Network analysis of DNA Repair phenotype using database of nano-biomimetic based single cell assay

Abolfazl Arab

Nano-biomimetic student,

Life Science Engineering,

University of Tehran

Advisors:

Dr. Faramarz Mehrnejad

Dr. Sama Goliaei

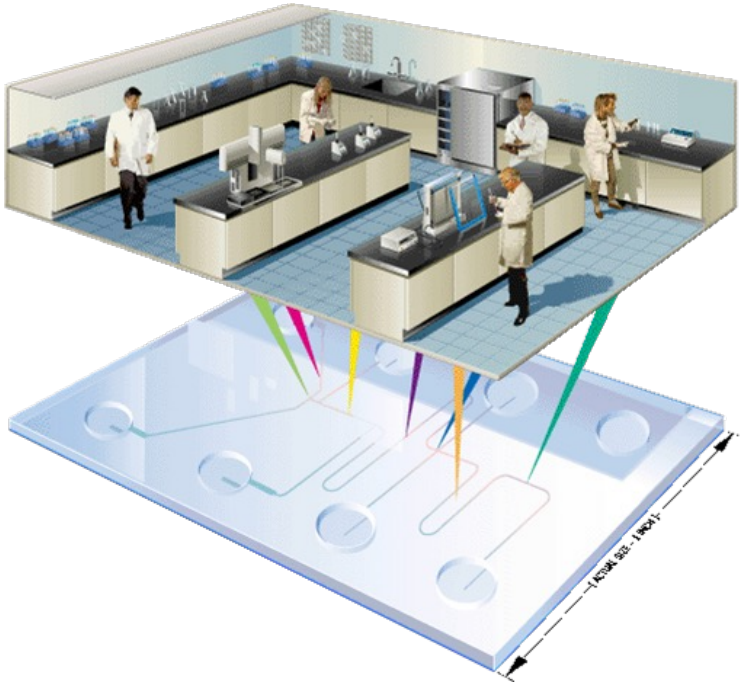
BIOLOGY IN SINGLE-CELL RESOLUTION

How and why technology enable high-throughput bio-assays in single cell resolution?

μ-TAS concept: miniaturized Total Analysis System

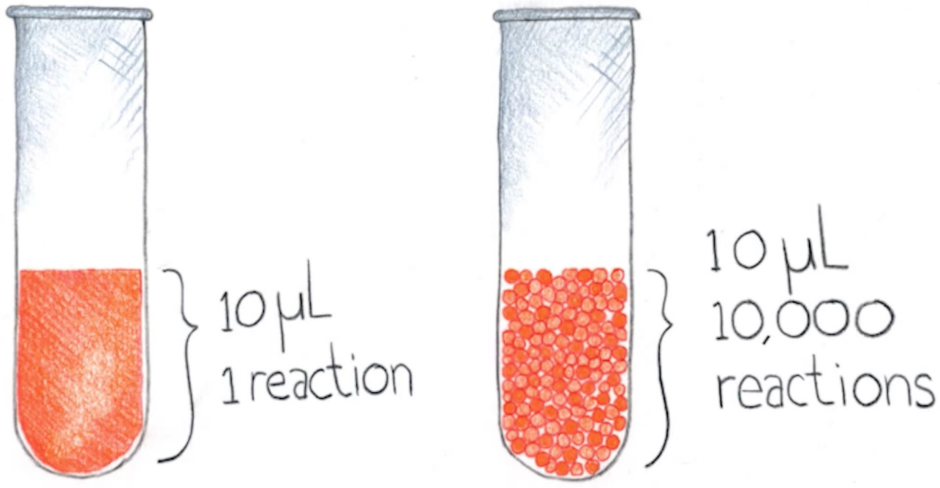
- If the device in question had characteristic dimensions on the microscale.
- A system that could automatically carry out all the functions required for analysis.

- Sampling
- Transport of the sample
- Any sample preparation steps
  - Ex. chemical reactions, separations, etc.
- Detection

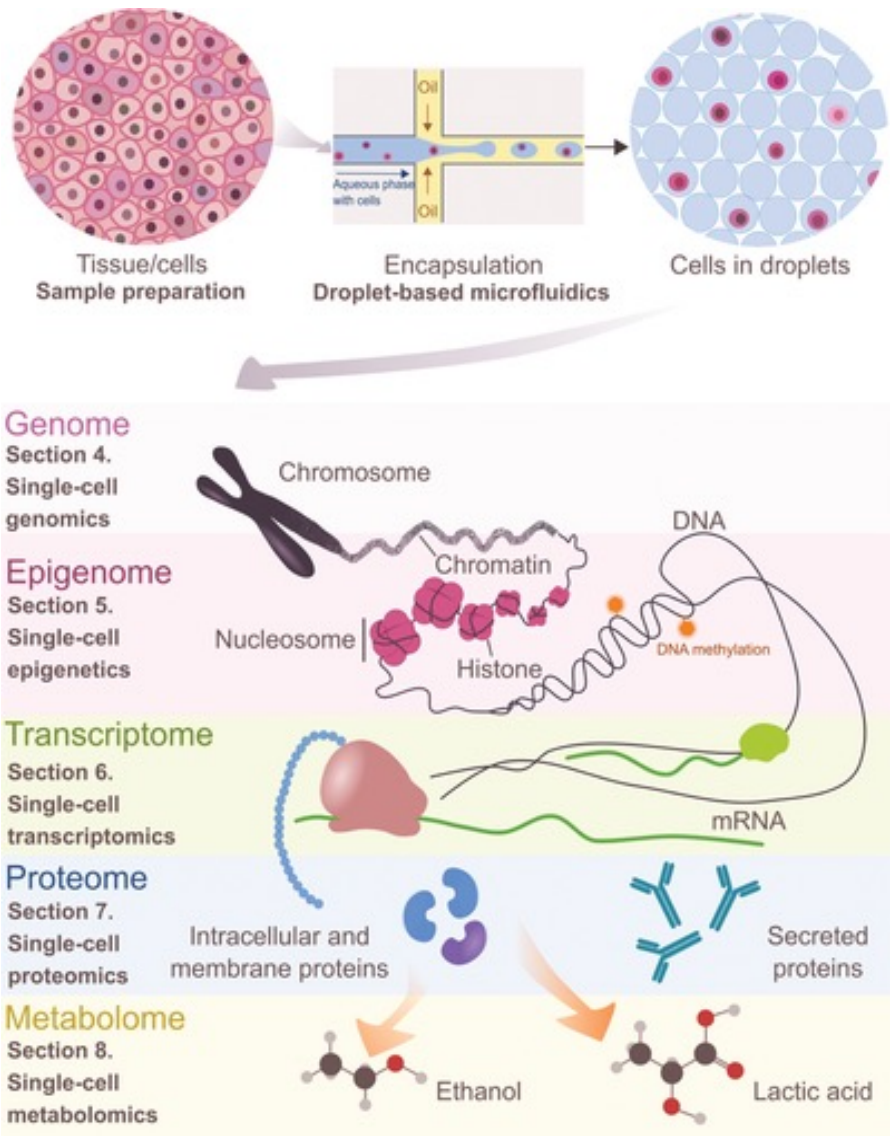


Manz et al. (1990)

Single-Cell Analysis Using Droplet Microfluidics

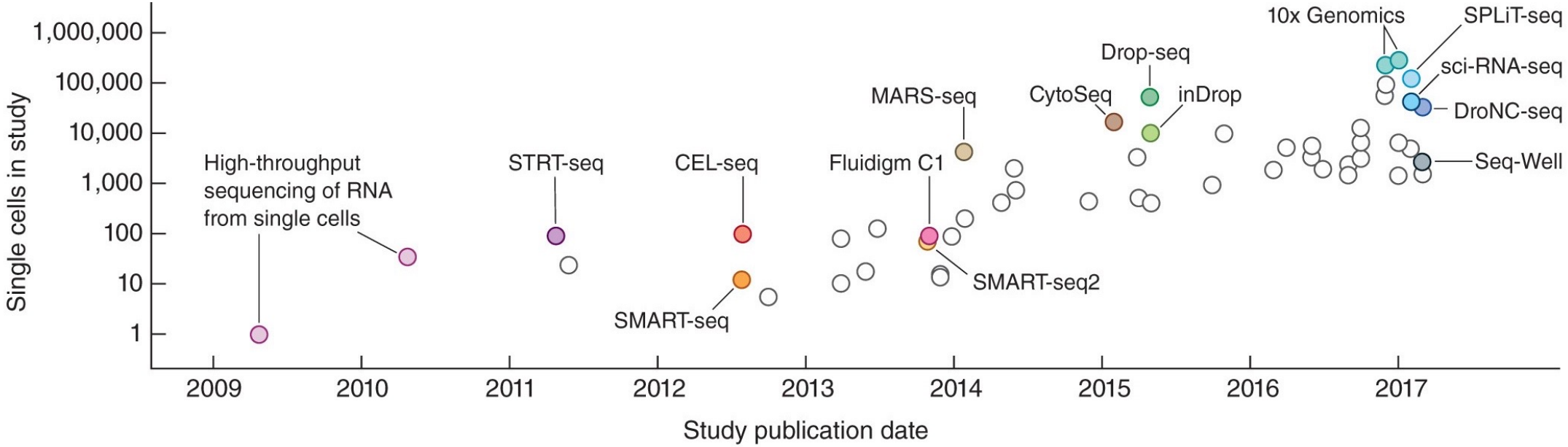


Macosko, E. et al. (2015)



Matuła, K., et.al (2020) | Review

Scaling of scRNA-seq experiments



Svensson, V., et al. (2018)

SINGLE CELL DNA-REPAIR MEASUREMENT

NGS & droplet microfluidic platforms enable high throughput measurement of biochemical phenotypes in single cells.

Nucleic Acids Research

Published online 14 April 2020

Nucleic Acids Research, 2020, Vol. 48, No. 10 ***e59***
doi: 10.1093/nar/gkaa240

Simultaneous measurement of biochemical phenotypes and gene expression in single cells

Amanda L. Richer[1,2], Kent A. Riemondy[3], Lakotah Hardie[1] and Jay R. Hesselberth[1,2,3,*]

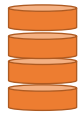
[1]Department of Biochemistry and Molecular Genetics, Aurora, CO 80045, USA, [2]Molecular Biology Program and [3]RNA Bioscience Initiative, University of Colorado School of Medicine, Aurora, CO 80045, USA

Received January 23, 2020; Revised March 16, 2020; Editorial Decision March 31, 2020; Accepted April 01, 2020

Method

Experimental molecular assay

Dataset

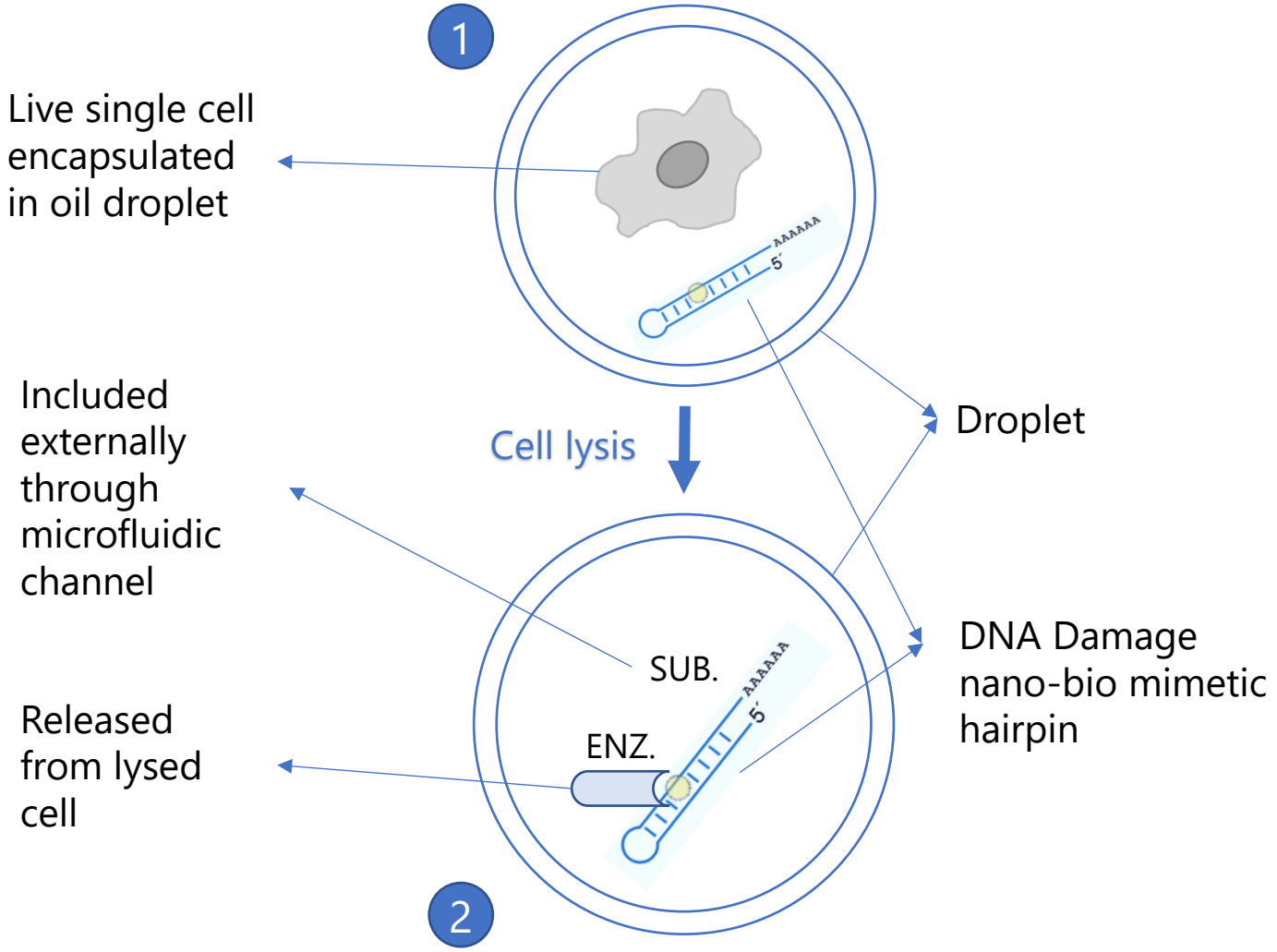


Several single-cell RNA-seq experiments

Proof of concept

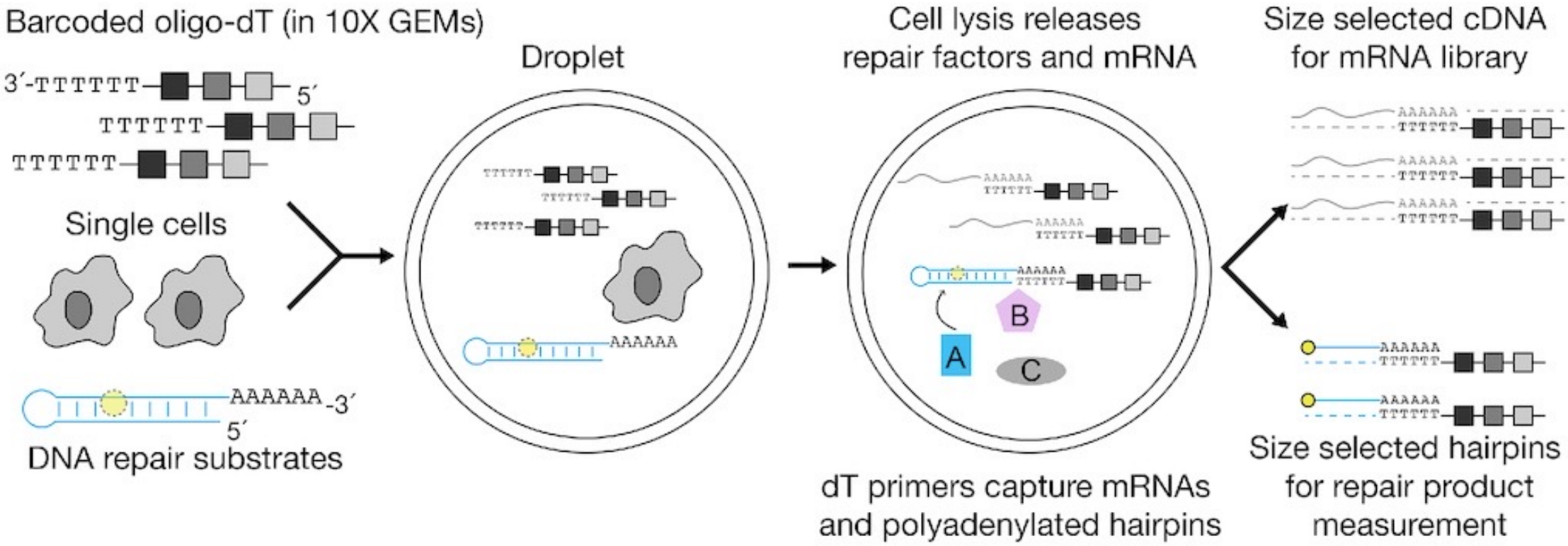
Nucleic-based ***nano-biomimetic*** probes are powerful paradigm to design novel molecular assays

Mimic DNA repair enzyme-substrate reaction inside droplet



- An external synthetic hairpin with a single lesion damage at certain position included into droplets using the same channel which cell loaded to the microfluidic chip
- This hairpin *mimic* substrate of DNA repair enzymes which released from the cell during in-droplet lysis
- Overall, this protocol made it possible to simply measure amount of enzymatic activity (i.e., number of strand incisions) alongside with mRNA abundance in single cell resolution

Measuring DNA-repair enzyme activity in single-cell resolution



Mixing and time series experiment

- KO cells were identified if counts at the repair site (position 44 for ribonucleotide and position 45 for uracil)
- After the emulsion was created, the sample was separated into 3 tubes and incubated for 15, 30, or 60 min at 37℃ prior to reverse transcription at 53℃.
- 800-1,500 cells were captured at each timepoint.
- DNA repair measurements determine ***cell types*** in a cell mixing experiment.
- Authors showed it fails to use UNG and RNASEH2C mRNA expression to determine cell types, but estimated repair activity clearly assign cell-types.

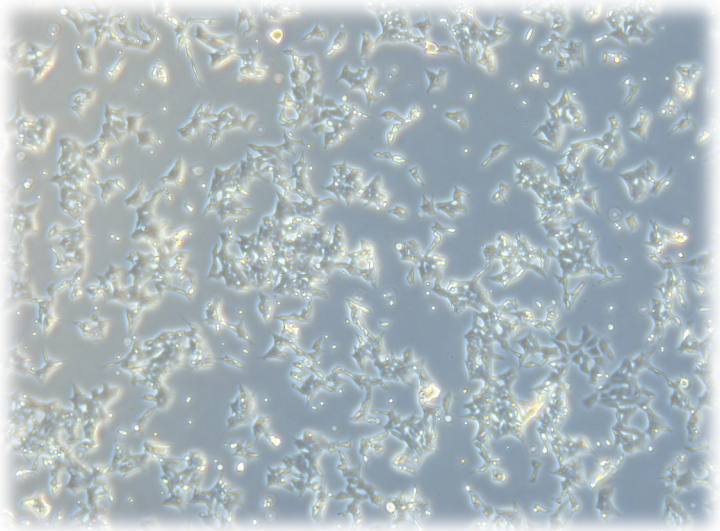
What else we can interpret from this experiment?

