Chromatin 3D Structure and Cancer Typing via Deep Learning

石毅 (Shi, Yi)

2017.06.21

Center for Systems Biomedicine
Shanghai Jiao Tong University

USyd-SJTU Joint Research Alliance
for Translational Medicine

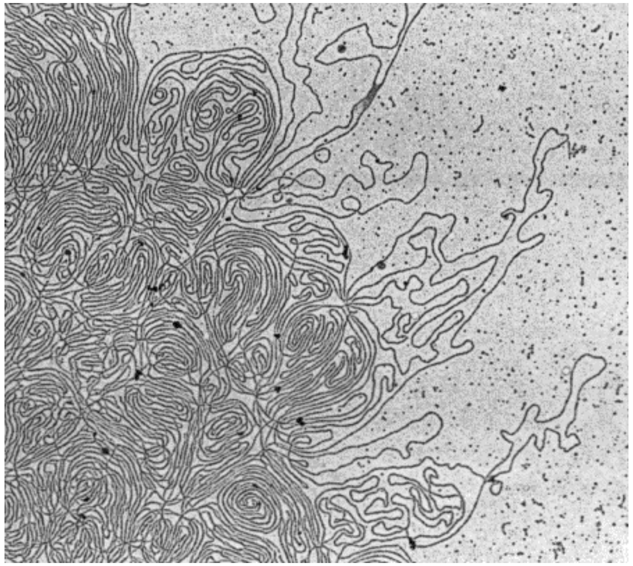
Outline

- Chromatin 3D Structure
- DNN-based Cancer Typing
- Discussion

Chromatin 3D Structure

Chromatin 3D Structure

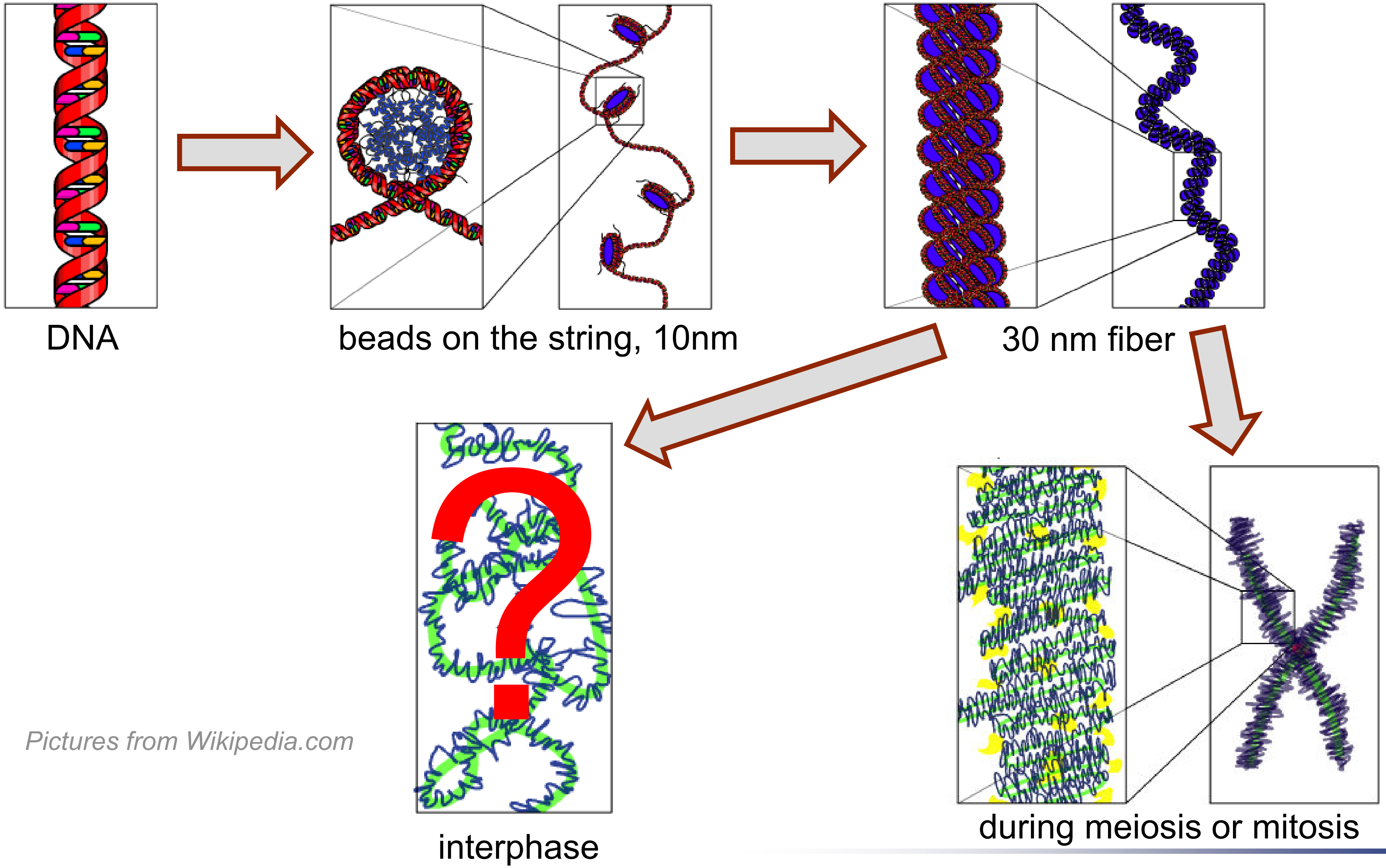
- Human chromatin from a single cell if unpacked and chained up: ~2 meters long
- Human nucleus: micron meter (10^{-6} m) scale in diameter



Chromatin structure illustration
Picture from users.rcn.com

Chromatin 3D Structure

From DNA to chromosome



DNA

beads on the string, 10nm

30 nm fiber

interphase

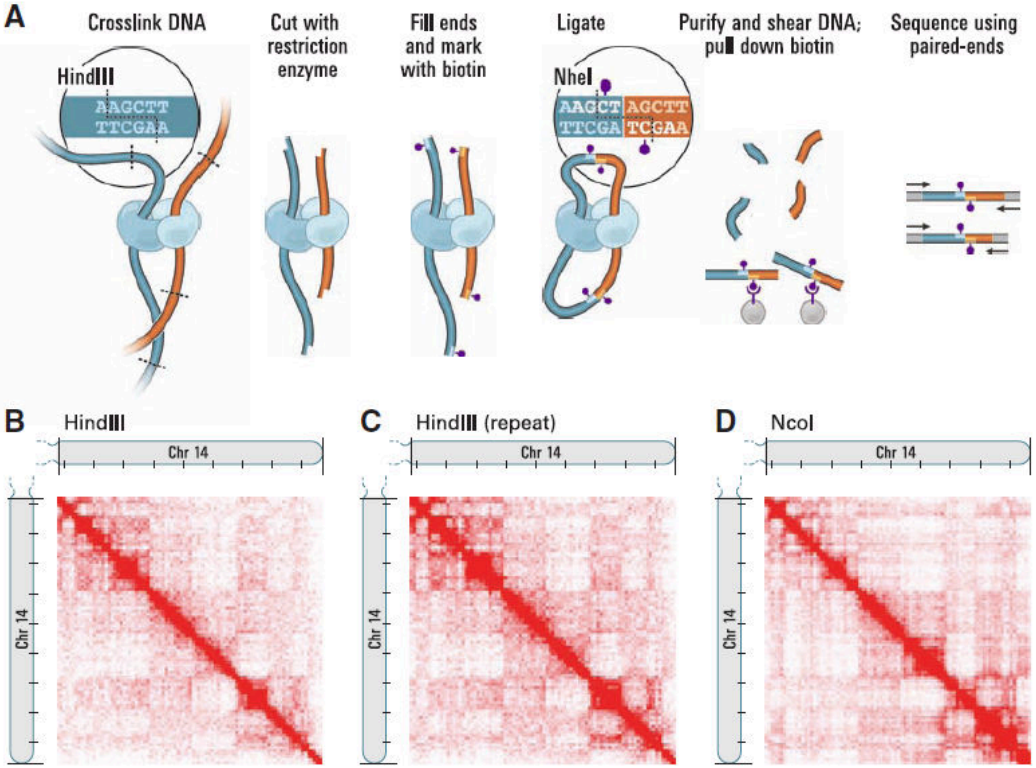
during meiosis or mitosis

Pictures from Wikipedia.com

Chromatin 3D Structure

Chromatin conformation capture technology (HiC)

- HiC provides genome-wise all-to-all chromatin contact profiling compared to the previous FISH (optical one-to-one), 3C (one-to-one) and 4C (one-to-all), and ChIA-PET (targeted all-to-all).


A
Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends
B
HindIII
Chr 14
C
HindIII (repeat)
Chr 14
D
NcoI
Chr 14

Lieberman-Aiden, E. *et al*. *Science*, 2009

Insights:

- Chromatin territories exist.
- Genome partitioned into 2 compartments, active and inactive, with high intra-compartment interaction and low inter-compartment interaction.
- The 2 compartment partitioning is correlated to epigenetic signals.
- There are more genes in active compartment and those genes are more active.

Chromatin 3D Structure

In Situ HiC

- Higher resolution, better insights.

