KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML) Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Best practice for epigenetic analysis

이동성 _ 서울대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로
제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우
발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

강의개요

Best practice for epigenetic analysis

이 강의는 대학원생 및 연구자, 그리고 생물정보학 초심자를 대상으로 하여 에피유전체 분석의 기본 원리와 실무적 접근 방식을 이해하고 적용할 수 있도록 설계되었습니다. 특히, 분석의 효율성을 높이고 반복 가능성을 보장하기 위해 Galaxy라는 웹 기반 플랫폼을 활용합니다.

이 강의를 통해 참가자는 Galaxy 플랫폼을 활용하여 DNA 메틸화, ChIP-Seq, 및 Hi-C 데이터를 분석하고 시각화하는 데 필요한 기술을 습득하게 됩니다.

주요 학습 목표

1. Galaxy 플랫폼의 활용 방법을 이해하고, 분석 워크플로를 구성할 수 있다.
2. DNA 메틸화, ChIP-Seq, 및 Hi-C 데이터를 분석하는 기본 원리와 실무 기법을 습득한다.
3. 공공 데이터베이스에서 데이터를 수집하고, 품질을 평가하며, 시각화 및 해석을 통해 생물학적 통찰을 얻는다.

강의내용

- Introduction to Galaxy
- DNA methylation data analysis
- ChIP-Seq data analysis
- Hi-C data analysis

* 참고강의교재: Galaxy (https://usegalaxy.org.au/)

* 교육생준비물: 노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

* 강의: 이동성 교수 (서울대학교 의과대학)

Curriculum Vitae

Speaker Name: Dongsung Lee, Ph.D.

▶ **Personal Info**

Name	Dongsung Lee
Title	Associate Professor
Affiliation	College of Medicine, Seoul National University

▶ **Contact Information**

Email	dongsung.lee@snu.ac.kr

Research Interest

Bioinformatics, Genomics, Epigenomics, Single Cell Genomics

Educational Experience

2010	B.S. in Life Science, Korea University, Korea
2015	Ph.D. in Medical Science, Seoul National University, Korea

Professional Experience

2016-2020	Post-doc research fellow, Salk Institute for Biological Studies, USA
2020-2024	Associate/Assistant Professor, Department of Life Science, University of Seoul, Korea
2024-	Associate Professor, College of Medicine, Seoul National University, Korea

Selected Publications (5 maximum)

1. Temporally distinct 3D multi-omic dynamics in the developing human brain. Nature (2024)
2. Foxp3 Orchestrates Reorganization of Chromatin Architecture to Establish Regulatory T Cell Identity. Nat Commun. (2023)
3. Structural variants drive context-dependent oncogene activation in cancer. Nature (2022)
4. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nature Methods (2019)
5. An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. Nature Commun. (2014).

KSBI | 한국생명정보학회

KSBi-BIML 2025

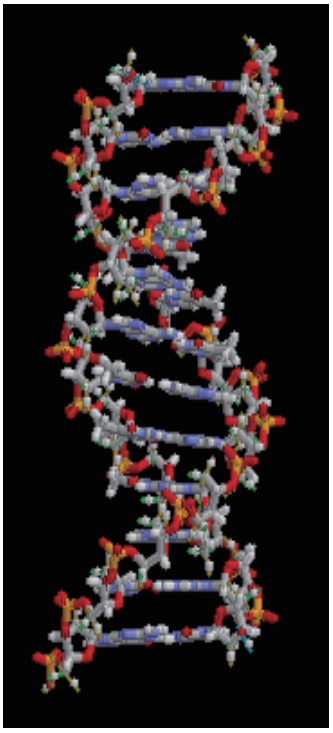
Best practice for epigenetic analysis

서울대학교
의과대학
이동성

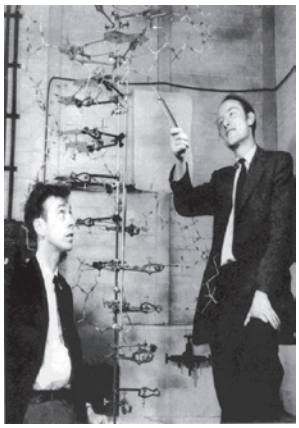
Contents

- Epigenetics의 기초
  - 의의
  - 종류
  - Profiling 기술
- 선행연구
  - iPSC reprogramming 과정동안의 epigenetic 변화
  - Single nucleus methyl 3C (snm3C)의 개발과 인간 뇌연구
- 실습
  - Galaxy 소개
  - Whole Genome Bisulfite sequencing
  - ChIP-seq

생명의 설계도, DNA



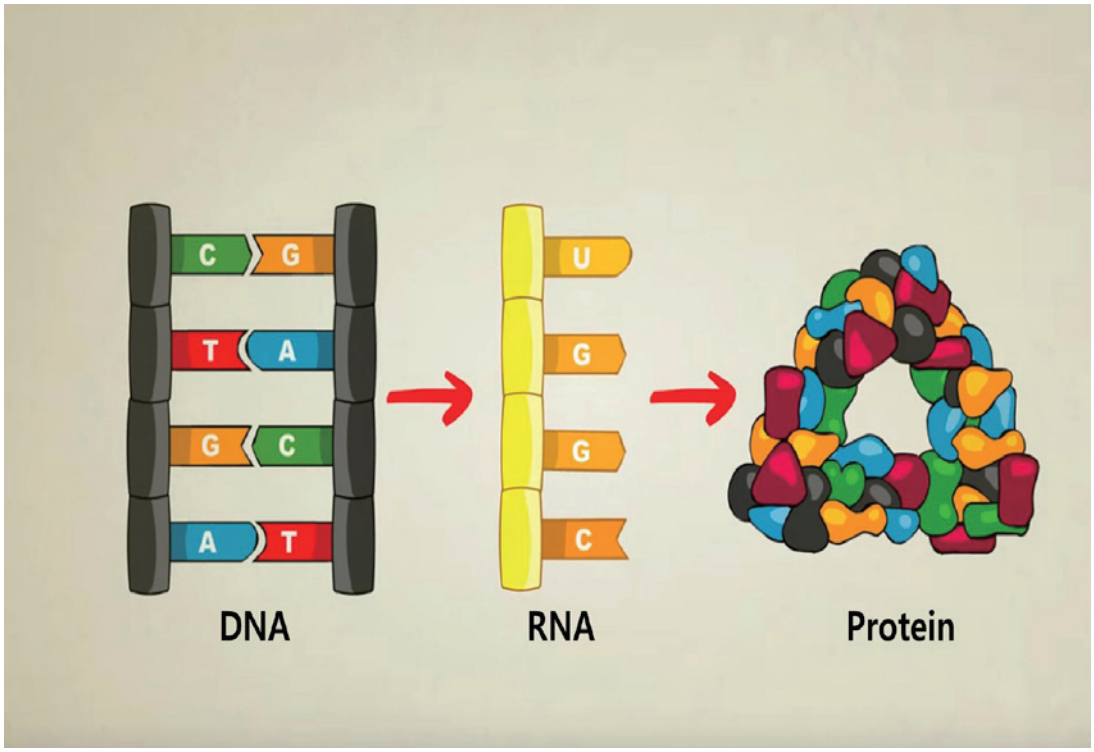
Johannes Friedrich Miescher



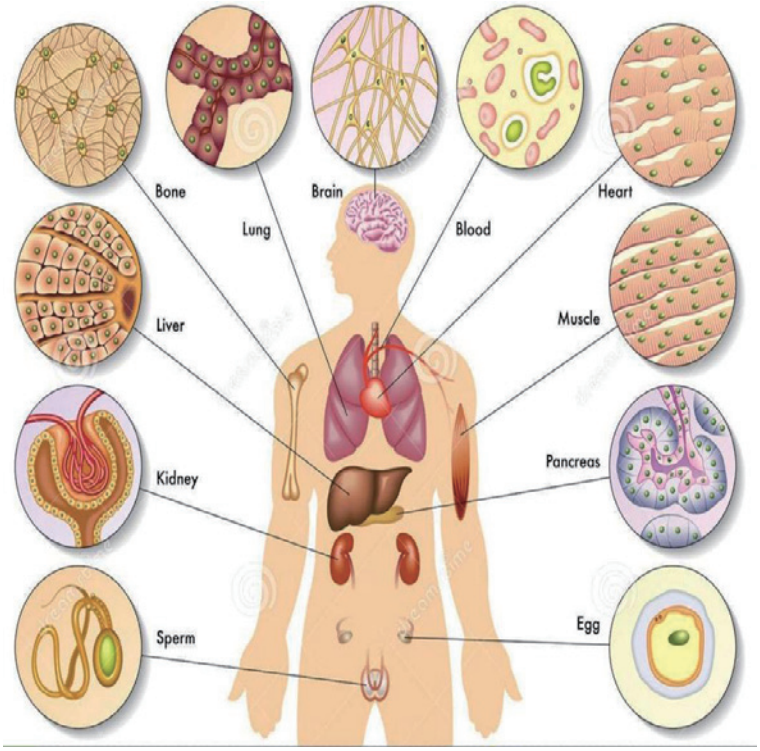
James Dewey Watson
Francis Harry Compton Crick

