

# Rambutan

Jacob Schreiber

Paul G. Allen School of Computer Science  
University of Washington



jmschreiber91



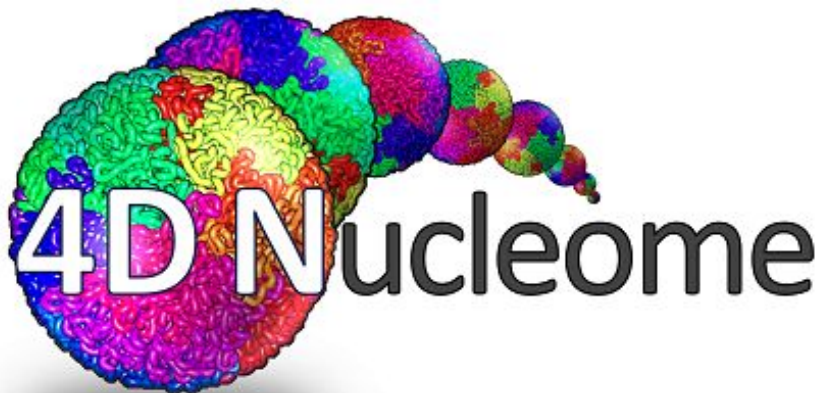
@jmschrei



@jmschreiber91



The structure of the genome is biologically important

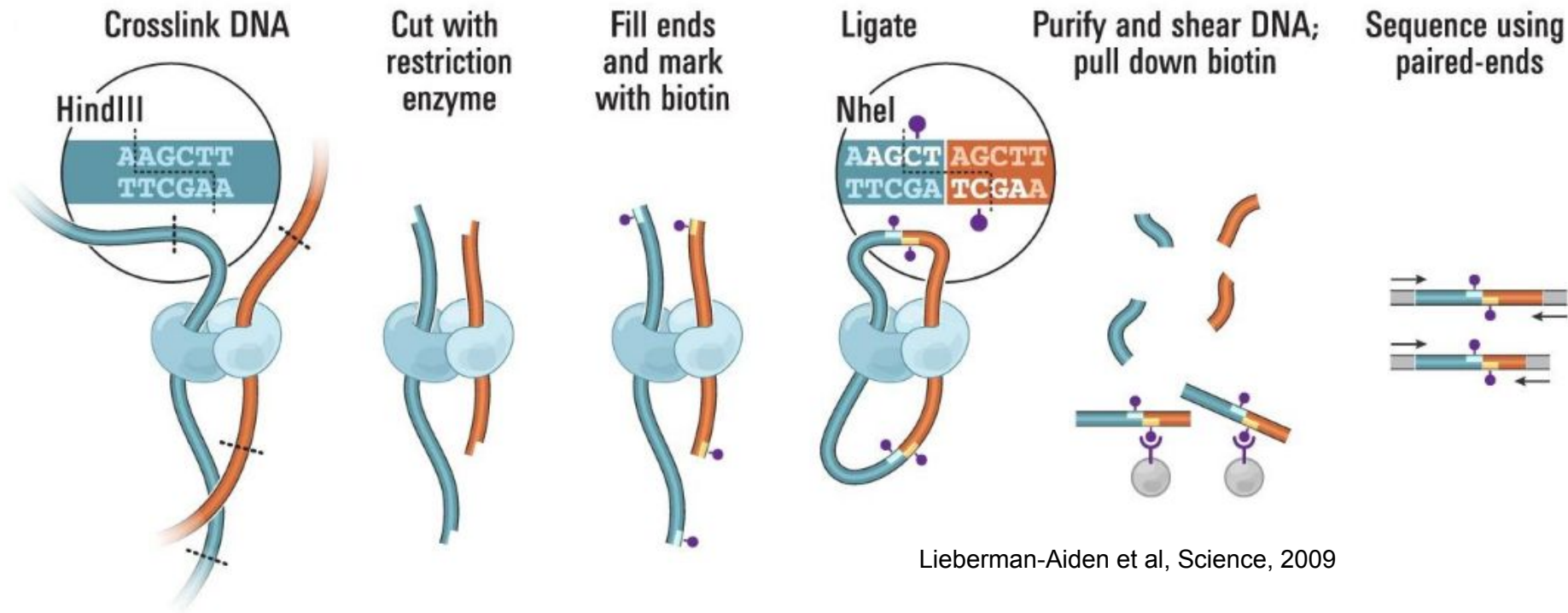


3D structure influences:

- Gene expression
- Replication timing
- Other processes?



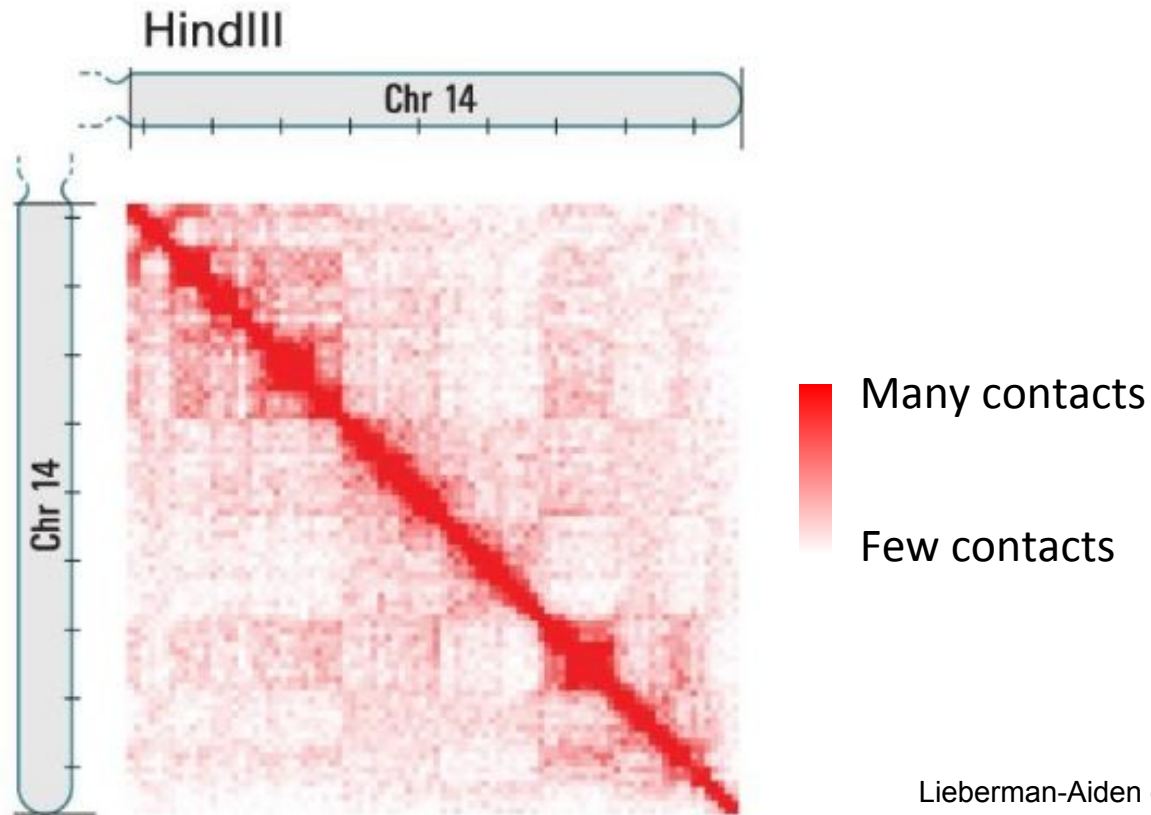
# Hi-C counts proximal pairs of DNA loci via cross-linking and sequencing



Lieberman-Aiden et al, Science, 2009



# Hi-C yields contact maps comprised of counts of pairwise contacts



Lieberman-Aiden et al, 2009

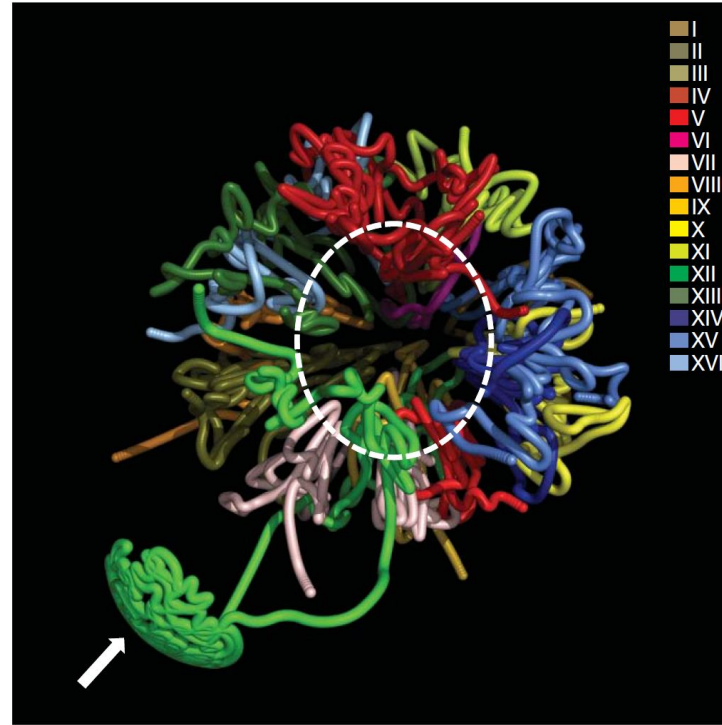
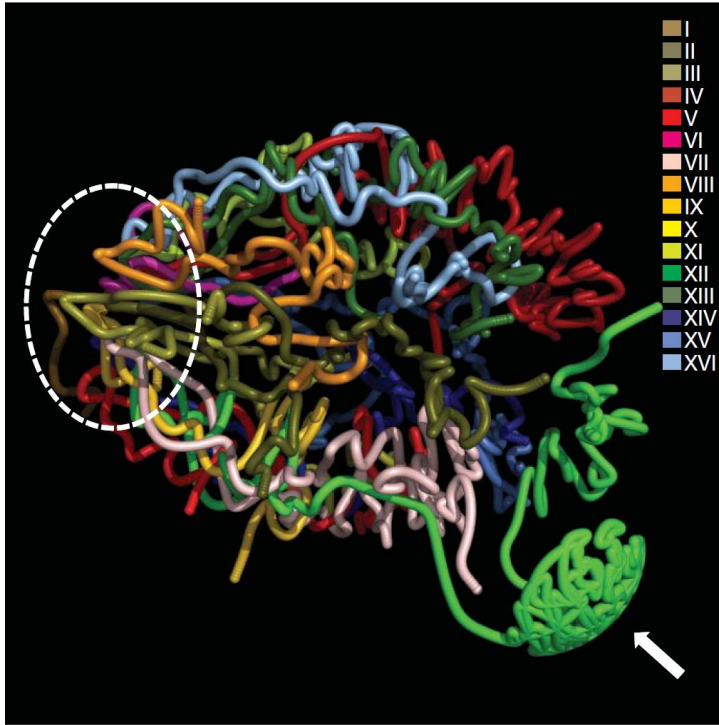


