Computational Genomics
Project 2 - Results

Team:
Agata Kaczmarek, Władysław Olejnik, Mateusz Stączek

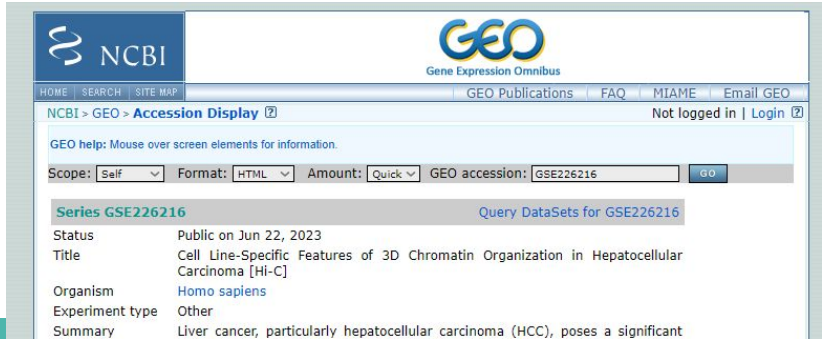
GITHUB: https://github.com/kaczmareka/CompGen_project2

Chosen approach

- Data
- Graph representation of interactions
- Arrowhead - ground truth
- Evaluation metrics
- Results

Datasets

- Requirements:
  - Small, below 0.5GB each
  - When loaded to Juicebox, should have easily visible target squares
  - From Hi-C experiments
- First file:
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE226216
  - GSE226216_HMEC_Res10_20_40_100_500kb.hic
  - human mammary epithelial cells (HMECs)
- Other files:
  - GSE226216_Huh1_Inter_Intra_Res10_20_40_100_500kb
  - GSE226216_SNU449_Inter_Intra_Res10_20_40_100_500kb
  - Both are hepatocellular carcinoma cell lines



Graph representation of interactions

- Interactions embedded as graph.
- Louvain Community Detection Algorithm
- To enhance performance of algorithm, we tried to erase interactions closest to main diagonal.
- We tried to find best parameters:
  - Available normalization algorithms
  - Best resolution parameter for LCDA
  - Leave diagonal as it is / erase main diagonal / erase main diagonal along with upper and lower diagonals

Arrowhead

- Algorithm which aims at finding contact domains.
- Highly automatic - you need to run three lines of code, including downloading the data and package.
- Output file contains information e.g. about:
  - Chromosome with the domain
  - Coordinates of the corner point of the domain
- The output from the algorithm is not perfect - as humans we would put some of the TADs in different places.
- Link to original repository: Arrowhead · aidenlab/juicer Wiki · GitHub

