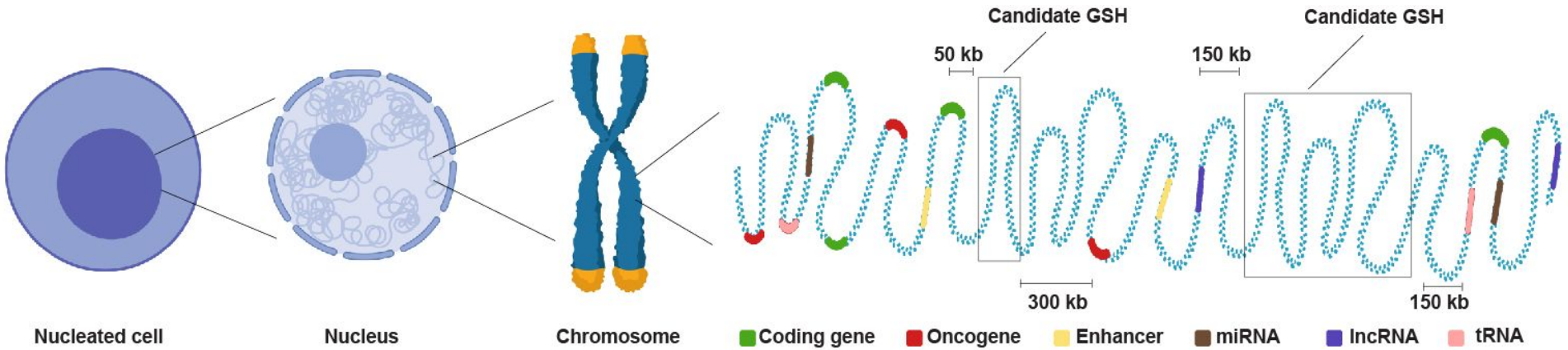
Computational identification of stable and active safe harbor sites in zebrafish

Hackathon project for ISMB 2025

What are genomic safe harbors (GSH)?

“Sites in the genome which can safely accommodate new genes without causing other, unintended changes in a cell’s genome”



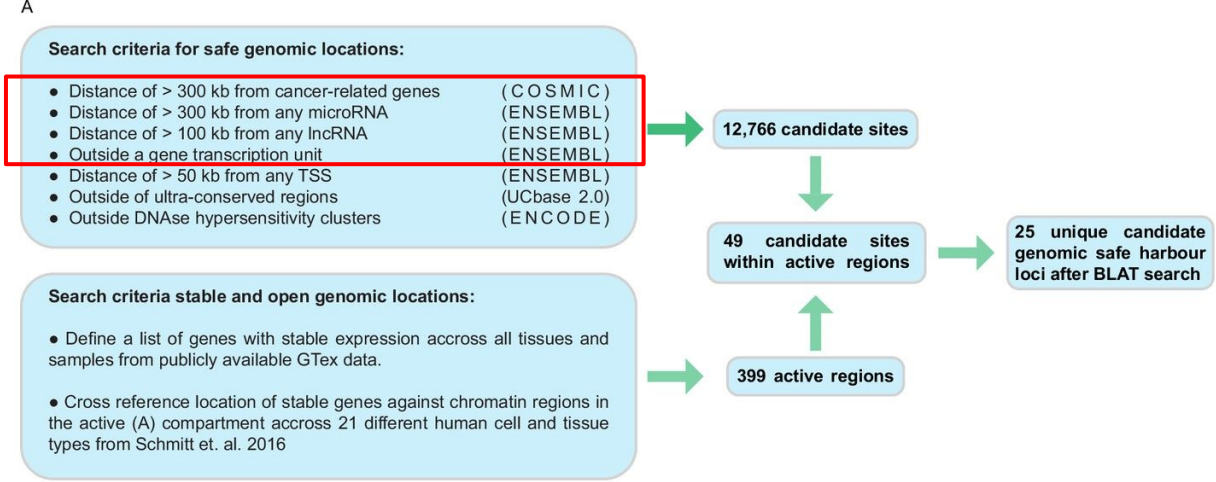
Aznauryan et al., 2022

Research Article

Stem Cells and Regenerative Medicine

Computationally defined and in vitro validated putative genomic safe harbour loci for transgene expression in human cells