# These maps typically have five columns

chr1	fragmentMid1	chr2	fragmentMid2	contactCount
chr17	41400500	chr17	41463500	330
chr17	41382500	chr17	41463500	171
chr17	41401500	chr17	41463500	143
chr17	41381500	chr17	41463500	137
chr14	106103500	chr14	106198500	91
chr16	25066500	chr16	25137500	106
chr10	27180500	chr10	27644500	65
chr17	43651500	chr17	44566500	54
chr19	53361500	chr19	53446500	96
chr3	195346500	chr3	195453500	79
chr19	23562500	chr19	23982500	48
chr19	15785500	chr19	15888500	69
chr17	41400500	chr17	41561500	63
chr6	29884500	chr6	29970500	86
chr1	31192500	chr1	31252500	68
chr1	46503500	chr1	46645500	58
chr15	75323500	chr15	75497500	54
chr17	41382500	chr17	41440500	64
chr6	37019500	chr6	37140500	55
chr17	41400500	chr17	41466500	61
chr19	21272500	chr19	21334500	56
chr1	121334500	chr1	121485500	52
chr19	10692500	chr19	10755500	60
chr14	22855500	chr14	22941500	53
chr1	46500500	chr1	46645500	47



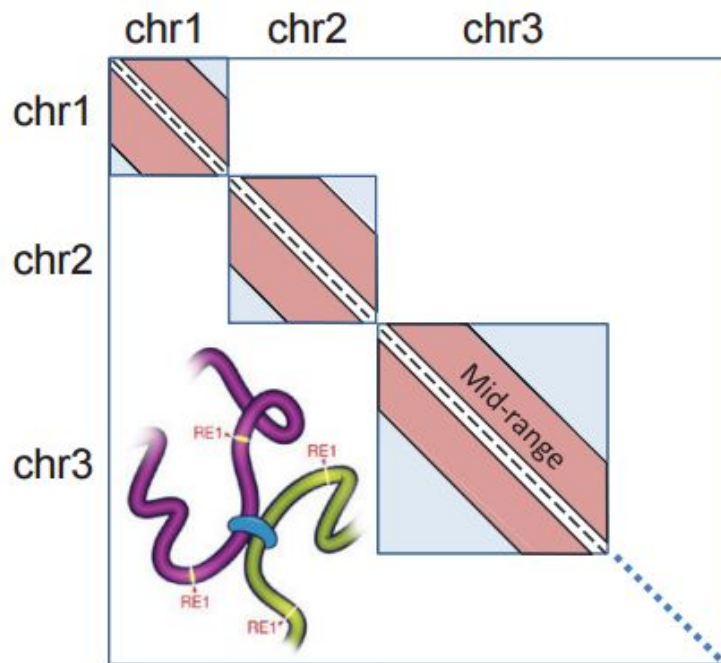
# 3D models of the genome can be built from Hi-C contact maps (yeast genome shown)



Duan et al, 2010



# Statistical significance can be assigned to the contacts in a Hi-C contact map with Fit-Hi-C



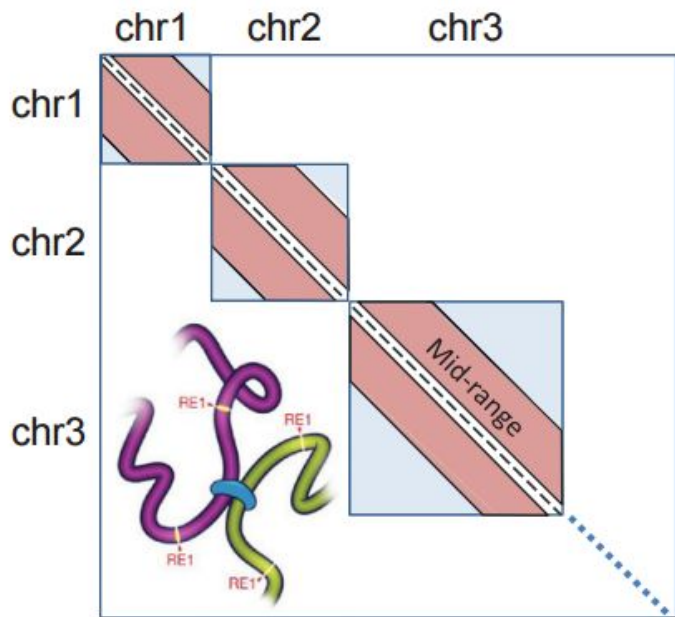
Ay et al, 2014

Genome-wide Hi-C contact map and mid-range contacts

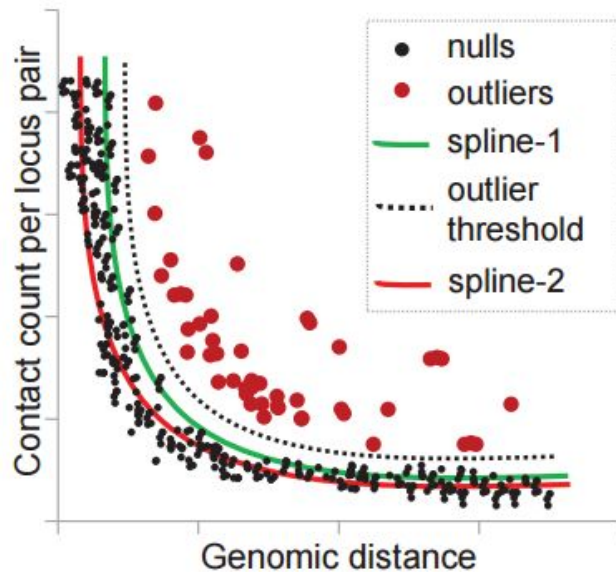




# Statistical significance can be assigned to the contacts in a Hi-C contact map with Fit-Hi-C



Genome-wide Hi-C contact map and mid-range contacts

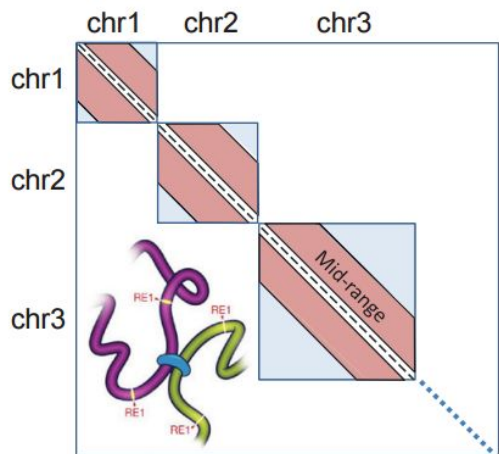


Outlier removal and spline fit to the refined null

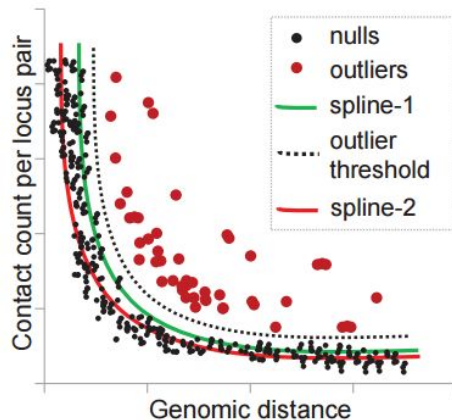




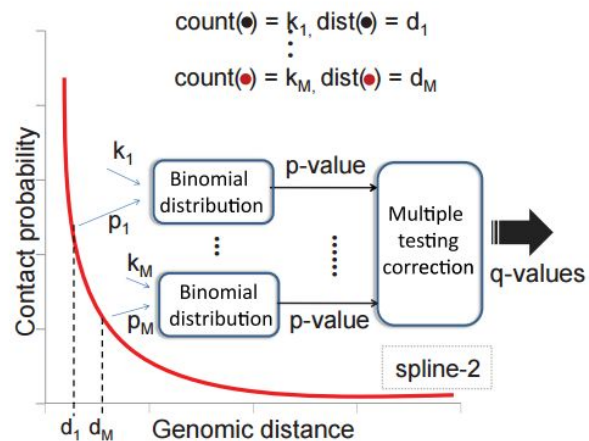
# Statistical significance can be assigned to the contacts in a Hi-C contact map with Fit-Hi-C