출처: wikipedia

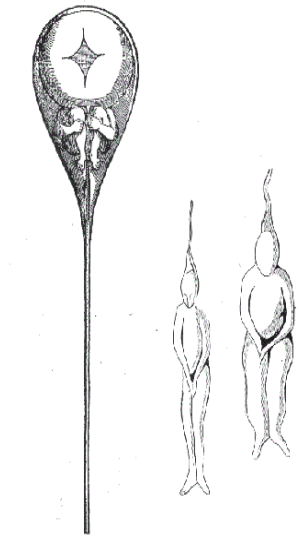
Central Dogma, 생명현상의 중심 원리



Tissues and cells in human body



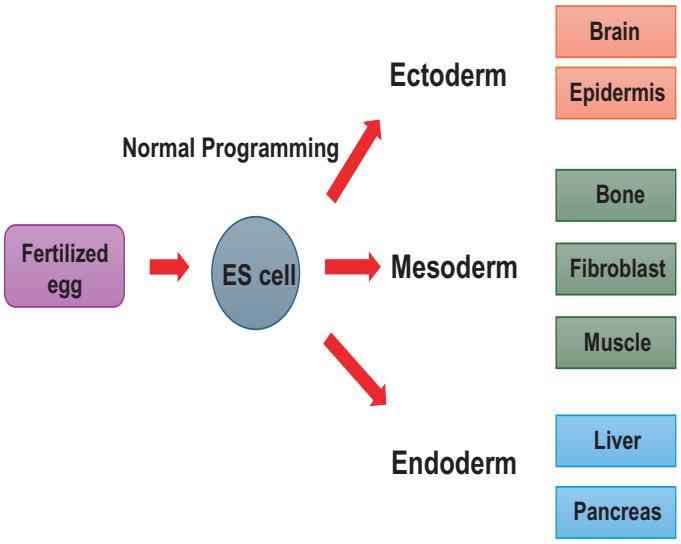
정선설, Preformationism



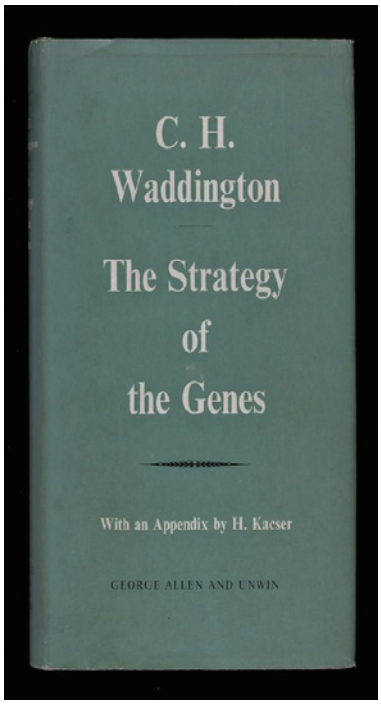
1695년 N. Hartsoecker에 의해
그려진 호문쿨루스(homunculus)

니콜라 말브랑슈(Nicolas Malebranche, 1638~ 1715):
Matryoshka과 같이 더 작은 배아를 무한히 가져야 한다.
"정자와 난자 안에는 식물과 동물의 무한한 시리즈가 존재하며, 이는 충분한 기술과 경험을 가진 자연학자에 의해서만 관찰될 수 있다."
"인간과 동물의 모든 기관은 이미 세상의 만들어짐과 동시에 형성되었다."고 주장하였다.

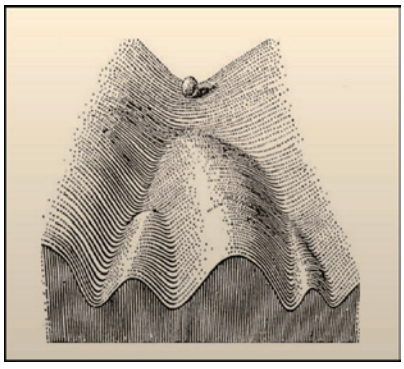
Normal Development (Epigenesis)

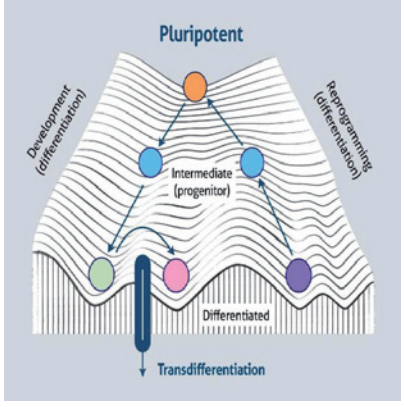


후성유전 지형(epigenetic landscape)

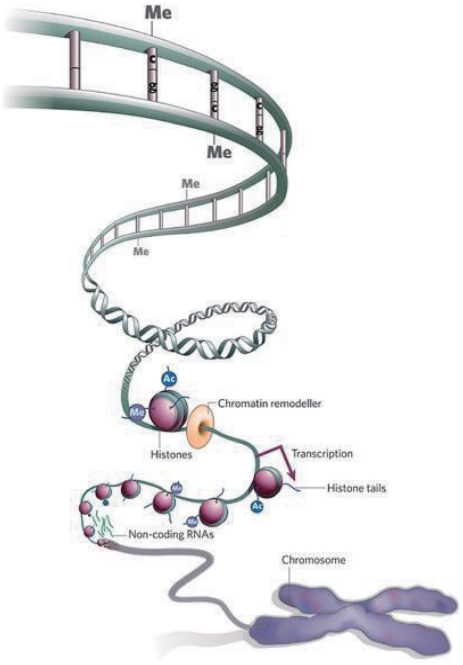


Conrad H. Waddington (1957)

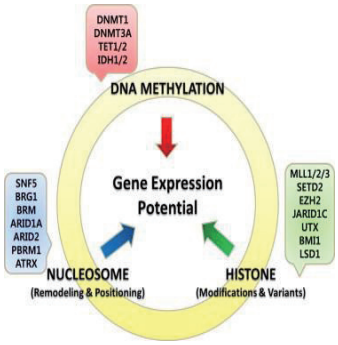




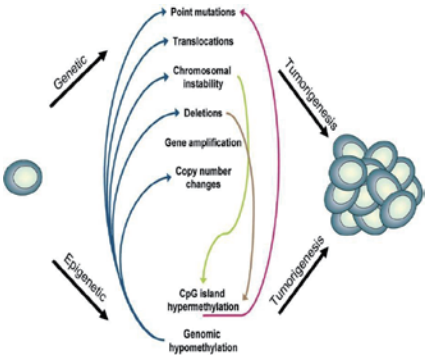
후성 유전학 (epigenetics)



Epigenome project. Nature 454, 711-5 (2008).

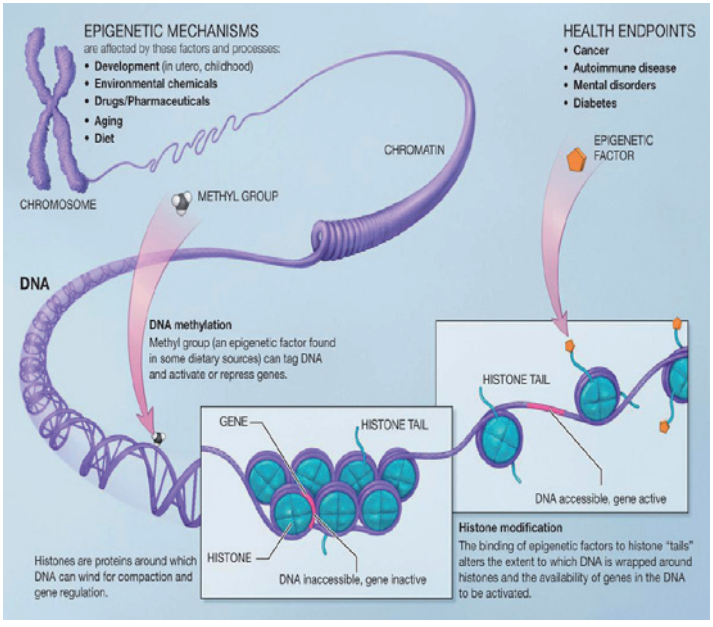


You, J.S. & Jones, P.A. Cancer Cell 22, 9-20 (2012)



Brena, R.M. & Costello, J.F. Hum Mol Genet 16 Spec No 1, R96-105 (2007).