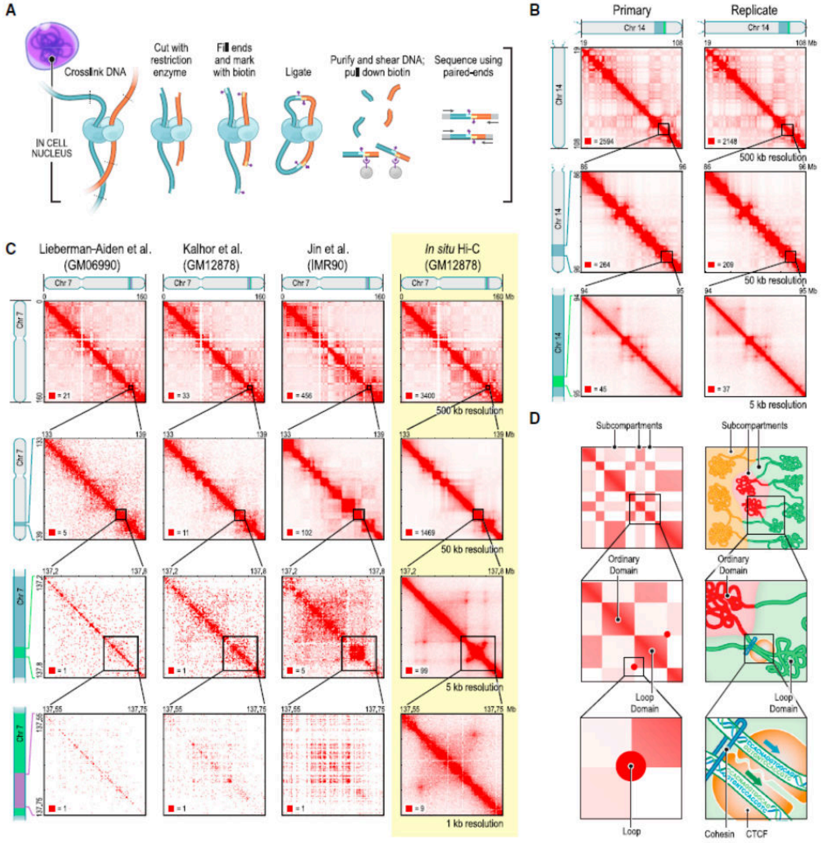


Figure 1. We Used In Situ Hi-C to Map over 15 Billion Chromatin Contacts across Nine Cell Types in Human and Mouse, Achieving 1 kb Resolution in Human Lymphoblastoid Cells

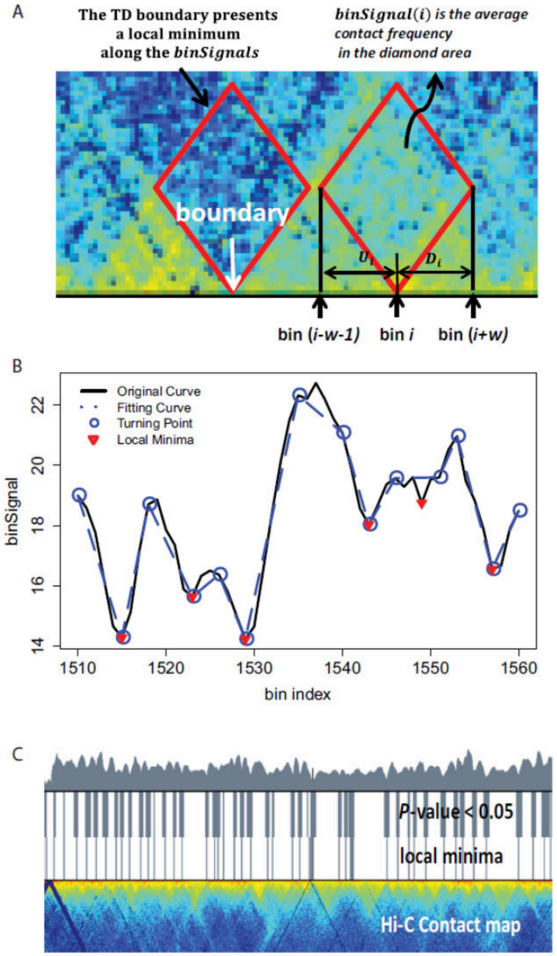
Picture from Rao et al. Cell, 2014

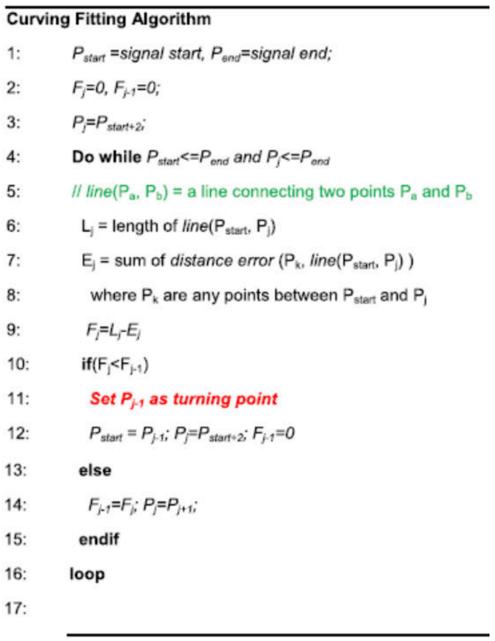
Insights:

- Six types of chromatins discovered.
- More certain about detailed looping.
- TAD and replication origin which is correlated to cancerous mutations.

Chromatin 3D Structure

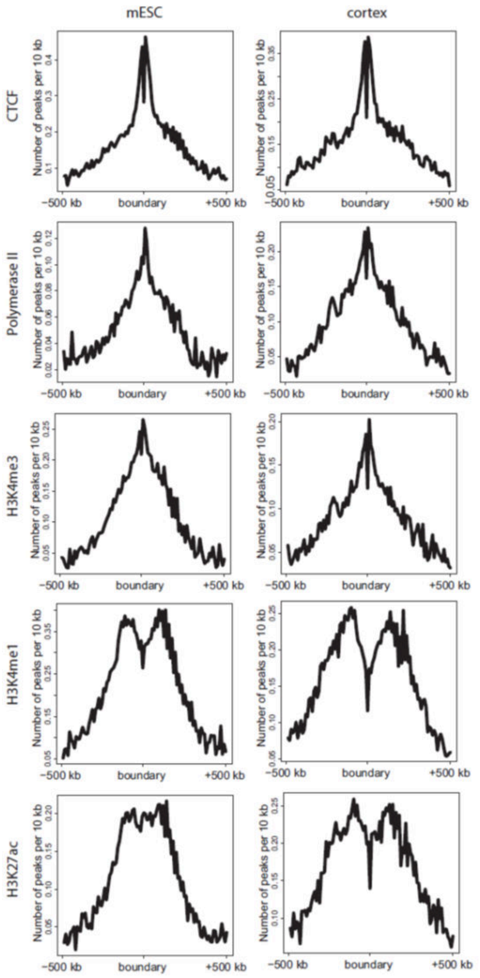
Topological domain (TD) identification





Curving Fitting Algorithm

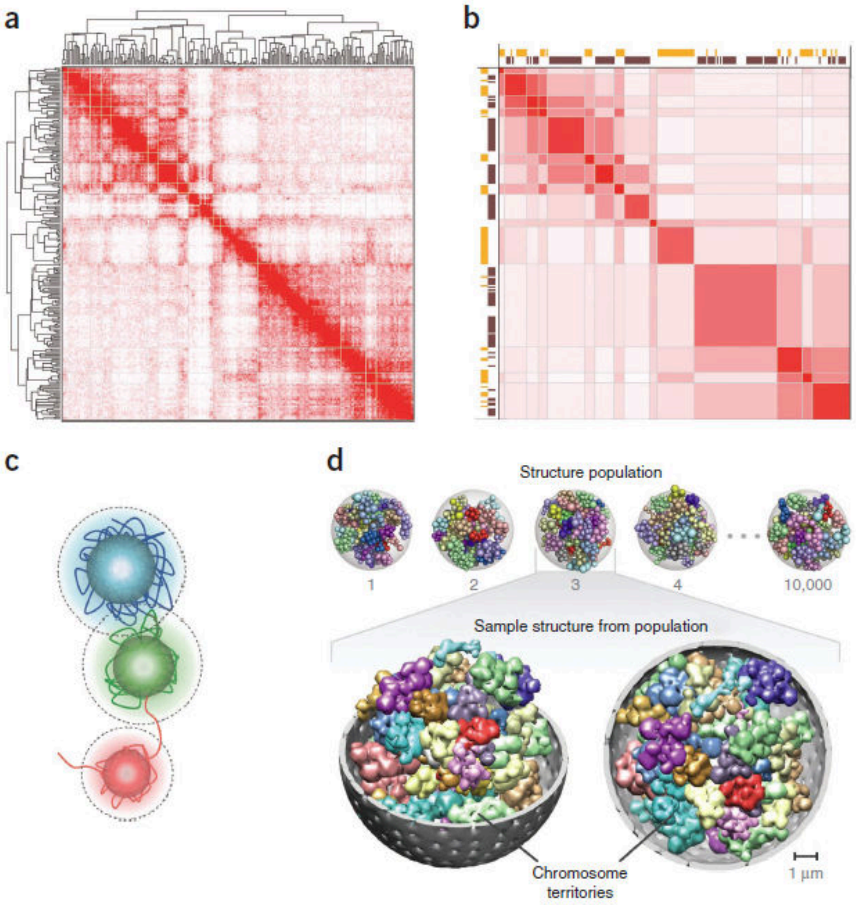
```
P_start = signal start, P_end = signal end;
F_j = 0, F_{j-1} = 0;
P_j = P_{start+2};
Do while P_start <= P_end and P_j <= P_end
  // line(P_a, P_b) = a line connecting two points P_a and P_b
    L_j = length of line(P_start, P_j)
    E_j = sum of distance error (P_k, line(P_start, P_j) )
      where P_k are any points between P_start and P_j
    F_j = L_j - E_j
    if(F_j < F_{j-1})
      Set P_{j-1} as turning point
      P_start = P_{j-1}; P_j = P_{start+2}; F_{j-1} = 0
    else
      F_{j-1} = F_j; P_j = P_{j+1};
    endif
loop
```