Genome-wide Hi-C contact map and mid-range contacts



Outlier removal and spline fit to the refined null



Statistical confidence estimation

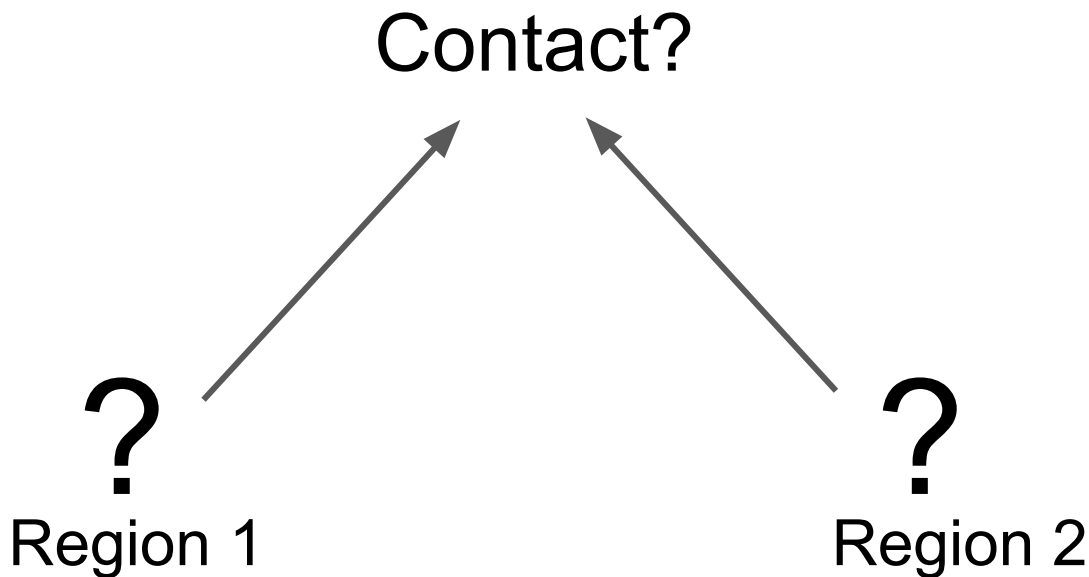


# After processing these maps now have 7 columns

chr1	fragmentMid1	chr2	fragmentMid2	contactCount	p-value	q-value
chr17	41400500	chr17	41463500	330	0.0	0.0
chr17	41382500	chr17	41463500	171	3.15385322826e-318	3.74721601131e-309
chr17	41401500	chr17	41463500	143	2.06334574749e-243	1.63436103126e-234
chr17	41381500	chr17	41463500	137	8.44843137138e-243	5.01895538791e-234
chr14	106103500	chr14	106198500	91	2.45145662377e-202	1.16506848566e-193
chr16	25066500	chr16	25137500	106	7.91004715473e-180	3.13274516421e-171
chr10	27180500	chr10	27644500	65	5.31998093699e-141	1.80596480415e-132
chr17	43651500	chr17	44566500	54	1.96084610385e-130	5.82439430775e-122
chr19	53361500	chr19	53446500	96	5.27621204289e-130	1.39308295213e-121
chr3	195346500	chr3	195453500	79	8.00442265288e-117	1.90207334047e-108
chr19	23562500	chr19	23982500	48	7.10417401876e-114	1.53468112603e-105
chr19	15785500	chr19	15888500	69	1.0168757619e-109	2.0136495797e-101
chr17	41400500	chr17	41561500	63	2.75946691286e-106	5.04404653862e-98
chr6	29884500	chr6	29970500	86	4.52065523647e-98	7.67309539777e-90
chr1	31192500	chr1	31252500	68	2.13253810283e-94	3.37833557675e-86
chr1	46503500	chr1	46645500	58	2.84313465789e-93	4.22254895095e-85
chr15	75323500	chr15	75497500	54	2.4278542551e-89	3.3936803759e-81
chr17	41382500	chr17	41440500	64	1.08726104061e-86	1.35980762344e-78
chr6	37019500	chr6	37140500	55	2.92387510113e-84	3.47397003001e-76
chr17	41400500	chr17	41466500	61	9.98510422997e-84	1.12987540236e-75
chr19	21272500	chr19	21334500	56	2.65934221052e-83	2.87242562426e-75
chr1	121334500	chr1	121485500	52	1.6191432774e-82	1.67284110401e-74
chr19	10692500	chr19	10755500	60	3.88896551421e-81	3.85052632118e-73
chr14	22855500	chr14	22941500	53	2.78497787752e-74	2.64715265765e-66
chr1	46500500	chr1	46645500	47	5.06587900171e-72	4.62997568924e-64
chr1	46501500	chr1	46645500	47	6.62227676568e-72	5.82828538832e-64
chr12	108957500	chr12	109011500	56	8.0435693286e-72	6.82634171027e-64



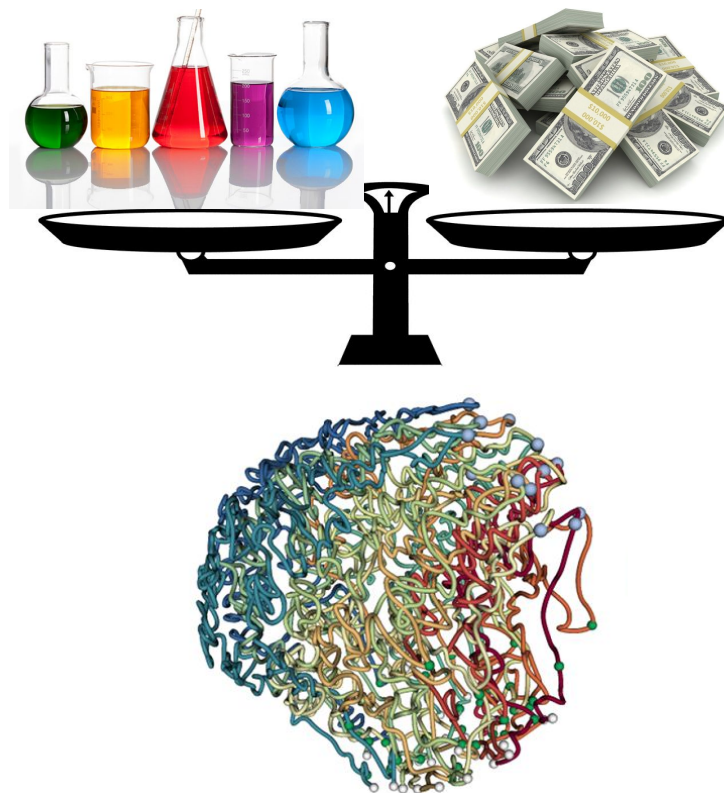
# Can we predict statistical significance directly?





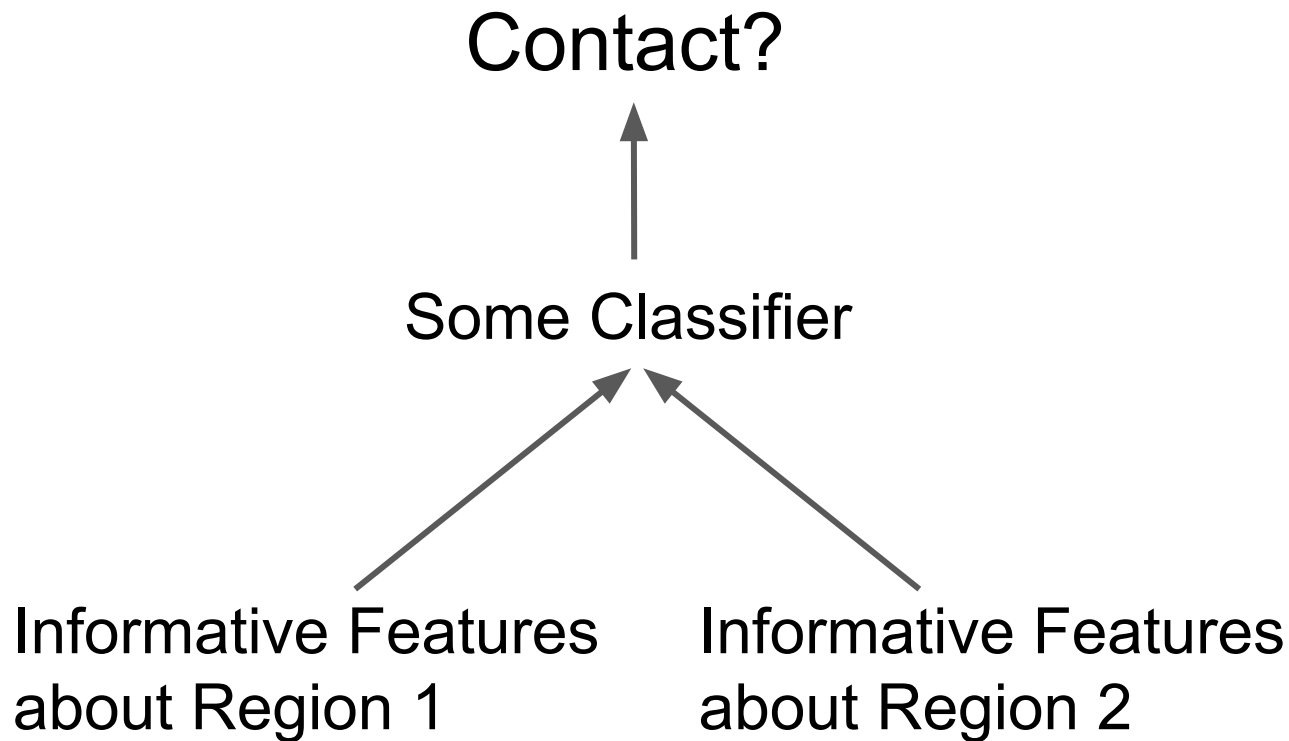
# Why bother trying to predict significance?

- Hi-C experiments are expensive, we can save money and time by predicting from cheaper data.
- An accurate model can reveal the genetic basis of 3D architecture.





Traditionally one would build a model with informative features





# Nucleotide sequence is certainly relevant

Contact?  
↑  
Some Classifier

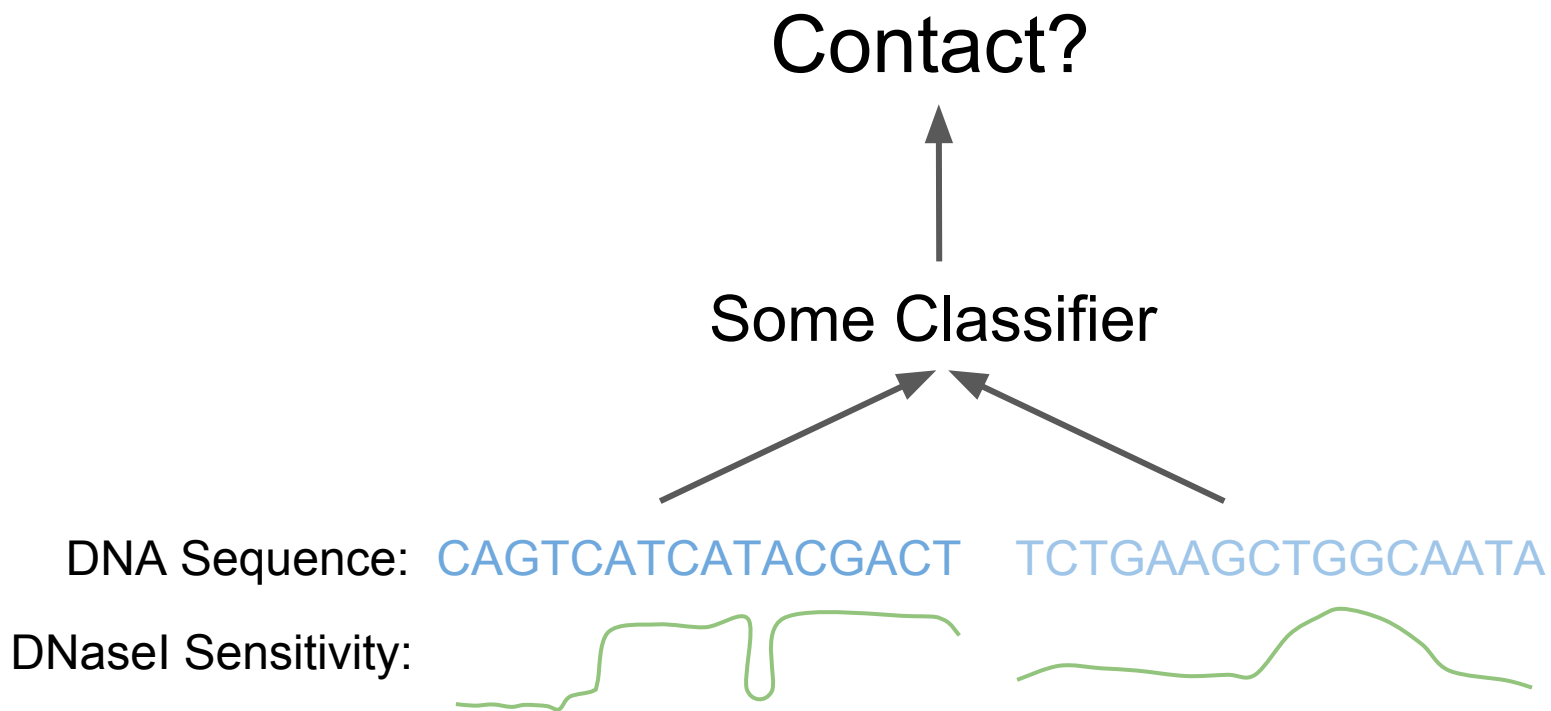
But the DNA  
sequence is the  
same in every  
cell type