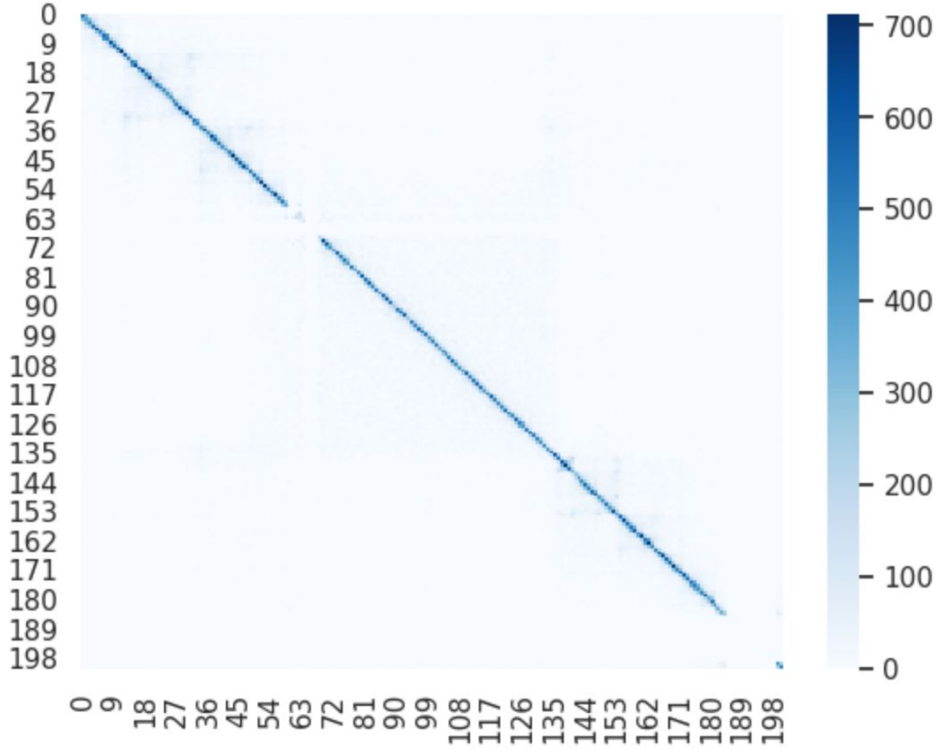
First results
For first file only

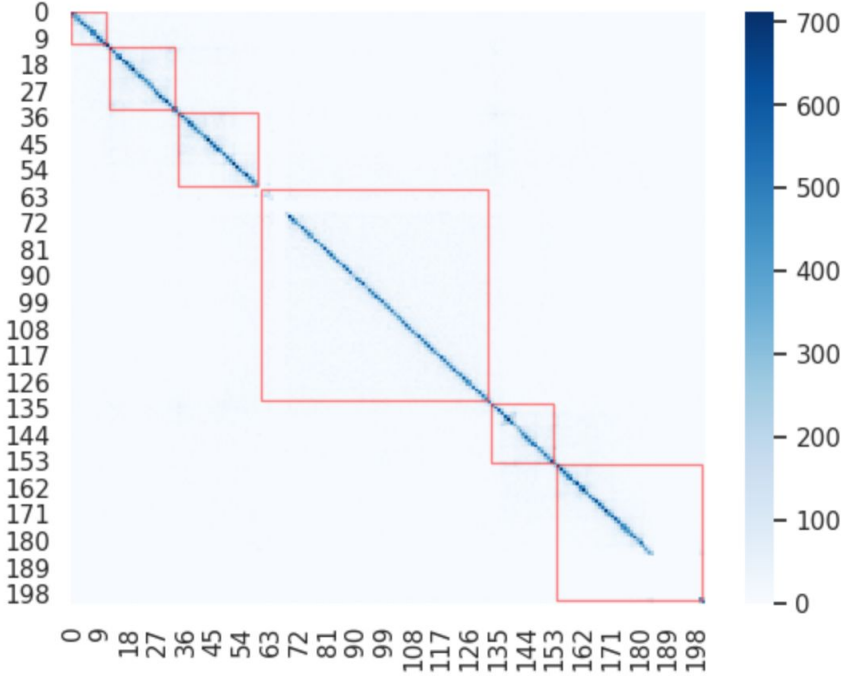
Analyzed for multiple metrics and settings of our algorithm