Shin & Shi. et al. NAR, 2015

Chromatin 3D Structure

- Structure Modeling
  - Generating reasonable structure decoys
- Structure based studies
  - Structure clustering to find cell states
  - Radius position and feature association
  - Proximity and feature associatior



Kalhor et al. Nature Biotech. 2011

Chromatin 3D Structure

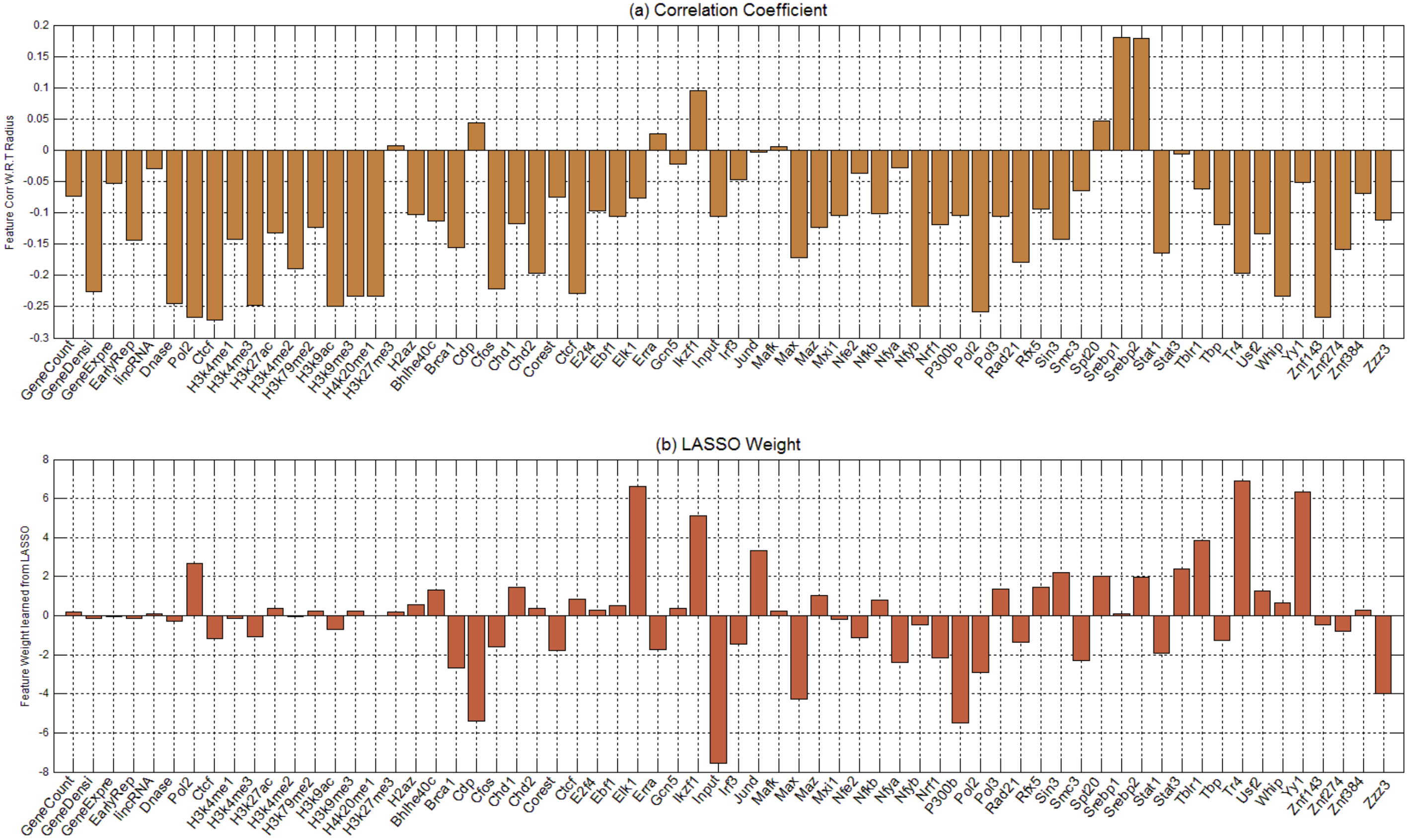
- Chromatin features and radius position

17 Features	66 Features
GeneDensi, GeneExpre, EarlyRep, lincRNA, Dnase, Pol2, Ctcf, H3k4me1, H3k4me3, H3k27ac, H3k4me2, H3k79me2, H3k9ac, H3k9me3, H4k20me1, H3k27me3, H2az	GeneCount, GeneDensi, GeneExpre, EarlyRep, lincRNA, Dnase, Pol2, Ctcf, H3k4me1, H3k4me3, H3k27ac, H3k4me2, H3k79me2, H3k9ac, H3k9me3, H4k20me1, H3k27me3, H2az, Bhlhe40c, Brca1, Cdp, Cfos, Chd1, Chd2, Corest, Ctcf, E2f4, Ebf1, Elk1, Erra, Gcn5, Ikzf1, Input, Irf3, Jund, Mafk, Max, Maz, Mxi1, Nfe2, Nfkb, Nfya, Nfyb, Nrf1, P300b, Pol2, Pol3, Rad21, Rfx5, Sin3, Smc3, Spt20, Srebp1, Srebp2, Stat1, Stat3, Tblr1, Tbp, Tr4, Usf2, Whip, Yy1, Znf143, Znf274, Znf384, Zzz3

Red: Histone Modification Markers
Green: TFs
Blue: Others

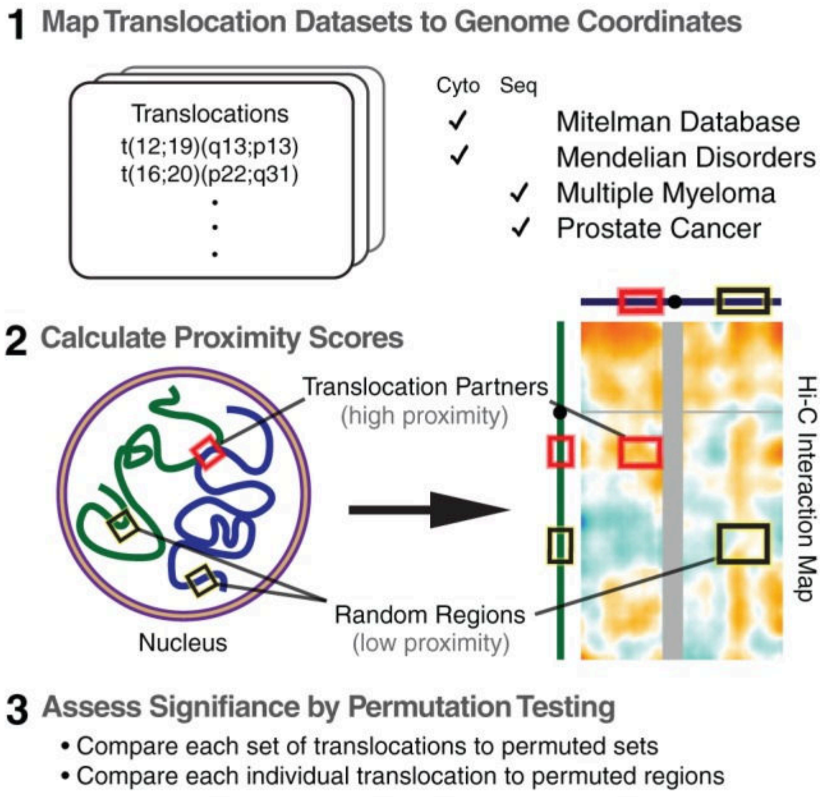
Chromatin 3D Structure

- Chromatin features and radius position



Chromatin 3D Structure

- Cancerous translocation and chromatin structure



1 Map Translocation Datasets to Genome Coordinates

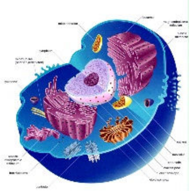
2 Calculate Proximity Scores

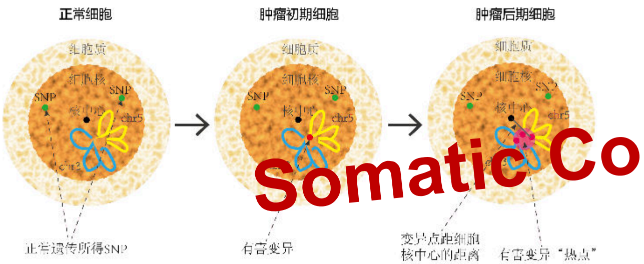
3 Assess Signifiance by Permutation Testing

- Compare each set of translocations to permuted sets
- Compare each individual translocation to permuted regions