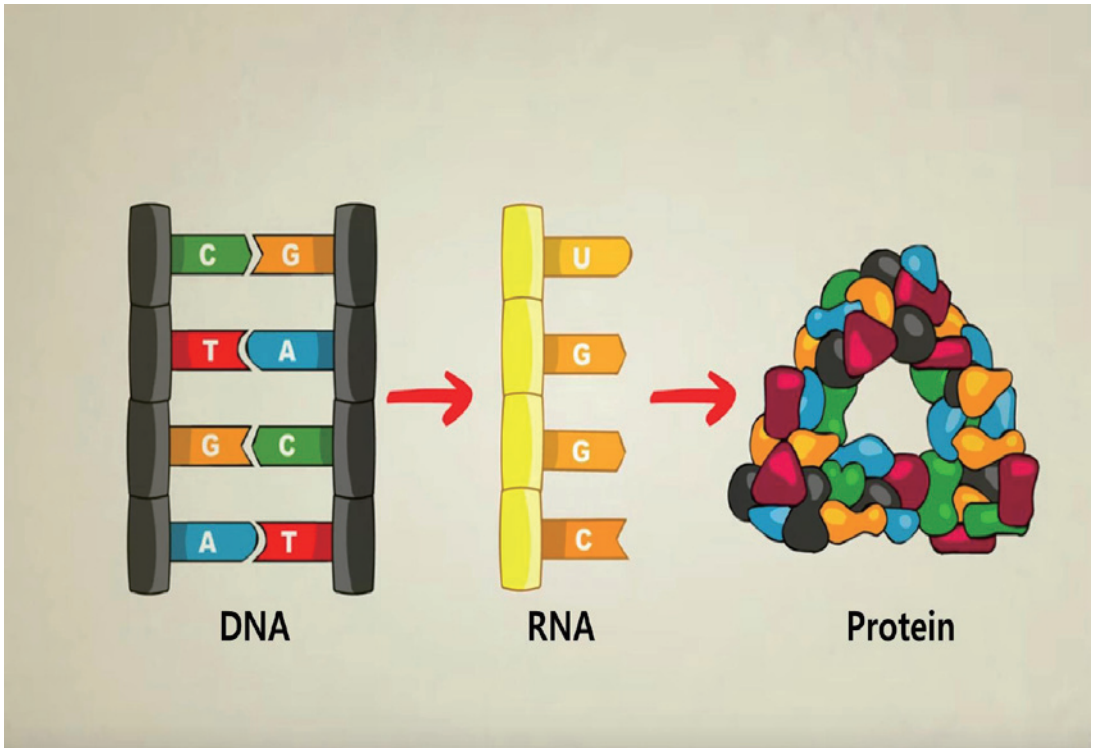
후성 유전학 (epigenetics)



"An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence"
-2008 Cold Spring Harbor Meeting

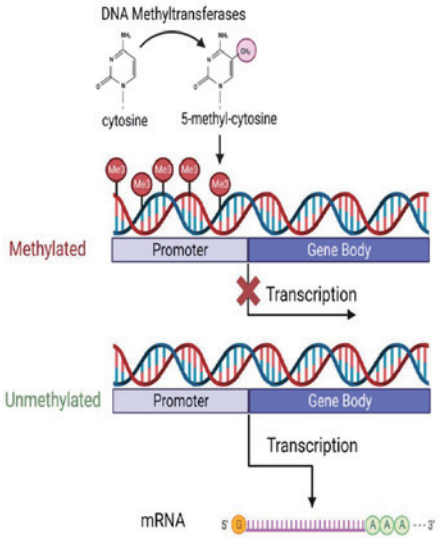
"유전적 변형없이 생기는 염색체 변화"
(1) 생체에 너지를 이용한 염색체 동력성을 유발하는 ATP-dependent chromatin remodeling (ATP 의존적 염색체 리모델링)
(2) 효소 반응을 이용한 염색체내의 DNA 및 histone(히스 톤)의 post-translational modification(번역후 변성)에 의해 이루어진다.
이로 인해 유전자 발현의 시공간적인 조절 및 특정 유전자만의 선별적 발현 조절, chromosomal replication(염색체 복제), recombination(재조합), DNA damage repair (DNA 손상 치료) 등 많은 생명현상들을 안정적으로 조절할 수 있게 된다. 최종적으로 이런 후성유전 정보는 염색체에 기억되며 이는 세포분열과정에서 다음 세대로 전달되게 된다.

Central Dogma, 생명현상의 중심 원리



DNA methylation

1948년 Rollin Hotchkiss에 의해 발견
미생물 DNA에서는 N6-methyladenine (m6A)과 5-methylcytosine (5mC)이 발견
포유류 DNA에서는 5mC만 발견되며, 이는 주로 CpG dinucleotide (5′-CG-3′)내에 주로 존재한다.
DNA 메틸화는 통상적으로 DNA내의 cytosine 메틸화를 의미하게 된다.



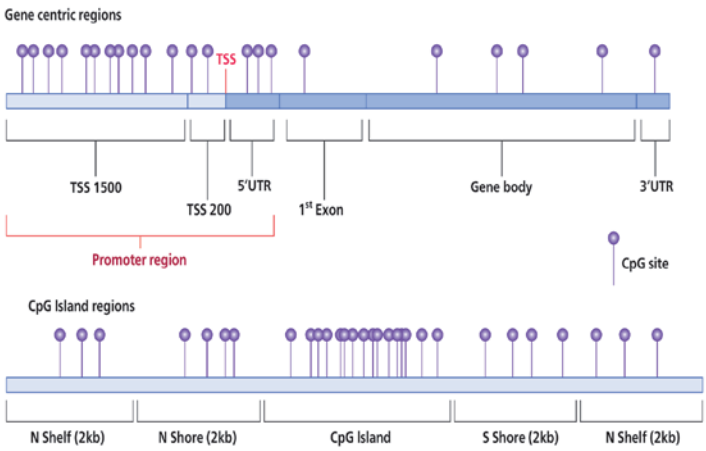


그림 7-3 CpG 섬 Resorts. CpG sites가 많이 존재하는 부분 중 특정 조건을 만족하는 부분을 칭한다. 최근에는 CpG 섬에서 2 kb 떨어져 있으면 shore, 그리고 거기서 다시 2 kb 떨어져 있으면 shelf라 정의하고 방향에 따라 N과 S를 붙이게 된다. 실제적으로 CpG 섬 안에 있는 DNA 메틸화 변화보다 shore 또는 shelf의 변화가 조직 특이적이라고 제안되기도 했다.

DNA methylation- writers

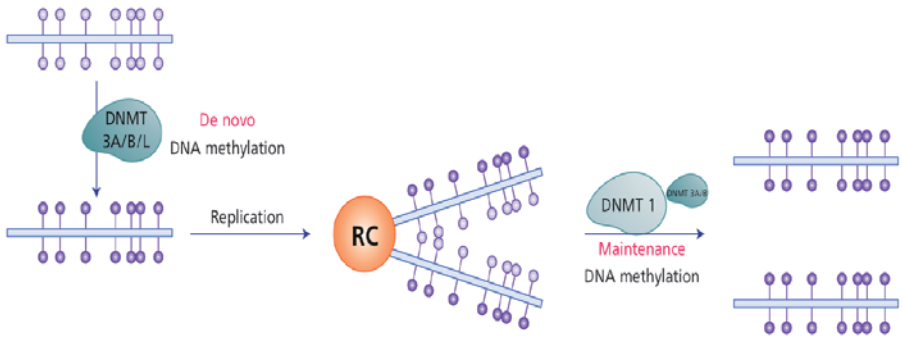


그림 7-8 DNA 메틸화 종류. De novo DNA 메틸화 패턴은 DNMT3A 및 DNMT3B에 의해 확립되는데 이 과정은 촉매적으로 비활성인 DNMT3L의 존재하에 향상된다. DNA 복제가 일어나는 동안 새롭게 생성되는 DNA 가닥에 원래 DNA 메틸화 패턴을 DNMT1의 활동에 의해 대부분 유지되며 DNMT3A 및 DNMT3B의 일부 참여가 있을 수 있다.