DNA Sequence: CAGTCATCATACGACT TCTGAAGCTGGCAATA



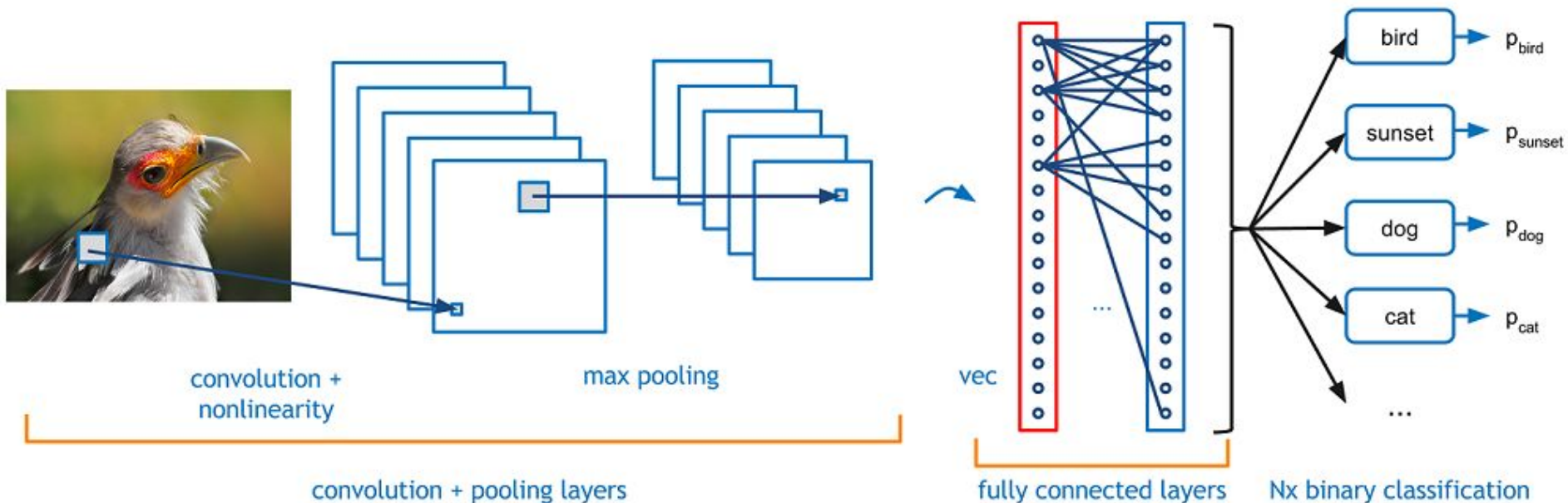


# DNaseI sensitivity can provide cell-type specific information





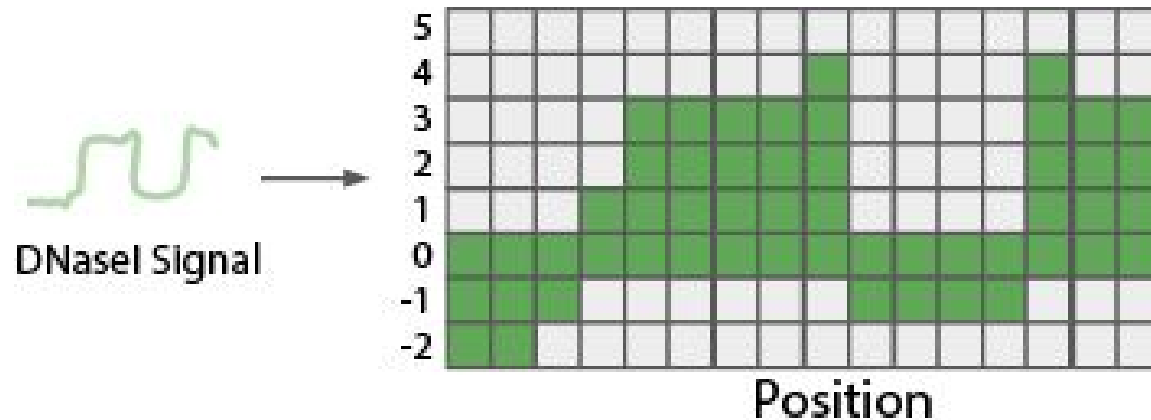
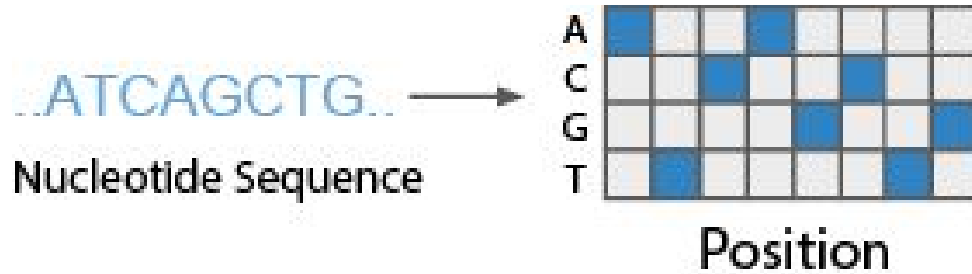
# Neural networks are good models when raw, structured, data is plentiful



<https://adeshpande3.github.io/adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/>