Differential expression analysis

Pathway enrichment analysis

Gene regulatory network analysis

Assess HAP1 cell line with Knock-out genes



Disease
Engineered

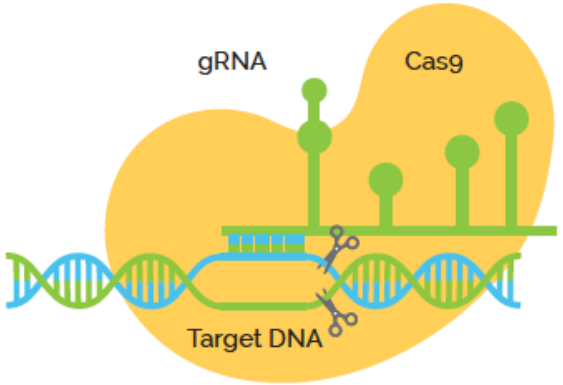
Disease Subtype
Chronic Myelogenous Leukemia (CML)

Lineage
Engineered Blood

Lineage Subtype
CML

Source
Horizon Discovery ™

Gender
Male



RNASEH2C KO

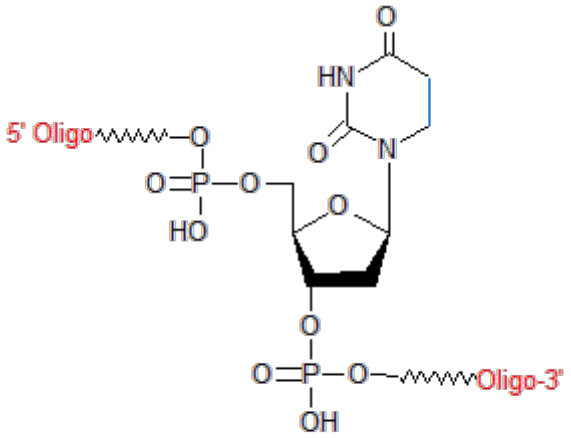
Ribonuclease H2 Subunit C
- 164 amino acids

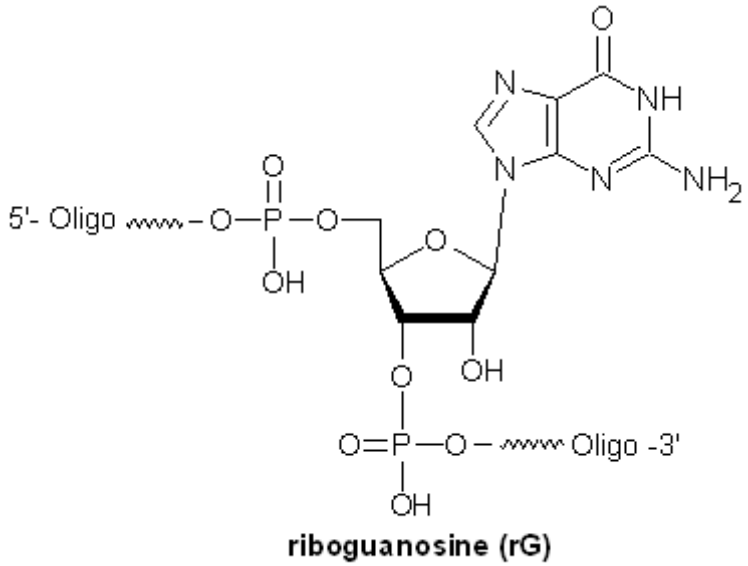
UNG KO

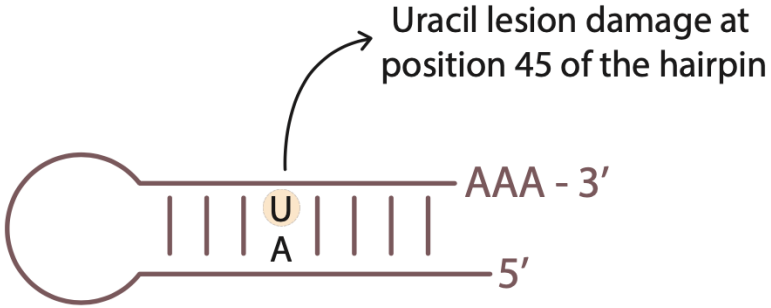
Uracil DNA Glycosylase
- 313 amino acids

Cellosaurus HAP1 RNASEH2C (-) | Cellosaurus HAP1 UNG (-)

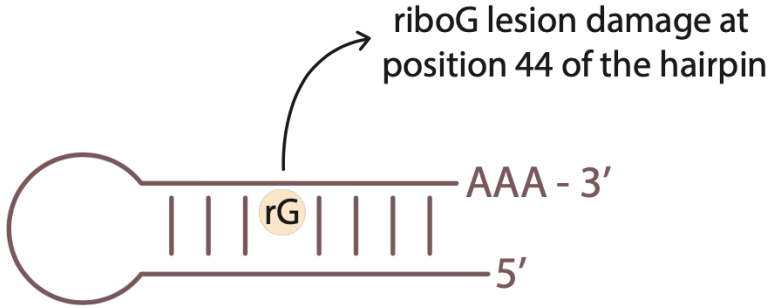
Synthetic Hairpins as the mimic of DNA repair enzyme substrate







dU-damaged hairpin



rG-damaged hairpin

RNASEH2C

Ribonuclease H2 Subunit C

UNG

Uracil DNA Glycosylase

RNAseH2C - Ribonuclease H2 Subunit C

- Non catalytic subunit of RNase H2, an endonuclease that specifically **degrades the RNA of RNA:DNA hybrids** and mediates the excision of single ribonucleotides from DNA:RNA duplexes.
- Participates in DNA replication, possibly by mediating the removal of lagging-strand Okazaki fragment RNA primers.
- Ribonucleotides are incorporated into DNA by the replicative DNA polymerases at frequencies of about 2 per kb which makes them by far the **most abundant form of potential DNA damage in the cell**.
- Their removal is essential for restoring a stable intact chromosome.

UNG - Uracil DNA Glycosylase

- Belongs to the uracil-DNA glycosylase (UDG) superfamily
- Excises uracil residues from the DNA which can arise as a result of misincorporation of **dUMP** residues by DNA polymerase or due to deamination of cytosine;
- UNG is the major **uracil-DNA glycosylase** in mammalian cells and is involved in both
  - Error-free base excision repair of genomic uracil
  - Mutagenic uracil-processing at the antibody genes.
- The regulation of UNG in these different processes is currently *not well understood*.

RNAseH2C - Ribonuclease H2 Subunit C

- Non catalytic subunit of RNase H2, an endonuclease that specifically **degrades the RNA of RNA:DNA hybrids** and mediates the excision of single ribonucleotides from DNA:RNA duplexes.
- Participates in DNA replication, possibly by mediating the removal of lagging-strand Okazaki fragment RNA primers.
- Ribonucleotides are incorporated into DNA by the replicative DNA polymerases at frequencies of about 2 per kb which makes them by far the **most abundant form of potential DNA damage in the cell**.
- Their removal is essential for restoring a stable intact chromosome.

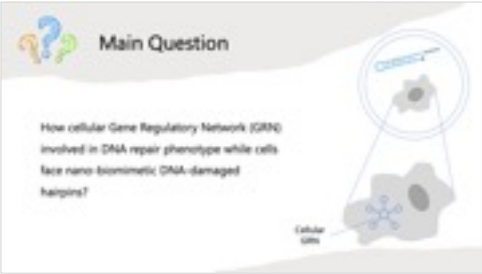
UNG - Uracil DNA Glycosylase

- Belongs to the uracil-DNA glycosylase (UDG) superfamily
- Excises uracil residues from the DNA which can arise as a result of misincorporation of **dUMP** residues by DNA polymerase or due to deamination of cytosine;

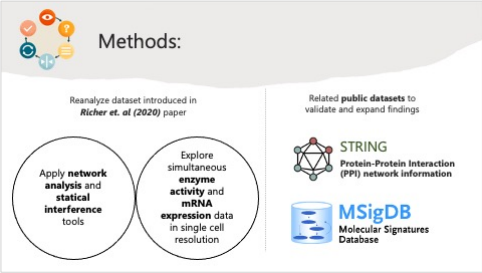
- UNG is the major **uracil-DNA glycosylase** in mammalian cells and is involved in both
  - Error-free base excision repair of genomic uracil
  - Mutagenic uracil-processing at the antibody genes.
- The regulation of UNG in these different processes is currently *not well understood*.

Thesis Overview

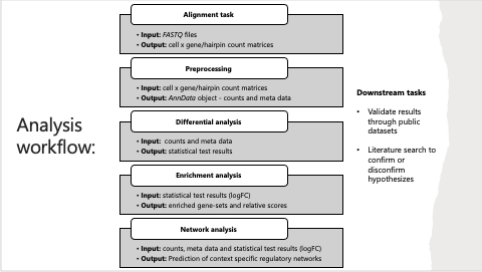
Main Question:



Methods:



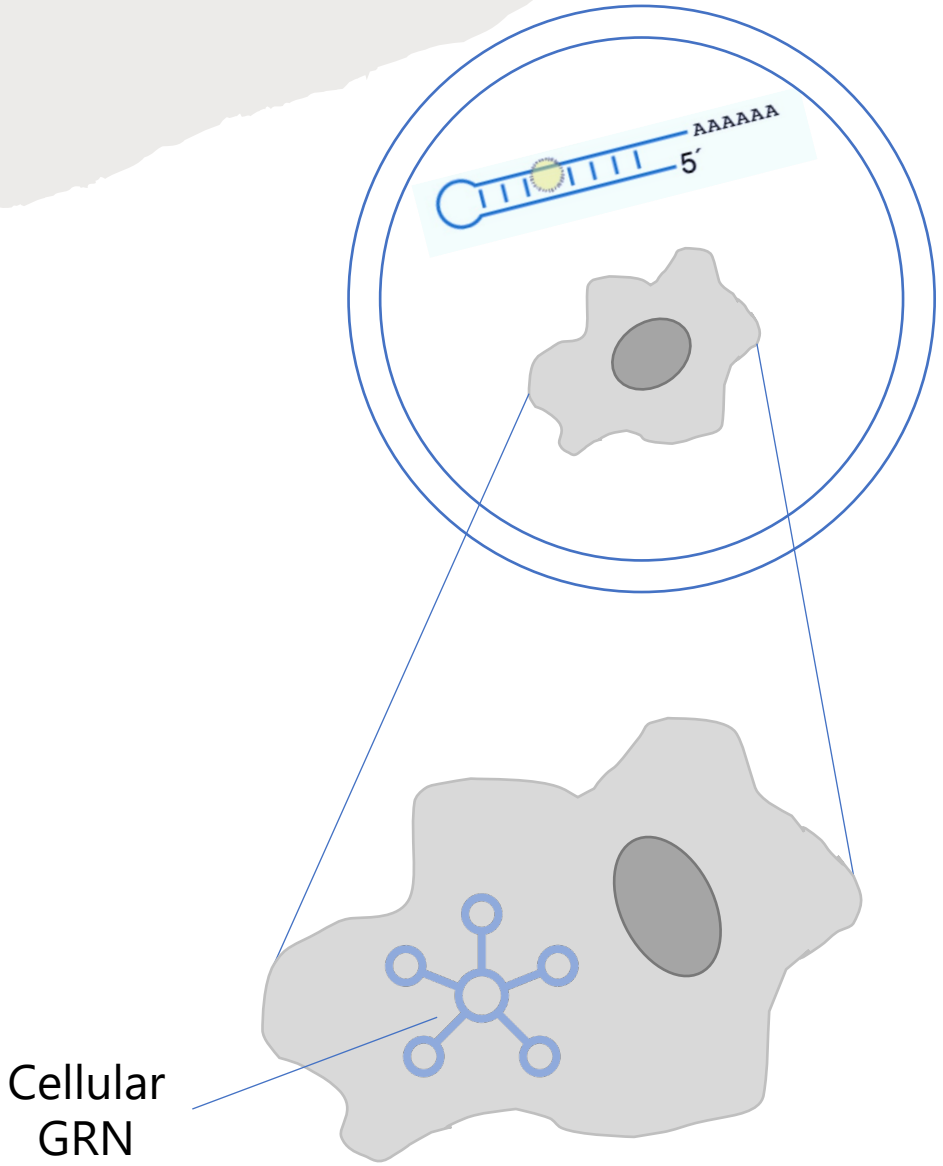
Workflows:



Results:

Main Question

How cellular Gene Regulatory Network (GRN) involved in DNA repair phenotype while cells face nano-biomimetic DNA-damaged hairpins?



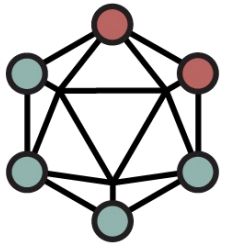
Methods:

Reanalyze dataset introduced in ***Richer et. al (2020)*** paper

Apply **network analysis** and **statical interference** tools

Explore simultaneous **enzyme activity** and **mRNA expression** data in single cell resolution

Related **public datasets** to validate and expand findings



STRING

Protein-Protein Interaction (PPI) network information

MSigDB

Molecular Signatures Database

Analysis workflow:

Alignment task

- **Input:** *FASTQ* files
- **Output:** cell x gene/hairpin count matrices

Preprocessing

- **Input:** cell x gene/hairpin count matrices
- **Output:** *AnnData* object - counts and meta data

Differential analysis

- **Input:** counts and meta data
- **Output:** statistical test results

Enrichment analysis

- **Input:** statistical test results (logFC)
- **Output:** enriched gene-sets and relative scores

Network analysis

- **Input:** counts, meta data and statistical test results (logFC)
- **Output:** Prediction of context specific regulatory networks

Downstream tasks

- Validate results through public datasets
- Literature search to confirm or disconfirm hypothesizes

Results:

Basic Analysis

- **Alignment task**
- **Preprocessing**
- Differential analysis
- Enrichment analysis
- Network analysis

Comparison Analysis

- Alignment task
- Preprocessing
- **Differential analysis**
- **Enrichment analysis**
- Network analysis

Network and Graph Analysis

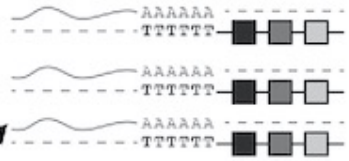
- Alignment task
- Preprocessing
- Differential analysis
- Enrichment analysis
- **Network analysis**

Basic Analysis

- **Alignment task**
- **Preprocessing**
- Differential analysis
- Enrichment analysis
- Network analysis

Alignment task

Size selected cDNA for mRNA library



Human genome

GENCODE

ENCODE, the Encyclopedia Of DNA Elements

A project to identify all functional elements in the human genome sequence.

Aligner algorithm

kallisto | bustools

A workflow for pre-processing single cell RNA-seq data.

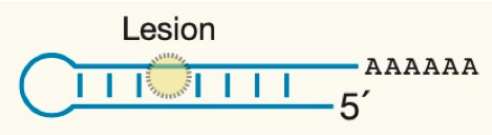
cell x gene matrix

Counts matrices + metadata

AnnData object

Size selected hairpins for repair product measurement

Hairpin (sudo-substrate)



*Hairpin sequences in **Fasta** format used to enable alignment task.*

Aligner algorithm

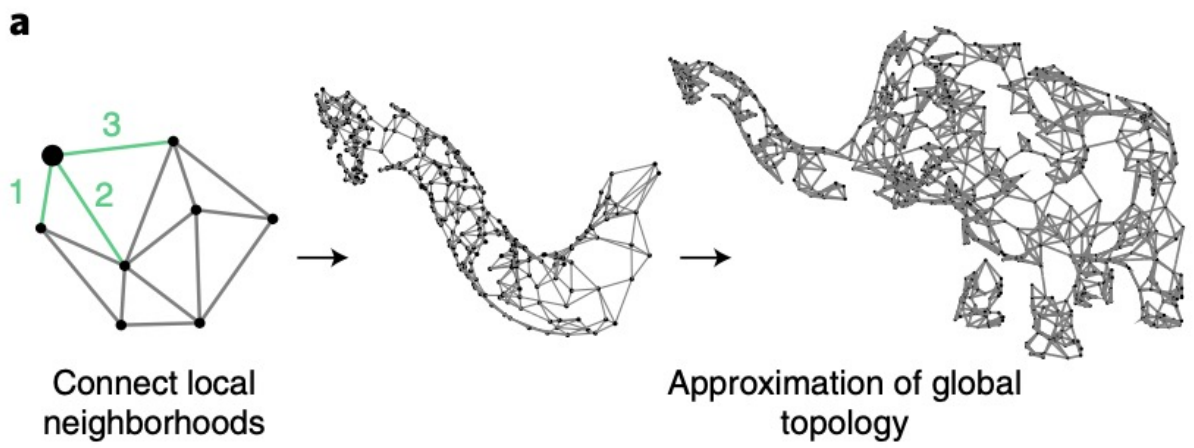
Snakemake

hesselberthlab / sc-haircut

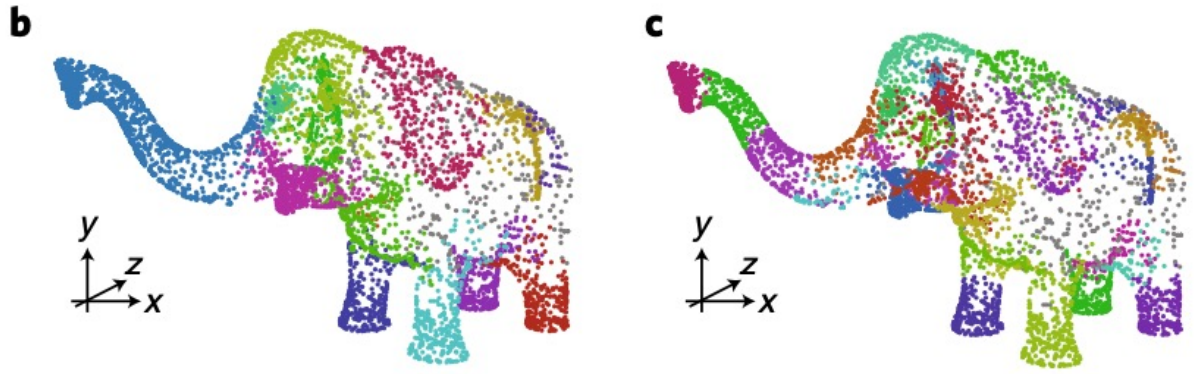
Snakemake pipeline to count functional data

cell x hairpin matrix

Approximating and partitioning complex manifolds



a, Complex, curved surfaces can be well approximated by neighborhood graphs. A simple graph connects each point with its *k* closest neighbors (kNN graph). As more points and regions are measured, the complex structure of the object can be revealed.



b, The elephant graph (in **a**) is clustered using the **Leiden clustering algorithm** (resolution $r = 0.5$). The resulting clusters are shown as colors on the 3D model (top) and *t*-SNE embedding (bottom) of the data.
c, Clustering resolution is arbitrary. Similar to **b**, the plots show clustering with increased resolution ($r = 3$). The clusters are smaller but capture equally valid anatomical elements.

Kharchenko, P.V. *Nat Methods* (2021).