현재 알려진 대표적인 methyltransferase들은 DNMT1, DNMT2, DNMT3A, DNMT3B, DNMT3L 임.
DNMT1은 maintenance, DNMT3는 De novo, DNMT2는 tRNA methylation에 관여하는 것으로 알려짐.

DNA methylation- readers

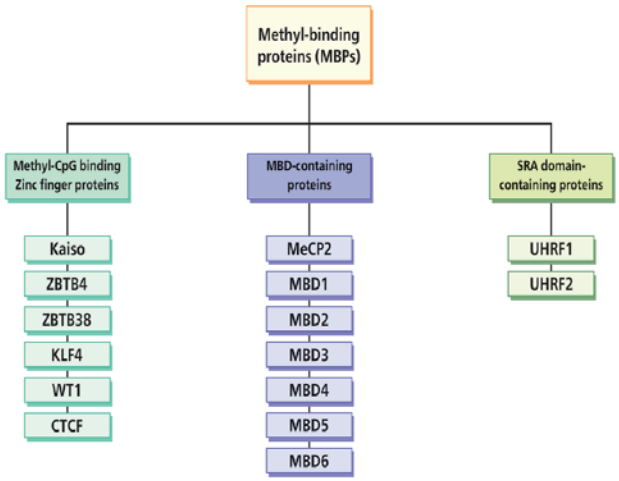


그림 7-9 메틸기 결합 단백질(MBP)의 분류. 메틸화된 DNA에 결합하는 데 사용되는 기능적 도메인에 따라 크게 세 가지 계열로 분류된다. 첫 번째 MBP 그룹이며 MBD 이외의 기능적 도메인의 존재에 따라 세 가지 하위 군(MeCP2-MBD, HMT-MBD 및 HAT-MBD)으로 추가로 분류된다. HMT-MBD 및 HAT-MBD 서브 패밀리의 구성원은 각각 단백질 메틸화효소 및 아세틸화효소 활성을 가지고 있다. 두 번째는 Zinc finger 모티프를 사용하여 메틸화된 영역에 결합할 수 있는 8개 이상의 구성원(Kaiso, ZBTB4, ZBTB38, ZFP57, KLF4, EGR1, WT1, CTCF)을 가지며, 세 번째 MBP 계열은 SRA 도메인을 가지고 있는 UHRF1 및 UHRF2 단백질로 나눌 수 있다.

MBD containing domain은 대칭적으로 메틸화된 DNA 에 결합하는 반면, SRA domain 단백질은 hemi-메틸화된 DNA에 더 잘 결합한다고 알려져 있다. MeCP2는 주로 pericentric 헤테로크로마틴에 존재하며, 메틸화된 CpG site에 MeCP2의 결합은 DNA 탈메틸화효소로부터 메틸 사이토신을 보호하여 전사적으로 침묵하고 있는 유전자의 재활성화를 막는다고 알려 져 있다.

DNA methylation- readers_continued

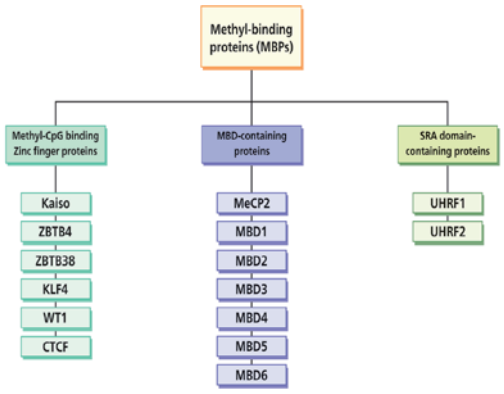


그림 7-9 메틸기 결합 단백질(MBP)의 분류. 메틸화된 DNA에 결합하는 데 사용되는 기능적 도메인에 따라 크게 세 가지 계열로 분류된다. 첫 번째 MBP 그룹이며 MBD 이외의 기능적 도메인의 존재에 따라 세 가지 하위 군(MeCP2-MBD, HMT-MBD 및 HAT-MBD)으로 추가로 분류된다. HMT-MBD 및 HAT-MBD 서브 패밀리의 구성원은 각각 단백질 메틸화효소 및 아세틸화효소 활성을 가지고 있다. 두 번째는 Zinc finger 모티프를 사용하여 메틸화된 영역에 결합할 수 있는 8개 이상의 구성원(Kaiso, ZBTB4, ZBTB38, ZFP57, KLF4, EGR1, WT1, CTCF)을 가지며, 세 번째 MBP 계열은 SRA 도메인을 가지고 있는 UHRF1 및 UHRF2 단백질로 나눌 수 있다.

CpG binding zinc finger 단백질: C terminal에 zinc finger motif를 갖고 DNA에 결합할 수 있는 단백질
Kaiso는 적어도 두 개의 메틸화된 CpG site가 필요하다고 알려져 있지만, ZBTB4 및 ZBTB38 단백질은 하나의 메틸화된 CpG에도 효율적으로 결합한다고 보고됨
이 계열 에 포함되는 단백질은 메틸화된 DNA에 결합하는 능력 외 에도, 메틸화되지 않은 consensus(컨센서스) 서열인 kasio binding sequence (kaiso 결합 부위, KBS, TCCTGCNA), E-box 모티프(CACCTG)와 상호작용할 수 있음이 보고되 어 있다. 이 계열 에 포함되는 단백질은 메틸화된 DNA에 결합하는 능력 외 에도, 메틸화되지 않은 consensus(컨센서스) 서열인 kasio binding sequence (kaiso 결합 부위, KBS, TCCTGCNA), E-box 모티프(CACCTG)와 상호작용할 수 있음이 보고되어 있다. 단순히 DNA 메틸화 reader 이외의 기능을 가 지고 있음을 시사한다.

DNA methylation- readers_continued

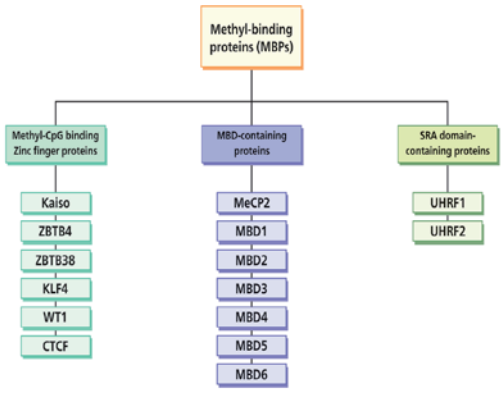


그림 7-9 메틸기 결합 단백질(MBP)의 분류. 메틸화된 DNA에 결합하는 데 사용되는 기능적 도메인에 따라 크게 세 가지 계열로 분류된다. 첫 번째 MBP 그룹이며 MBD 이외의 기능적 도메인의 존재에 따라 세 가지 하위 군(MeCP2-MBD, HMT-MBD 및 HAT-MBD)으로 추가로 분류된다. HMT-MBD 및 HAT-MBD 서브 패밀리의 구성원은 각각 단백질 메틸화효소 및 아세틸화효소 활성을 가지고 있다. 두 번째는 Zinc finger 모티프를 사용하여 메틸화된 영역에 결합할 수 있는 8개 이상의 구성원(Kaiso, ZBTB4, ZBTB38, ZFP57, KLF4, EGR1, WT1, CTCF)을 가지며, 세 번째 MBP 계열은 SRA 도메인을 가지고 있는 UHRF1 및 UHRF2 단백질로 나눌 수 있다.

SRA domain containing 단백질은 UHRF1 과 UHRF2가 있다. 이 두 단백질은 세포 증식에서 근본적인 역할을 한다고 보고되어 있으며, 단백질이 가지고 있는 domain으로 히스톤에 유비퀴틴을 붙이는 고유 한 효소 활성을 갖고 있다고 보고되었다.