Evaluation metrics

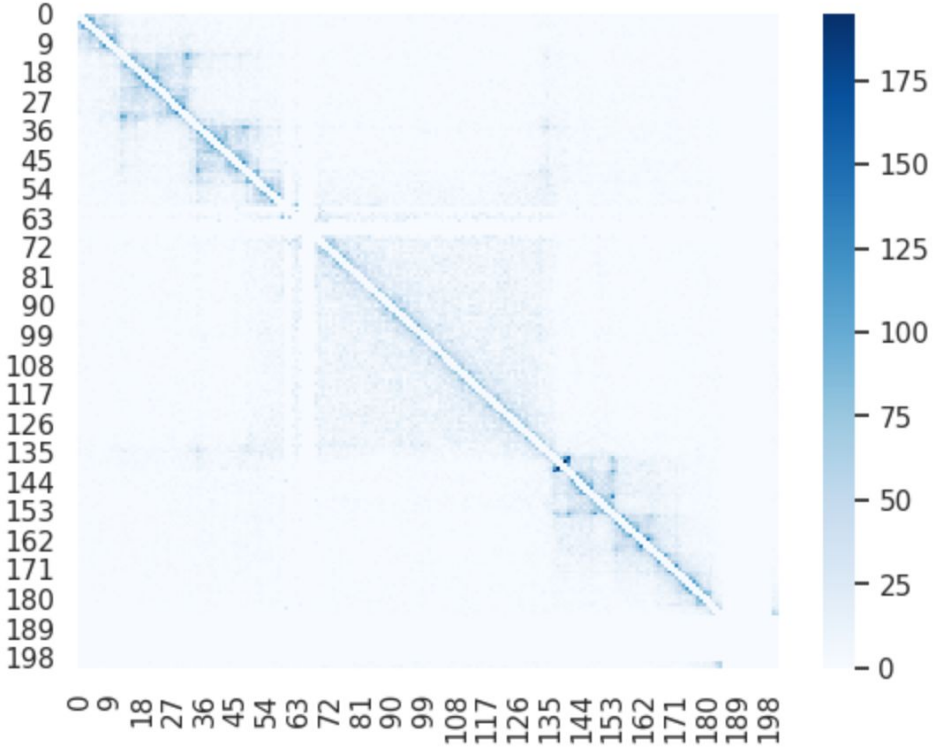
- The ground truth for our measures is the results we got from Arrowhead.
- Confusion matrix and calculated:
  - Accuracy
  - Precision
  - Recall
  - F1
  - Balanced Accuracy
- Histogram of detected TADs sizes - comparison between our method and Arrowhead.

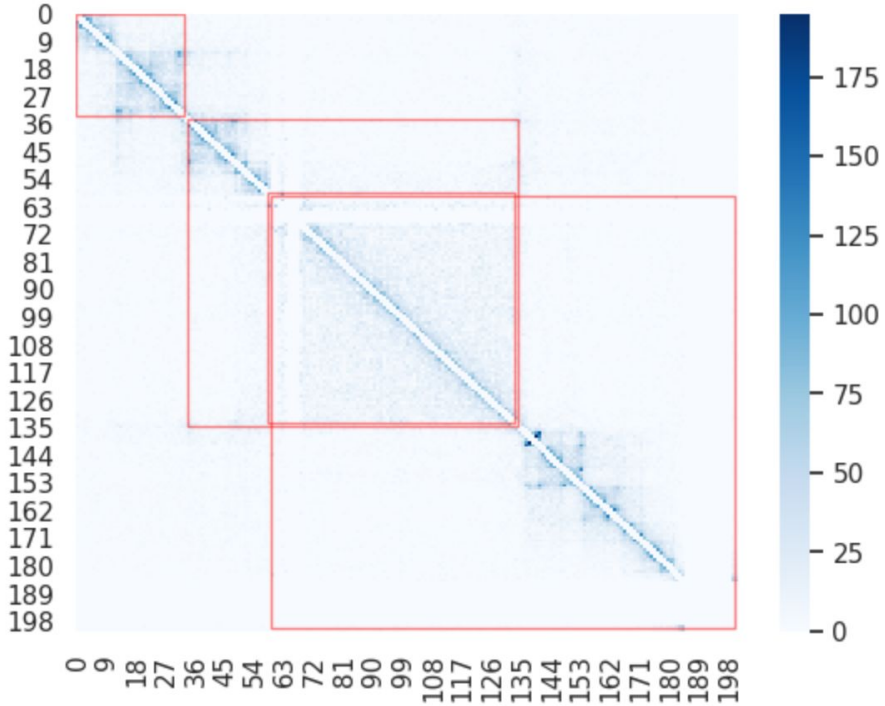
Example - normalization | w/o erasing diagonal | res=1