IDEA

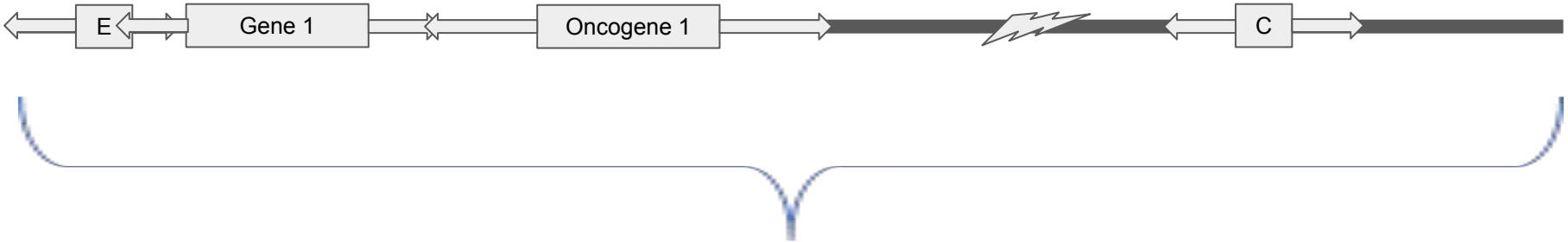


1. RNA-Seq multiple tissues: regions of the chromosome involved in active transcription
   a. Find genes actively transcribed at same levels across multiple tissues a.k.a low-variance housekeeping genes
   b. Cross-referenced these list of genes against open chromatin regions
2. Hi-C data from multiple tissues: open chromatin regions
   a. Chromatin regions located in open compartments in multiple tissues

Autio et al., 2024

Computational identification of stable and active safe harbor sites in zebrafish

Genes | Onco genes | miRNA genes | lncRNA genes | tRNA genes | Enhancer regions | Gaps