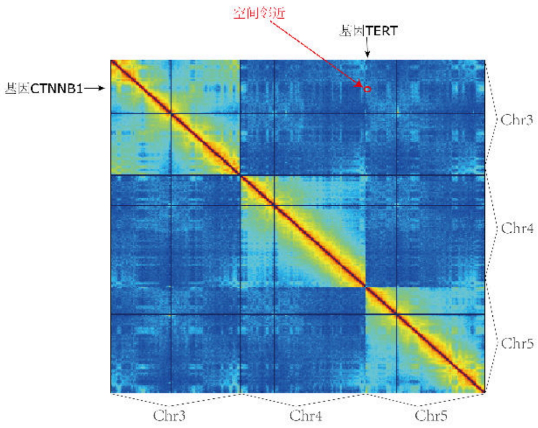
Engreitz et al. PLoS One, 2013

Chromatin 3D Structure

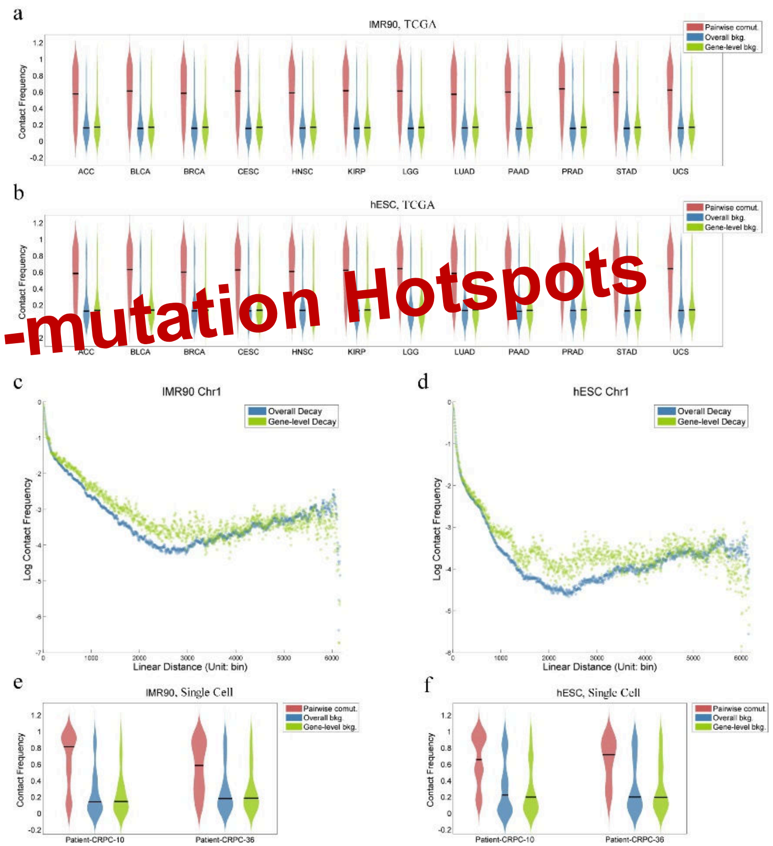




(a)



(b)

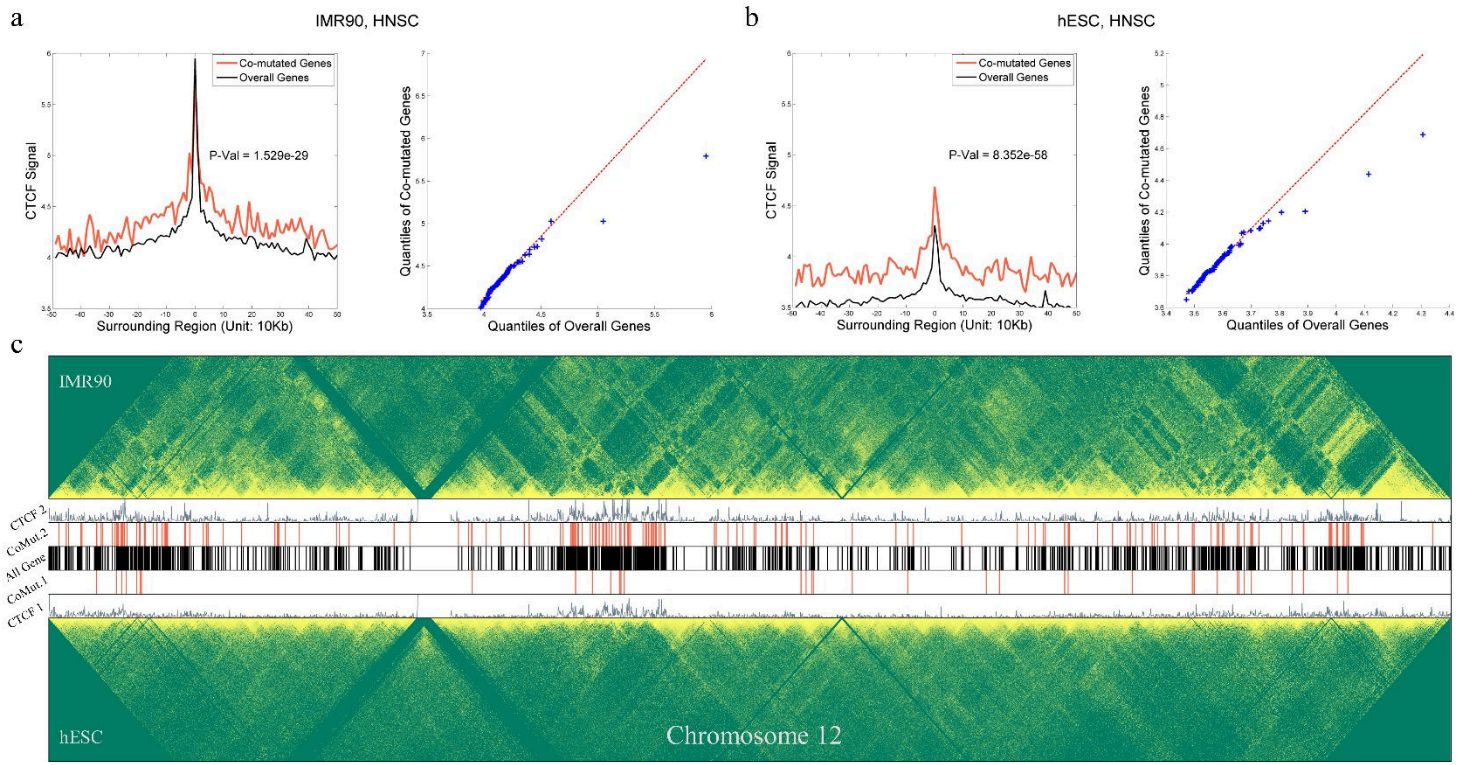


Somatic Co-mutation Hotspots

Shi. *et al*. *Scientific Reports*, 2016

Chromatin 3D Structure

- CTCF enriched in “hotspots”

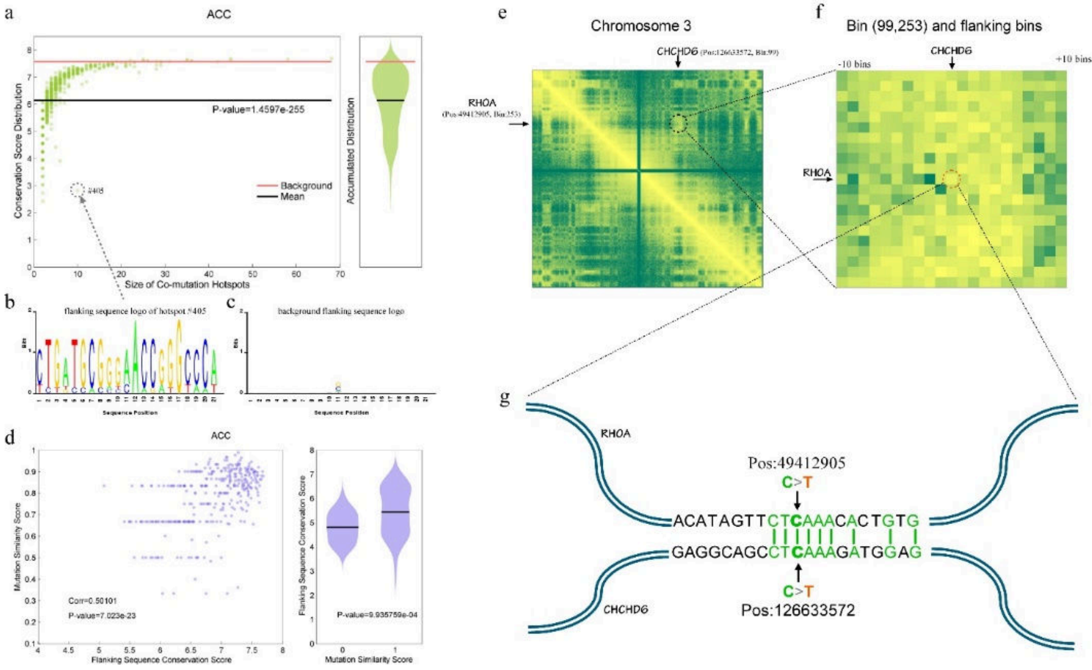


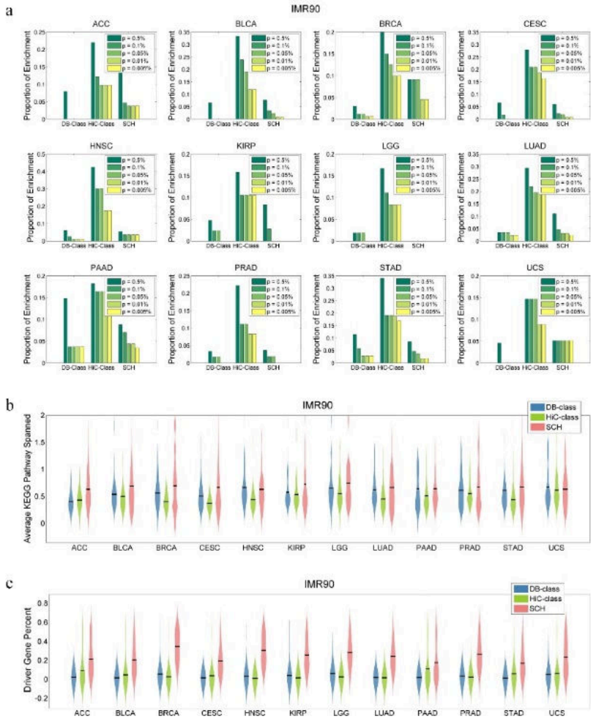
Shi. *et al*. *Scientific Reports*, 2016