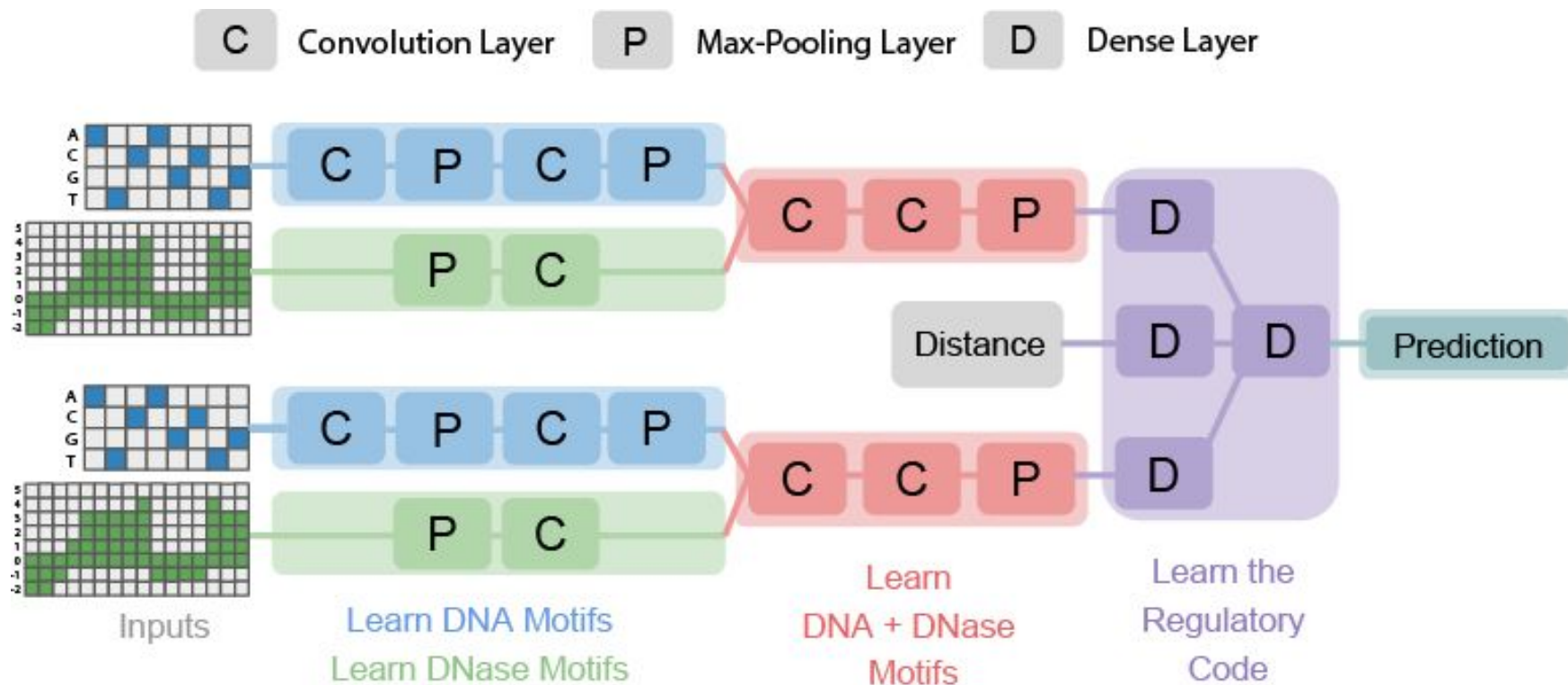


# Data is prepared by bit encoding it



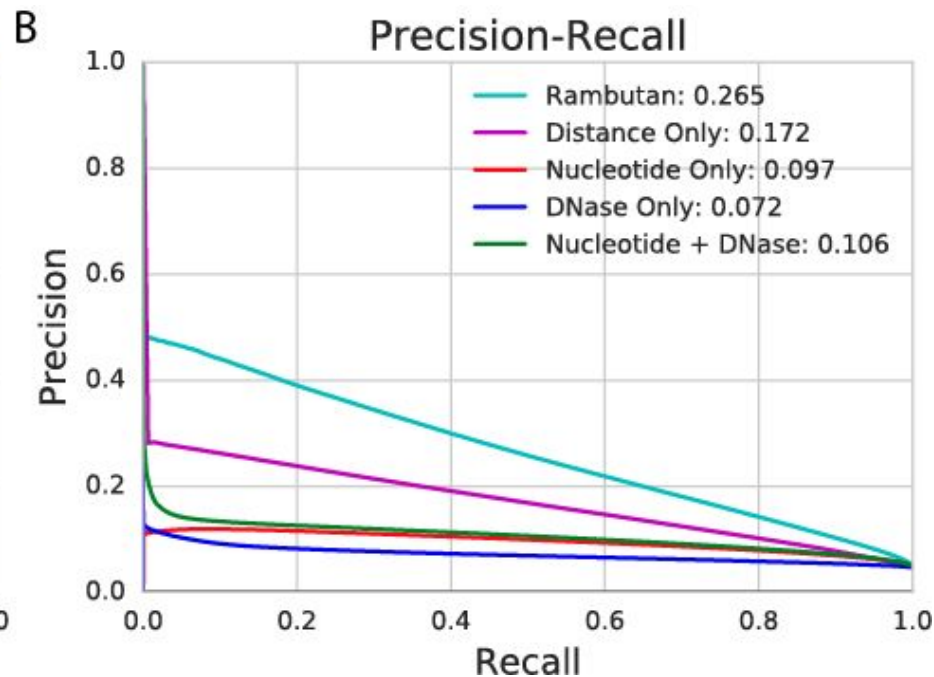
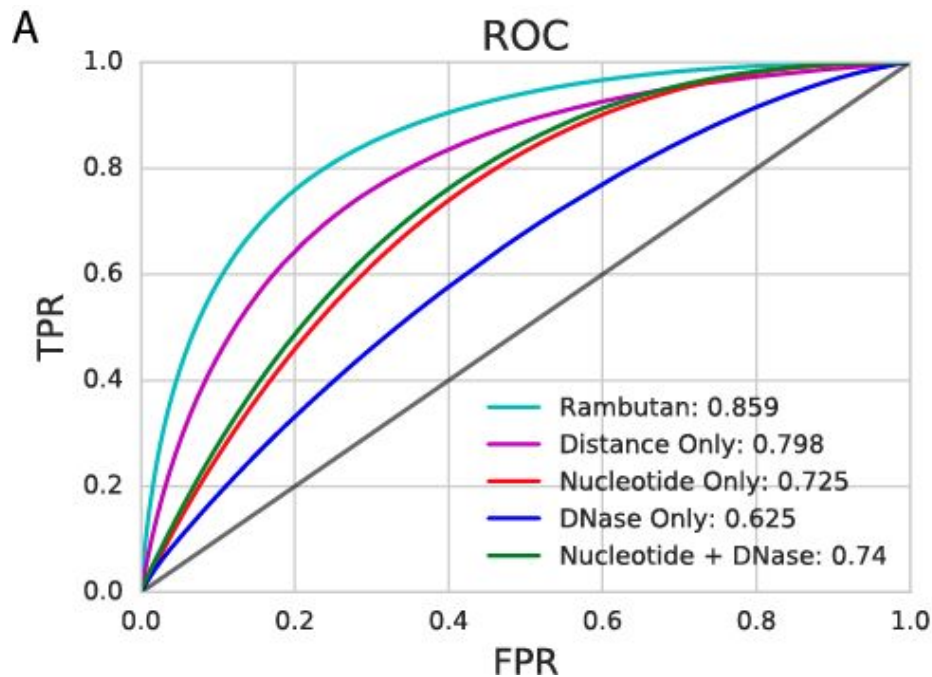


# Rambutan is a convolutional neural network that predicts statistical significance directly

















# Rambutan performs well at cross-chromosomal predictions in GM12878





# We only have 1kb resolution data for GM12878

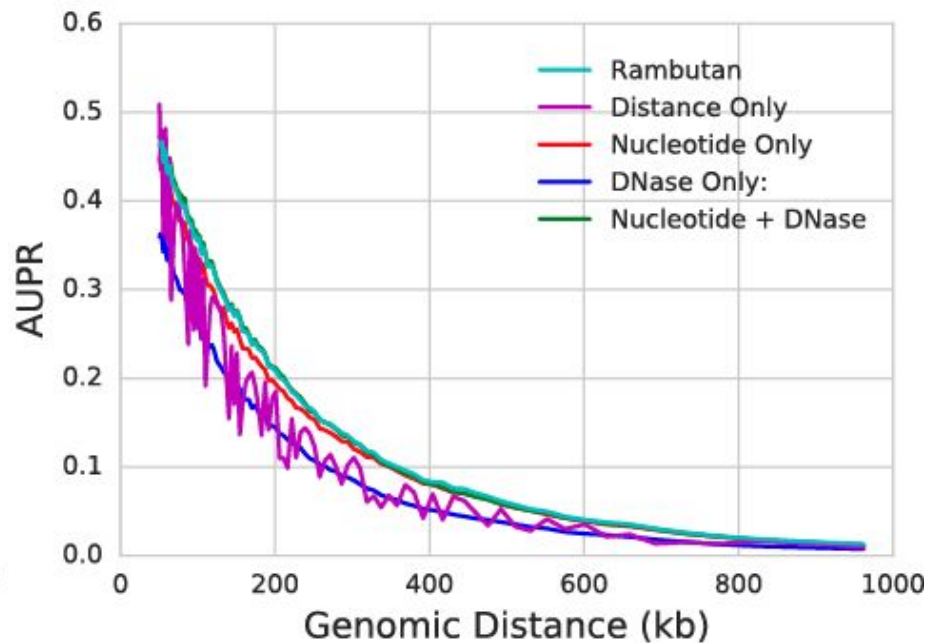
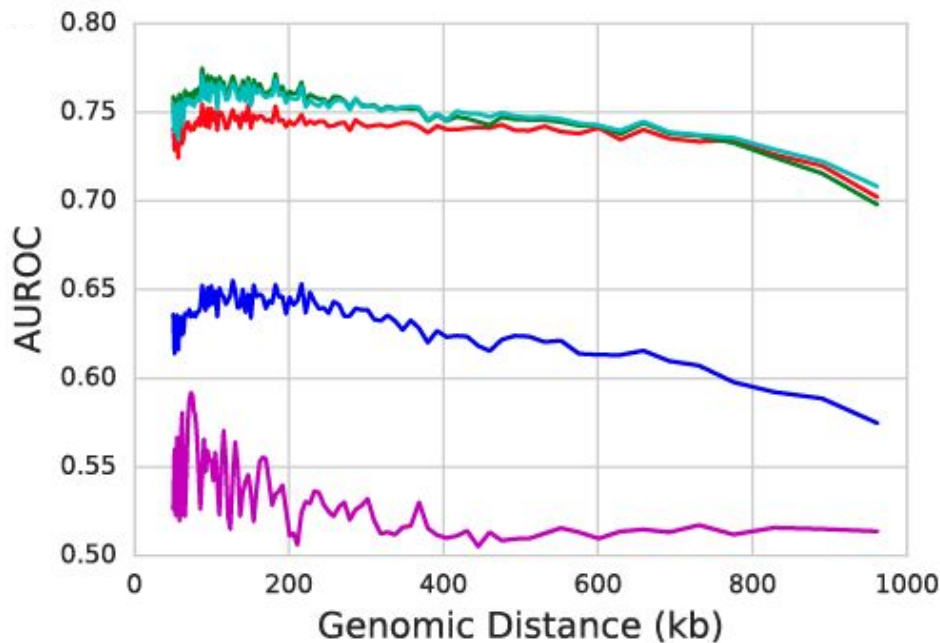
	GM12878	K562	IMR90	HUVEC	HMEC	NHEK
5kb Res.						
1kb Res.						

Solution: Train in GM12878, make predictions in other cell types at 1 kb resolution, and then convert to 5 kb resolution