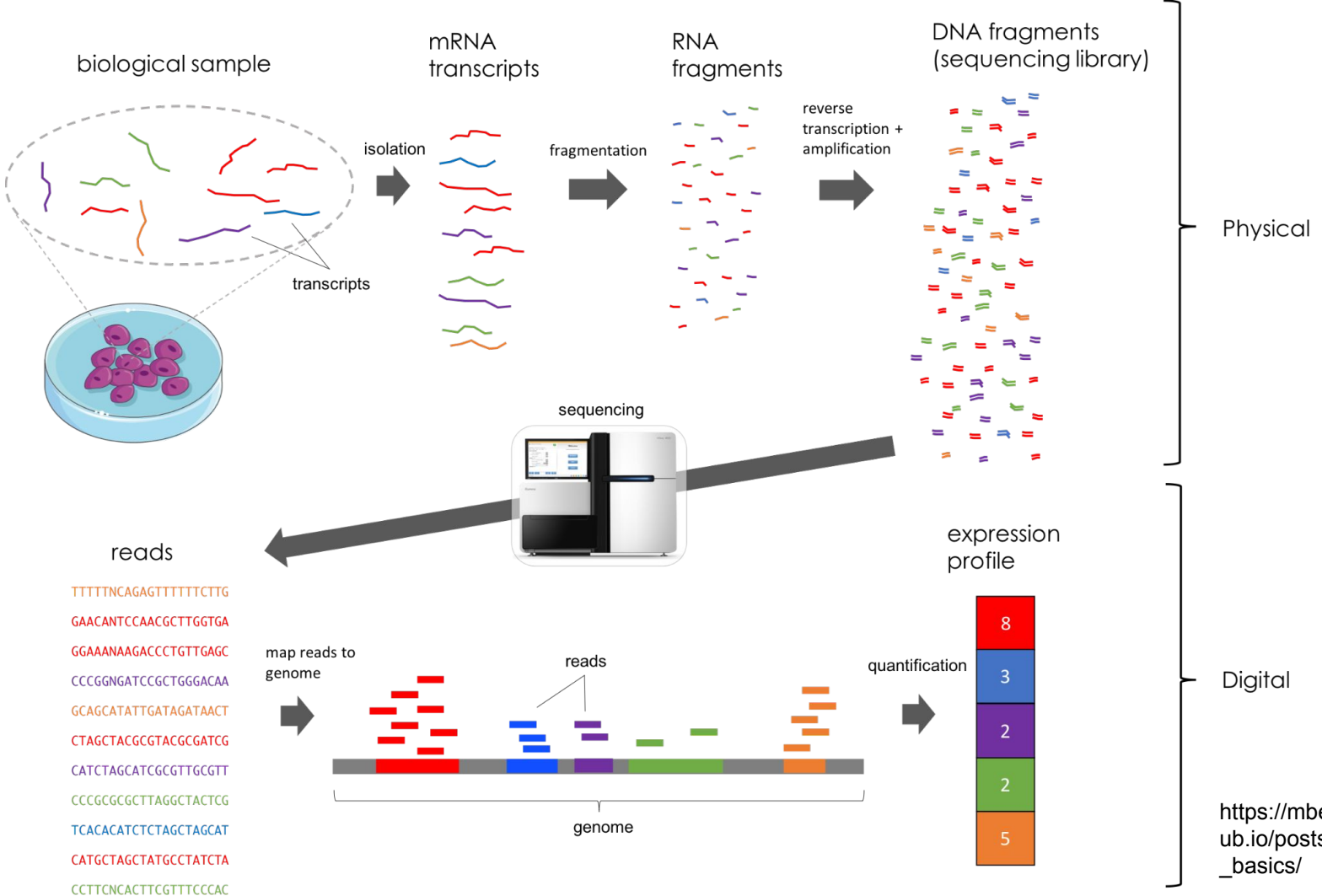


Safe harbor regions

Computational identification of stable and active safe harbor sites in zebrafish

1. Find which areas of the genome contain genes ubiquitously expressed across 4 different developmental timepoints

RNA-Seq


biological sample
mRNA transcripts
RNA fragments
DNA fragments (sequencing library)
isolation
fragmentation
reverse transcription + amplification
transcripts
Physical
sequencing
reads
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
CCCGCGCGCTTAGGCTACTCG
TCACACATCTCTAGCTAGCAT
CATGCTAGCTATGCCTATCTA
CCTTCNCACTTCGTTTCCCAC
map reads to genome
reads
genome
quantification
expression profile
8
3
2
2
5
Digital
https://mbernste.github.io/posts/rna_seq_basics/

Computational identification of stable and active safe harbor sites in zebrafish

A high-resolution mRNA expression time course of embryonic development in zebrafish.

White et al., 2017

We will provide…

1. Gene-level TPMs for 4 developmental timepoints in zebrafish

You will find out…

1. Genes with low TPM variability across different timepoints (expression levels do not change significantly)

Computational identification of stable and active safe harbor sites in zebrafish

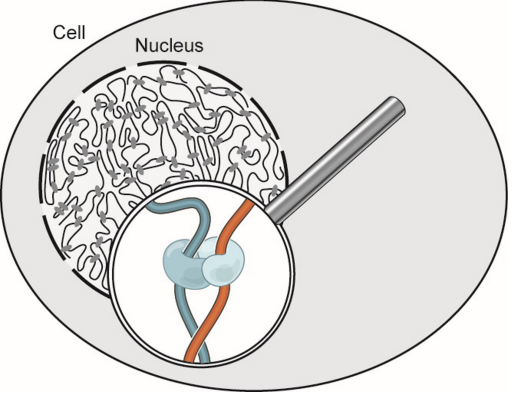
2. Find which areas of the chromatin are open and active in 4 developmental timepoints