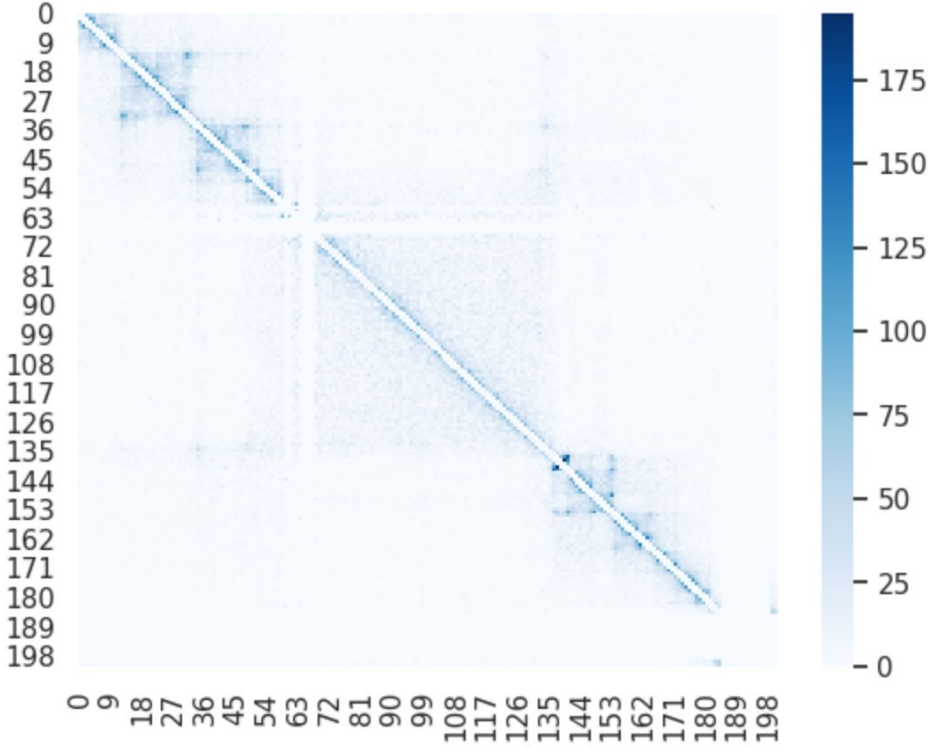


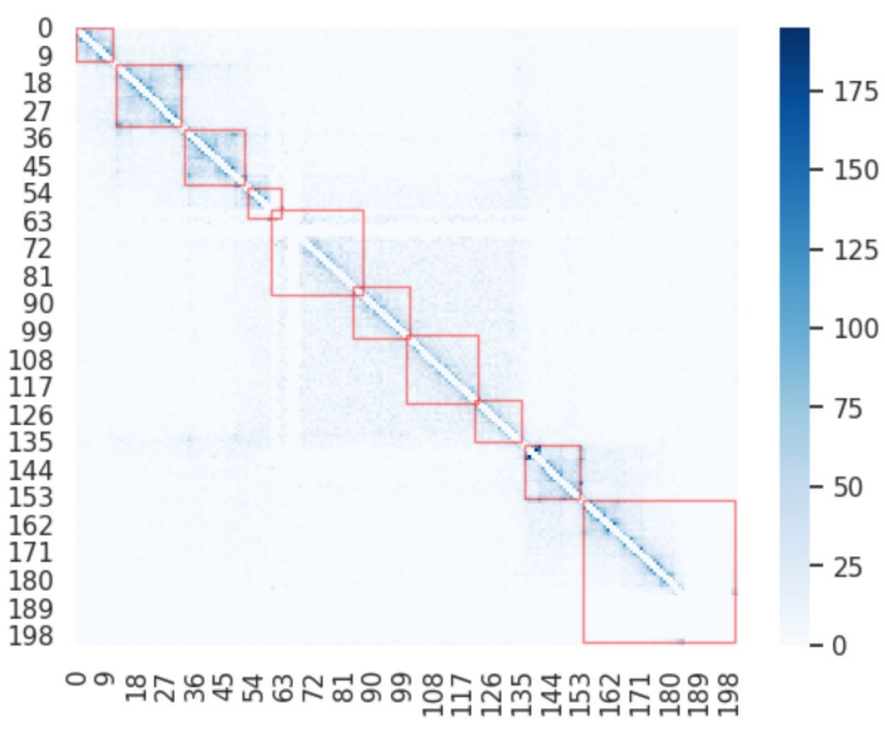
Example - normalization | erasing diagonal | res=1



