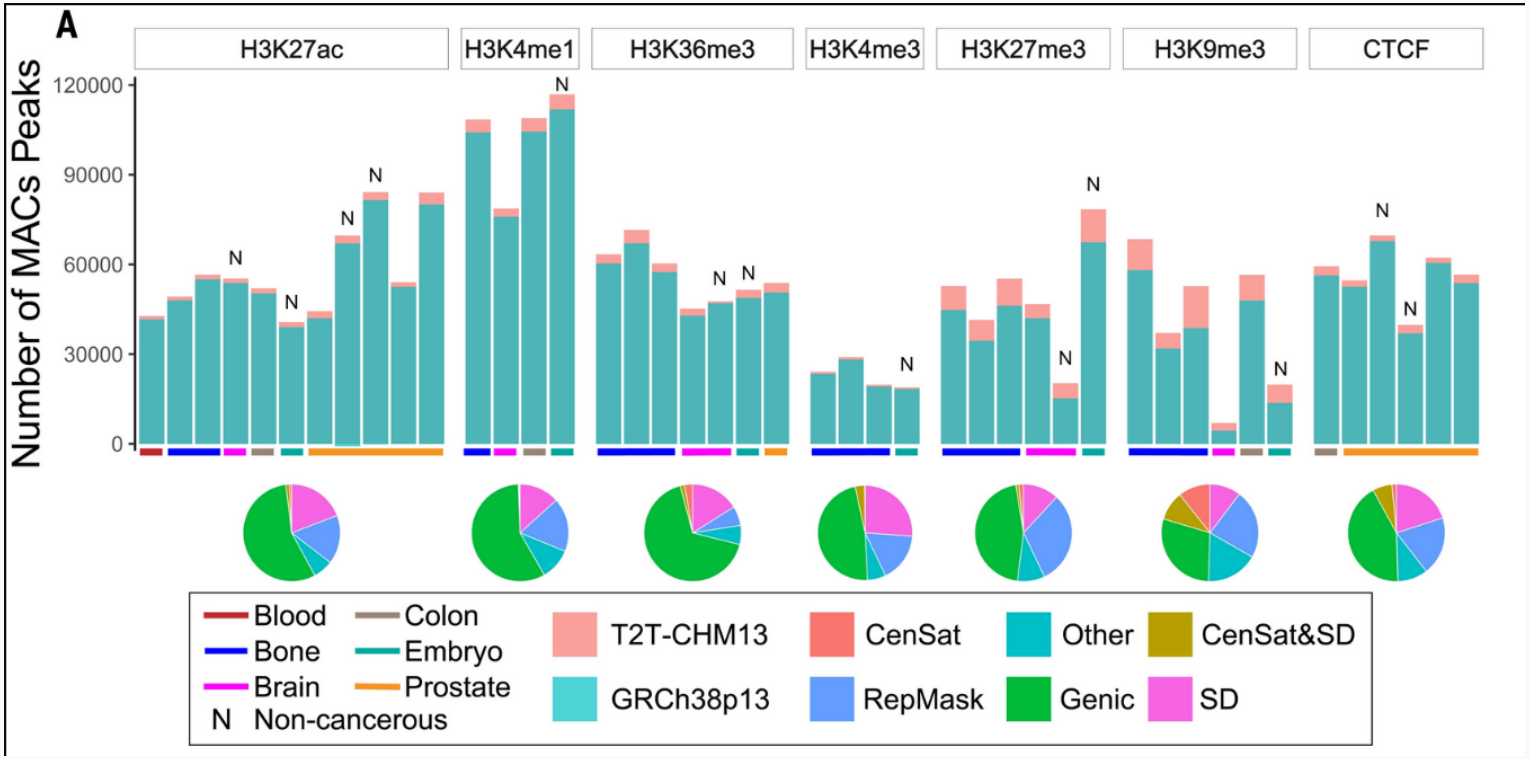



**Figure 10:** (Gershman et al. 2022)

# PINTS - different patterns of signals captured by TSS and NT Assays

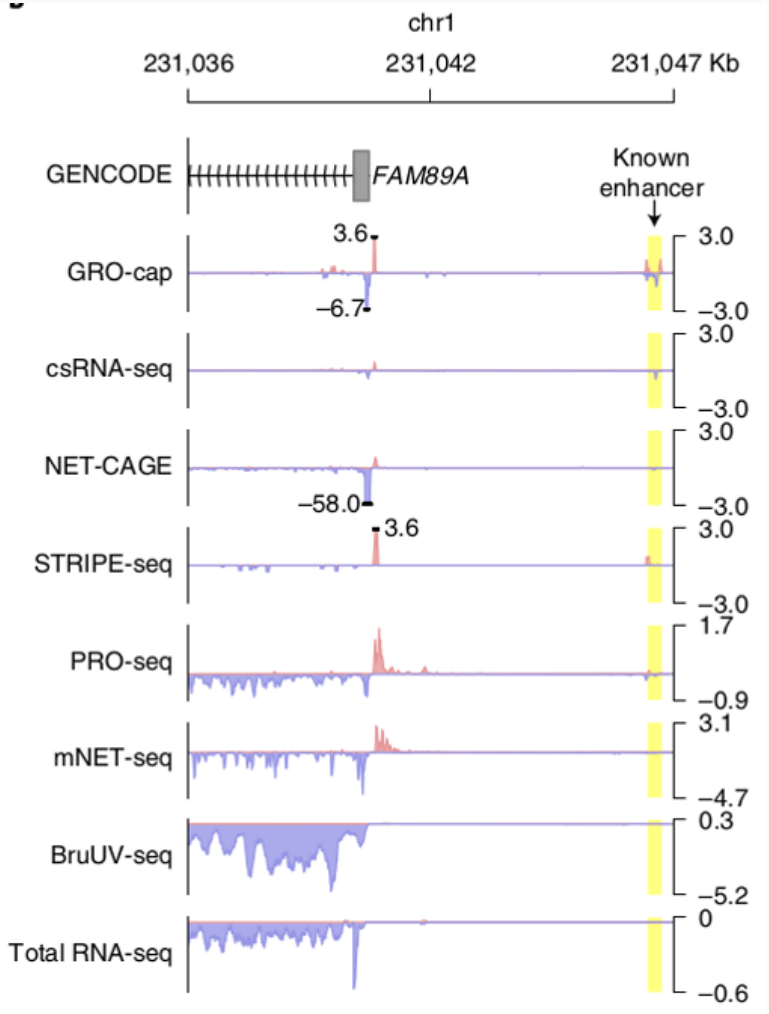



**Figure 11:** (Yao et al. 2022)

# PINTS - NT vs TSS assays

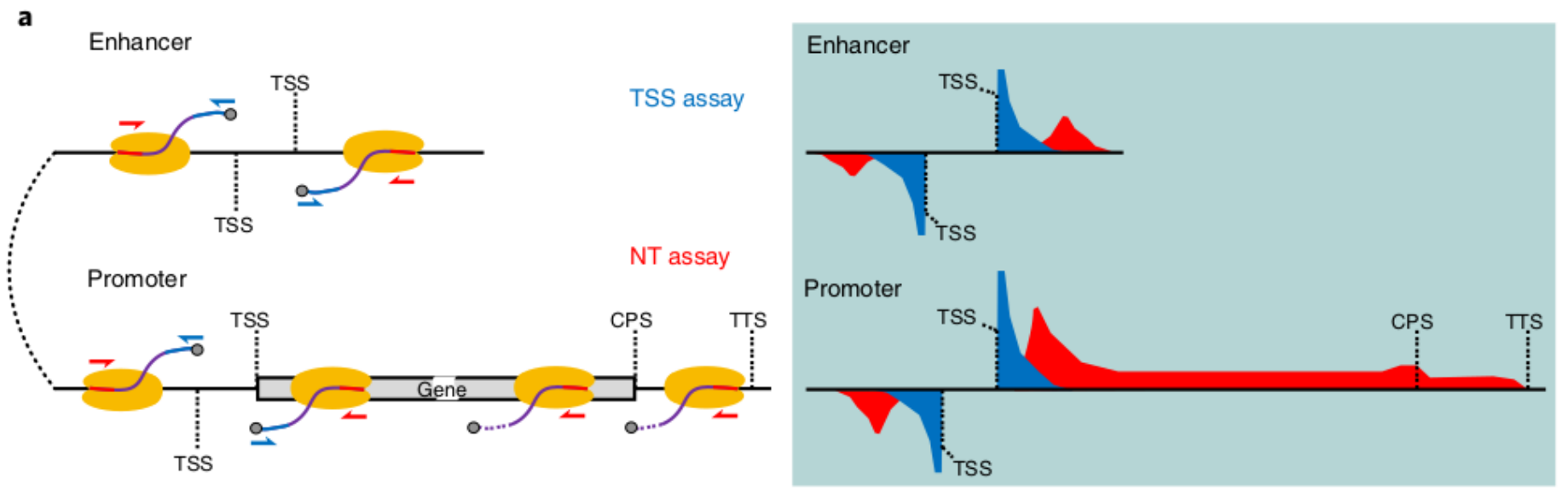



**Figure 12:** (Yao et al. 2022)

# PINTS

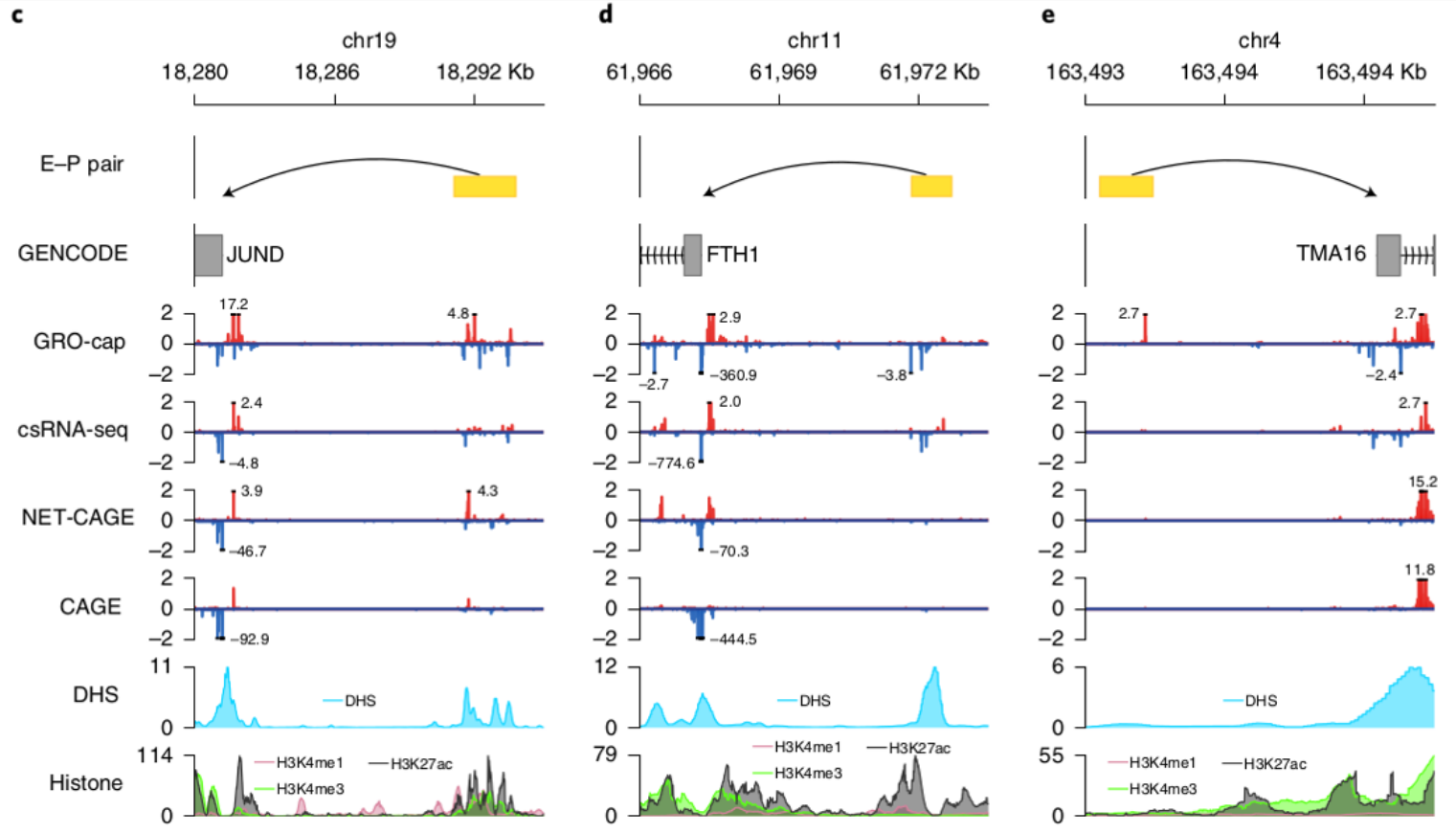