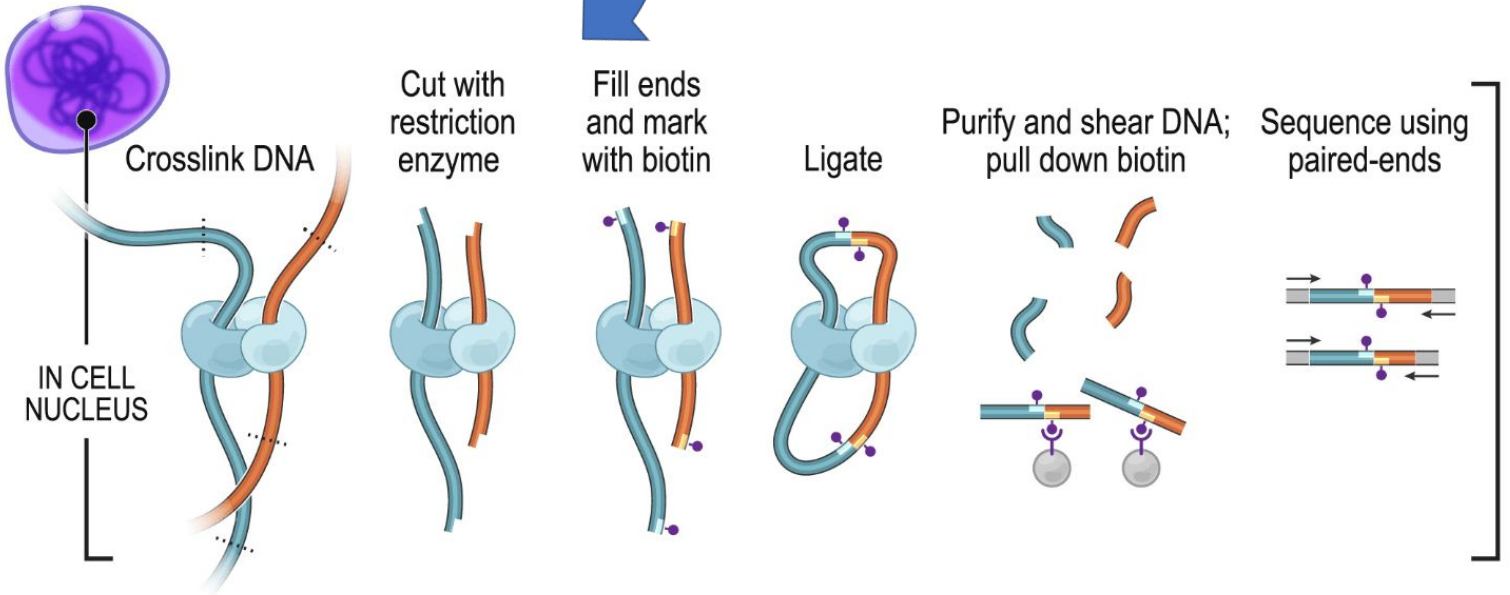
Hi-C

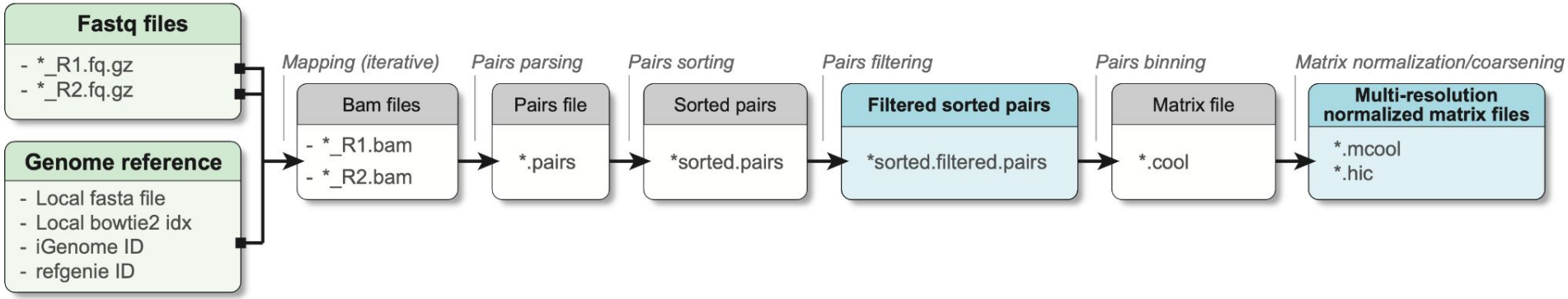
› Genome Res. 2021 Jun;31(6):981-994. doi: 10.1101/gr.269860.120. Epub 2021 May 18.