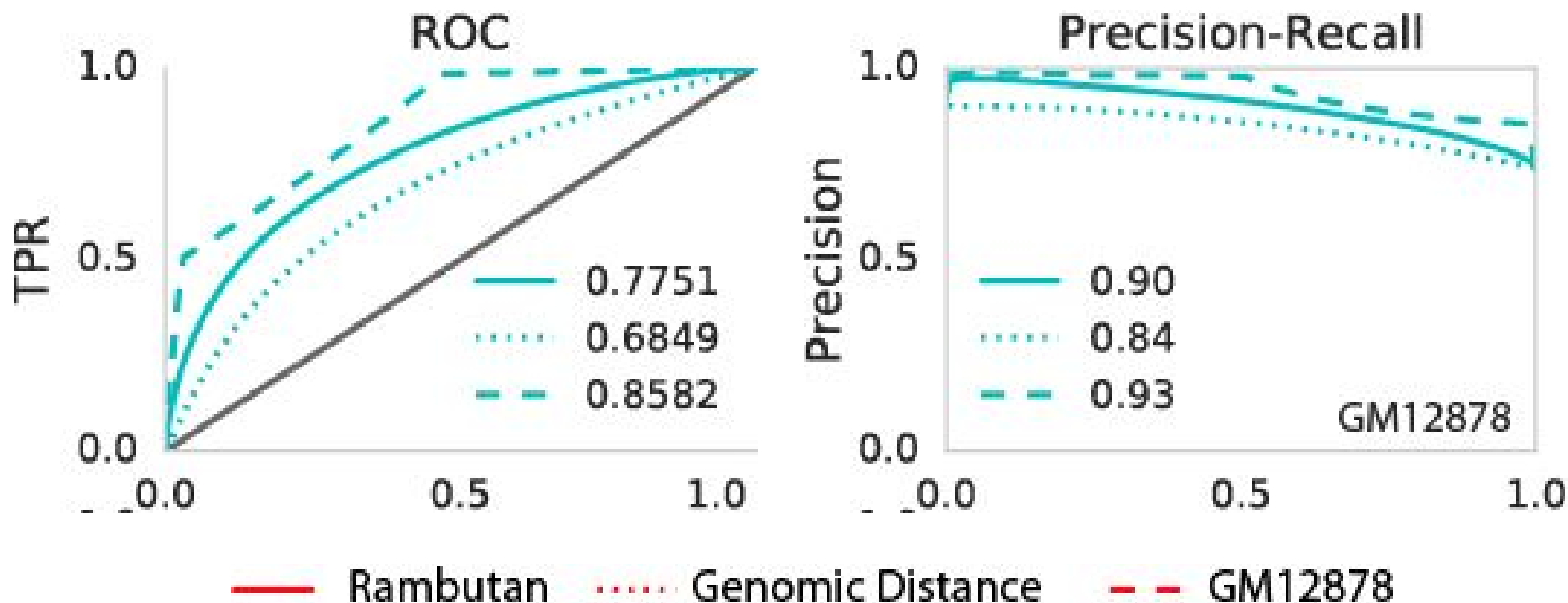


# Rambutan performs well at all genomic distances



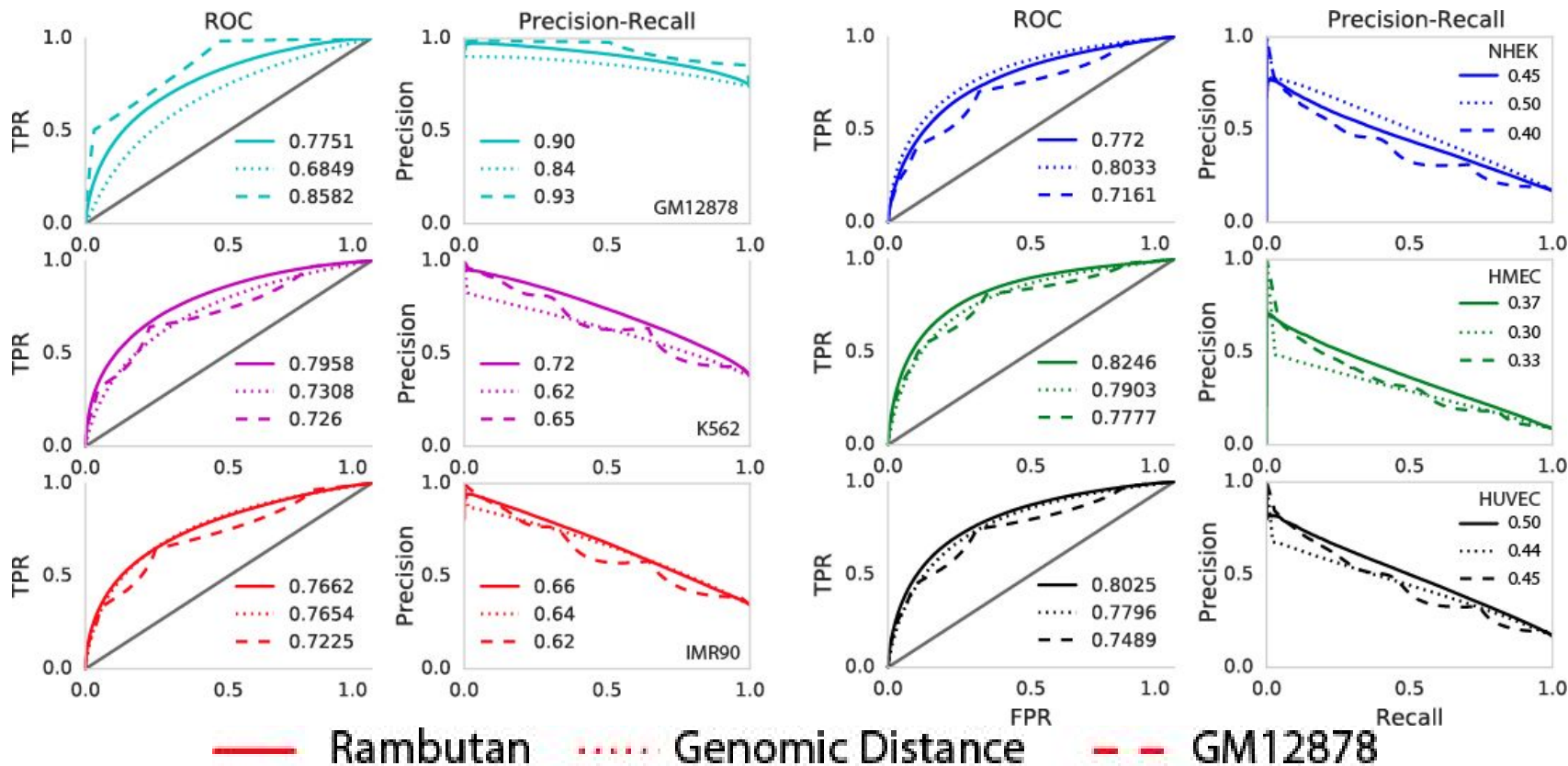


# Rambutan performs well across cell types at 5kb resolution



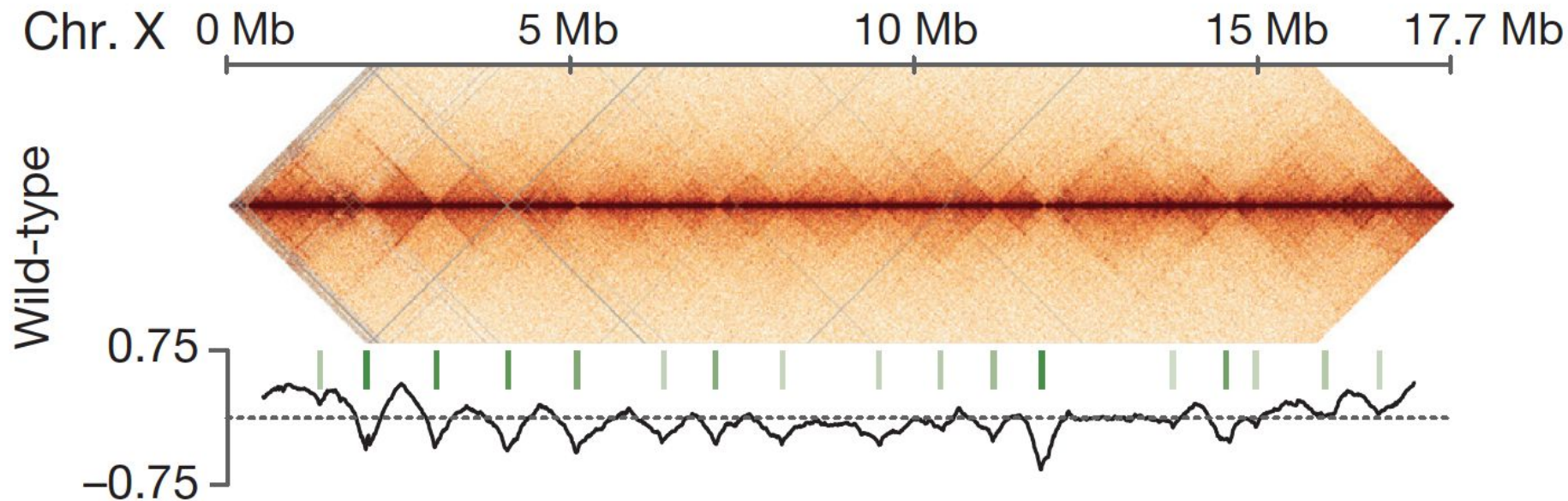


# Rambutan performs well across cell types at 5kb resolution





Insulation score is a measurement of local connectivity