DNA methylation- erasers

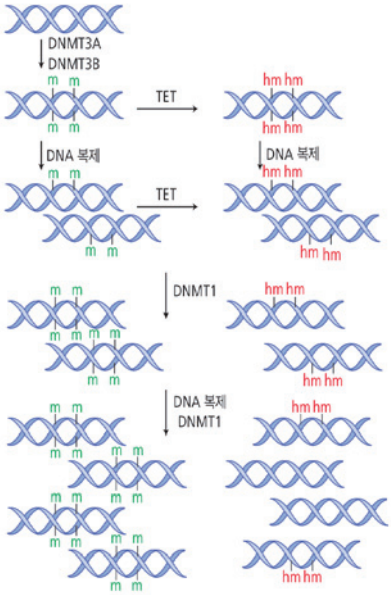


그림 8-2 수동적 DNA 탈메틸화. DNA가 복제되는 동안, DNMT1은 hemimethylated DNA를 인식하여 반대쪽 가닥의 DNA를 메틸화시킴으로써 DNA 메틸화 패턴을 유지한다. TET에 의해 생성된 hemihydroxymethylated DNA는 DNMT1에 의한 메틸화가 유지되지 못하고 DNA 복제과정에서 수동적인 탈메틸화가 일어난다. 출처: (Tutorial Review) Chem. Soc. Rev., 2012, 41, 6916-6930

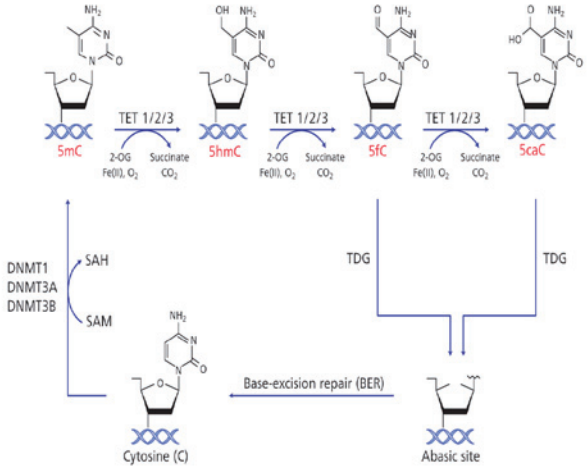


그림 8-3 능동적 DNA 탈메틸화. TET은 5mC를 5hmC (5-hydroxymethylcytosine)로 산화시킬 뿐만 아니라, 5fC (5-formylcytosine)과 5caC (5-carboxylcytosine)로도 산화시키며, 5fC와 5caC은 TDG (thymine-DNA-glycosylase)에 의해 base excision(염기 절단)이 일어난다. 염기가 사라진 abasic site는 BER (base excision repair)에 의해 원래의 cytosine 상태가 되어 DNA 탈메틸화가 일어난다. 출처: Genes Dev. 2016; 30(7): 733-750.

TET family 단백질 중에서, TET1은 주로 초기 embryo (배아), ESC (embryonic stem cell, 배아줄기세포), 그리고 PGC (primordial germ cells, 시원생식세포)에서 발현되는 반면, TET2와 TET3는 다양한 세포에서 발현된다. TET family 단백질들은 상호 보완적이며 포유류의 발생과 분화에 매우 중요한 역할을 한다

Epitranscriptome-RNA modifications

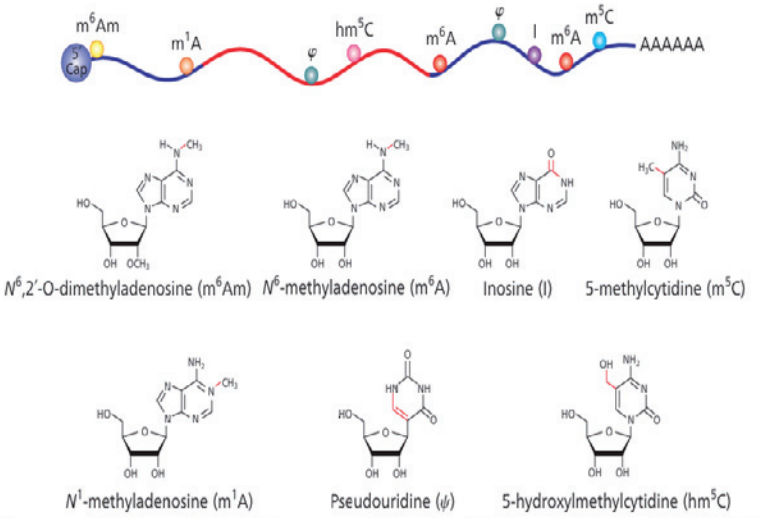


그림 9-1 mRNA에 존재하는 대표적인 화학 변형의 종류. A, U, G, C의 염기가 화학적으로 변형되어 mRNA의 다양한 위치에 여러 종류의 후성전사체 변화가 일어날 수 있다. 출처: Nature Methods volume 14, pages 23–31(2017).

Histone modifications

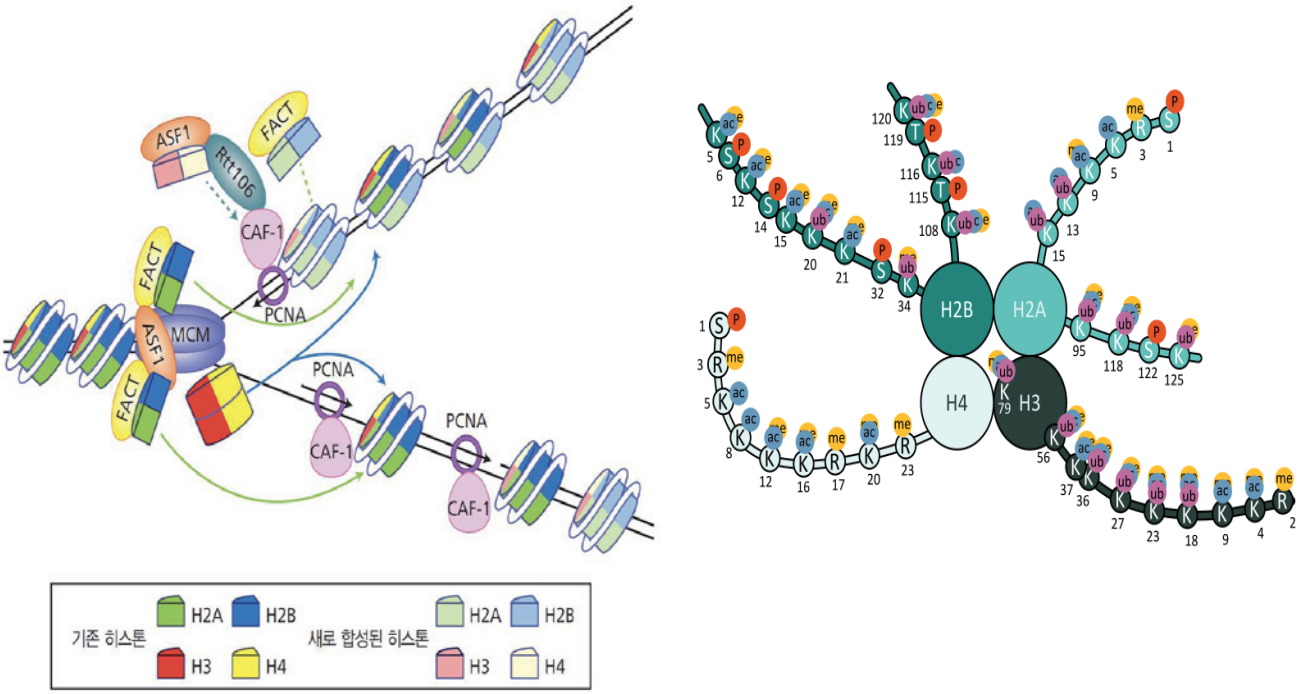
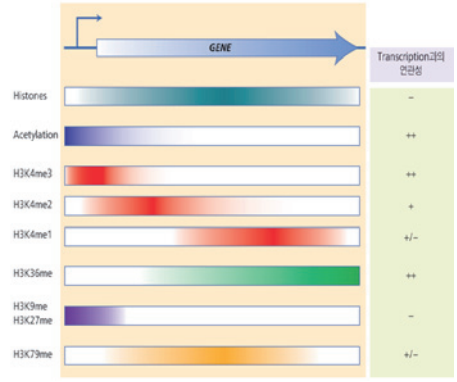


그림 6-1 DNA 복제 과정에서 일어나는 뉴클레오솜 해체 및 조립. DNA가 복제되는 동안 뉴클레오솜 구조는 해체되었다가 다시 조립되는 과정을 거친다. 복제 이후 기존 DNA 양의 2배가 되기 때문에 기존 히스톤 단백질과 새로 합성된 히스톤 단백질이 함께 조립되며, 샤페론 단백질이 뉴클레오솜 해체 및 조립 과정을 돕는다.