Chromatin 3D Structure

- Similar mutation type and flanking sequence conservation in “hotspots”

- Pathway enrichment





Shi. *et al*. *Scientific Reports*, 2016

DNN-based Cancer Typing

DNN-based Cancer Typing

Traditional cancer diagnosis

- Morphological appearance:
  - Pathological section (golden standard)
  - Imaging techniques

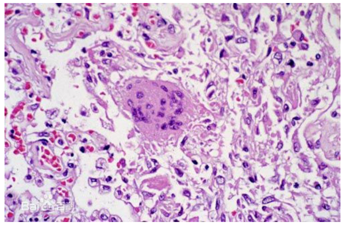


Image from baidu.com

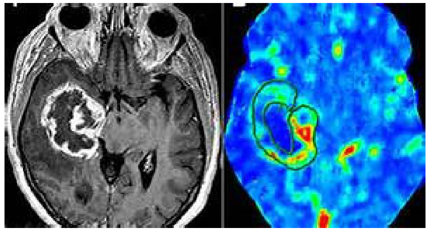


Image from radiology.med.nyu.edu

- Gene or protein expression

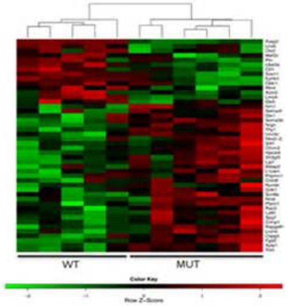


Image from well.ox.ac.uk

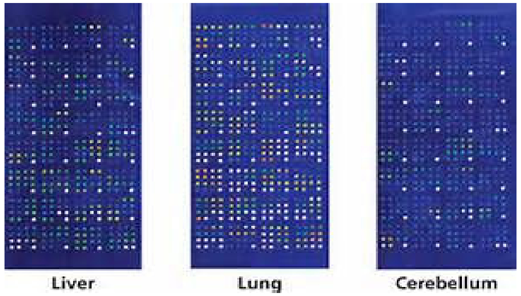


Image from sigmaaldrich.com

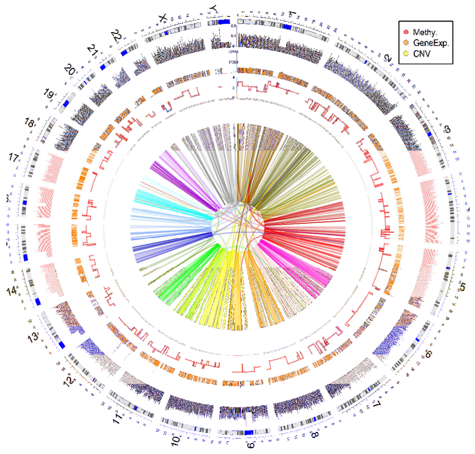
SHANGHAI JIAO TONG UNIVERSITY

DNN-based Cancer Typing

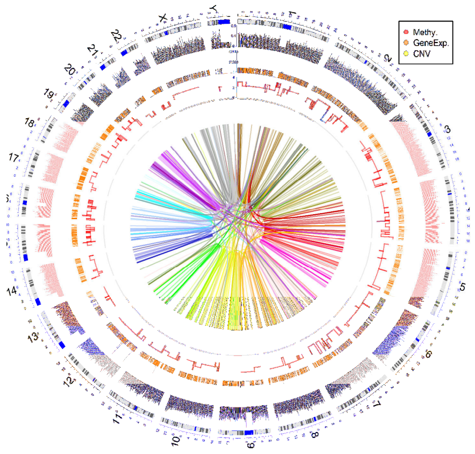
Inside drives

- Somatic point mutations
- Insertions and deletions (INDELs)
- Chromatin translocations
- Copy number abnormalities

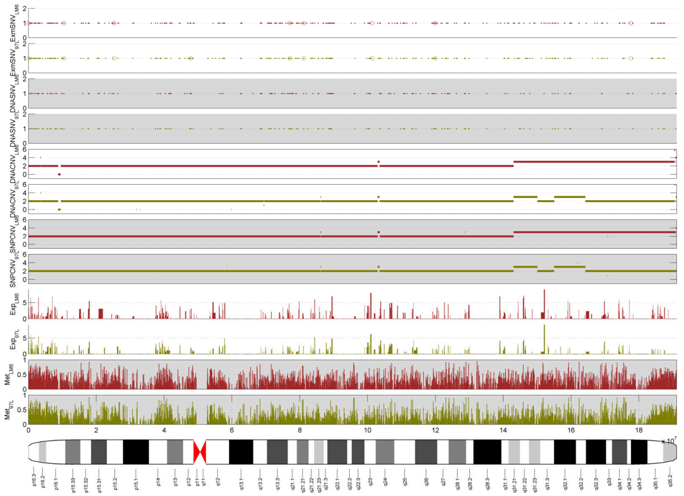
97L Trakcs



LM6 Trakcs



Chromosome 4 Exome-SNV, DNA-SNV, SNP-CNV, DNASeq-CNV, GeneExp, Methy

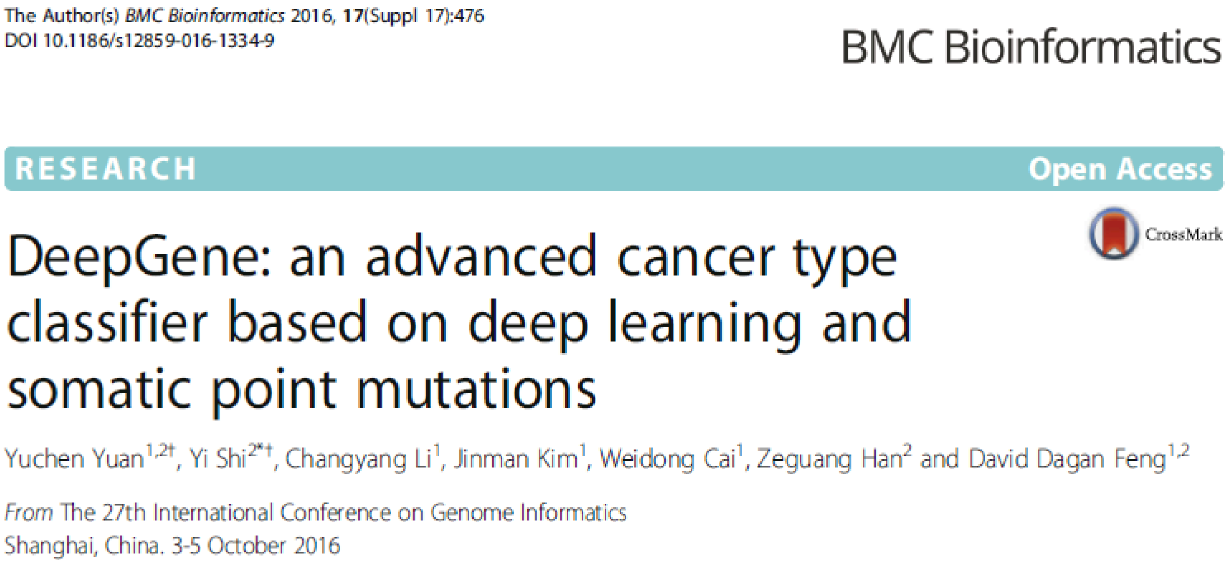


DNN-based Cancer Typing

UNIVERSITY OF ALBERTA
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last – a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO
CONSERVATION
SONGBIRDS À LA CARTE
Illegal harvest of millions of Mediterranean birds
PAGE 452
RESEARCH ETHICS
SAFEGUARD TRANSPARENCY
Don't let openness backfire on individuals
PAGE 459
POPULAR SCIENCE
WHEN GENES GOT 'SELFISH'
Dawkins's calling card forty years on
NATURE.COM/NATURE
Diagonal Line Node
Face Node
Cat Node
快科技 KKJ.CN

Applications of deep neural network (DNN) learning

DNN-based Cancer Typing



The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 17):476
DOI 10.1186/s12859-016-1334-9