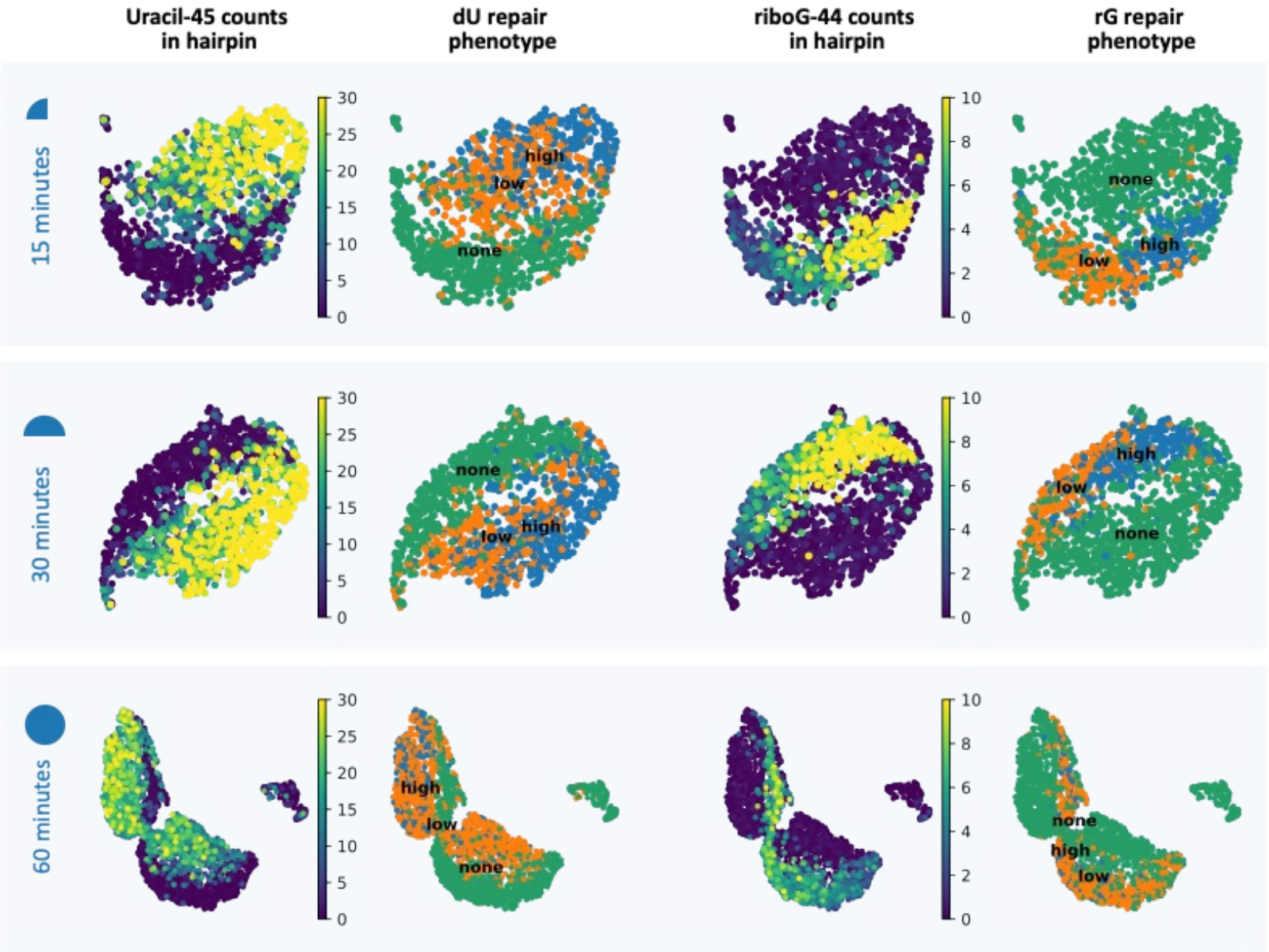
Single-Cell Analysis in Python

scanpy

UMAP plots

counts of hairpins with lesion damage (left)

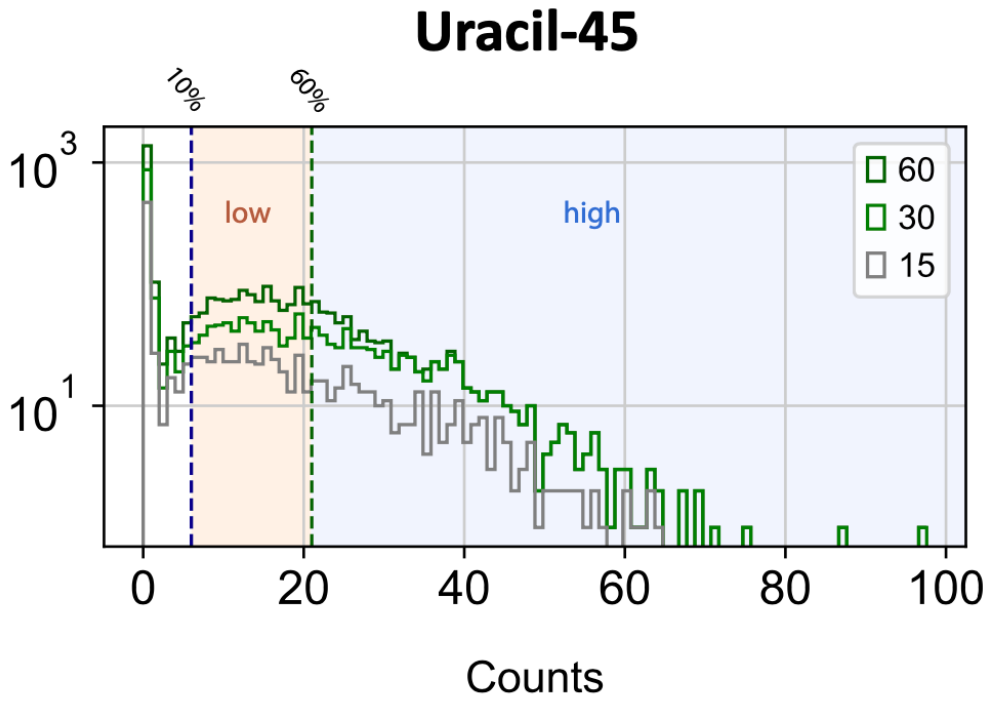
label group of cells for repair phenotype (right)

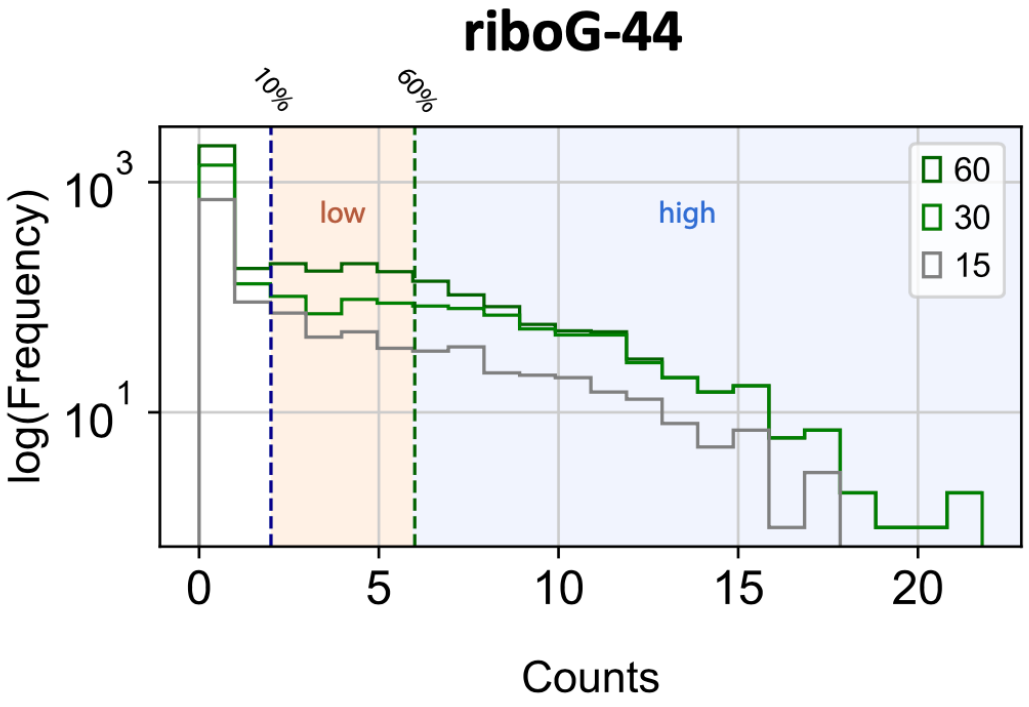


15′ -> 1187
30′ -> 1301
60′ -> 2377

Total cells: 4865

Define binary label for repair phenotypes



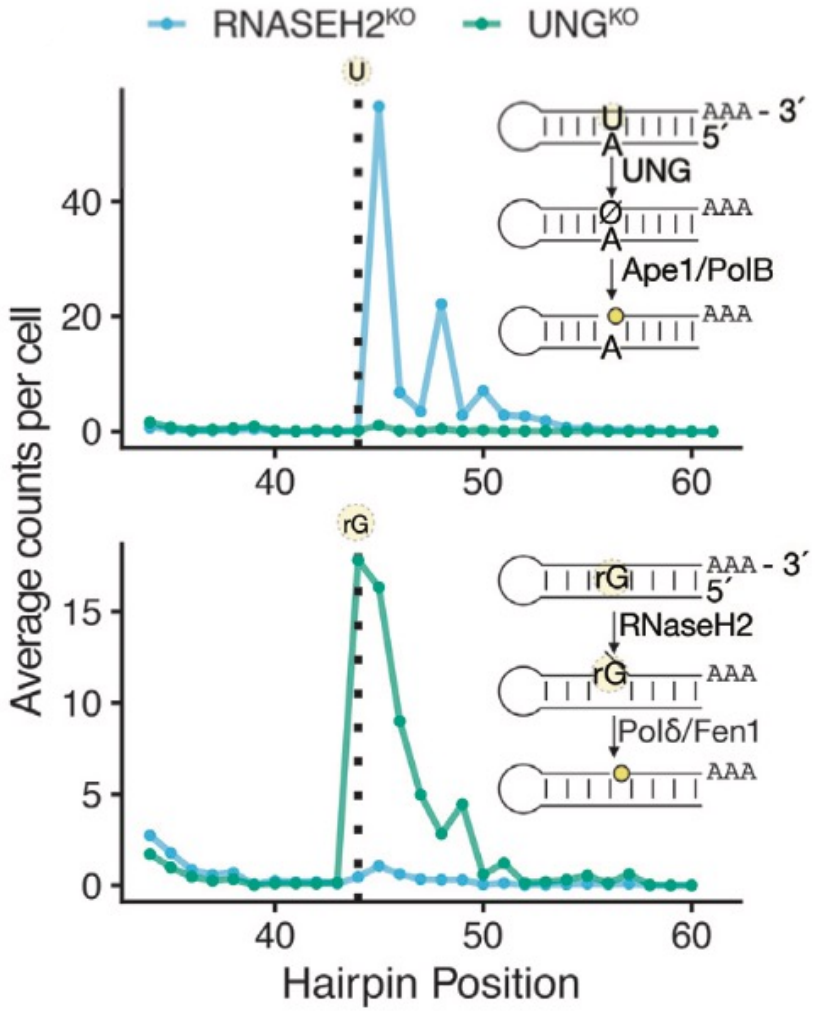


Total cells: 4865

none 3134
low 1072
high 659

none 2267
low 1617
high 981

What are labels representing?



Uracil-44 count:

high	High **dU** count, high **dU** repair phenotype
low	Low **dU** count, low **dU** repair phenotype
none	**UNG^{KO}** cells

RNASEH2C^{KO} cells *fail* to incise ribonucleotide damage

riboG-45 count:

high	High **rG** count, high **rG** repair phenotype
low	Low **rG** count, low **rG** repair phenotype
none	**RNASEH2C^{KO}** cells

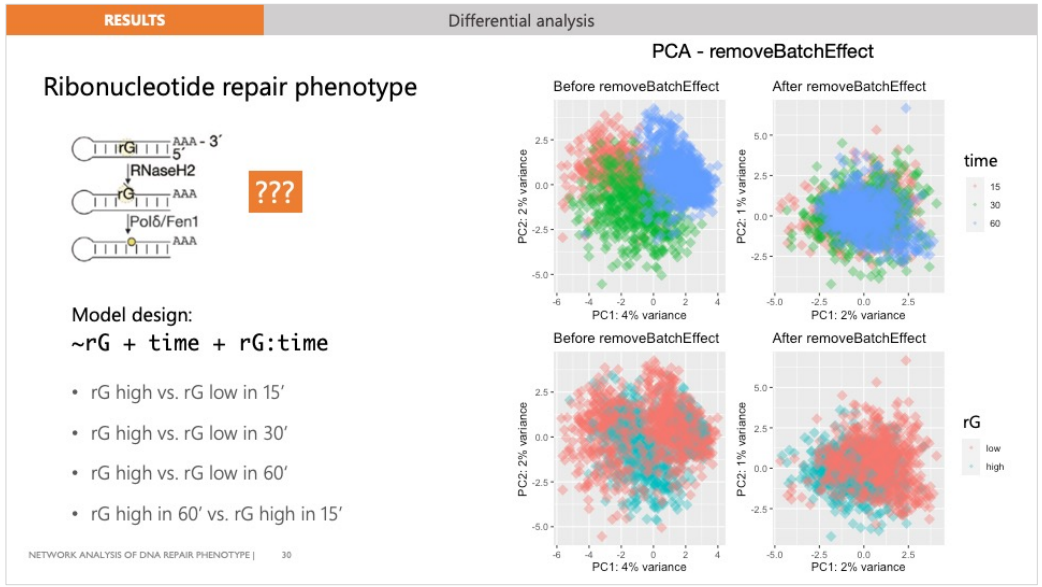
UNG^{KO} cells *fail* to incise uracil damage

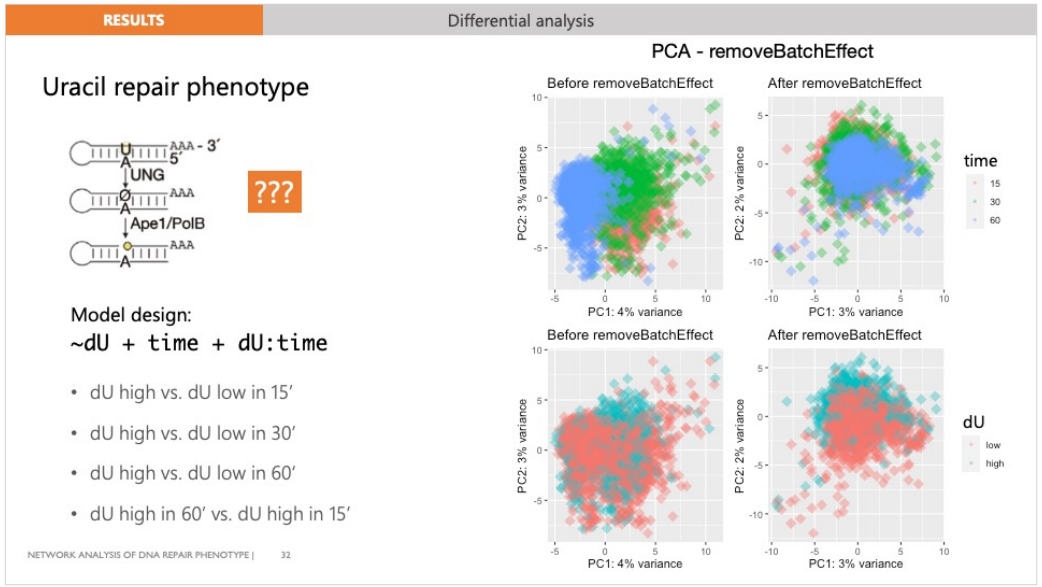
Comparison Analysis

- Alignment task
- Preprocessing
- **Differential analysis**
- **Enrichment analysis**
- Network analysis

The model formula and design matrices

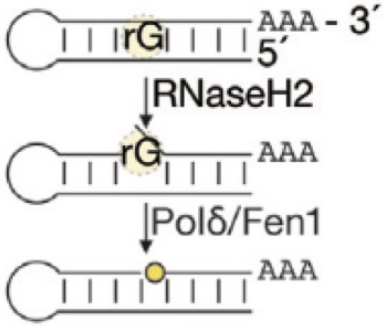
- We aim to test and report multiple comparisons in our dataset:





Variables:
- dU (High / Low / ~~None~~)
- rG (High / Low / ~~None~~)
- time (15, 30, 60)

Ribonucleotide repair phenotype



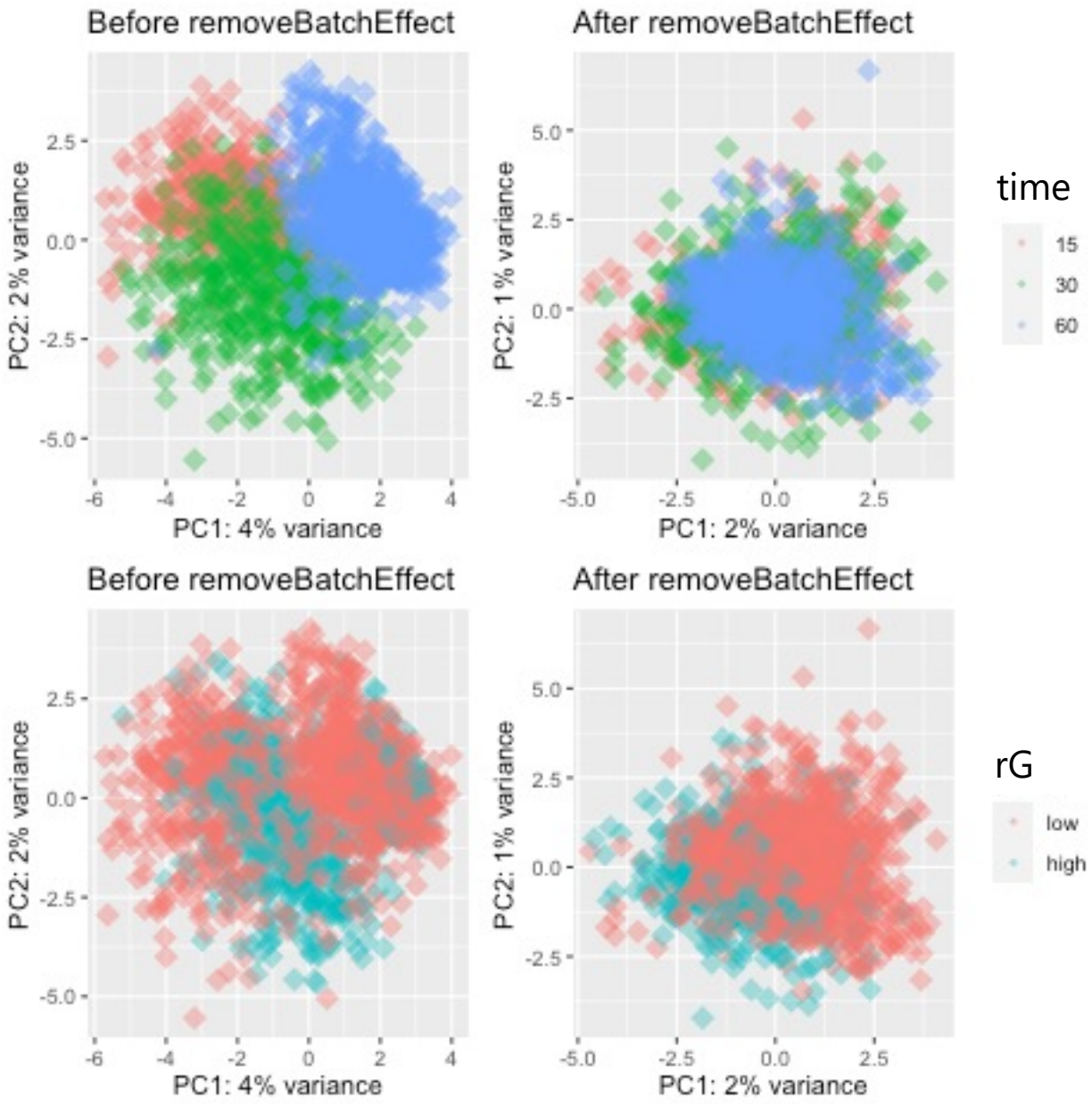
???