Chromatin architecture transitions from zebrafish sperm through early embryogenesis

Wike et al., 2021


Cell
Nucleus


Crosslink DNA
IN CELL NUCLEUS
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends


Fastq files
- *_R1.fq.gz
- *_R2.fq.gz
Genome reference
- Local fasta file
- Local bowtie2 idx
- iGenome ID
- refgenie ID
Mapping (iterative)
Bam files
- *_R1.bam
- *_R2.bam
Pairs parsing
Pairs file
*.pairs
Pairs sorting
Sorted pairs
*sorted.pairs
Pairs filtering
Filtered sorted pairs
*sorted.filtered.pairs
Pairs binning
Matrix file
*.cool
Matrix normalization/coarsening
Multi-resolution normalized matrix files
*.mcool
*.hic

Processing paired-end reads will give a .pairs file

dummy.pairs

```
## pairs format v1.0
#sorted: chr1-chr2-pos1-pos2
#columns: readID chr1 pos1 chr2 pos2 strand1 strand2
#chromsize: chr1 389
. chr1 162 chr1 172 . .
. chr1 180 chr1 192 . .
. chr1 183 chr1 254 . .
. chr1 221 chr1 273 . .
. chr1 254 chr1 298 . .
```

dummy.pairs

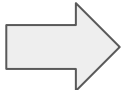
```
## pairs format v1.0
#sorted: chr1-chr2-pos1-pos2
#columns: readID chr1 pos1 chr2 pos2 strand1 strand2
#chromsize: chr1 389
. chr1 162 chr1 172 . .
. chr1 180 chr1 192 . .
. chr1 183 chr1 254 . .
. chr1 221 chr1 273 . .
. chr1 254 chr1 298 . .
```

bins.bed

<chr>	<pos>	<bin>
chr1	1	100
chr1	101	200
chr1	201	300
chr1	301	389

<chr1>	<pos1>	->	<bin1>
chr1	162	->	2
chr1	180	->	2
chr1	183	->	2
chr1	221	->	3
chr1	254	->	3

<chr2>	<pos2>	->	<bin2>
chr1	172	->	chr1 2
chr1	192	->	chr1 2
chr1	254	->	chr1 3
chr1	273	->	chr1 3
chr1	298	->	chr1 3



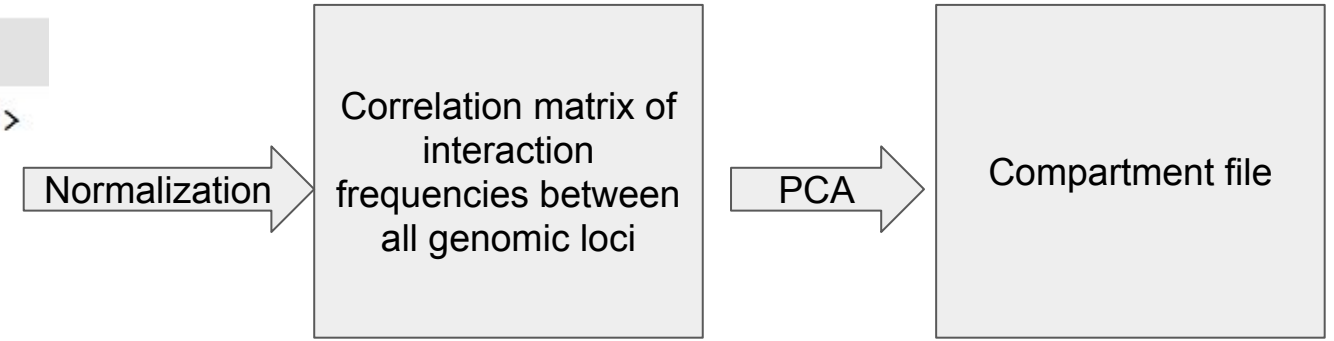
<bin1>	<bin2>
2	2
2	2
2	3

count.matrix

<bin1>	<bin2>	<count>
2	2	2
2	3	1
3	3	2

count.matrix

<bin1>	<bin2>	<count>
2	2	2
2	3	1
3	3	2



We will provide…

1. Compartments file for 4 developmental timepoints in zebrafish

You will find out…

1. Active (open) chromatin regions across different timepoints

Put together 1 and 2

Cross-reference genomic coordinates of ubiquitously expressed low-variance genes with regions of open chromatin

This will give…

List of consistently active chromatin regions where it would be ideal to insert a foreign gene. Which of our identified safe harbor sites fall in these regions?