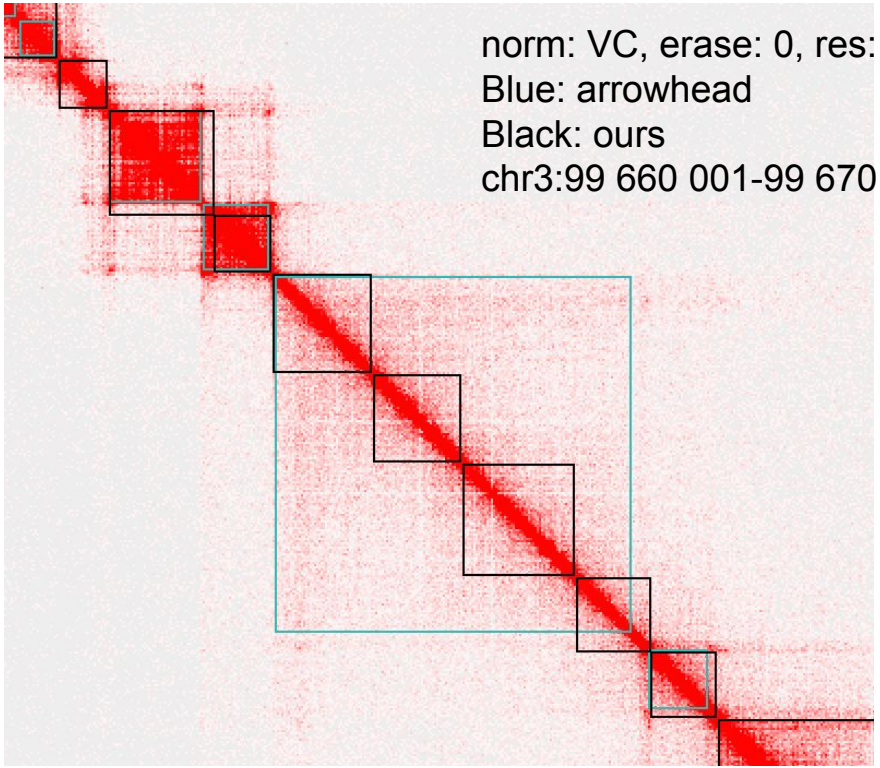


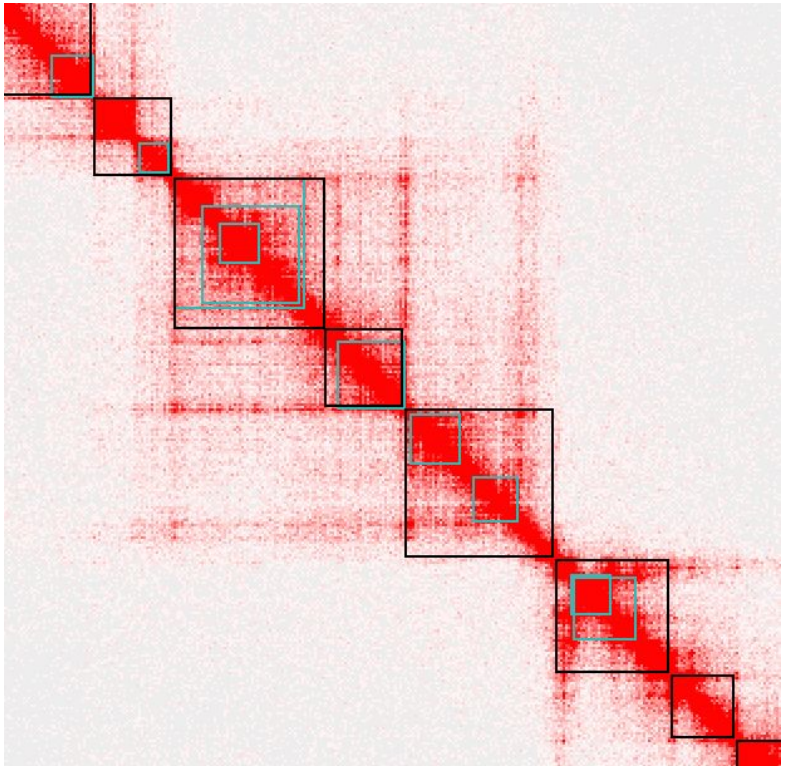
Example - normalization | erasing diagonal | res=2