Model design:

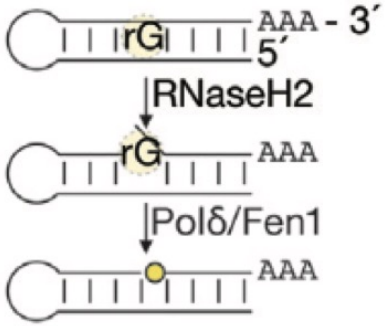
```
~rG + time + rG:time
```

- rG high vs. rG low in 15′
- rG high vs. rG low in 30′
- rG high vs. rG low in 60′
- rG high in 60′ vs. rG high in 15′

PCA - removeBatchEffect



Ribonucleotide repair phenotype

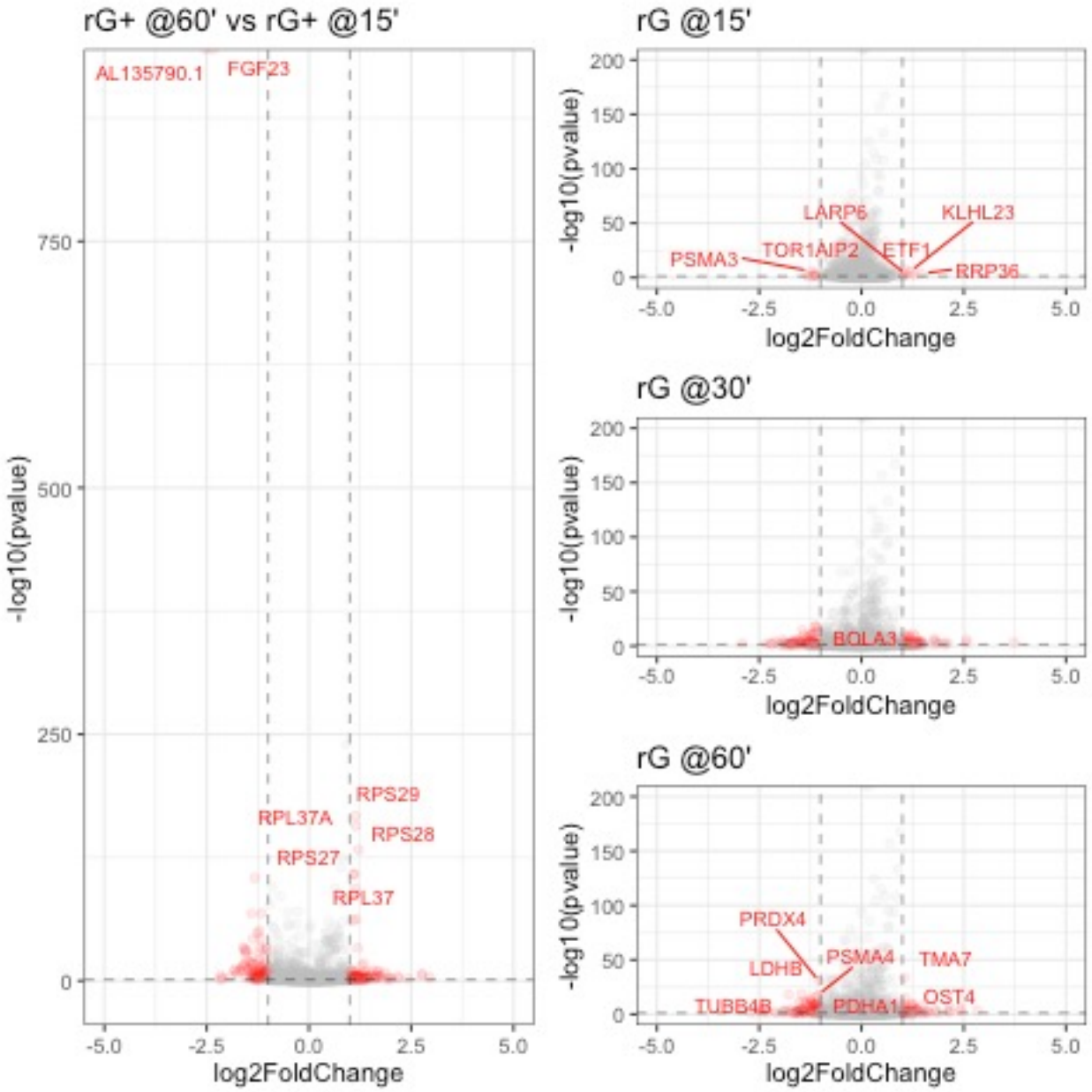


???

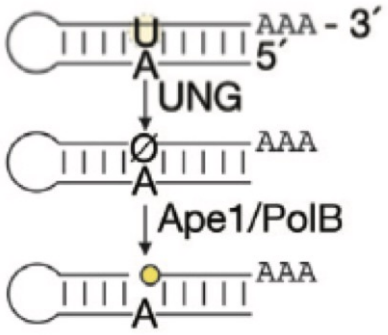
Model design:

`~rG + time + rG:time`

- rG high vs. rG low in 15′
- rG high vs. rG low in 30′
- rG high vs. rG low in 60′
- rG high in 60′ vs. rG high in 15′



Uracil repair phenotype



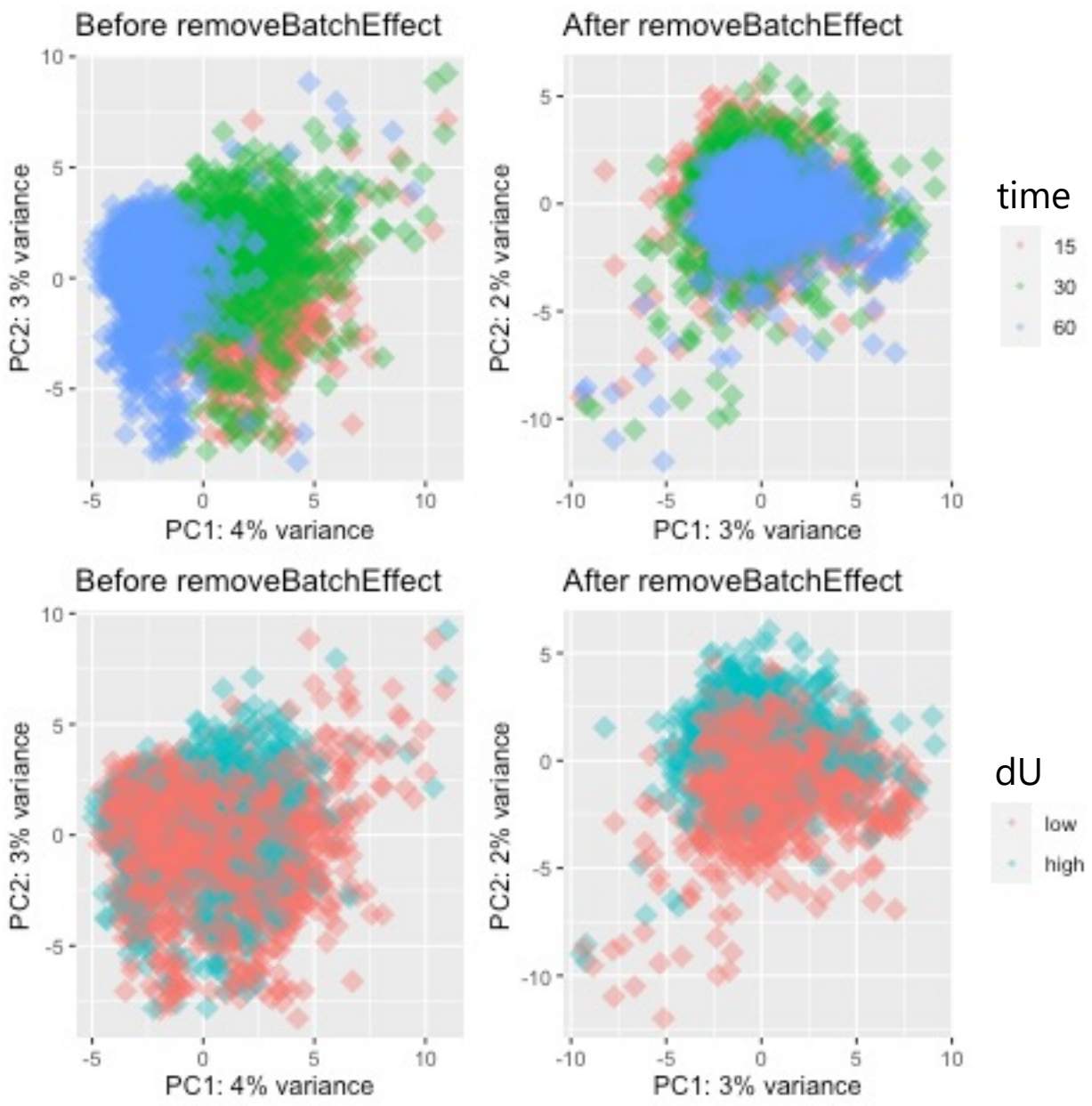
???

Model design:

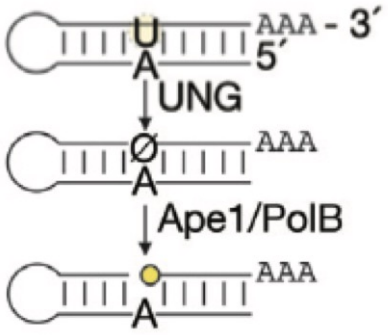
```
~dU + time + dU:time
```

- dU high vs. dU low in 15′
- dU high vs. dU low in 30′
- dU high vs. dU low in 60′
- dU high in 60′ vs. dU high in 15′

PCA - removeBatchEffect



Uracil repair phenotype

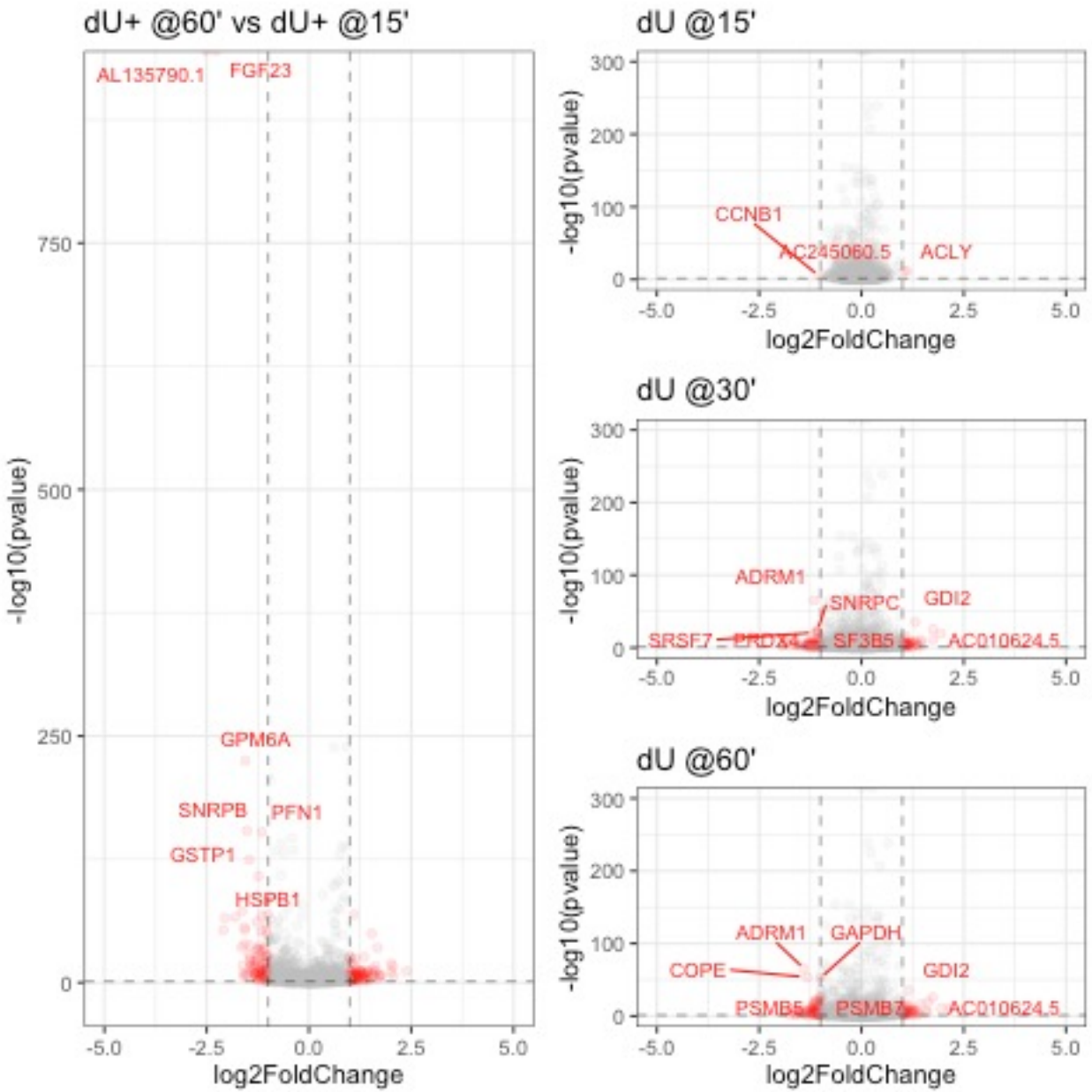


???

Model design:

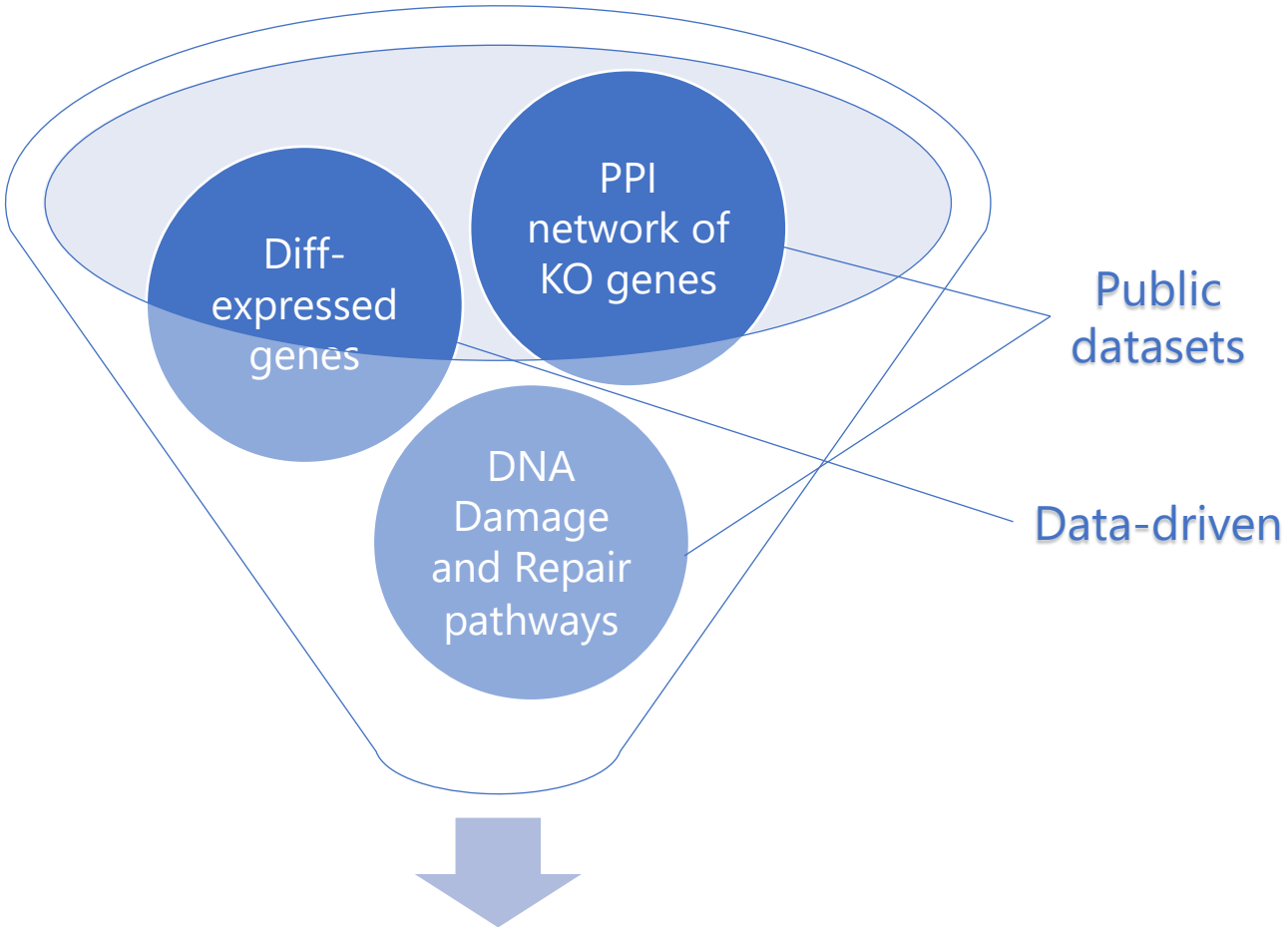
```
~dU + time + dU:time
```

- dU high vs. dU low in 15′
- dU high vs. dU low in 30′
- dU high vs. dU low in 60′
- dU high in 60′ vs. dU high in 15′

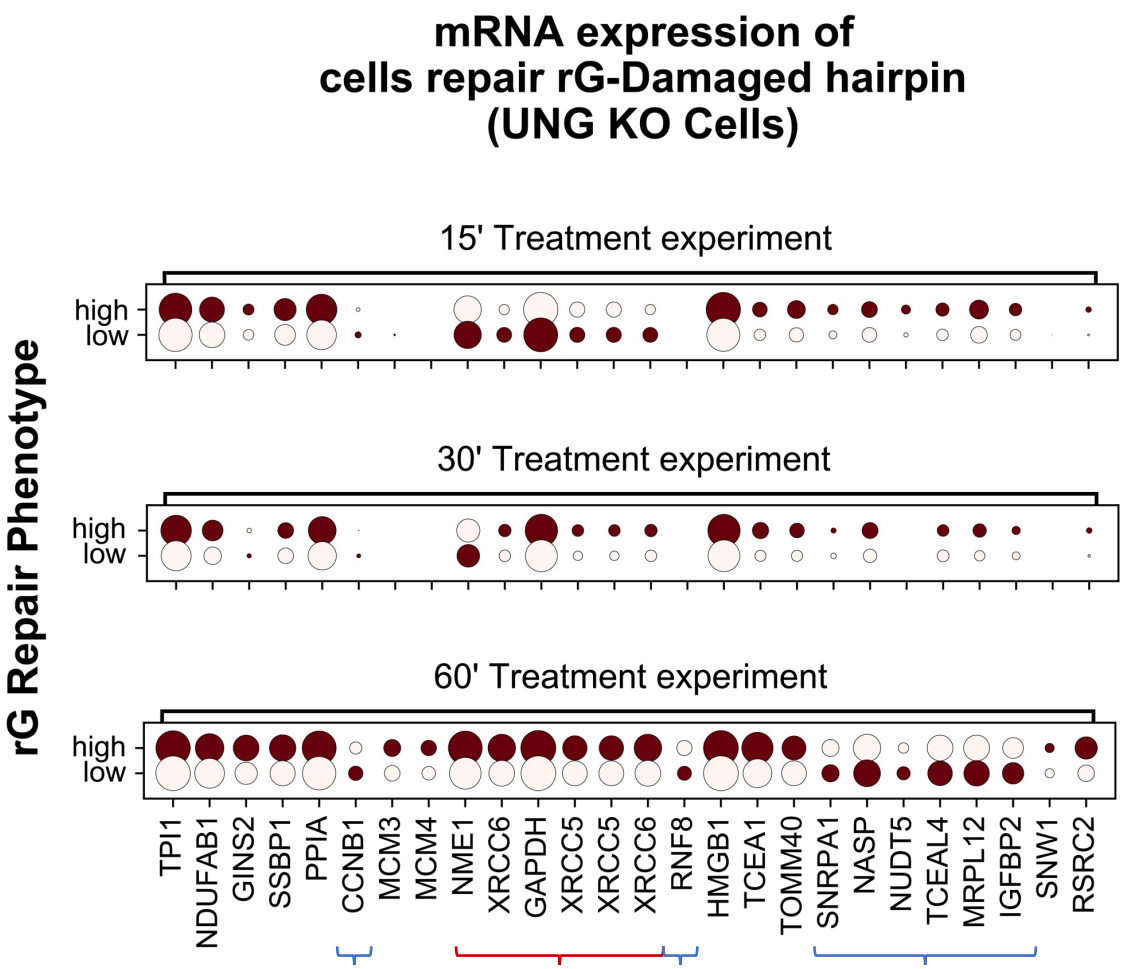


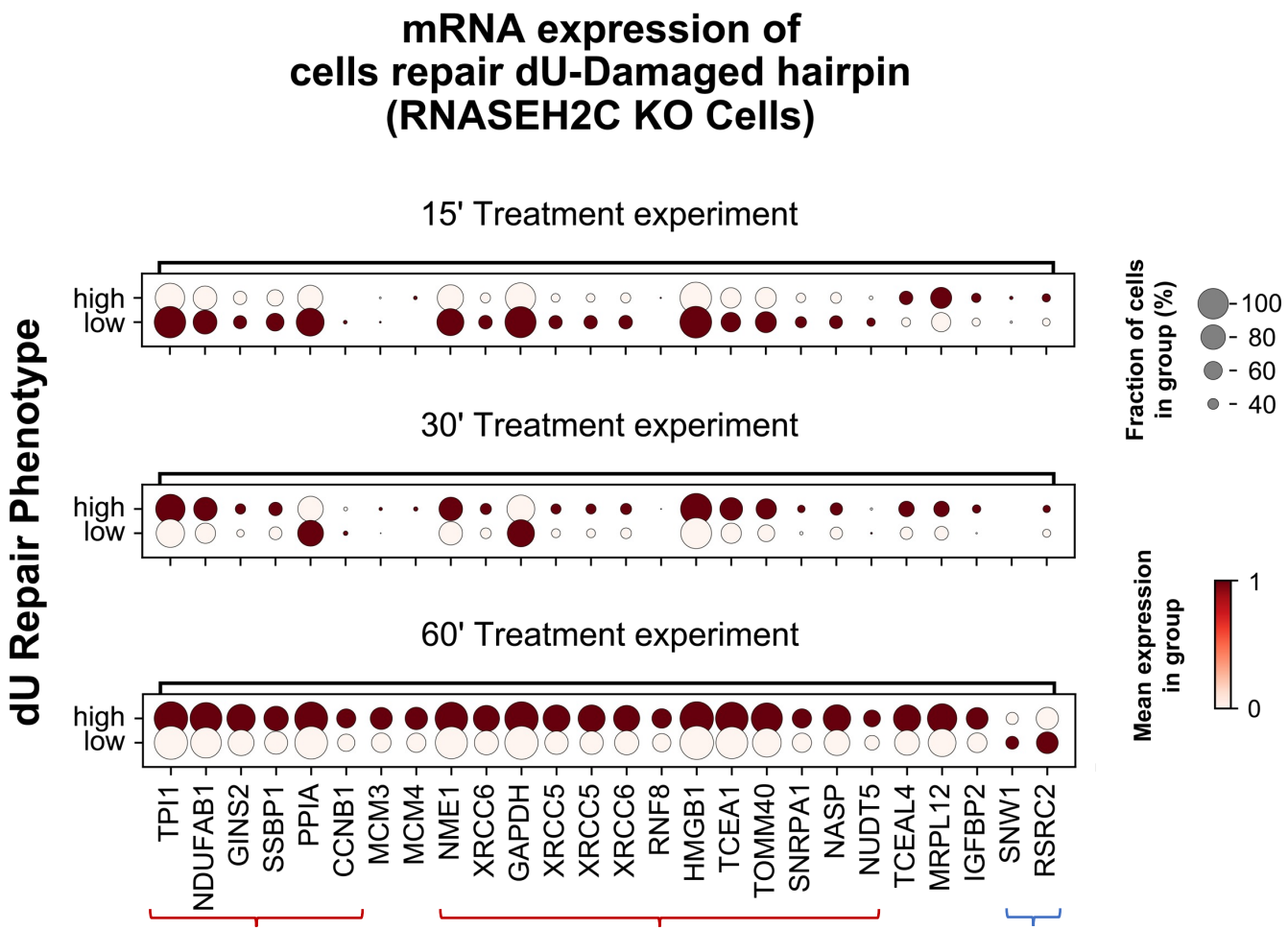
Long list of investigated genes

Find genes with expression alteration over time



Manually select altered genes


mRNA expression of
cells repair rG-Damaged hairpin
(UNG KO Cells)
rG Repair Phenotype
15' Treatment experiment
30' Treatment experiment
60' Treatment experiment
high
low
TPI1
NDUFAB1
GINS2
SSBP1
PPIA
CCNB1
MCM3
MCM4
NME1
XRCC6
GAPDH
XRCC5
XRCC5
XRCC6
RNF8
HMGB1
TCEA1
TOMM40
SNRPA1
NASP
NUDT5
TCEAL4
MRPL12
IGFBP2
SNW1
RSRC2