**Figure 13:** (Yao et al. 2022)

# PINTS

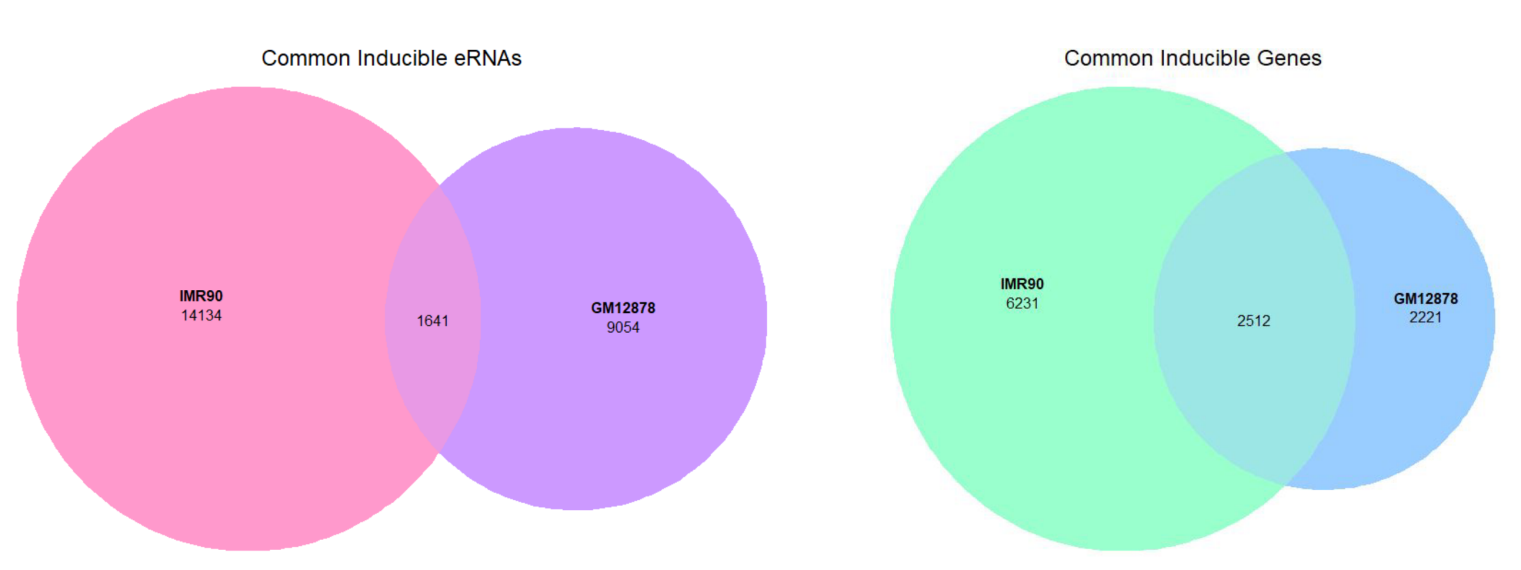
- Can we use PINTS for Nascent Transcript Assays?
- Can we swap naive method of selecting for Histone modications with PINTS indentified transcriptional regulatory elements (TREs)?
- Can we indetify full length transcripts from the Nascent Transcript assays?

# Aim 3 Compare eRNA dynamics between cell lines

# IMR and GM

- GM12878 - lymphoblastoid(immune response) cell line produced from the blood of a female donor (ENCODE Tier 1)
- IMR90 - fibroblasts(connective tissue) isolated from the normal lung tissue(ENCODE Tier 2.5)

# IMR and GM


Common Inducible eRNAs
IMR90
14134
1641
GM12878
9054
Common Inducible Genes
IMR90
6231
2512
GM12878
2221

# Cell-Type Specific eRNAs

| | |
|---|---|
| Common inducible genes: | 2512 |
| Common inducible genes with cell-type specific eRNAs: | 2053 |

- 81.7% Common inducible genes had cell-type specific eRNAs
- While they may have a common gene expression each cell line had their own unique way of solving the need.