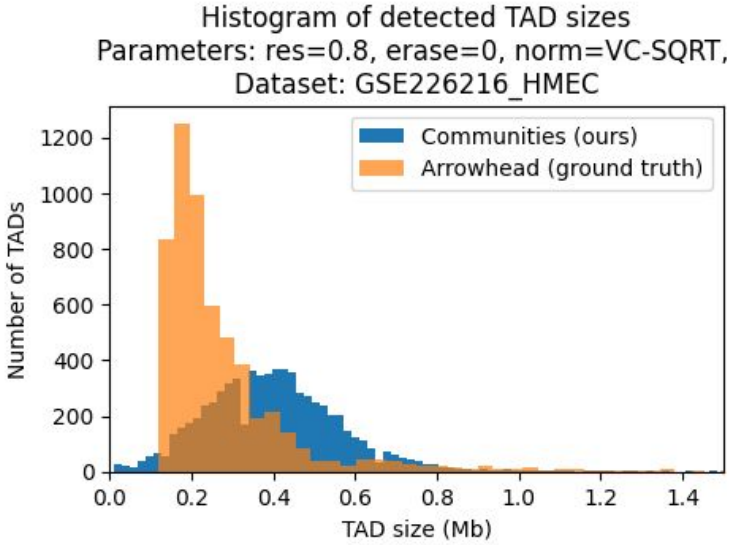


Example - normalization | w/o erasing diagonal | res=0.8





Results from first dataset




```
confusion_matrix / all_datapoints
✓ 0.0s
```

	Predicted TAD	Predicted non-TAD
Arrowhead TAD	0.489836	0.003116
Arrowhead non-TAD	0.428607	0.078442

```
metrics_df
✓ 0.0s
```

	Accuracy	Precision	Recall	F1	Balanced accuracy
0	0.5683	0.5333	0.9937	0.6941	0.5742

Final results

Analyzed for multiple files

- only specific metrics
- one setting of our algorithm,
- Added new metric - coverage

Among applied transformations, we used VC_SQRT normalization and erasing values close to the diagonal.

New metrics

New Metric algorithm:

- for each chromosome:
  - create a vector of ones and zeros, with ones where TADs were detected
  - multiply element-wise the vectors from 2 sources: our results and arrowhead results
  - for each of the source files:
    - for each community:
      - calculate the fraction of ones from this community that are in the product vector
      - it add to the current sum of products
    - divide the total sum of products by the number of communities in the source file

This way we obtain a metric that calculates the average fraction of elements from each community that are also in another results file.