mRNA expression of
cells repair dU-Damaged hairpin
(RNASEH2C KO Cells)
dU Repair Phenotype
15' Treatment experiment
30' Treatment experiment
60' Treatment experiment
high
low
TPI1
NDUFAB1
GINS2
SSBP1
PPIA
CCNB1
MCM3
MCM4
NME1
XRCC6
GAPDH
XRCC5
XRCC5
XRCC6
RNF8
HMGB1
TCEA1
TOMM40
SNRPA1
NASP
NUDT5
TCEAL4
MRPL12
IGFBP2
SNW1
RSRC2
Fraction of cells in group (%)
100
80
60
40
Mean expression in group
1
0

Network and Graph Analysis

- Alignment task
- Preprocessing
- Differential analysis
- Enrichment analysis
- **Network analysis**

GRN - Gene Regulatory Networks

pySCENIC

A lightning-fast python implementation of the SCENIC pipeline (**Single-Cell rEgulatory Network Inference and Clustering**)

Enables biologists to infer from scRNA-seq data

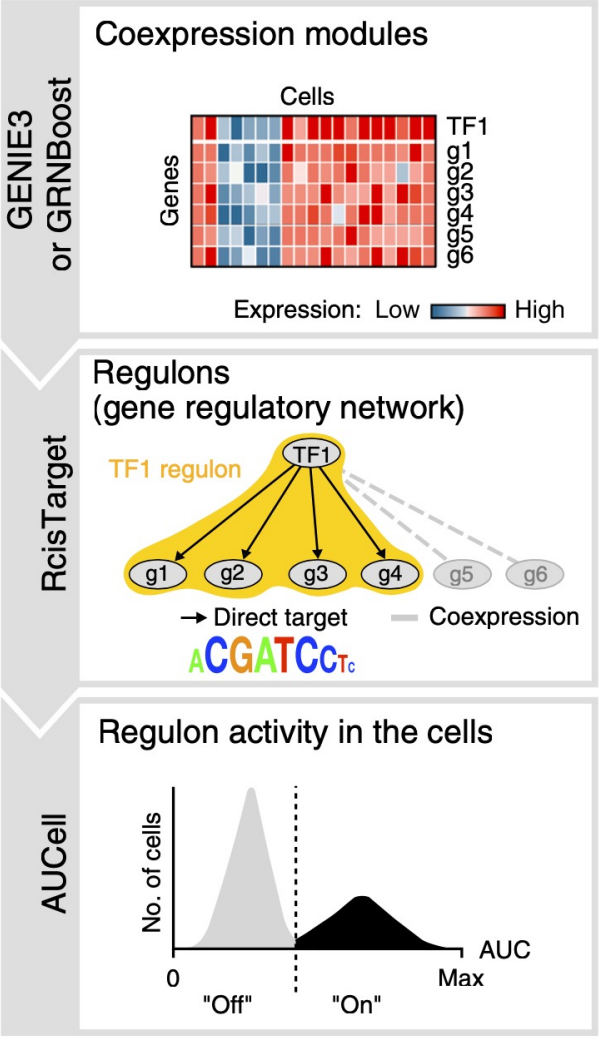
Transcription factors (TFs)

Gene Regulatory Networks (GRNs)

Cell types

https://scenic.aertslab.org/

pySCENIC workflow



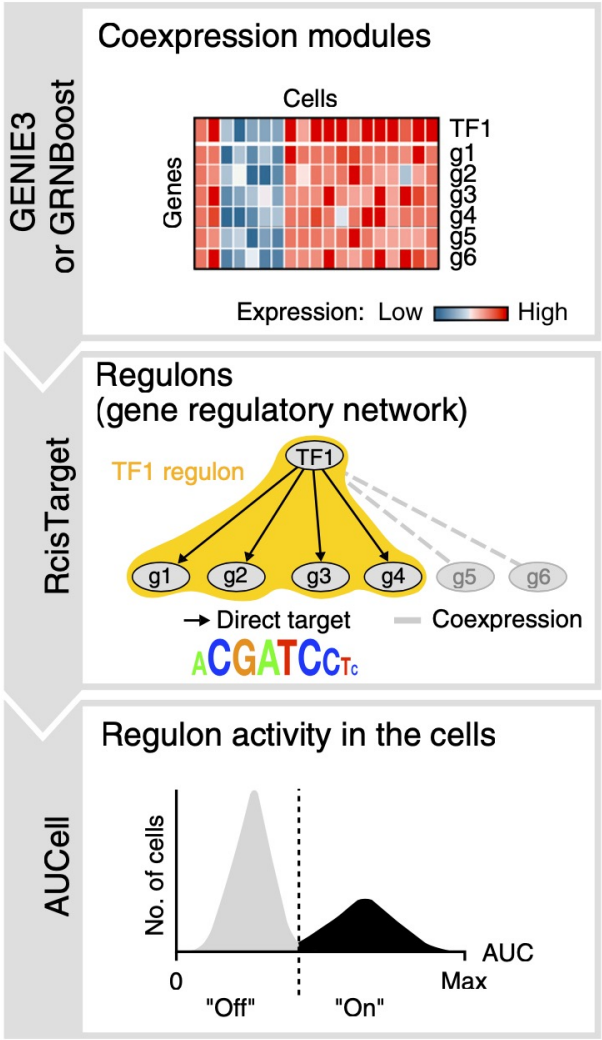
1. Sets of genes that are coexpressed with TFs are identified using GENIE3

2. Since the GENIE3 modules are only based on coexpression, they may include many **false positives** and indirect targets. ⇒ To identify putative direct-binding targets, each coexpression module is subjected to *cis*-regulatory motif analysis using RcisTarget.

3. Estimate AUC score as regulons activity representation among cells.

https://scenic.aertslab.org/

GRN analysis results



(24,484,108) co-expression profile found

(4,652,523) TF/target interaction found

Columns: TF Target importance

(4,865) Cells in 3 time points and 4 unique conditions considered.

(268) Regulons' functionality for each cells evaluated.

igraph

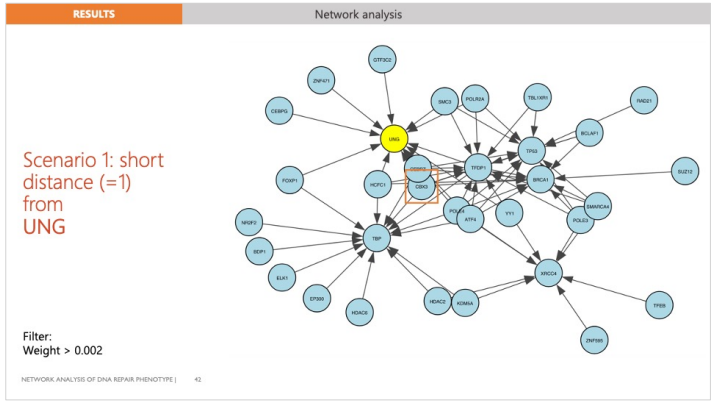
I aim to create weighted and directed graph toward further analysis (finding context specific master regulators)

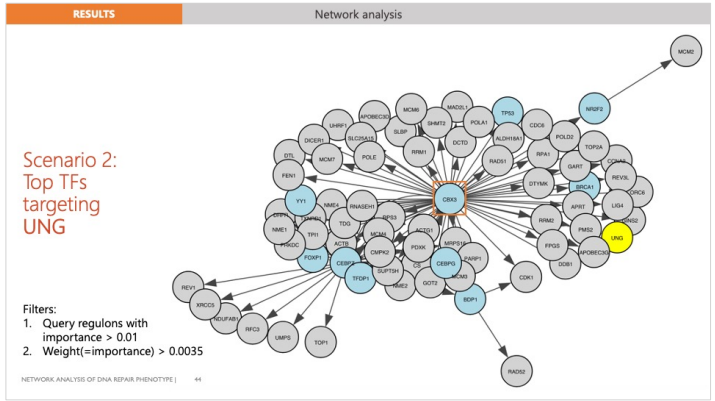
Build and analyze **context-specific** networks

1. Make the large context-specific GRN network
   - Number of vertices in the graph: 18,292
   - Number of edges in the graph 4,652,523
2. Create sub-networks contain nodes from PPI network of KO genes
   1. UNG
      - Number of vertices in the graph: 447
      - Number of edges in the graph 48,703
   2. RNASEH2C
      - Number of vertices in the graph: 281
      - Number of edges in the graph 3,482

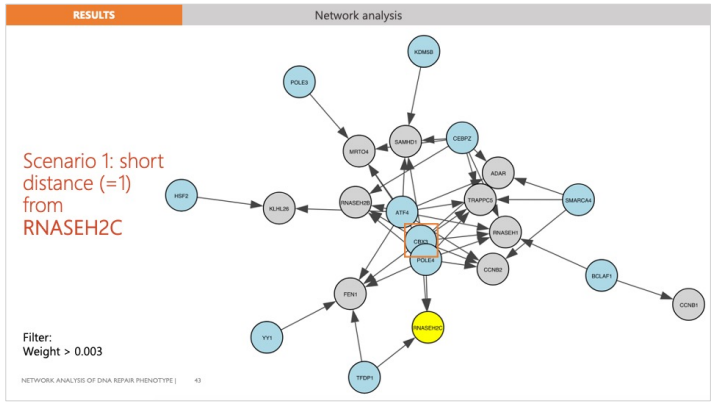
Different scenarios to explore KO subnetworks