Histone modifications



Modification	Genomic distribution	Functional association
H3K4me1	poised enhancers, promoters	enhancer priming, transcriptional activation
H3K4me3	active promoters	transcriptional activation
H3K27me3	inactive promoters	transcriptional repression
H3K27ac	active enhancers, promoters	transcriptional activation,
H3K9me1	heterochromatin, enhancers	transcriptional repression, heterochromatin formation
H3K9me3	heterochromatin	transcriptional repression, heterochromatin formation
H3K36me3	gene bodies	transcriptional activation, RNA splicing
H4K16ac	gene bodies, enhancers, promoters	chromatin de-condensation, transcriptional activation

Steinbach, N. (2017)

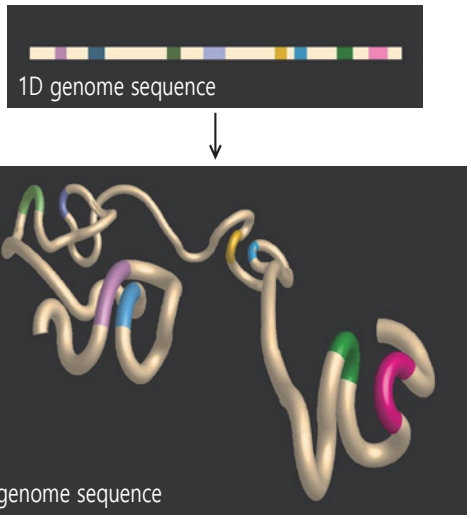
Modification	Histone	Residue	Effects of transcription
Acetylation	H2A	K5	Activation
	H2B	K5, K12, K15, K20	Activation
	H3	K4, K9, K14, K18,	Activation
	H3	K23, K36,	Activation
	H3	K56	DNA repair, histone deposition
	H4	K5, K8, K16	Activation
	H4	K12	Activation, histone deposition
	H4	K91	Histone deposition
Methylation	H3	K4, K79	Activation
	H3	K9, K27	Repression
	H3	R2, R8, R17, R26	Activation
	H3	K36	Elongation
	H4	R3	Activation
	H4	K20	Repression
Phosphorylation	H2A	S1, T120	Mitosis
	H2AX	S139	DNA repair
	H2B	S14	Apoptosis
	H3	T6	Activation
	H3	T3, S10, T11, S28	Mitosis, DNA repair
	H3	T45	DNA replication
	H4	S1	Mitosis, activation
Ubiquitination	H2A	K119	Repression
	H2B	K120	Elongation
	H3	K23	Maintenance of DNA methylation

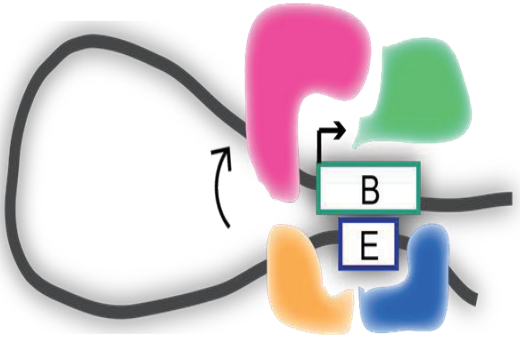
K lysine, *R* arginine, *S* serine, *T* threonine

Thomas, Elizabeth. (2017)

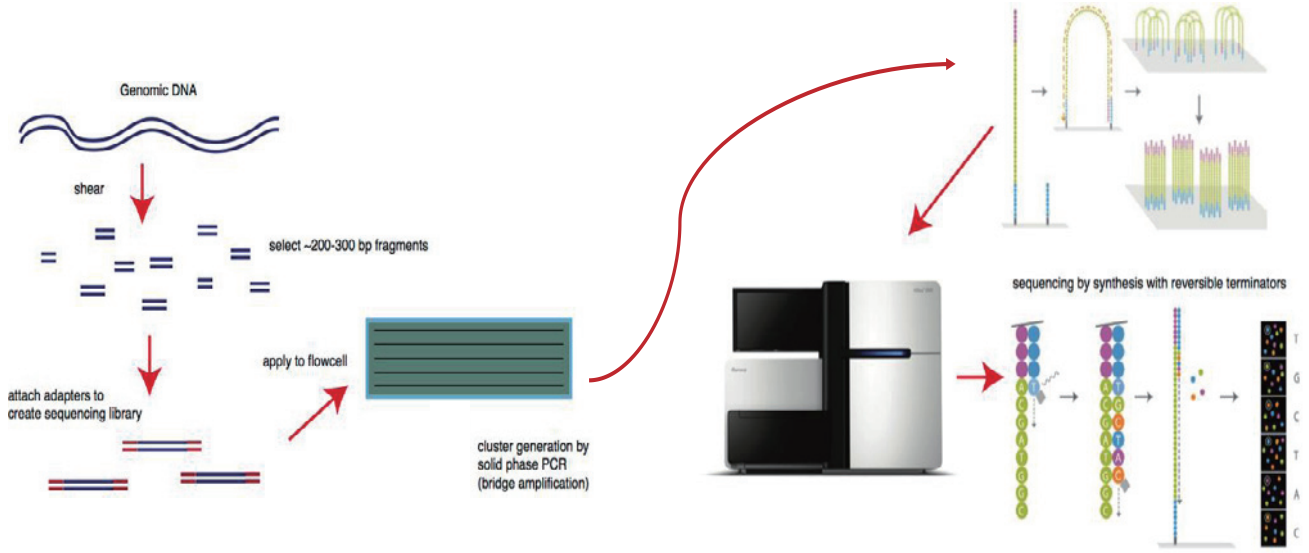
3D genome Organization
Chromatin Conformation

GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT
TTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAAGC
AGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGG
CGGAAAATGAGAAAATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCT
TTCGCCATCAACTAACGATTCTGTCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGC
TATGCTTGGCACGTTCGTCAAGGACTGGTTTAGATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATT
TTAAAAGAGCGTGGATTACTATCTGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCG





Next Generation Sequencing (NGS)

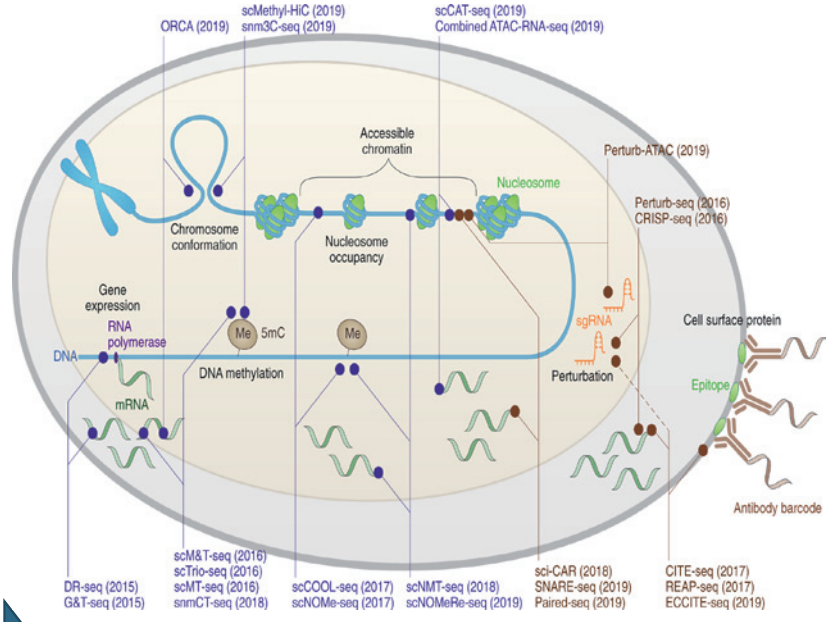


NGS applications

Genome (DNA)		Whole Genome Seq
		Exome Seq
Transcriptome (RNA)		RNA seq
Epigenome	DNA methylation	Bisulfite Seq
	Protein binding	ChIP-seq
	Chromatin accesibility	ATAC-seq
	Chromatin conformation	Hi-C seq