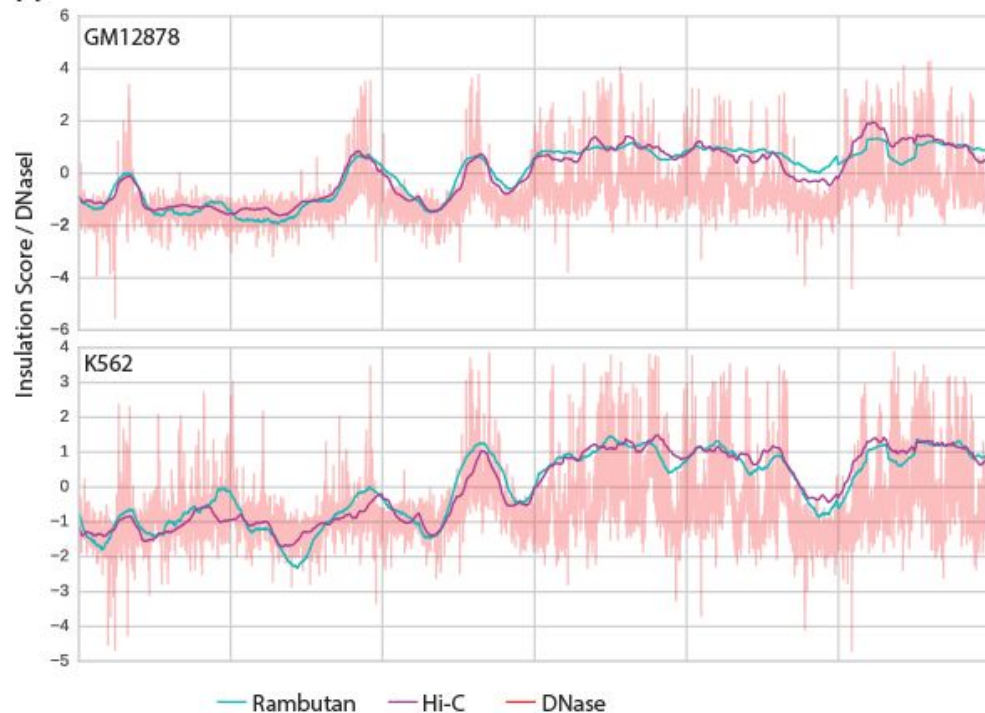


Crane et al *Nature* 2015

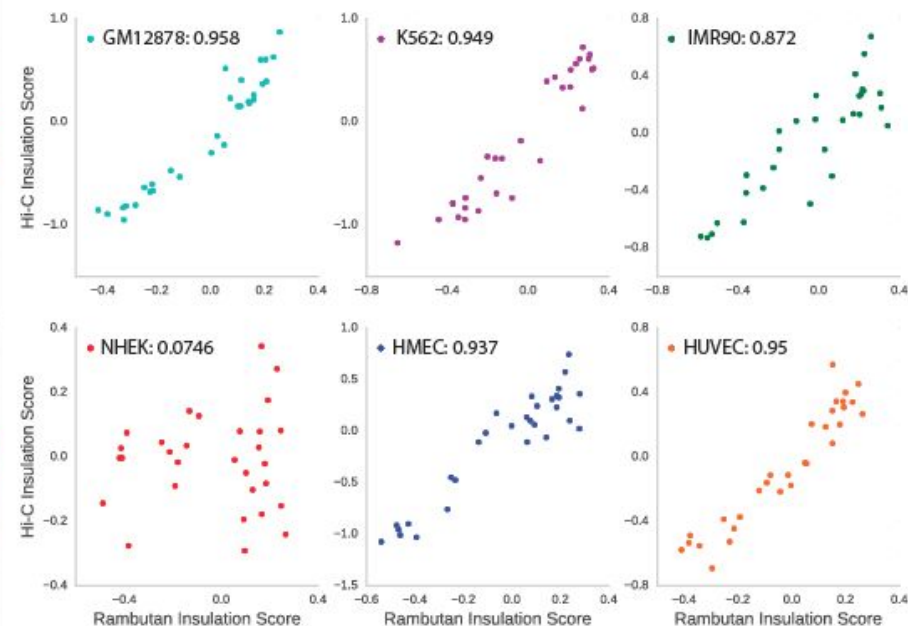


# Rambutan predictions can recreate insulation score

A

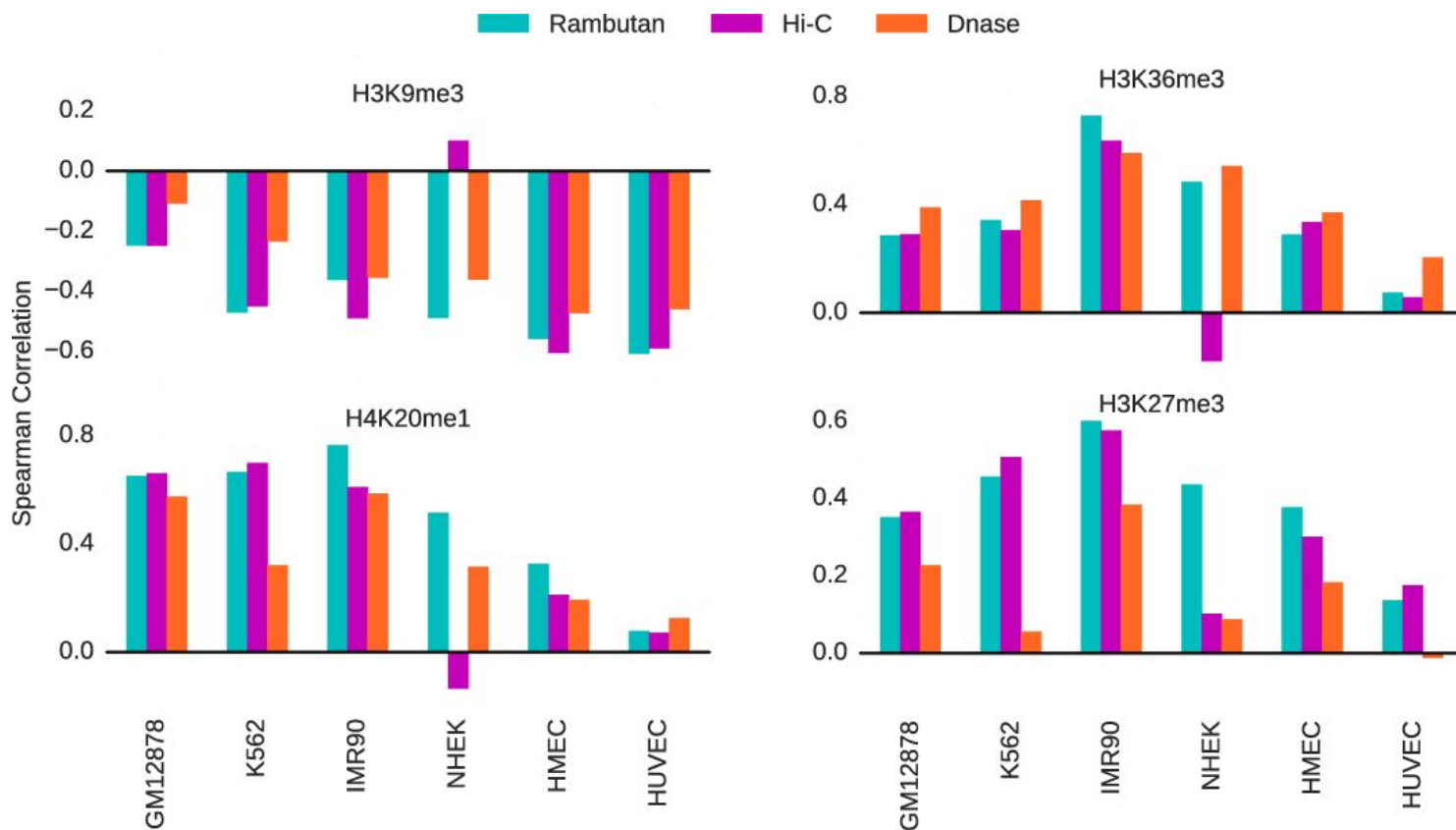


B





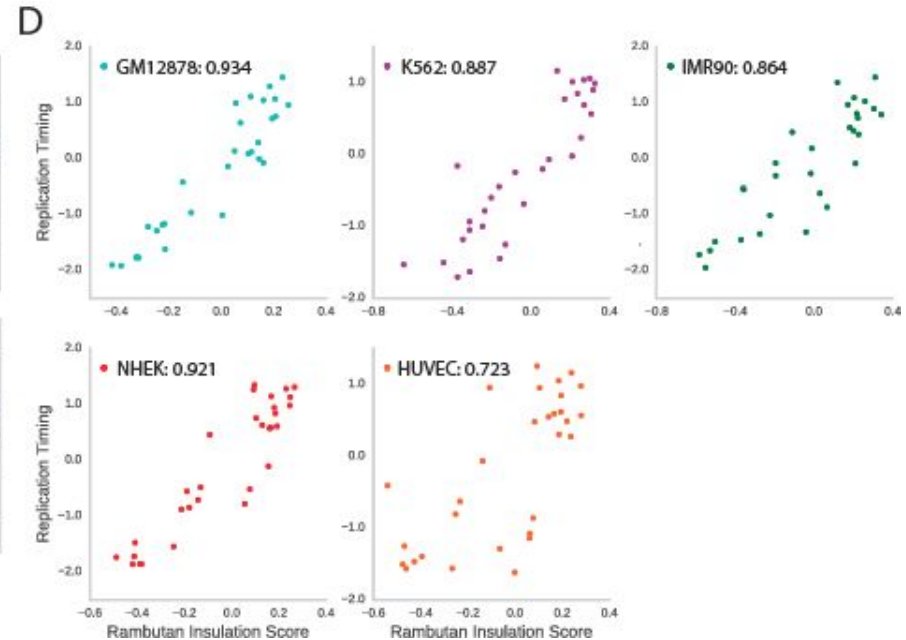
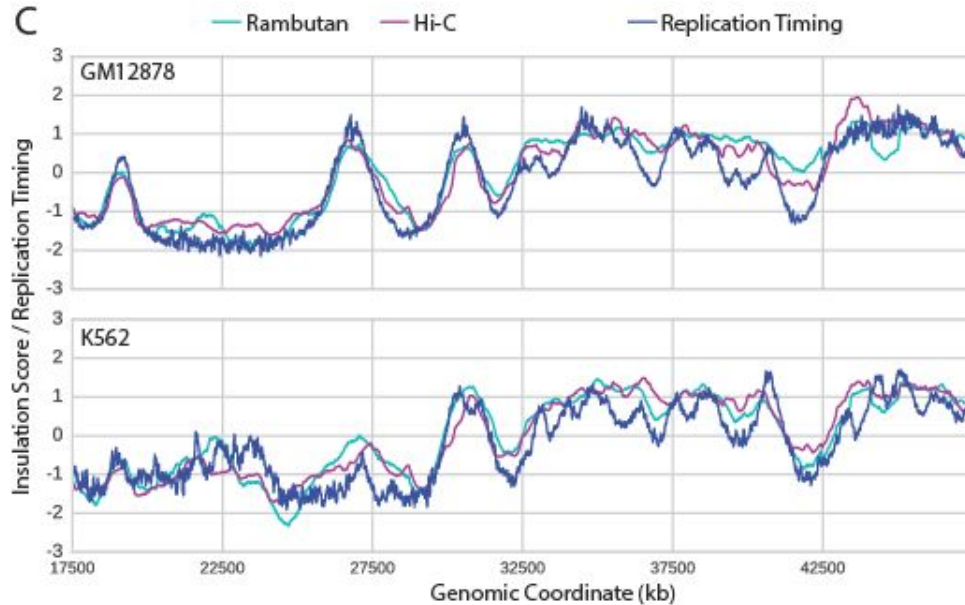
# Insulation score correlates with histone modifications that regulate genomic activity