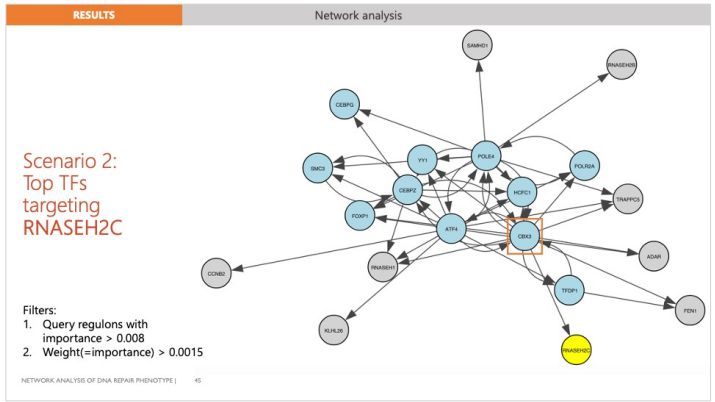
UNG Sub-net



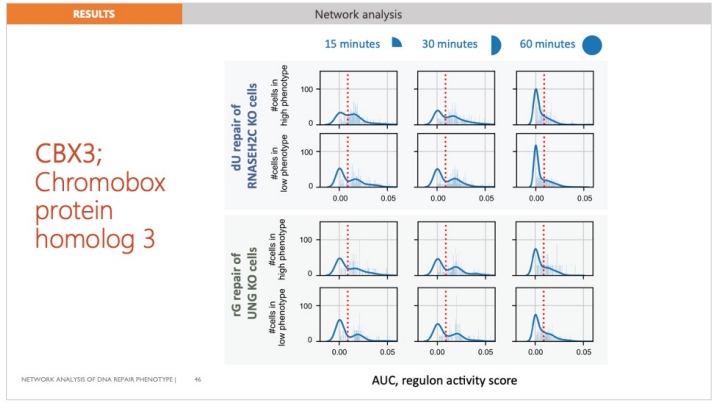


RNASEH2C Sub-net

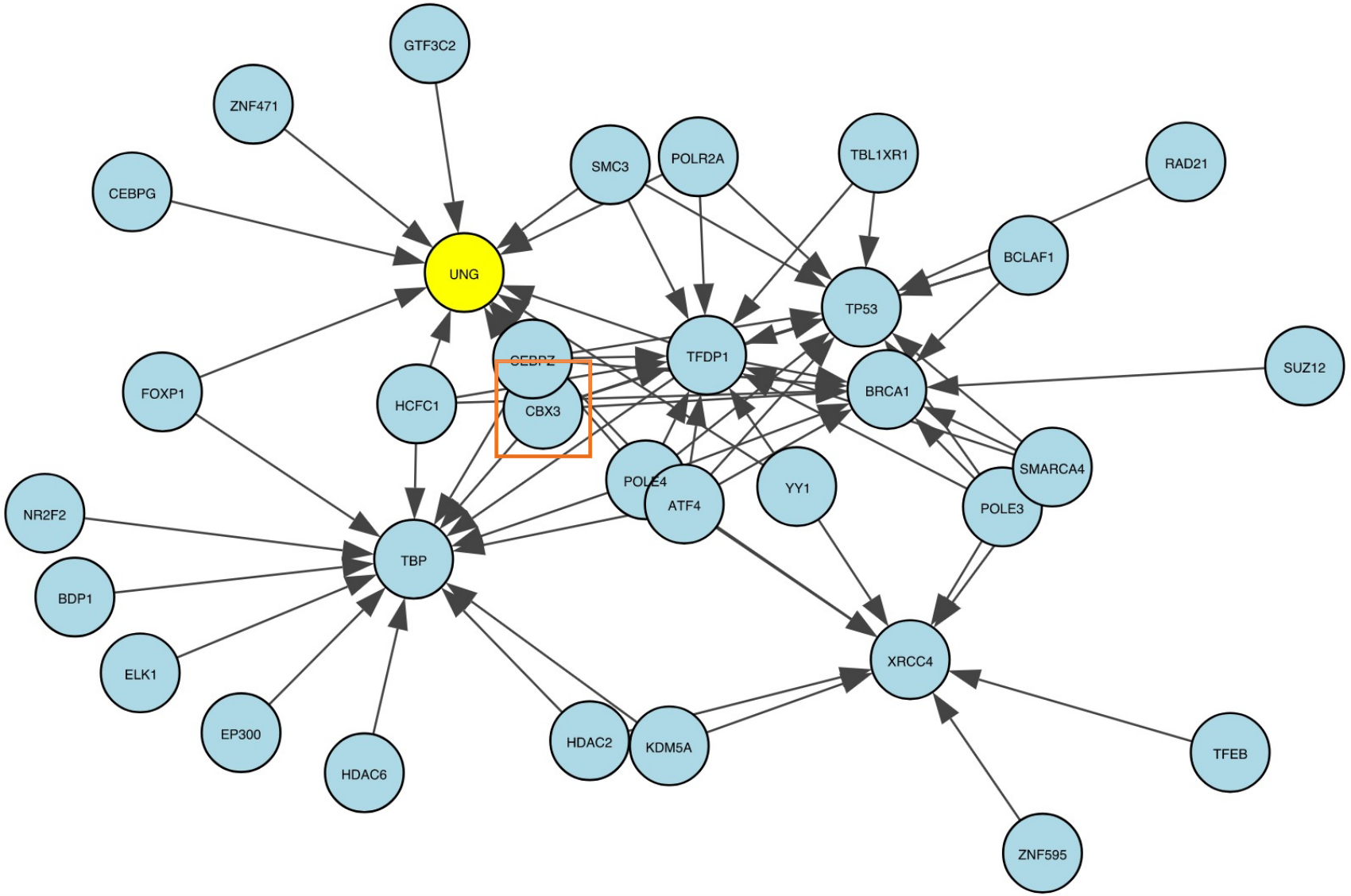




Candidate a regulon with dynamic activity over time

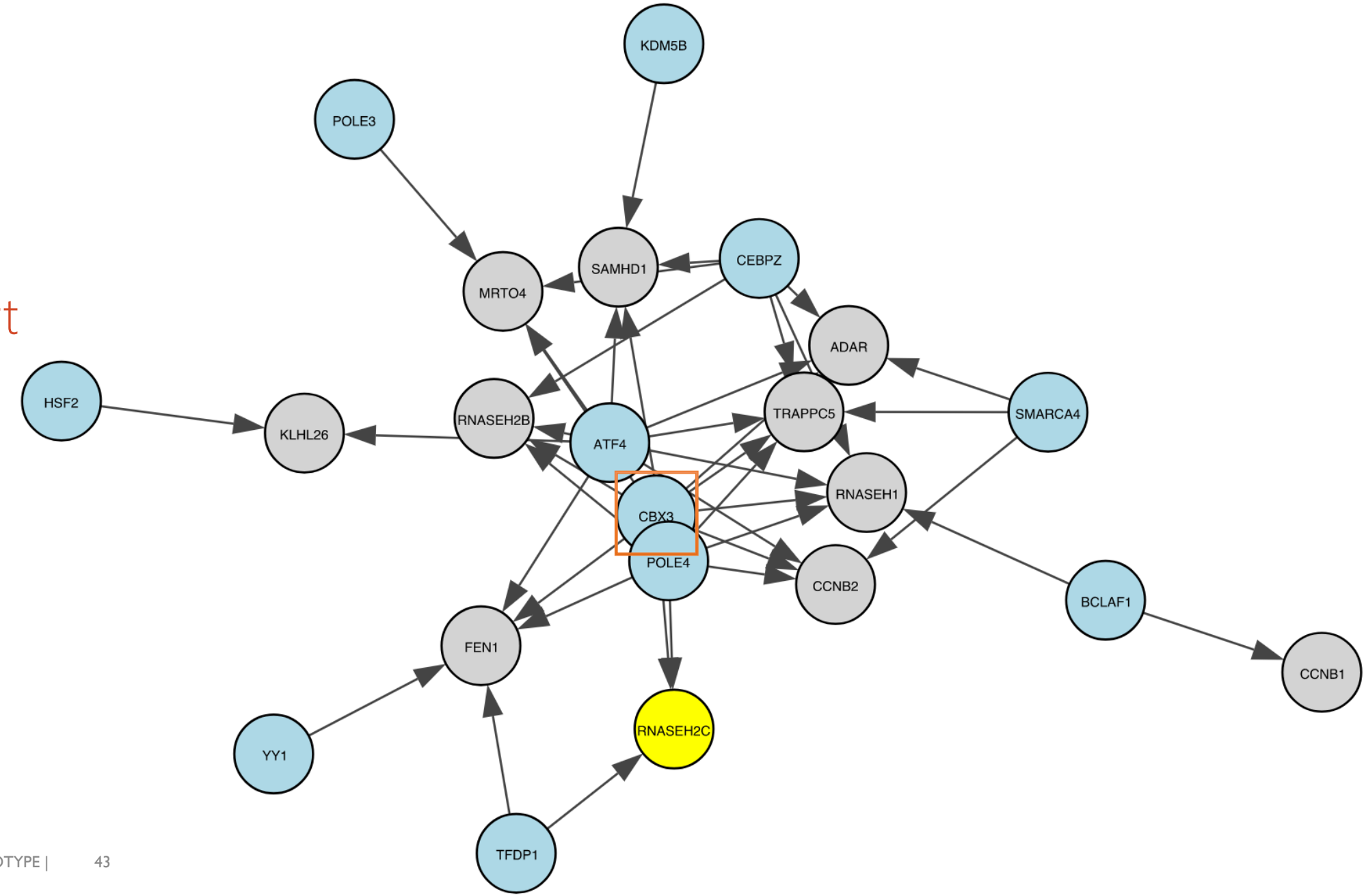


Scenario 1: short distance (=1) from UNG



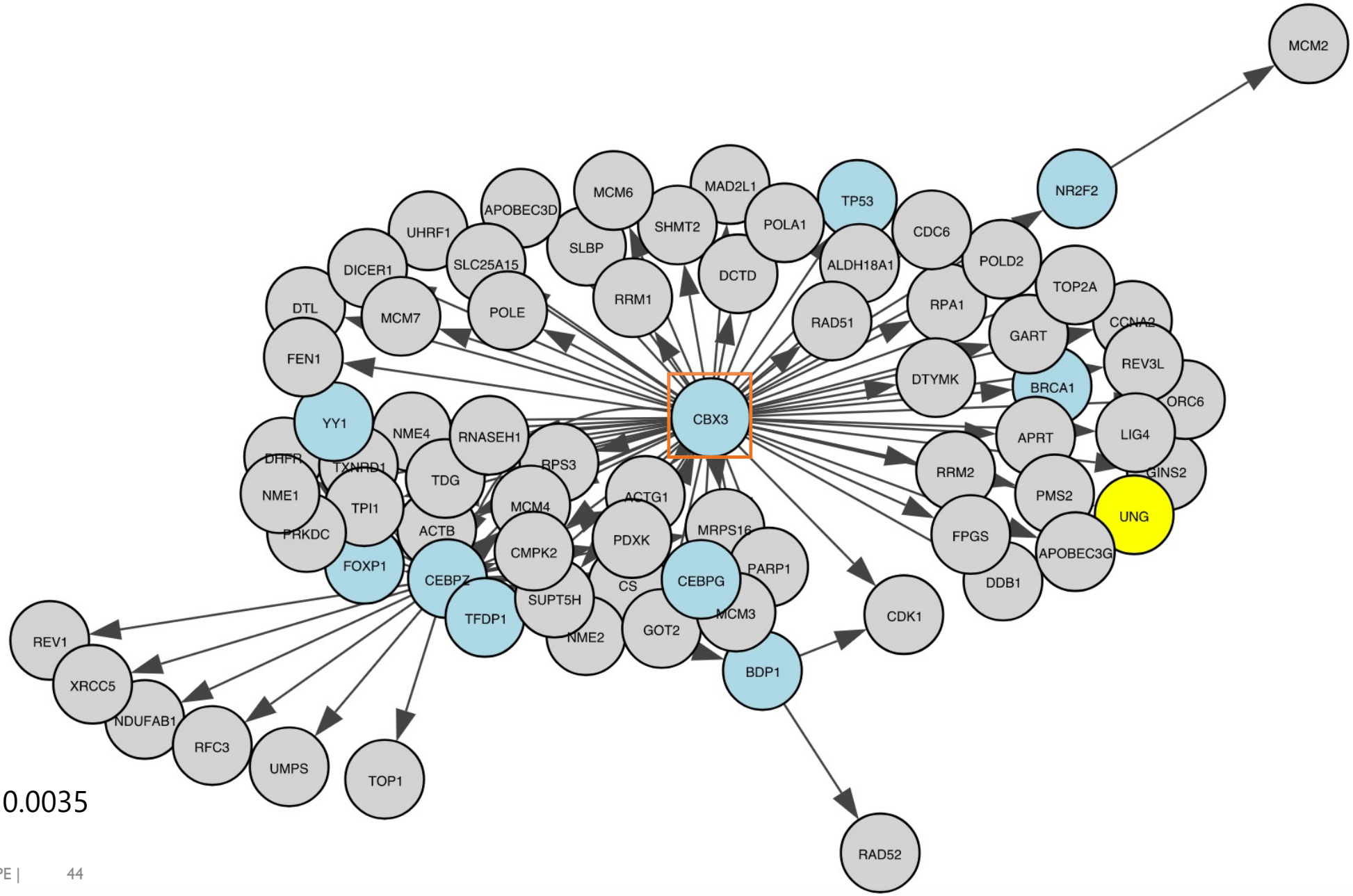
Filter:
Weight > 0.002

Scenario 1: short distance (=1) from RNASEH2C



Filter:
Weight > 0.003

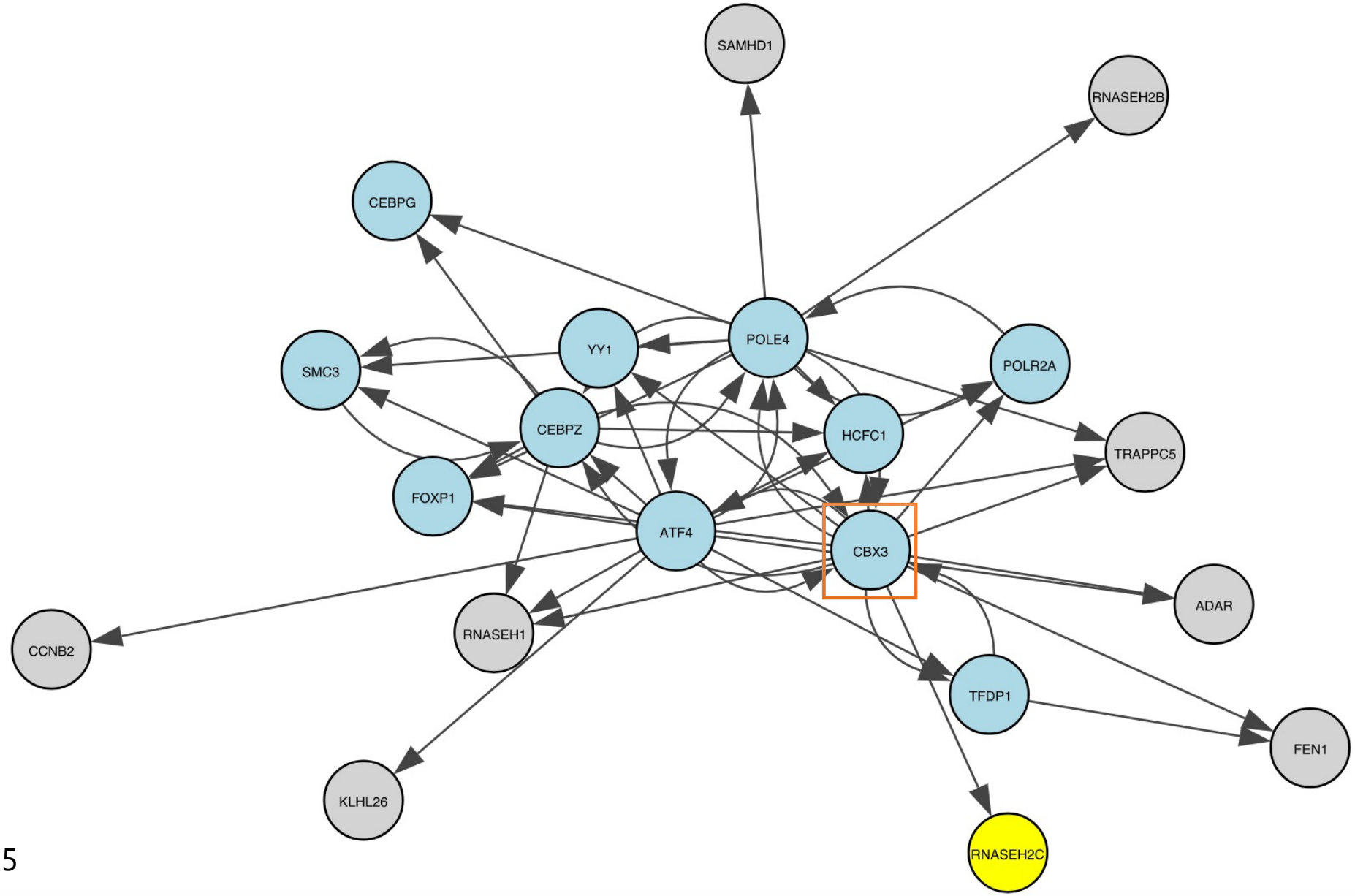
Scenario 2: Top TFs targeting UNG



Filters:

1. Query regulons with importance > 0.01
2. Weight(=importance) > 0.0035

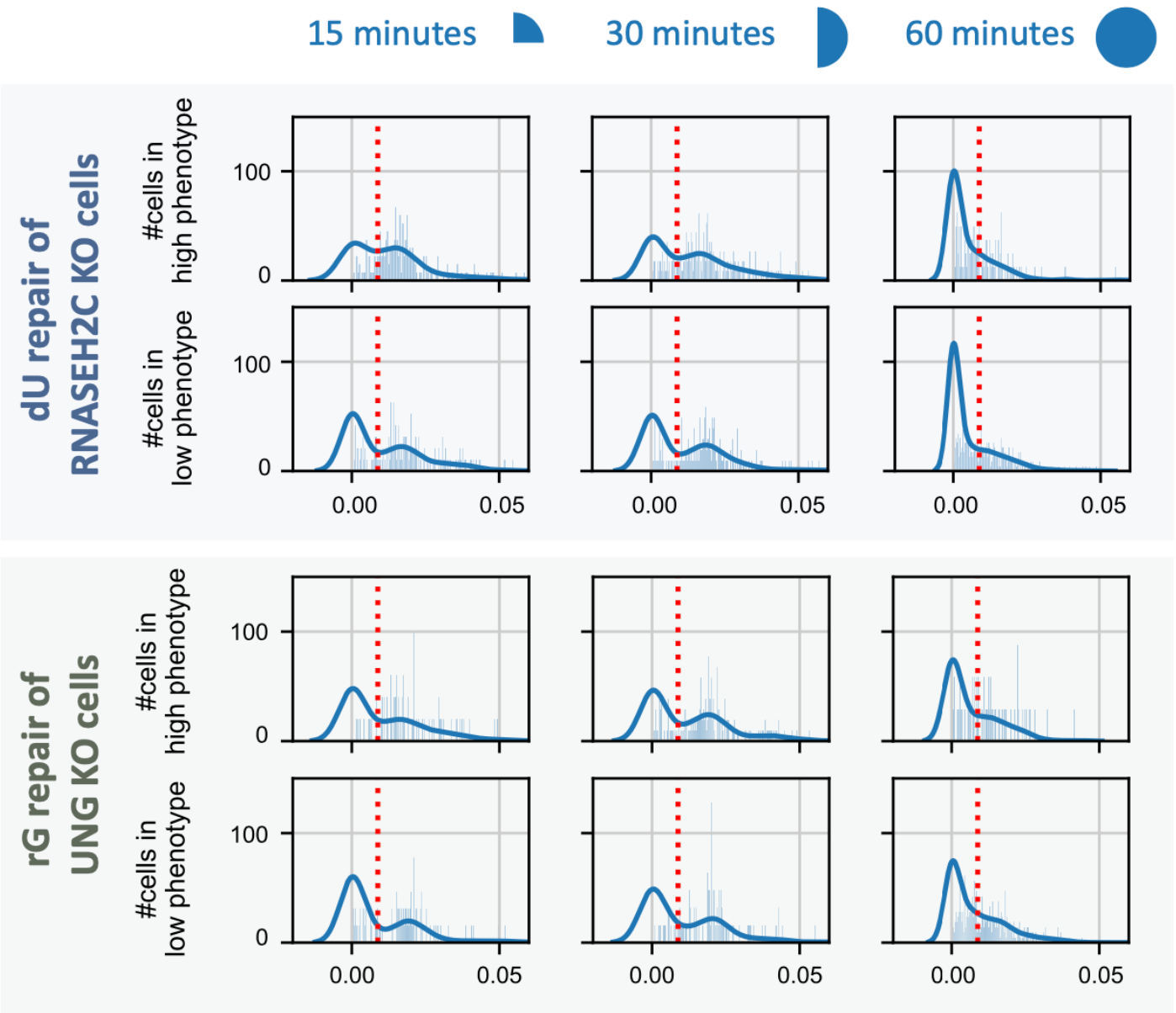
Scenario 2: Top TFs targeting RNASEH2C



Filters:

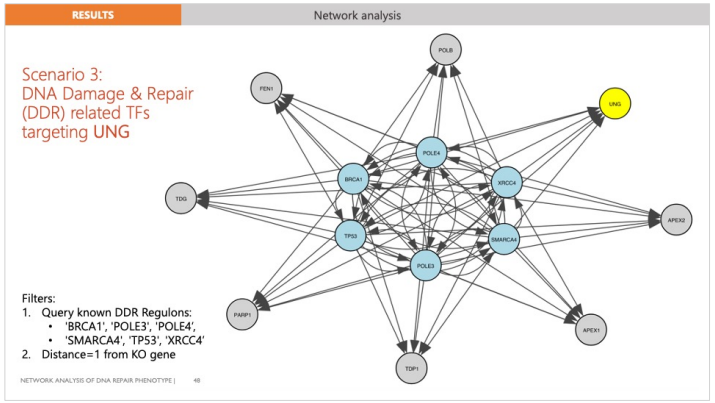
1. Query regulons with importance > 0.008
2. Weight(=importance) > 0.0015

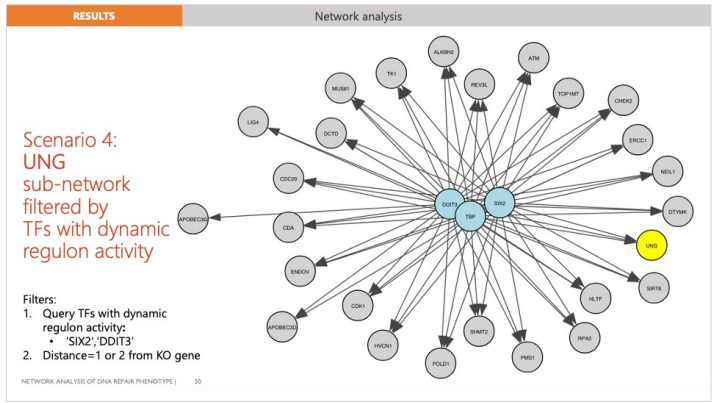
CBX3; Chromobox protein homolog 3



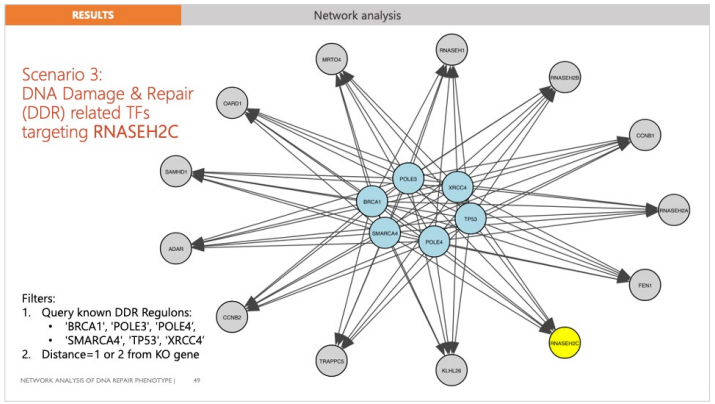
Different scenarios to explore KO subnetworks