Single cell Multi-modal omics

NGS applications

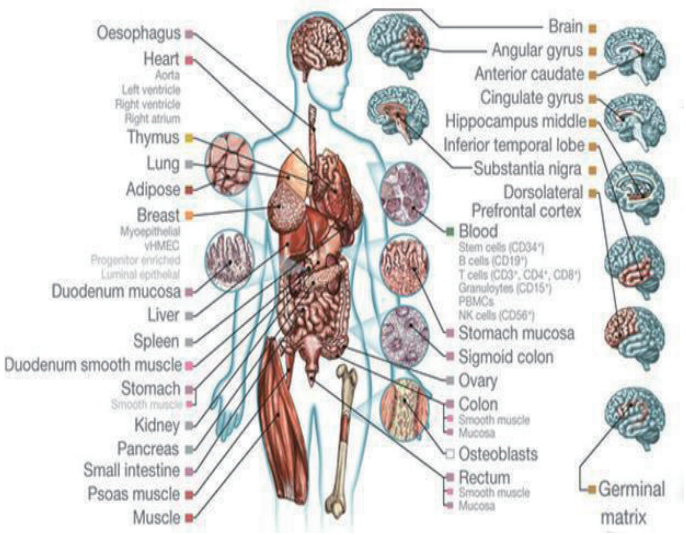


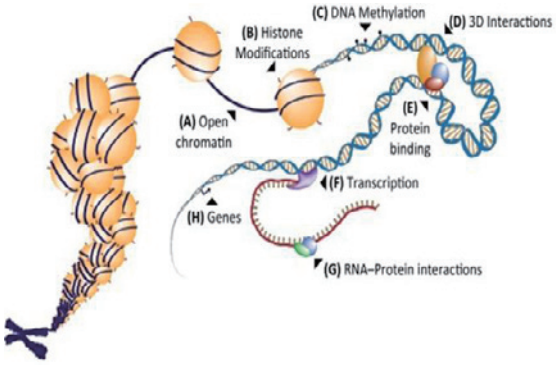
Single cell Multi-modal omics

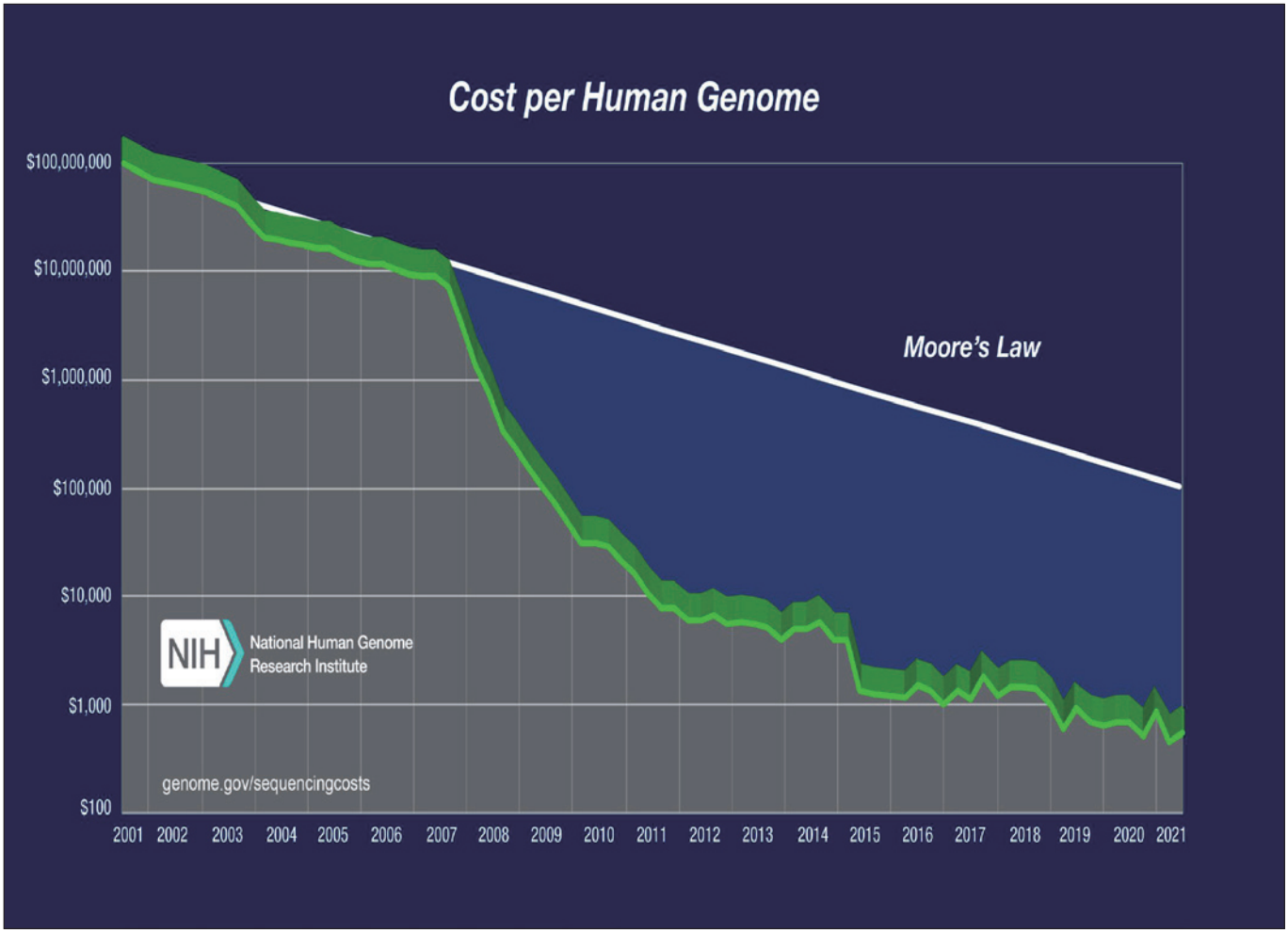
후성유전학 Projects

ENCODE

ROADMAP epigenomics PROJECT







[연구배경] 최신 NGS 동향

ULTIMA GENOMICS

UG-100

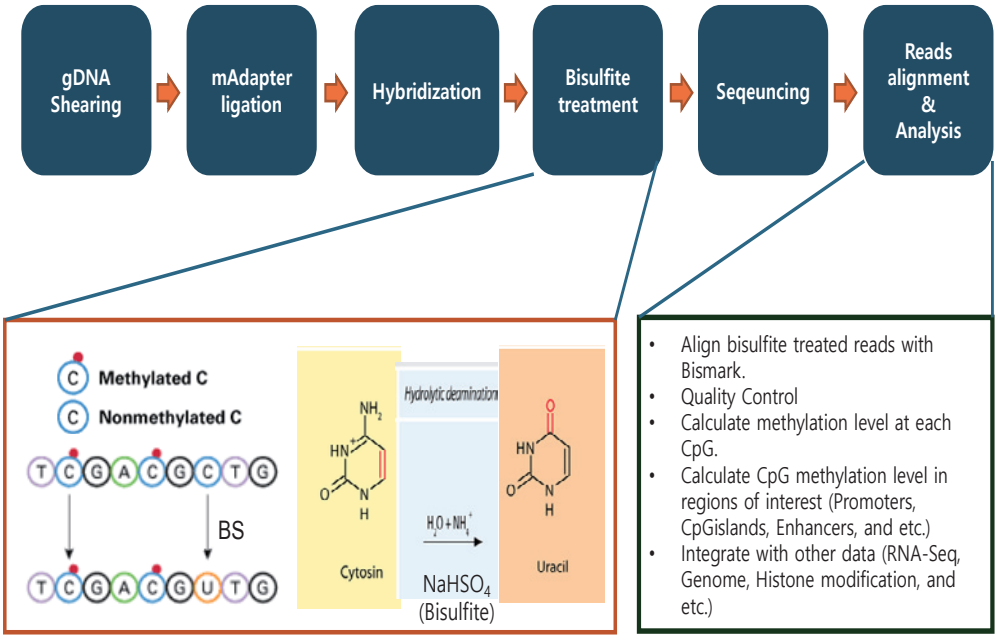
$100

Ultima Genomics is launching high-end DNA sequencers that can read a human genome for as little as $100.

STAT

Ultima Genomics사는 최근 $100에 한사람의 genome을 30x로 시퀀싱할 수 있는 UG-100을 개발, 상용화 하였음.