BMC Bioinformatics

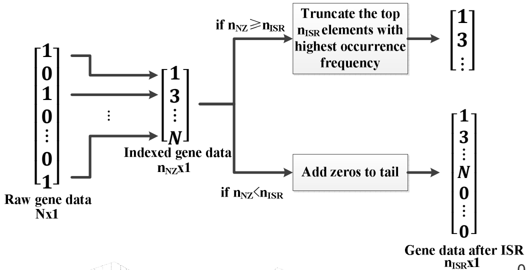
RESEARCH Open Access

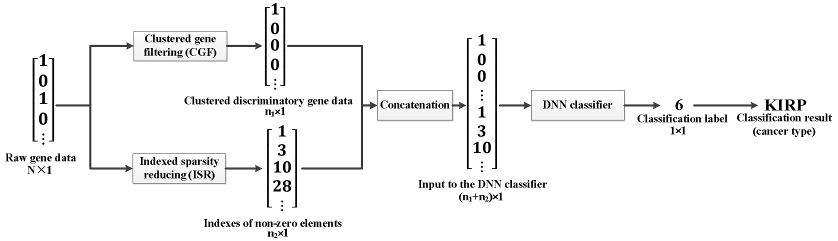
CrossMark

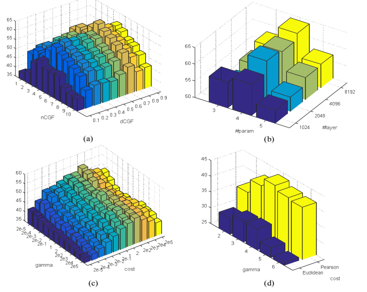
DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations

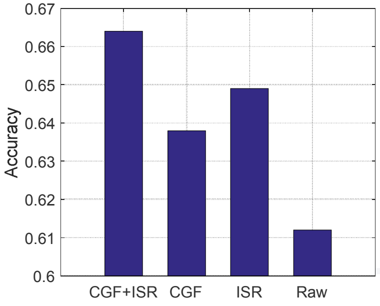
Yuchen Yuan[1,2†], Yi Shi[2*†], Changyang Li[1], Jinman Kim[1], Weidong Cai[1], Zeguang Han[2] and David Dagan Feng[1,2]

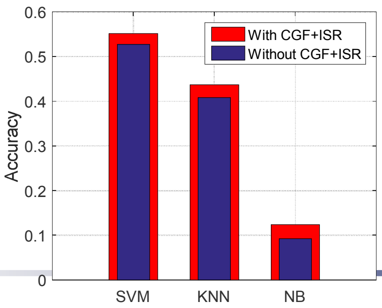
From The 27th International Conference on Genome Informatics
Shanghai, China. 3-5 October 2016




Raw gene data
N×1
Clustered gene filtering (CGF)
Clustered discriminatory gene data
n_1×1
Indexed sparsity reducing (ISR)
Indexes of non-zero elements
n_2×1
Concatenation
Input to the DNN classifier
(n_1+n_2)×1
DNN classifier
6
Classification label
1×1
KIRP
Classification result
(cancer type)

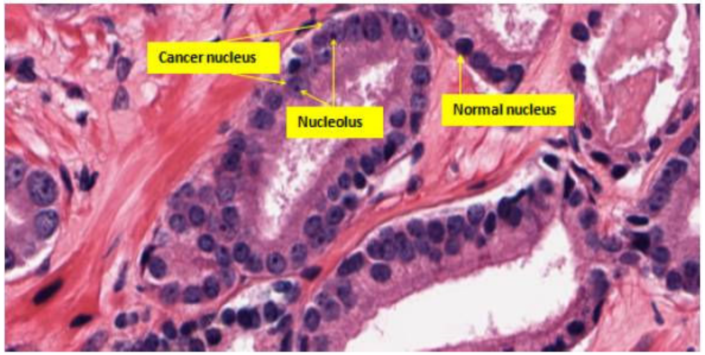






DNN-based Cancer Typing

- Why CNA:?
  - Links between aneuploidy and cancer have long been recognized.
  - CNA is the major form of chromosomal instability, affecting a larger fraction of the genome in cancers.
  - The technologies of profiling genome-wide CNV is more developed than before, from DNA microarray based to whole-genome DNA sequencing based to exome sequencing based.



Picture from biometrics.cse.msu.edu

DNN-based Cancer Typing

Data preprocessing

- The CNA data is first empirically clipped into the interval [0, 10].
- The clipped data is then zero-padded at tail to have the desired length that fits the input of the subsequent neural networks.
- For 2D CNN, the CNA samples are then reshaped into 176*176*1, just like single-layered images.

DNN-based Cancer Typing

- 1D CNN

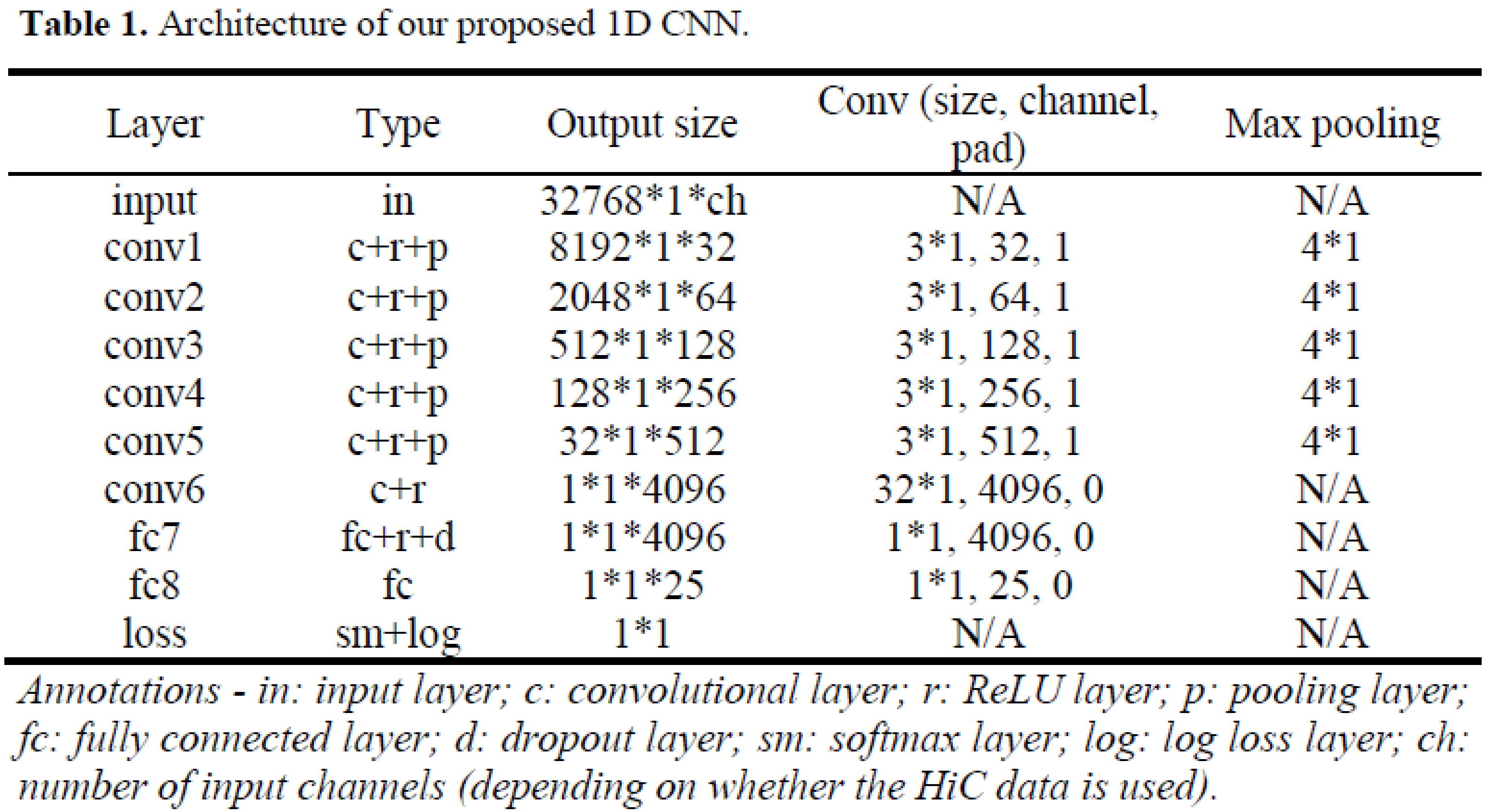


Table 1. Architecture of our proposed 1D CNN.

Layer	Type	Output size	Conv (size, channel, pad)	Max pooling
input	in	32768*1*ch	N/A	N/A
conv1	c+r+p	8192*1*32	3*1, 32, 1	4*1
conv2	c+r+p	2048*1*64	3*1, 64, 1	4*1
conv3	c+r+p	512*1*128	3*1, 128, 1	4*1
conv4	c+r+p	128*1*256	3*1, 256, 1	4*1
conv5	c+r+p	32*1*512	3*1, 512, 1	4*1
conv6	c+r	1*1*4096	32*1, 4096, 0	N/A
fc7	fc+r+d	1*1*4096	1*1, 4096, 0	N/A
fc8	fc	1*1*25	1*1, 25, 0	N/A
loss	sm+log	1*1	N/A	N/A

Annotations - in: input layer; c: convolutional layer; r: ReLU layer; p: pooling layer; fc: fully connected layer; d: dropout layer; sm: softmax layer; log: log loss layer; ch: number of input channels (depending on whether the HiC data is used).