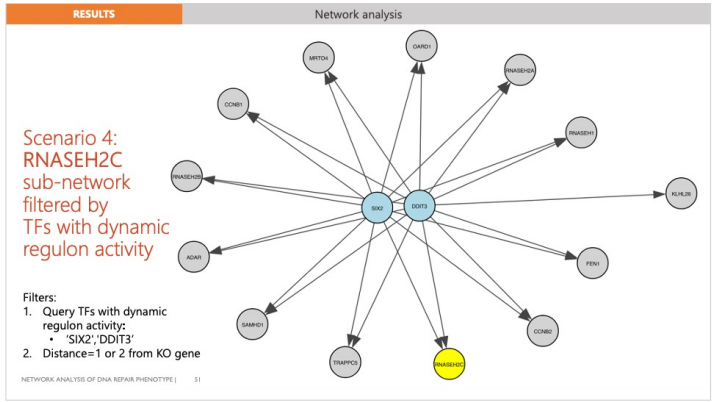
UNG Sub-net



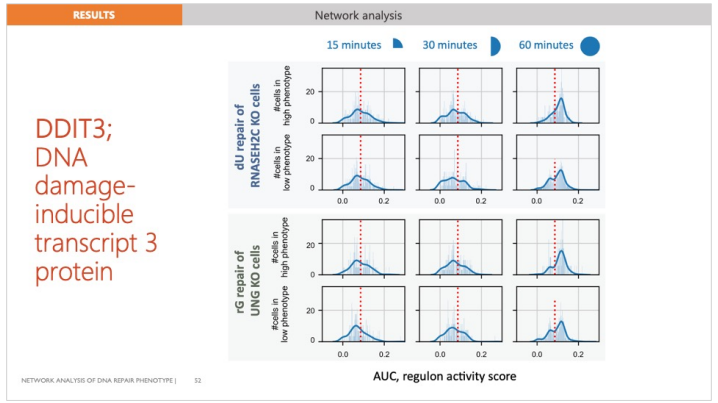


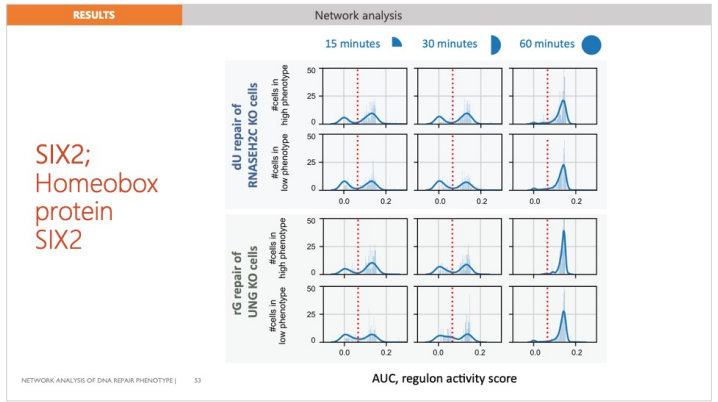
RNASEH2C Sub-net



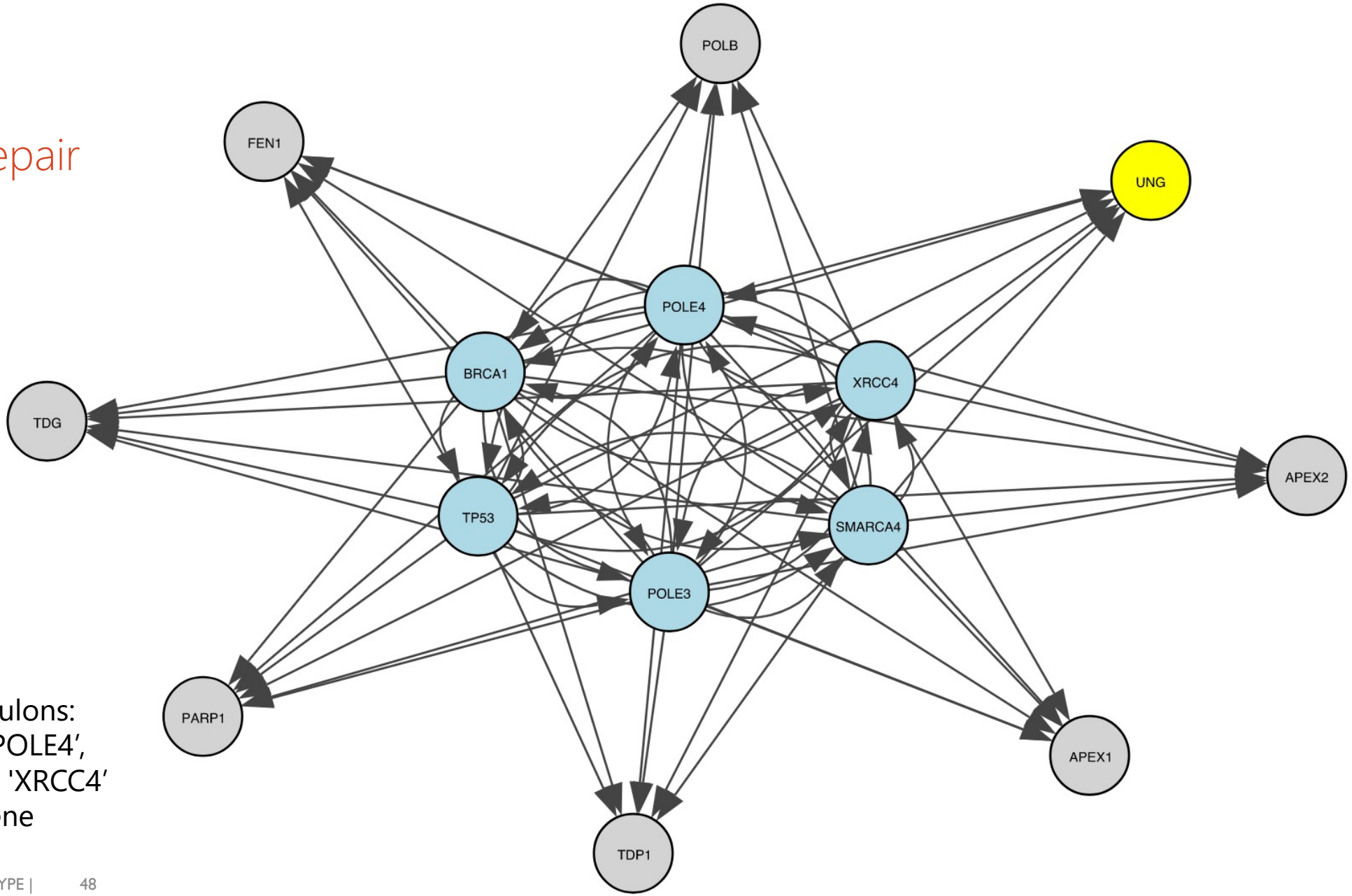


Candidate regulons with dynamic activity over time





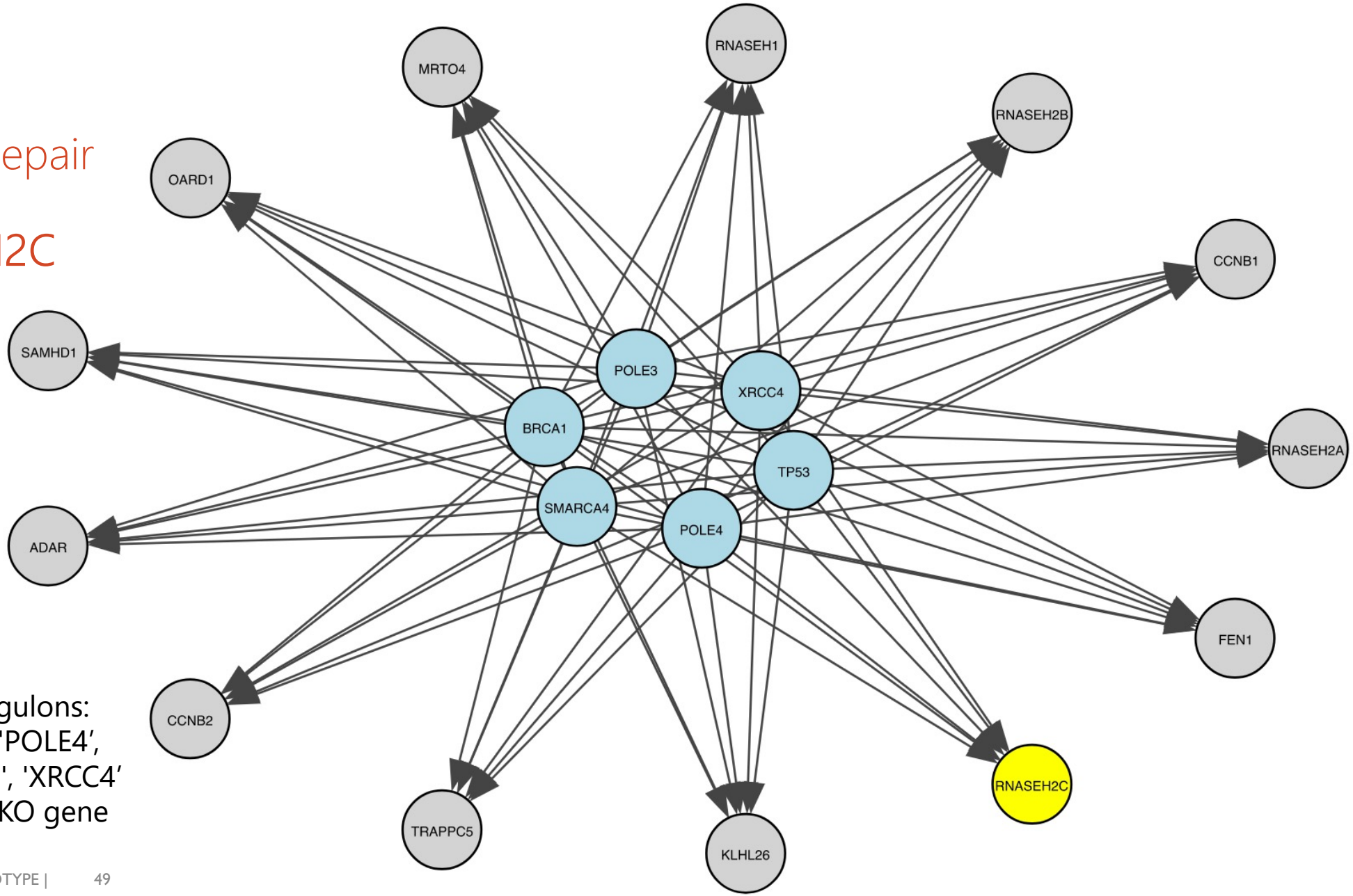
Scenario 3: DNA Damage & Repair (DDR) related TFs targeting **UNG**



Filters:

1. Query known DDR Regulons:
   - 'BRCA1', 'POLE3', 'POLE4',
   - 'SMARCA4', 'TP53', 'XRCC4'
2. Distance=1 from KO gene

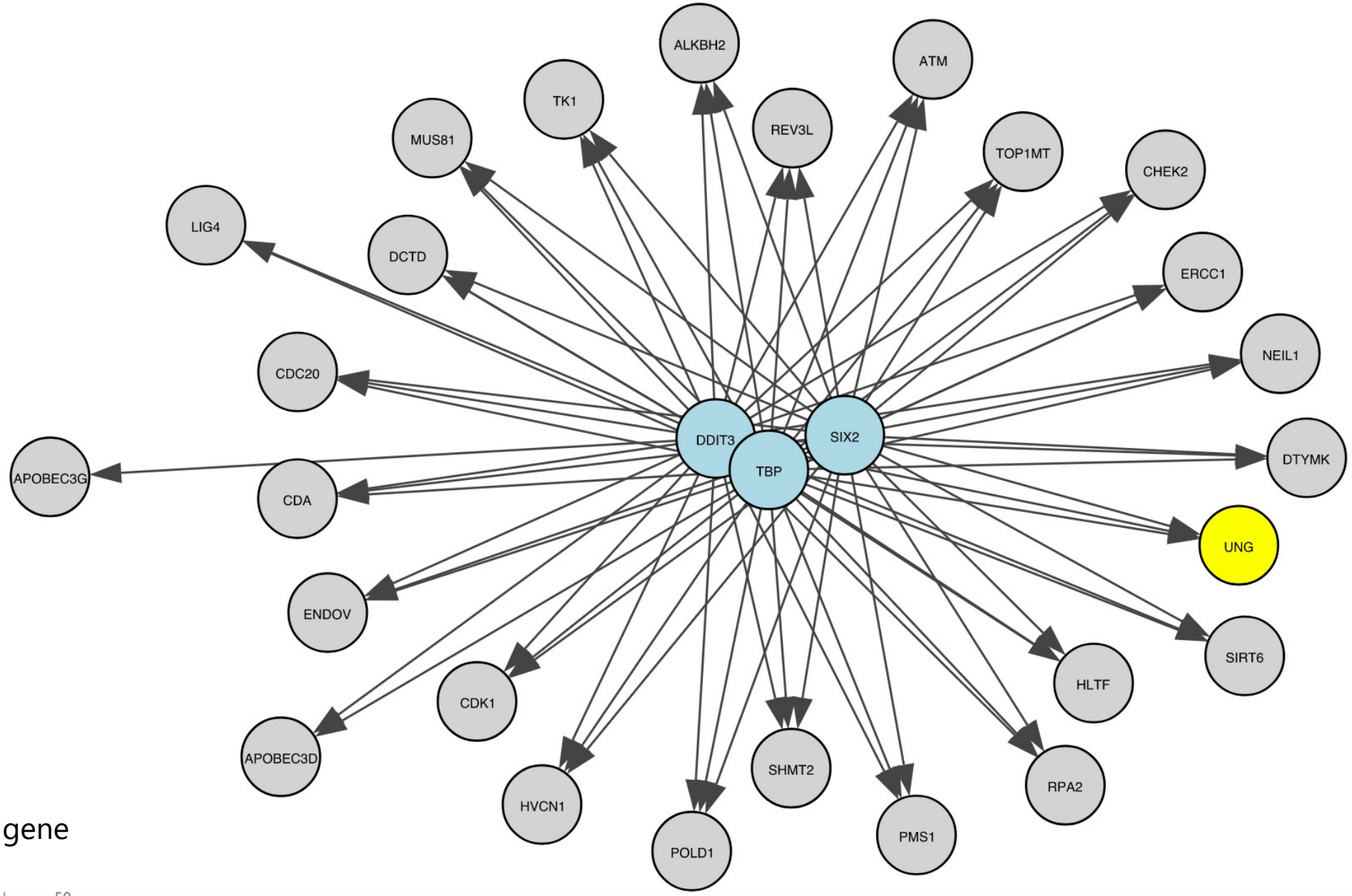
Scenario 3: DNA Damage & Repair (DDR) related TFs targeting **RNASEH2C**



Filters:

1. Query known DDR Regulons:
   - 'BRCA1', 'POLE3', 'POLE4',
   - 'SMARCA4', 'TP53', 'XRCC4'
2. Distance=1 or 2 from KO gene

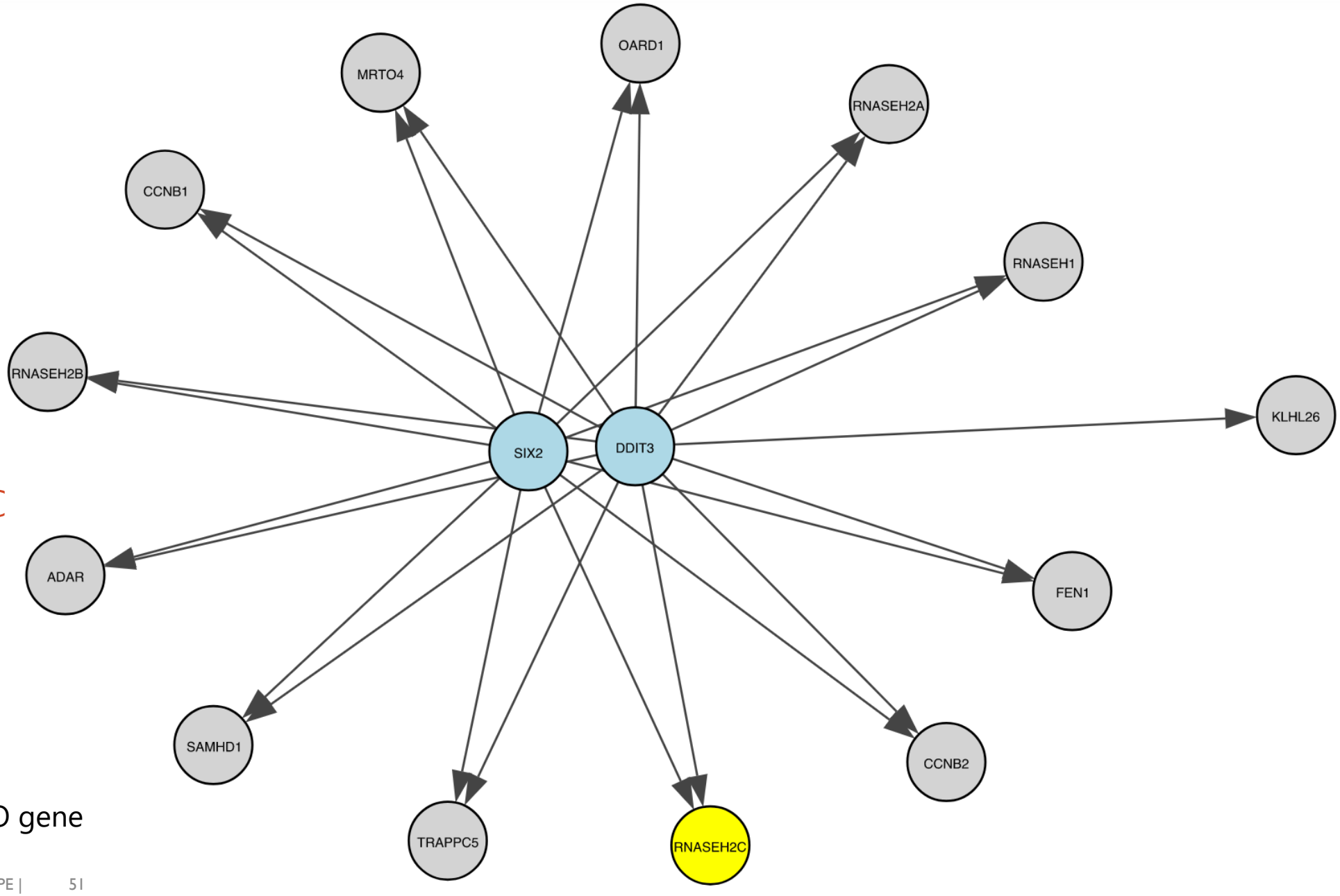
Scenario 4: UNG sub-network filtered by TFs with dynamic regulon activity



Filters:

1. Query TFs with dynamic regulon activity**:**
   - 'SIX2','DDIT3'
2. Distance=1 or 2 from KO gene

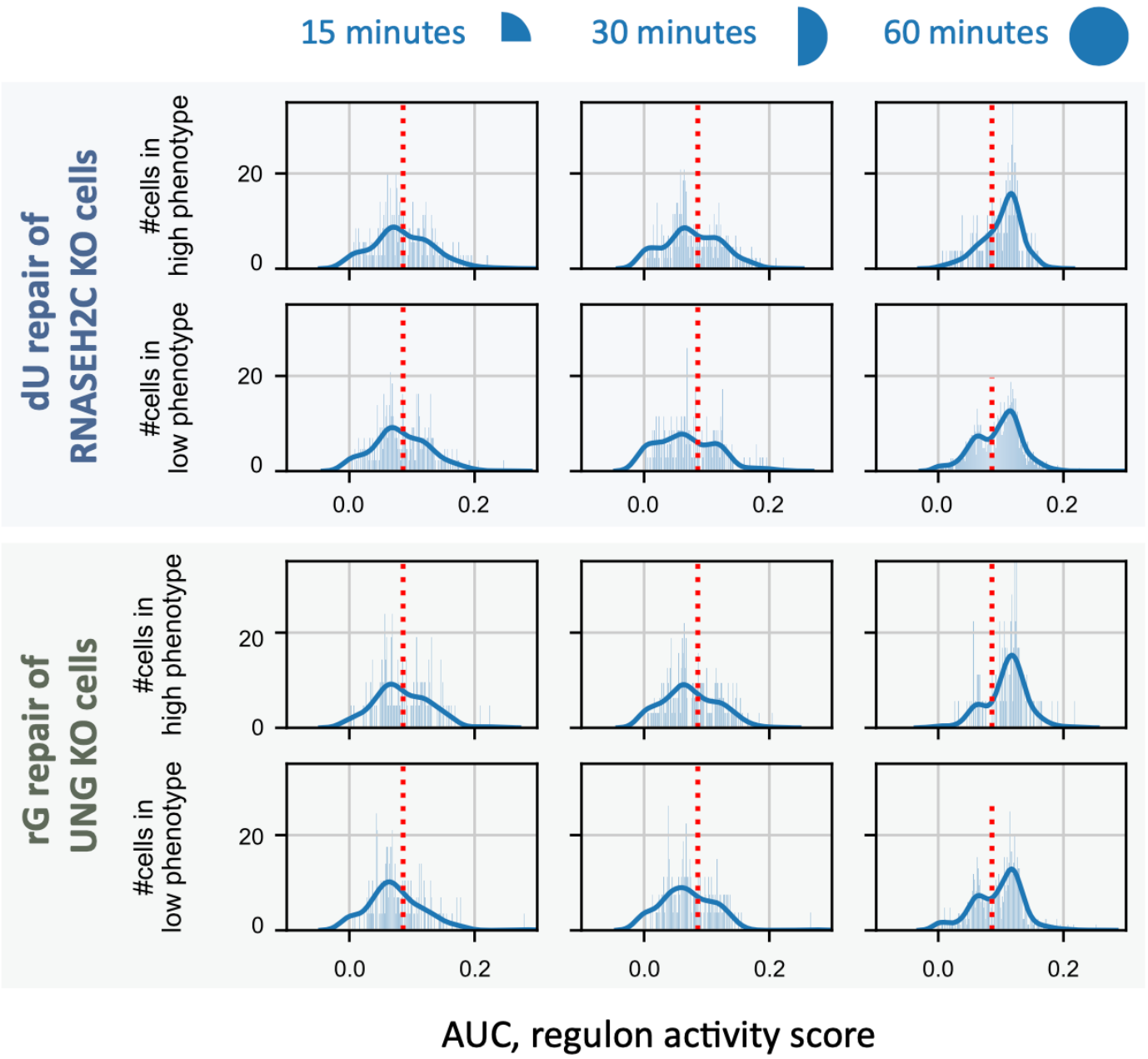
Scenario 4: RNASEH2C sub-network filtered by TFs with dynamic regulon activity