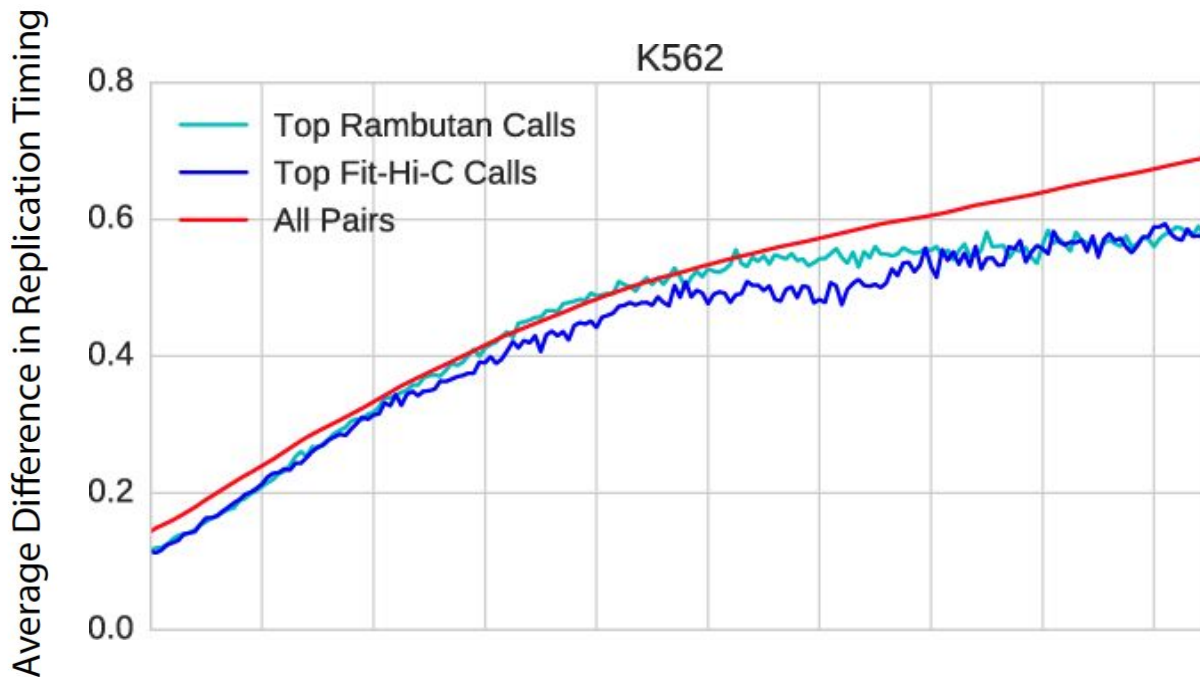


# Insulation score is connected to replication timing



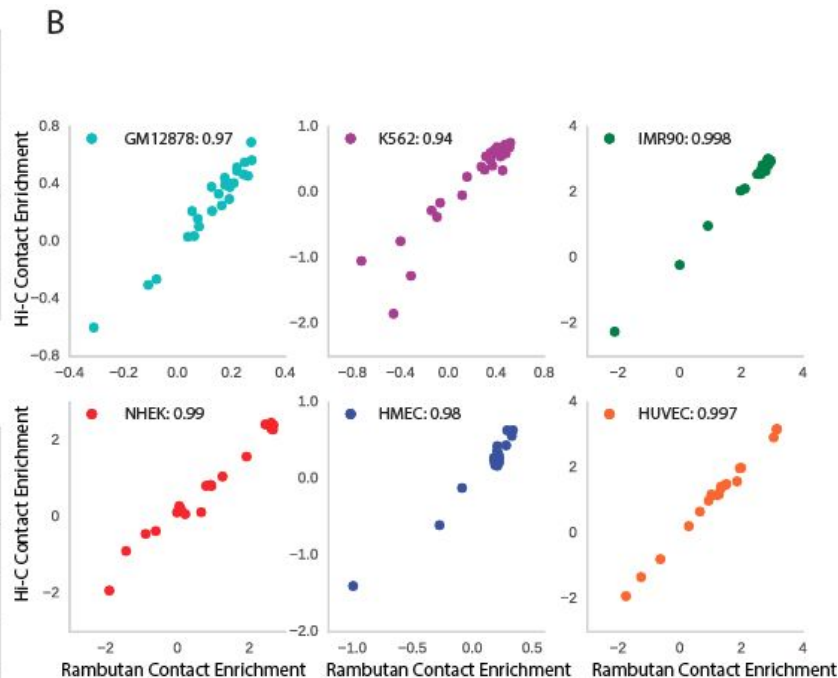
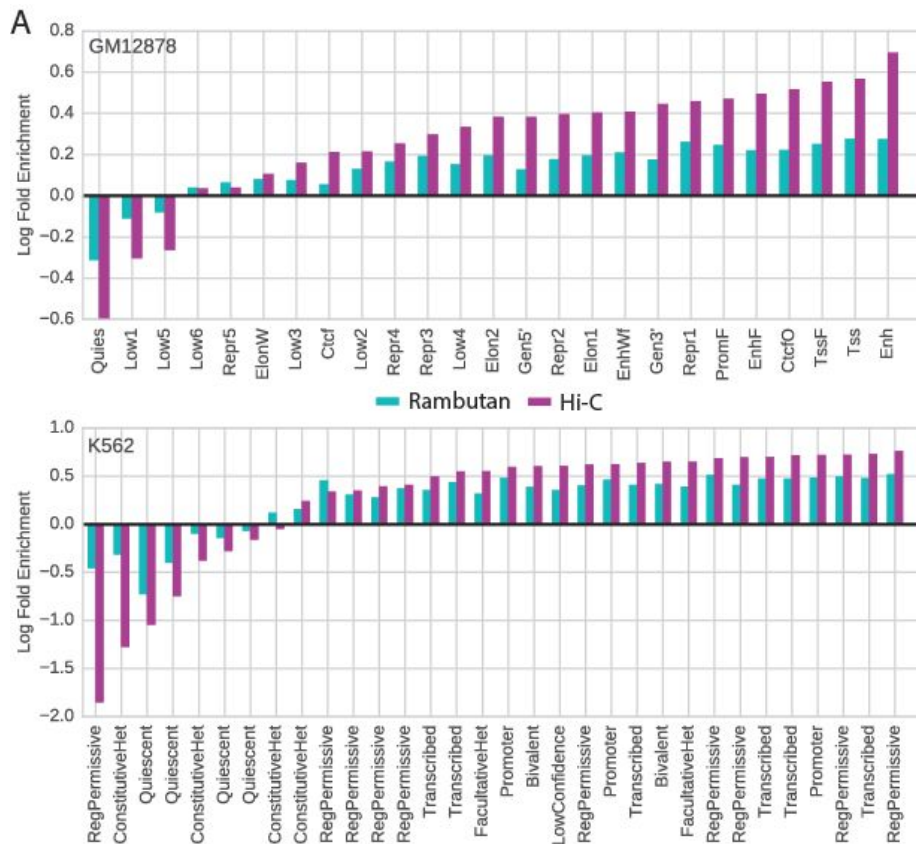


# Rambutan predictions link synchronously replicating regions



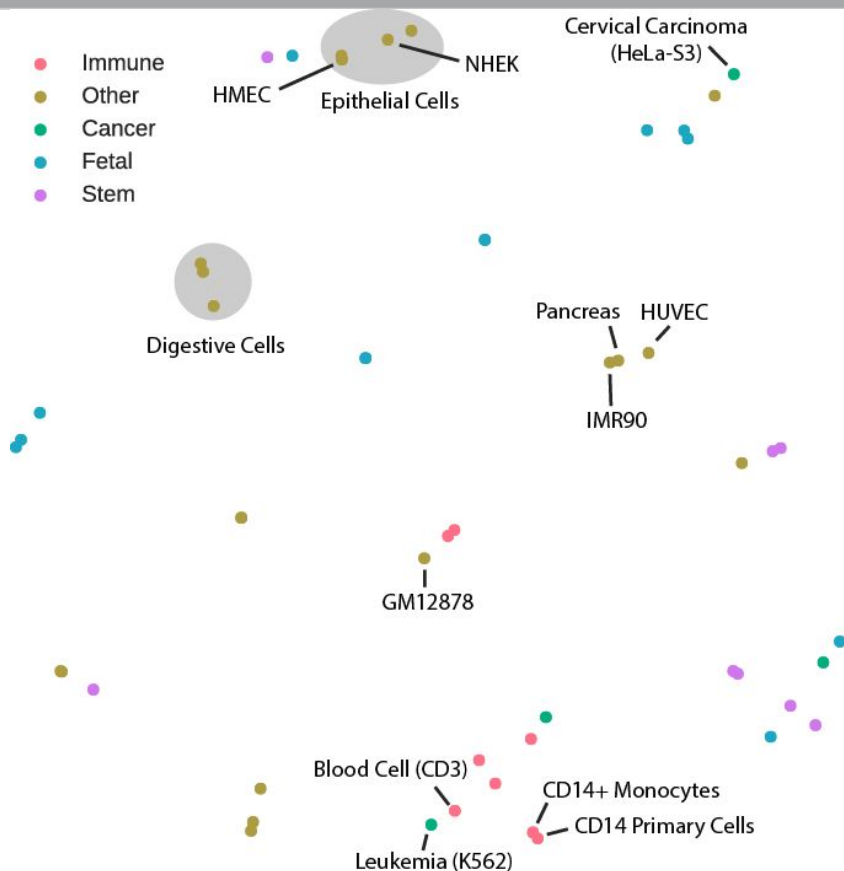


# Rambutan shows similar enrichment of contacts by functional element as Hi-C data





# Genome structure predictions for 53 cell types cluster by cellular function





# Code and paper figures available on GitHub

jmschrei committed on GitHub [MRG] Documentation and usability improvements to `fit` (#7) ... Latest commit 0fa2aef 4 days ago		
docs	FIX installation doc	5 days ago
rambutan	[MRG] Documentation and usability improvements to `fit` (#7)	4 days ago
tests	[MRG] Unit tests (#6)	4 days ago
.travis.yml	[MRG] Unit tests (#6)	4 days ago
Biological_Validation.ipynb	ENH documentation, final model, figures	6 months ago
LICENSE	Initial commit	2 years ago
README.md	Update README.md	4 days ago
dev-requirements.txt	[MRG] Documentation and usability improvements to `fit` (#7)	4 days ago
rambutan-0025.params	ENH documentation, final model, figures	6 months ago
rambutan-symbol.json	ENH documentation, final model, figures	6 months ago
setup.py	[MRG] Documentation and usability improvements to `fit` (#7)	4 days ago

## README.md

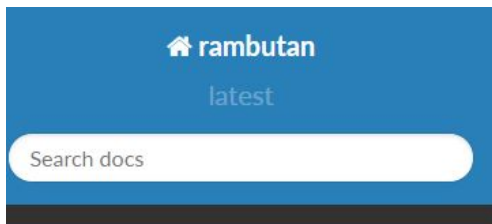
## Rambutan

build passing docs latest

Rambutan is a deep convolutional neural network which predicts 3D chromatin architecture using only nucleotide sequence and DNaseI sensitivity. Specifically it predicts whether a pair of 1kb loci engage in a statistically significant contact with



# Documentation and API reference on ReadTheDocs



## Home

- Installation
- Frequently Asked Questions
- Rambutan
- API Reference
- Data Generators
- Utilities

[Docs](#) » [Home](#)

[Edit on GitHub](#)

## Home

Rambutan is a package for the prediction of the 3D structure of human cell types. It focuses on the prediction of Hi-C contact maps, but rather than trying to predict the number of contacts that a pair of loci engage in, it instead predicts whether the contact is statistically significant with respect to their genomic distance. This genomic distance effect is extremely important as pairs of loci that are close together are very likely to be in contact simply due to physics as opposed to biological importance, whereas long-range contacts are typically enriched for important biological interactions. The predictions are made using a convolutional neural network that takes in nucleotide sequence and DNase-seq sensitivity from two loci spanning 1000 nucleotides. The goal is



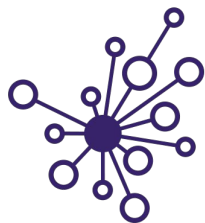


# Acknowledgements



**Bill Noble**

**Jeffrey Bilmes**



**eScience Institute**

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS



**National Science Foundation**

WHERE DISCOVERIES BEGIN