DNN-based Cancer Typing

- 2D CNN

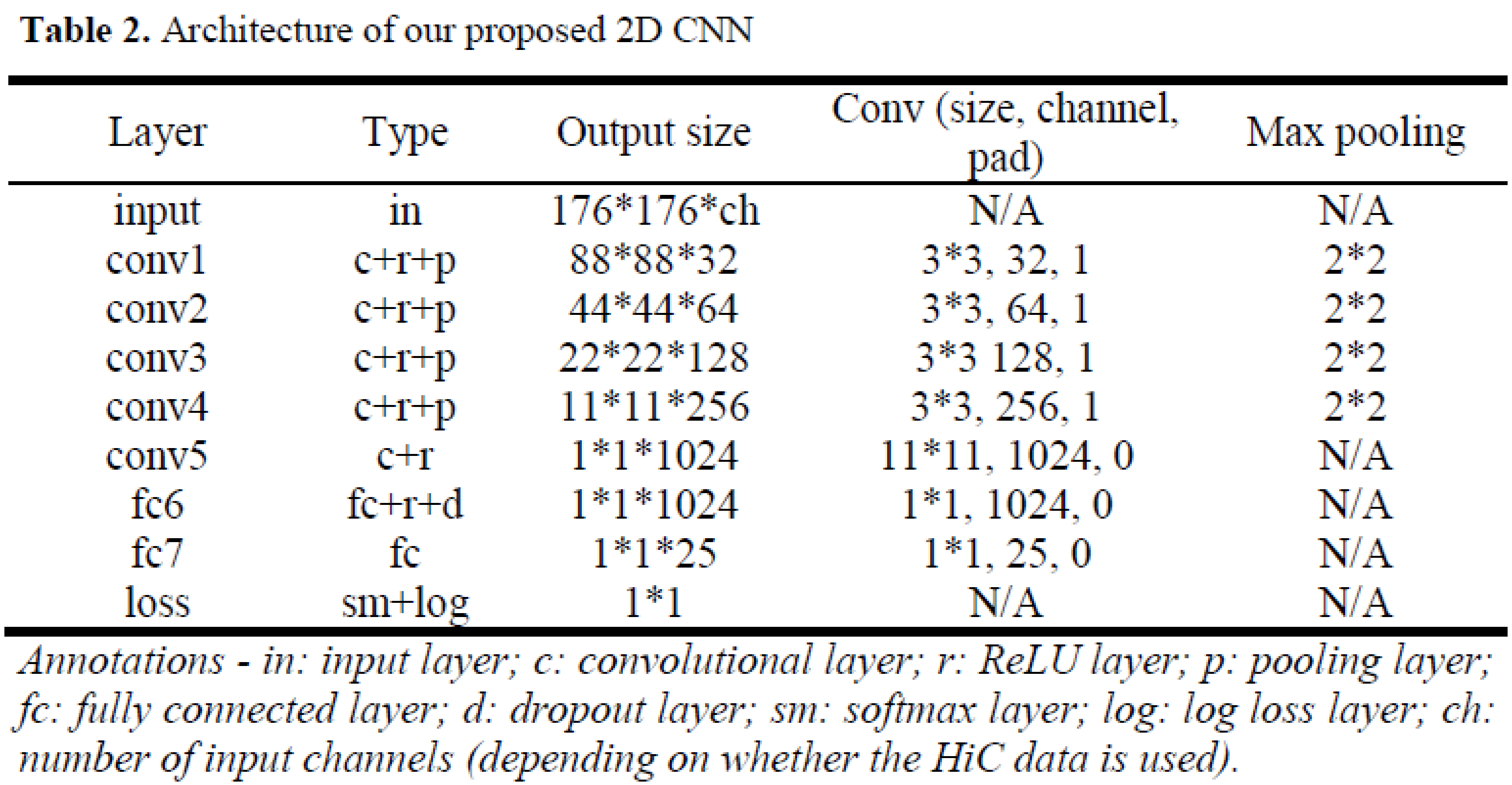


Table 2. Architecture of our proposed 2D CNN

Layer	Type	Output size	Conv (size, channel, pad)	Max pooling
input	in	176*176*ch	N/A	N/A
conv1	c+r+p	88*88*32	3*3, 32, 1	2*2
conv2	c+r+p	44*44*64	3*3, 64, 1	2*2
conv3	c+r+p	22*22*128	3*3 128, 1	2*2
conv4	c+r+p	11*11*256	3*3, 256, 1	2*2
conv5	c+r	1*1*1024	11*11, 1024, 0	N/A
fc6	fc+r+d	1*1*1024	1*1, 1024, 0	N/A
fc7	fc	1*1*25	1*1, 25, 0	N/A
loss	sm+log	1*1	N/A	N/A

Annotations - in: input layer; c: convolutional layer; r: ReLU layer; p: pooling layer; fc: fully connected layer; d: dropout layer; sm: softmax layer; log: log loss layer; ch: number of input channels (depending on whether the HiC data is used).

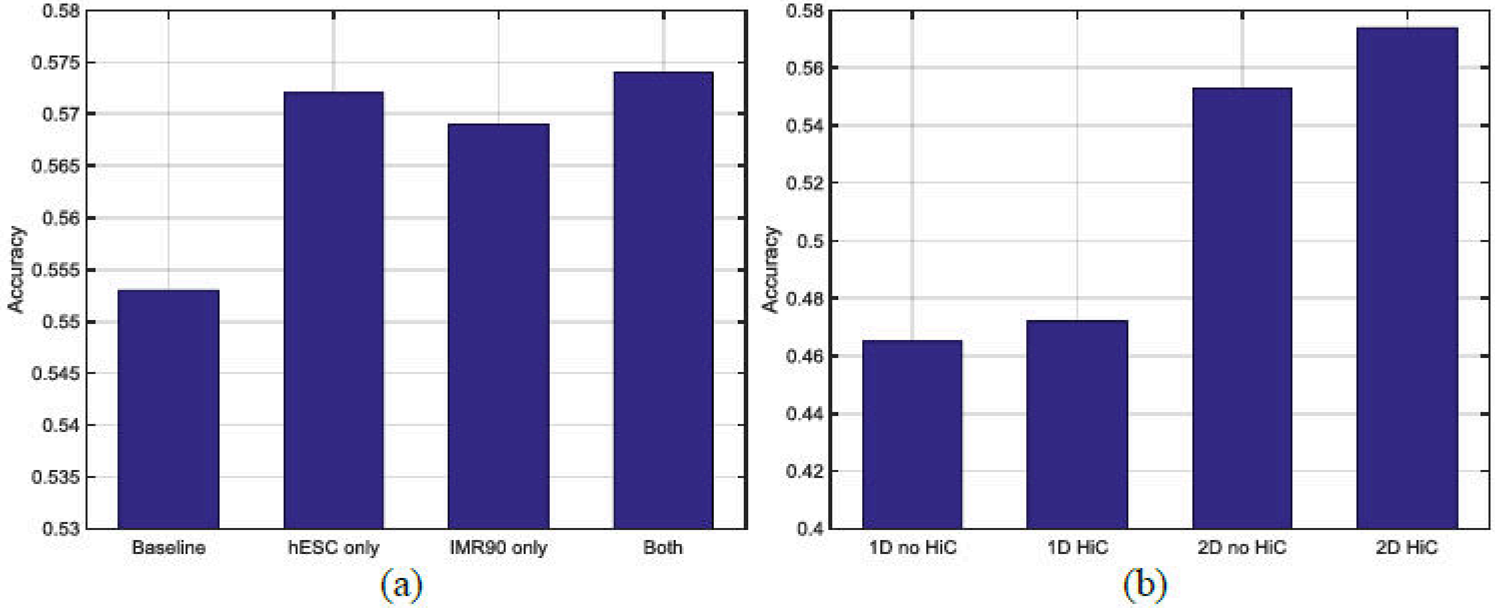
DNN-based Cancer Typing

Implementation details

- Both the 1D CNN and the 2D CNN are implemented in Python under the Caffe framework, which is an open source framework for CNN training and testing.
- The machine used for our experiments is a PC with Intel 6-Core i7-5820K 3.3GHz CPU, 64GB RAM, GeForce GTX TITAN X 12GB GPU, and 64-bit Ubuntu 14.04.3 LTS.
- Software dependencies include CUDA 8.0 and cuDNN 5.1.

DNN-based Cancer Typing

- Proposed method in different design options



(a)

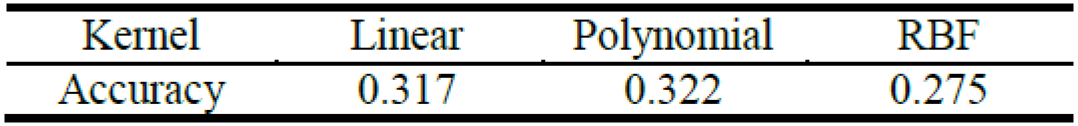
(b)

Fig. 1. Performances of our proposed method with different design options. (a) With different HiC data configurations. From left to right: baseline model (2D CNN); baseline with hESC only; baseline with IMR90 only; baseline with both types of HiC data. The last configuration leads to the optimal performance. (b) With different network and HiC combinations. From left to right: 1D CNN without HiC data; 1D CNN with HiC data; 2D CNN without HiC data; 2D CNN with HiC data. The last configuration leads to the optimal performance.