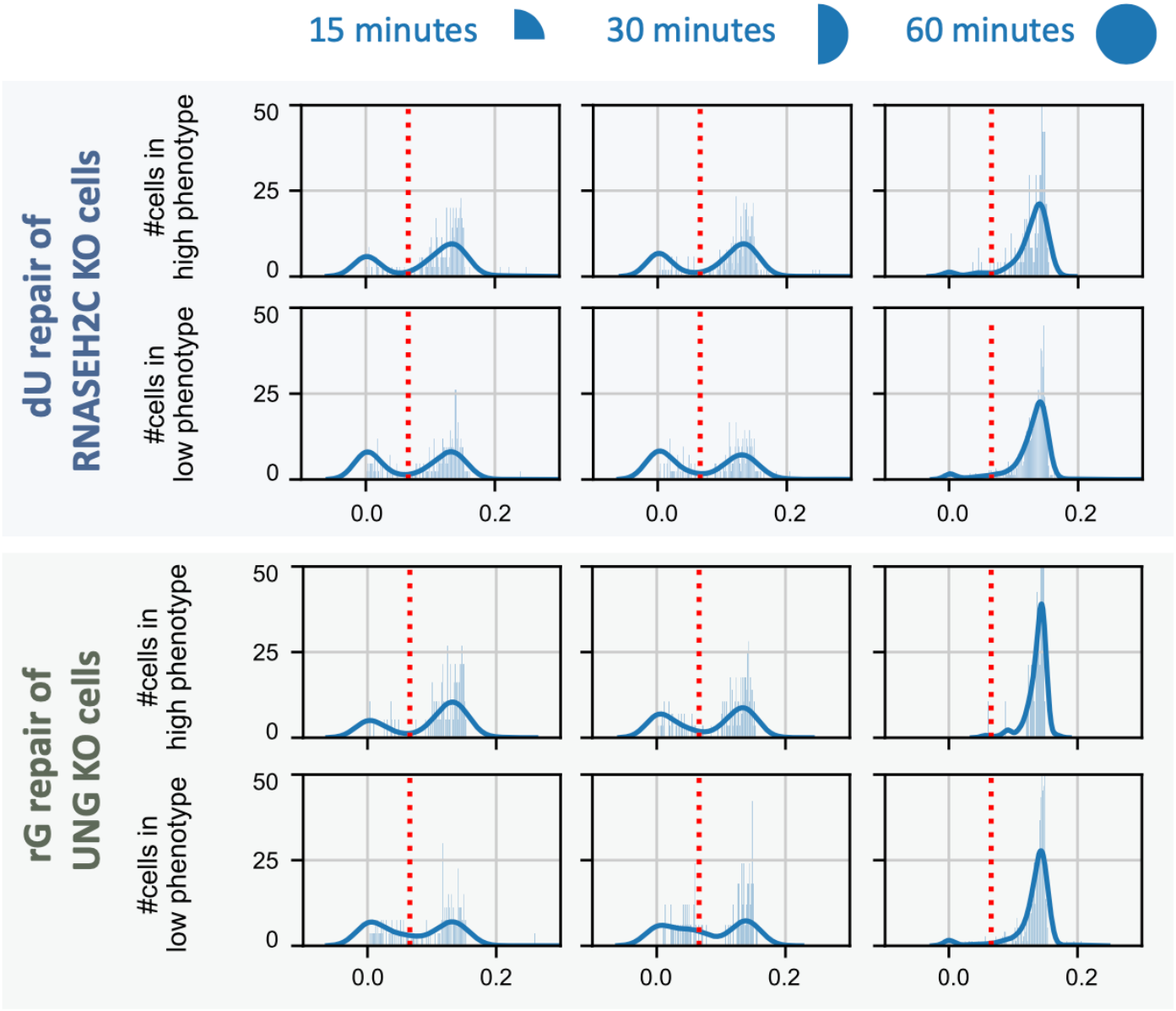
Filters:

1. Query TFs with dynamic regulon activity**:**
   - ‘SIX2','DDIT3’
2. Distance=1 or 2 from KO gene

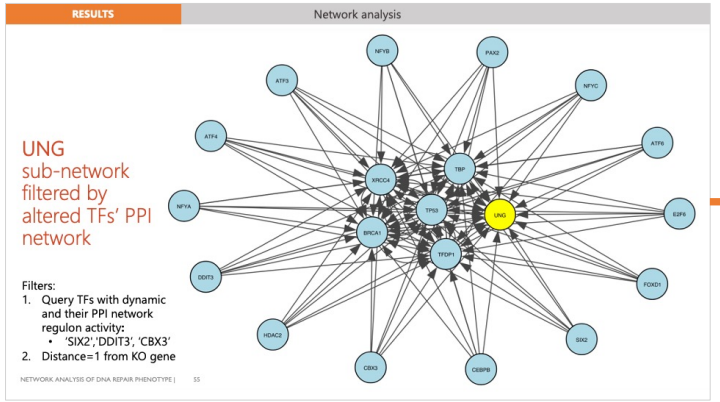
DDIT3; DNA damage-inducible transcript 3 protein

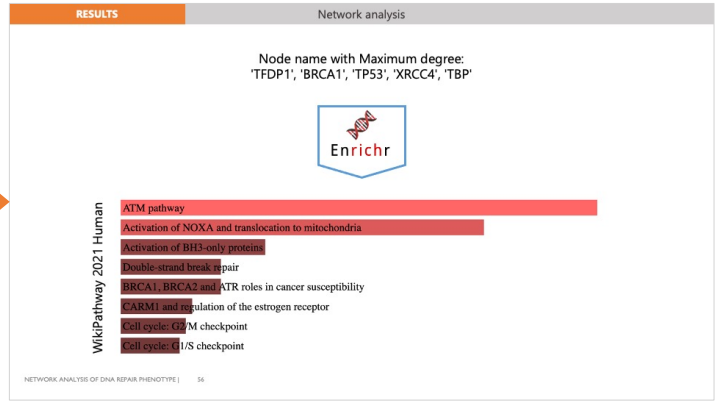


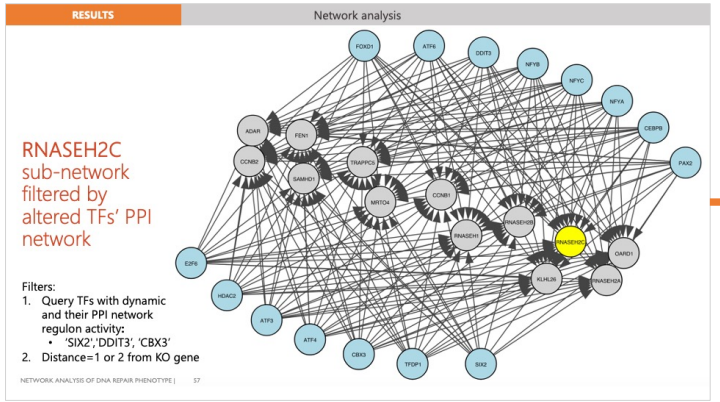
SIX2; Homeobox protein SIX2

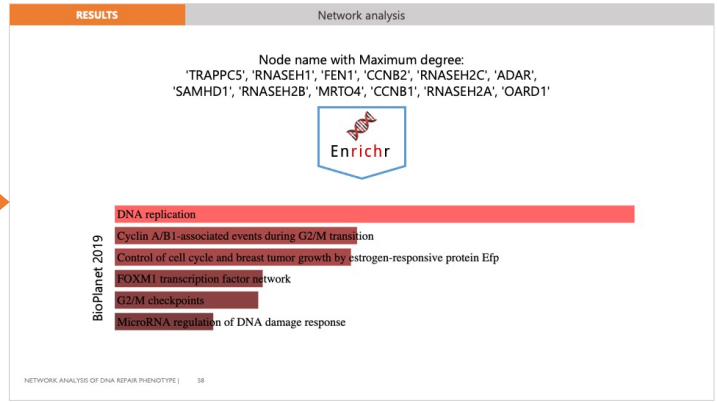


Filter KO sub-network by altered TFs' PPI network

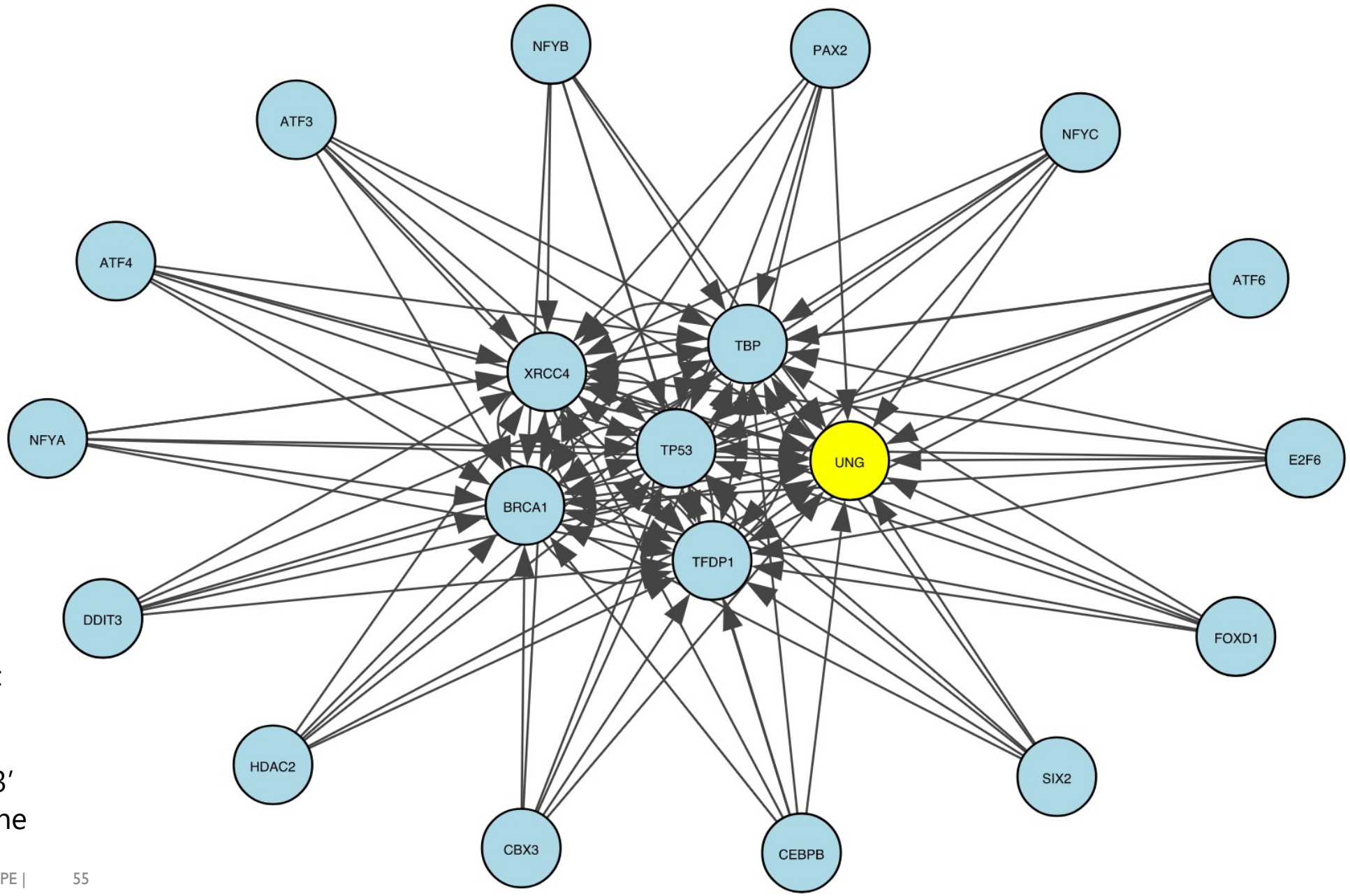








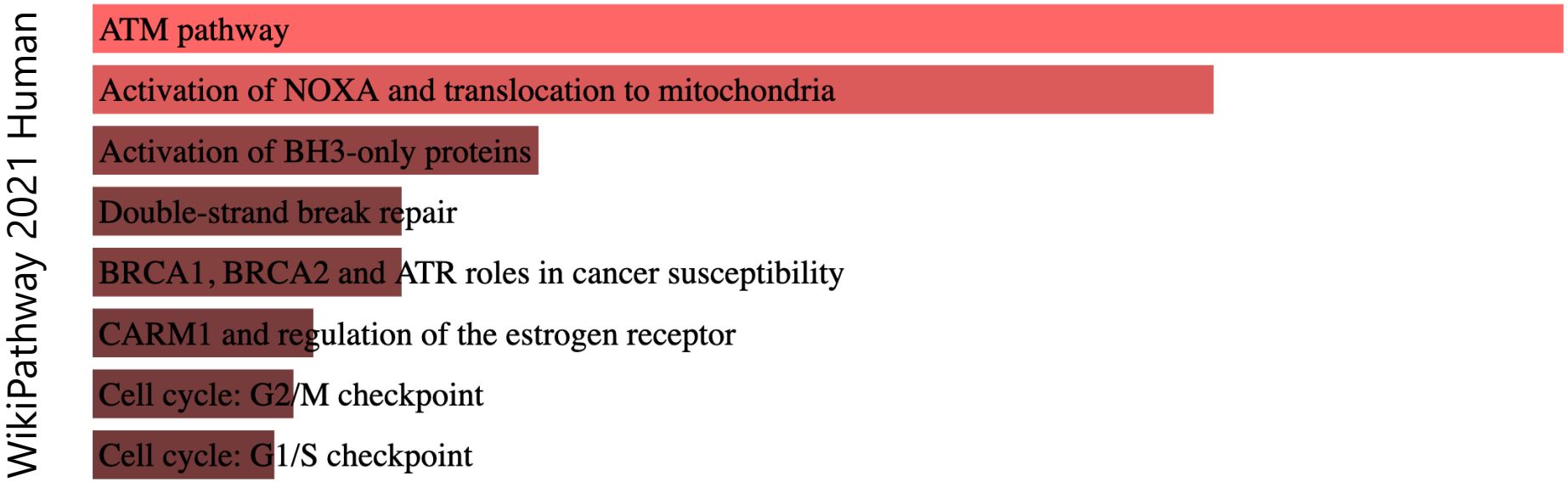
UNG sub-network filtered by altered TFs' PPI network



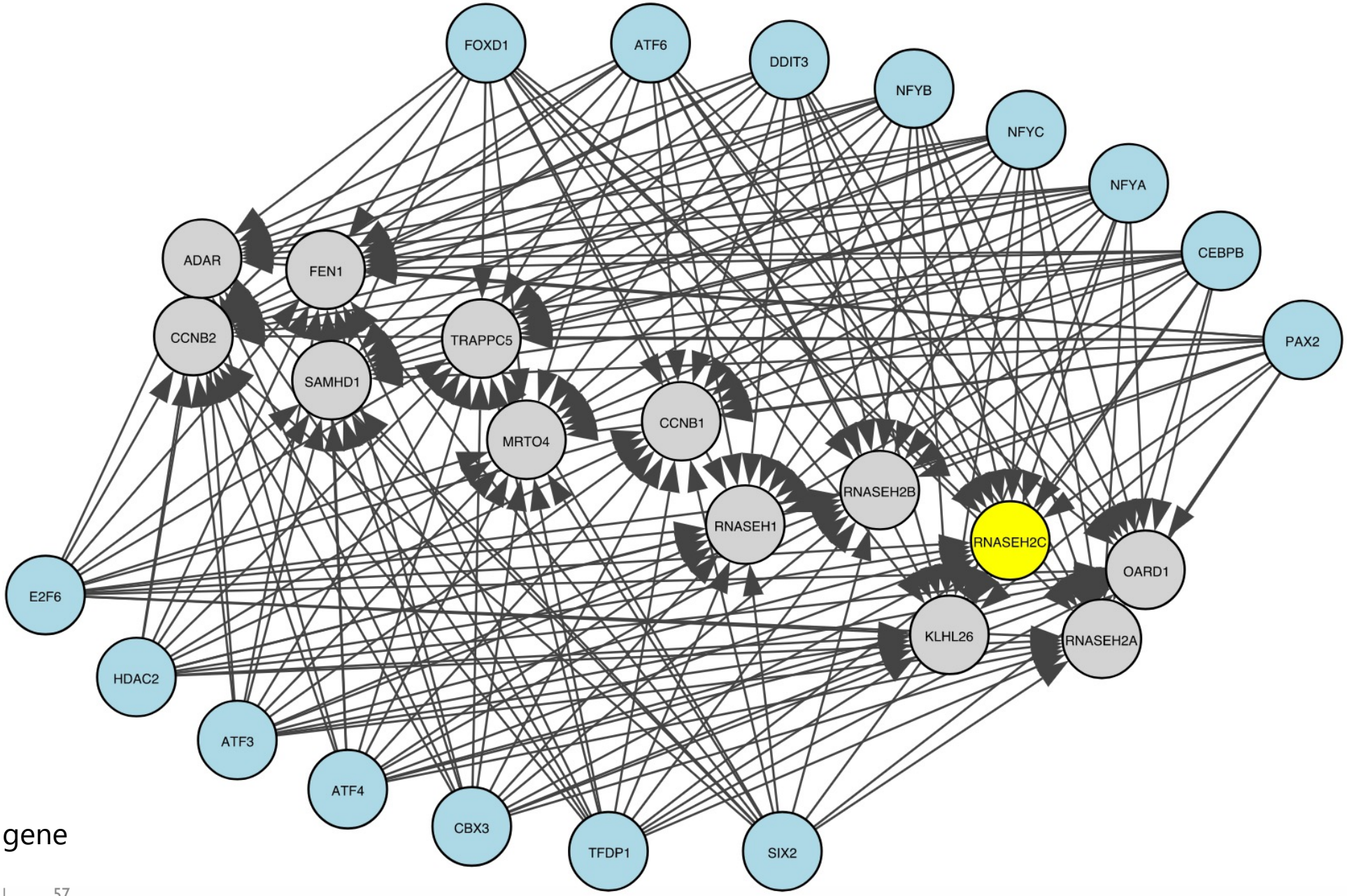
Filters:

1. Query TFs with dynamic and their PPI network regulon activity**:**
   - 'SIX2','DDIT3', 'CBX3'
2. Distance=1 from KO gene

Node name with Maximum degree:
'TFDP1', 'BRCA1', 'TP53', 'XRCC4', 'TBP'



RNASEH2C sub-network filtered by altered TFs' PPI network

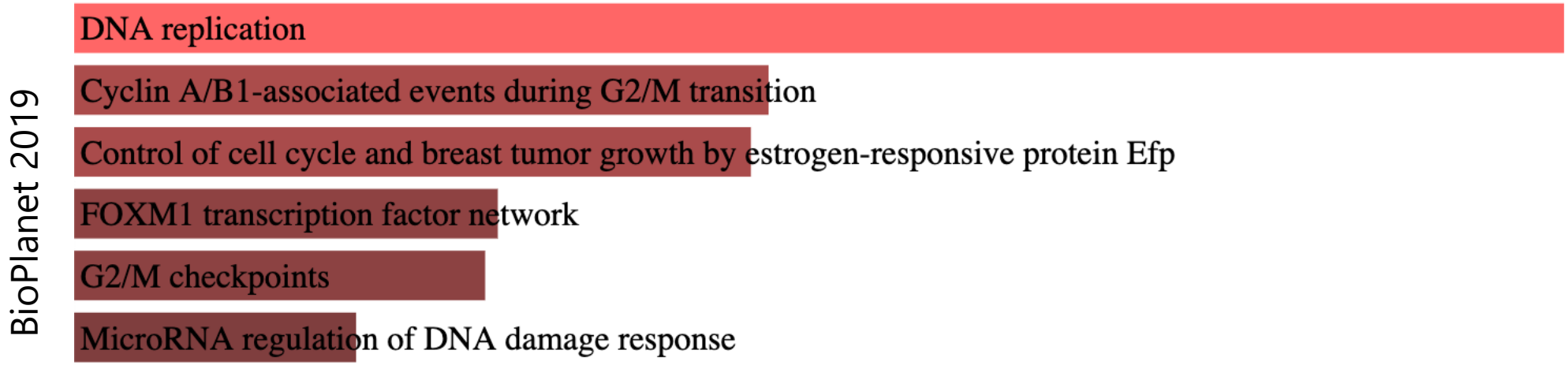


Filters:

1. Query TFs with dynamic and their PPI network regulon activity**:**
   - 'SIX2','DDIT3', 'CBX3'
2. Distance=1 or 2 from KO gene

Node name with Maximum degree:
'TRAPPC5', 'RNASEH1', 'FEN1', 'CCNB2', 'RNASEH2C', 'ADAR', 'SAMHD1', 'RNASEH2B', 'MRTO4', 'CCNB1', 'RNASEH2A', 'OARD1'

Enrichr



CONCLUSIONS

- Nano-bio-mimetic DNA damaged hairpins (DNA repair enzyme substrate) induce alterations in cellular gene regulatory network through changing some TF activities, and gene expression over time.
- We observed *CCNB1* over-represented in cells with high dU repair at 60' although it's opposite (over-represented in cells with low phenotype) in rG repair and earlier time repairing dU.
- It suggests potential dynamics of cell cycle due to the presence of DNA damage stimulus.
- RNASEH2C^{KO} Cells with high dU-repair might forbidden to replicate through a cell cycle check point. On the other hand, rG damage might skip the check point.
- Our analysis suggests *SIX2* and *DDIT3* TFs' activity increase by time due to the stimulus.
- *CBX3* is a TF with high centrality in the main context-specific network and subnetworks. It seems its activity decrease by time due to the stimulus.