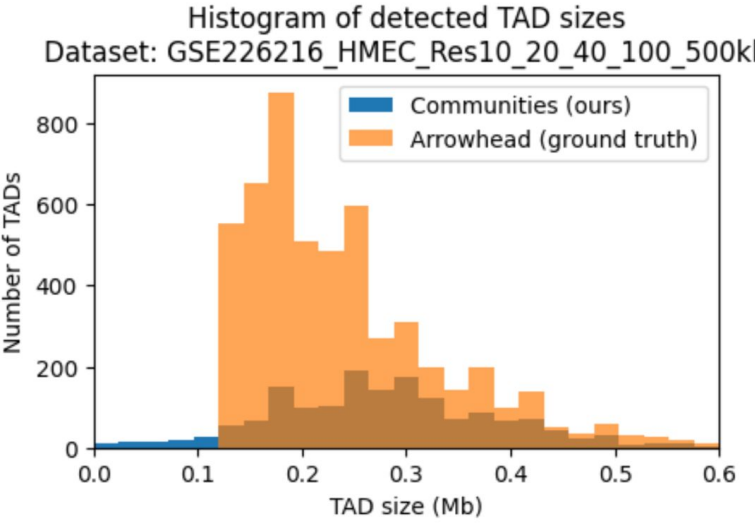
Results

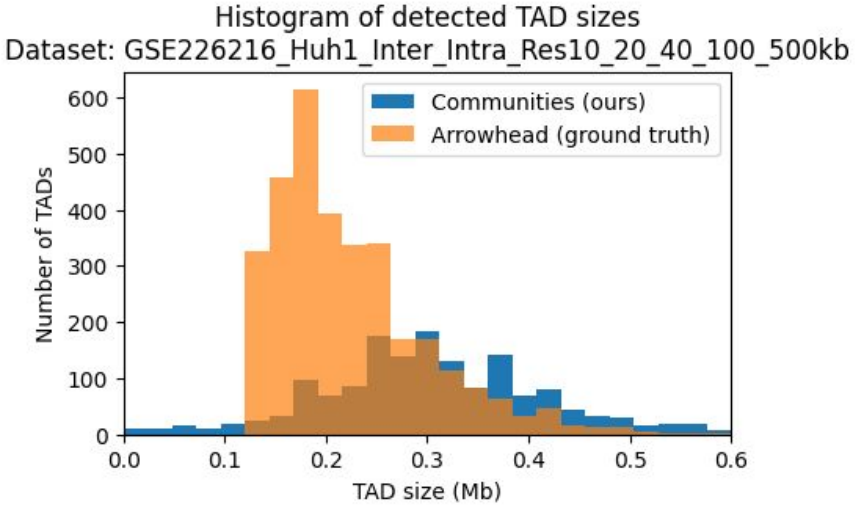
Filename	Balanced accuracy	Jaccard index	Average match for our communities	Average match for arrowhead
GSE226216_HMEC	0.4982	0.0352	0.142463	0.034653
GSE226216_Huh1_Inter_Intra	0.4987	0.0375	0.107103	0.040056
GSE226216_SNU449_Inter_Intra	0.4988	0.0391	0.110612	0.042513

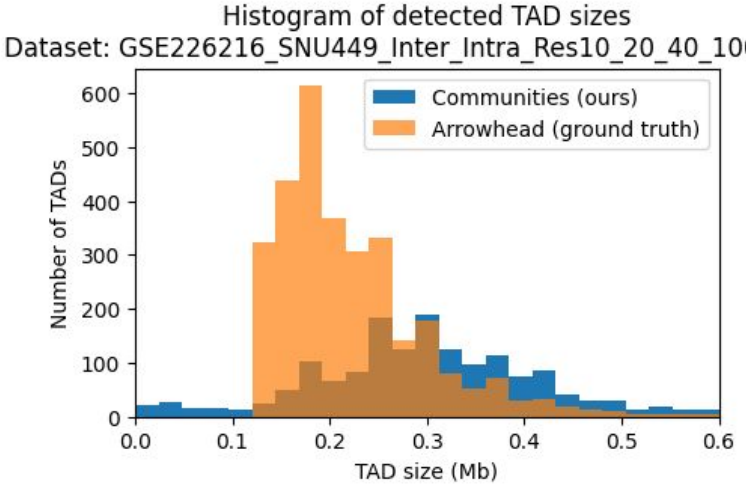
Results presented above are for all 3 files. Unfortunately, in all cases, our algorithm marked much fewer TADs than Arrowhead, and as such, the metrics are poor.

Our new metric shows that our results rarely overlap with Arrowhead.

Distributions of TAD sizes







Conclusions

Our implementation does:

- Marks TADs based on community detection algorithm
- Obtains different results from Arrowhead ground truth

Based on comparing our results with Arrowhead results on 3 file:

- TADs marked by our algorithm have similar sizes to those marked by Arrowhead, by we found either more or much fewer than correct,

In the future, we could fine tune more parameters, and try preprocessing too.

Thank you!

Any questions/ comments?