DNN-based Cancer Typing

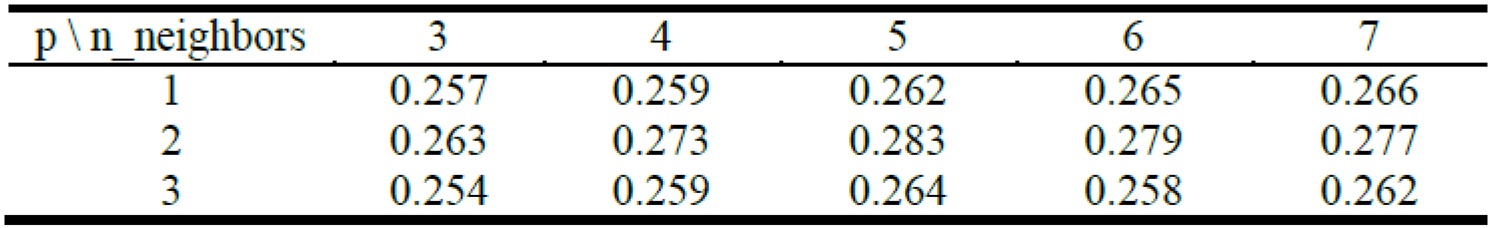
Other classifiers

Table 3. Evaluation of SVM with different kernel types.



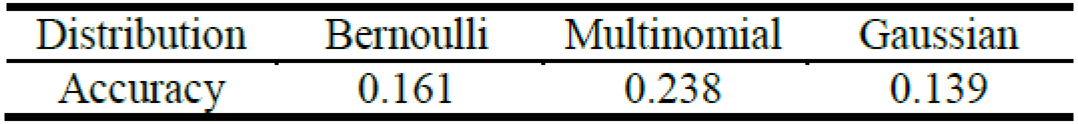
Kernel	Linear	Polynomial	RBF
Accuracy	0.317	0.322	0.275

Table 4. Evaluation of KNN with different number of neighbors and p value.



p \ n_neighbors	3	4	5	6	7
1	0.257	0.259	0.262	0.265	0.266
2	0.263	0.273	0.283	0.279	0.277
3	0.254	0.259	0.264	0.258	0.262

Table 5. Evaluation of NB with different data distribution assumptions



Distribution	Bernoulli	Multinomial	Gaussian
Accuracy	0.161	0.238	0.139

DNN-based Cancer Typing

Comparing with other classifiers

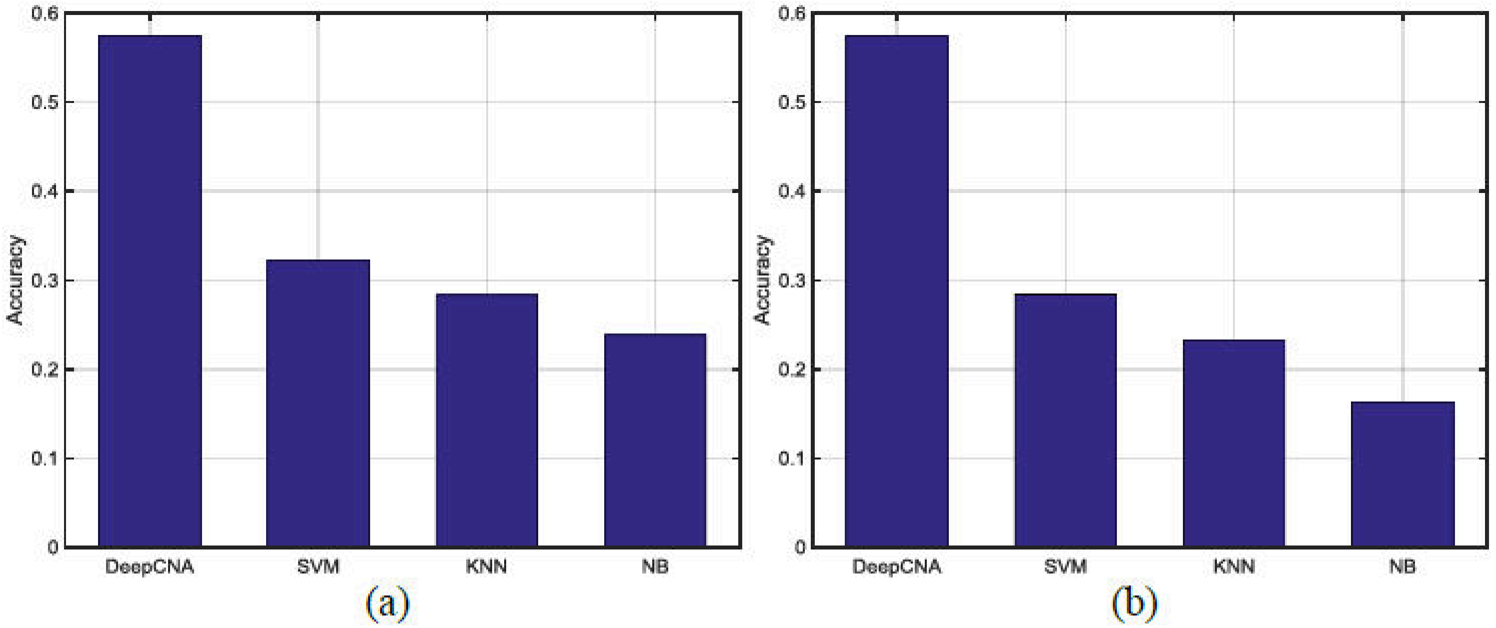


Fig. 2. Performances of our proposed method against three widely adopted data classifiers. (a) The comparison methods use raw CNA input data (without HiC). From left to right: Our method, SVM (polynomial kernel), KNN (number of neighbors = 5 and p = 2) and NB (multinomial distribution). Our method shows significant advantage against the comparison methods. (b) The comparison methods use both CNA and HiC as input data. From left to right: Our method, SVM (polynomial kernel), KNN (number of neighbors = 5 and p = 2) and NB (multinomial distribution). Our method shows even greater advantage against the comparison methods.

Discussion

- Further investigation
  - Integrating heterogeneous mutation data together, e.g. SNV, INDEL, CNV, translocation
  - What feature (gene) combinations contribute to better prediction accuracy? Why?
- How this can help real diagnosis?
  - Applying to CTC or ctDNA for early diagnosis, subtyping, locating.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Acknowledgement

Prof. Ze-guang Han

Prof. David Feng

Dr. Yuchen Yuan & I

Prof. Tom Cai

THE UNIVERSITY OF SYDNEY

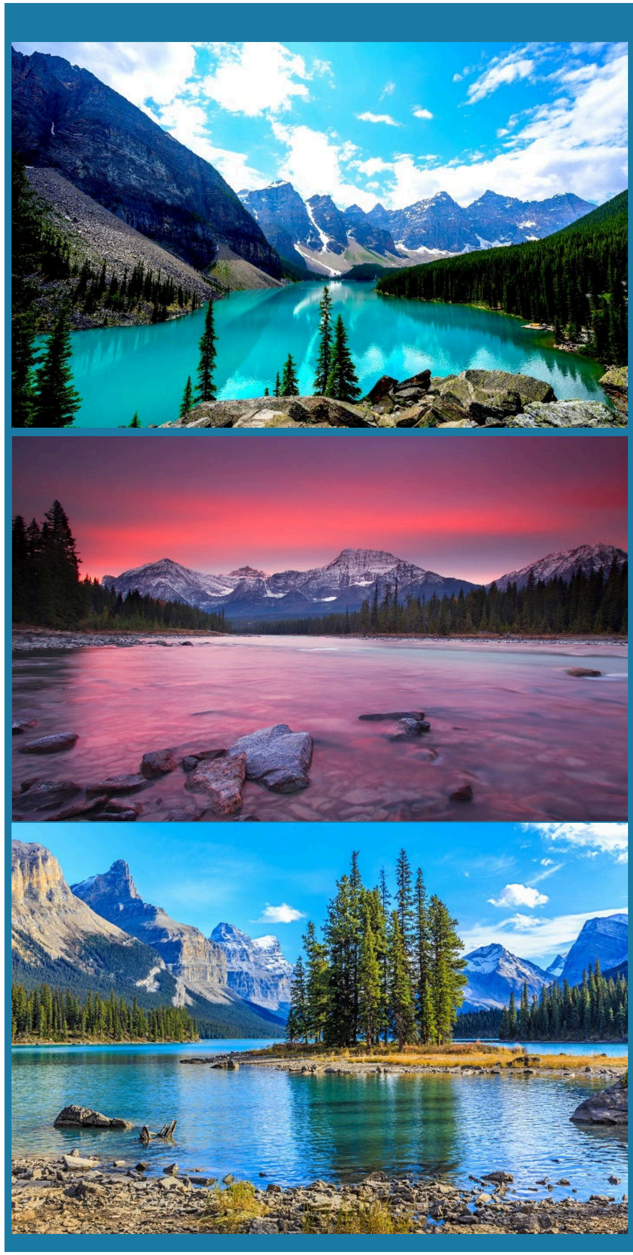
SHANGHAI JIAO TONG UNIVERSITY 1896

USyd-SJTU Joint Research Alliance for Translational Medicine

NSFC
National Natural Science Foundation of China

上海科技
上海市科学技术委员会主办
www.stcsm.gov.cn

浦江人才计划
Pujiang Scholar



Questions & Comments?

고맙습니다!

ありがとう!

谢谢!

Thank you!

2017.06.21