Bisulfite Sequencing (WGBS)

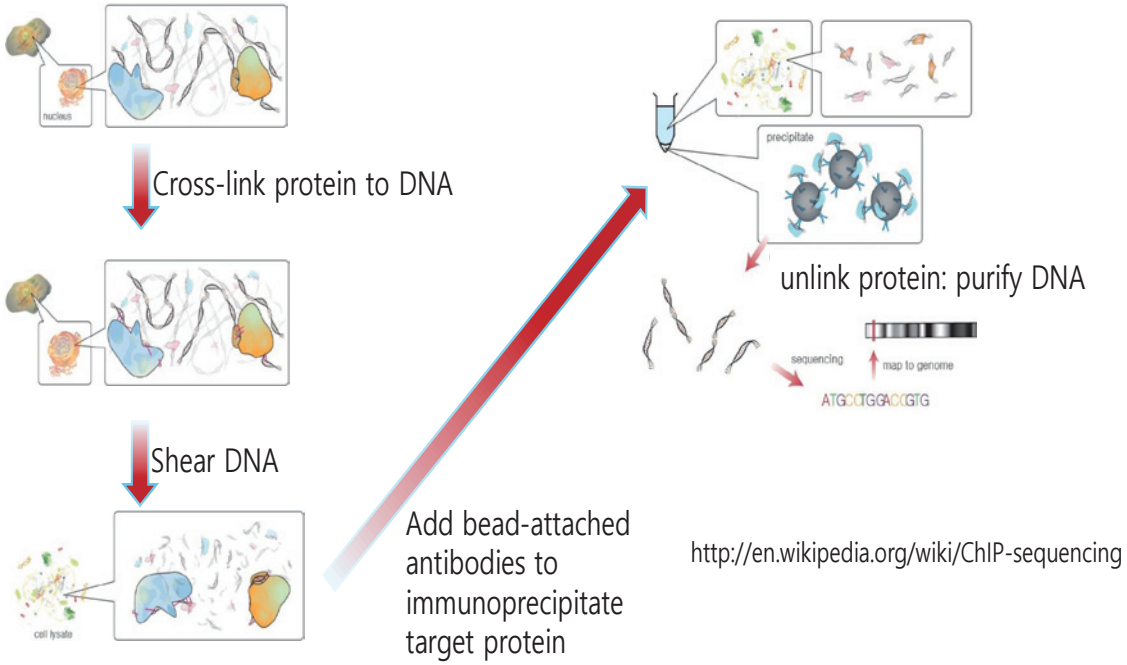


- Align bisulfite treated reads with Bismark.
- Quality Control
- Calculate methylation level at each CpG.
- Calculate CpG methylation level in regions of interest (Promoters, CpGislands, Enhancers, and etc.)
- Integrate with other data (RNA-Seq, Genome, Histone modification, and etc.)

GMI-SNU

Chromatin immunoprecipitation sequencing

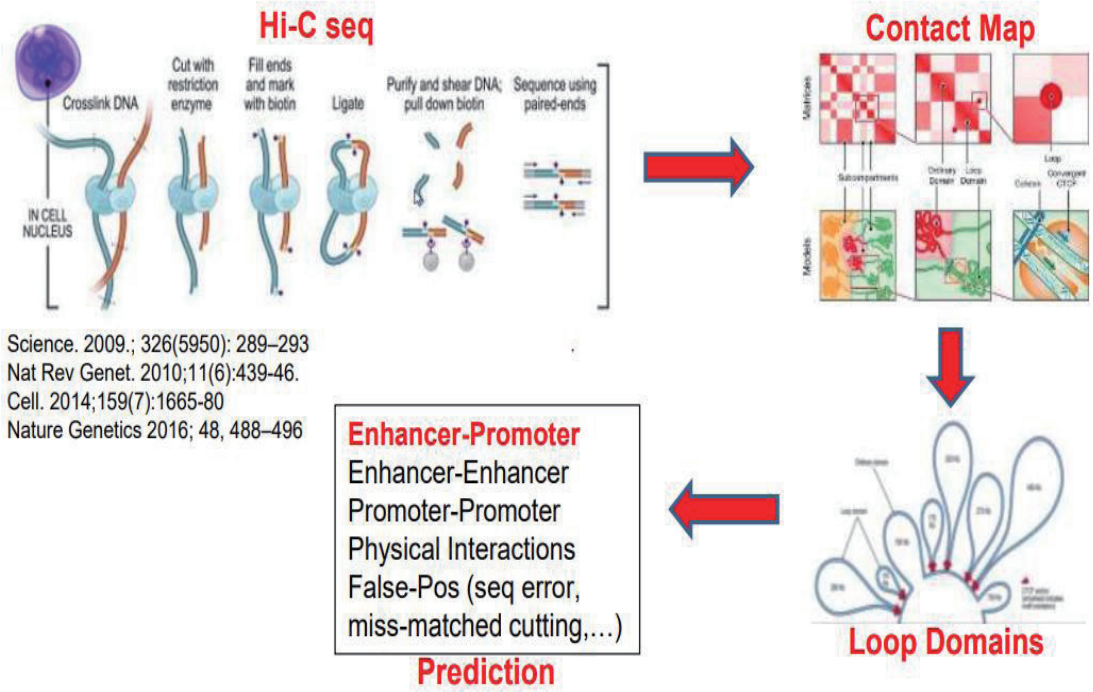
(Chip-Seq): for study histone marks (H3K4/K27/K36me3)



http://en.wikipedia.org/wiki/ChIP-sequencing

High throughput chromatin conformation capture (Hi-C) sequencing

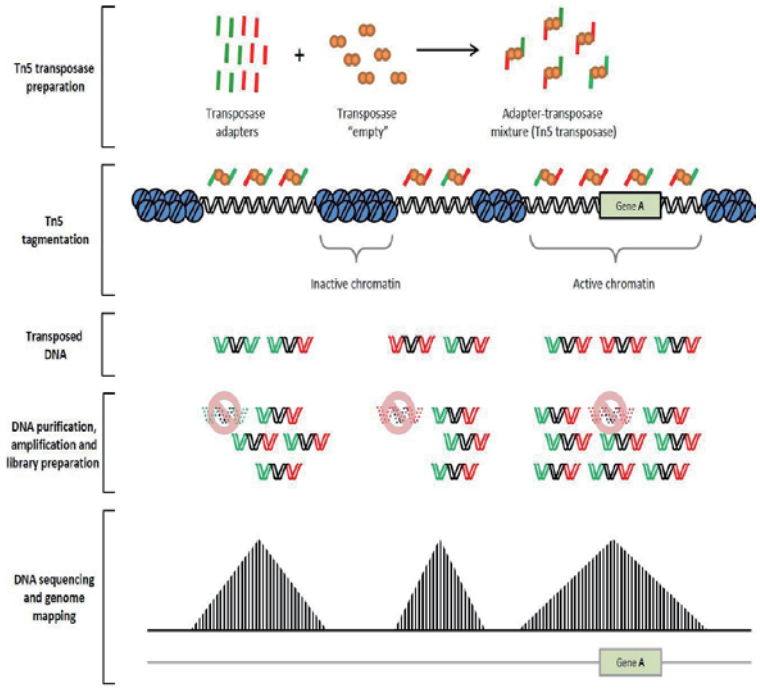
: for study genome 3d genome interaction

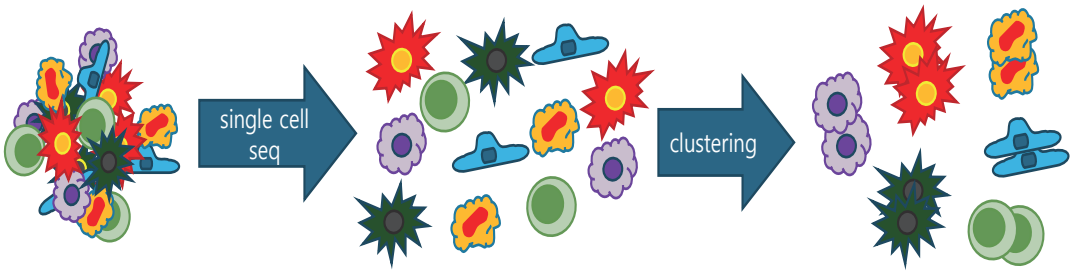


Science. 2009.; 326(5950): 289–293
Nat Rev Genet. 2010;11(6):439-46.
Cell. 2014;159(7):1665-80
Nature Genetics 2016; 48, 488–496

ATAC-seq

(**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing)

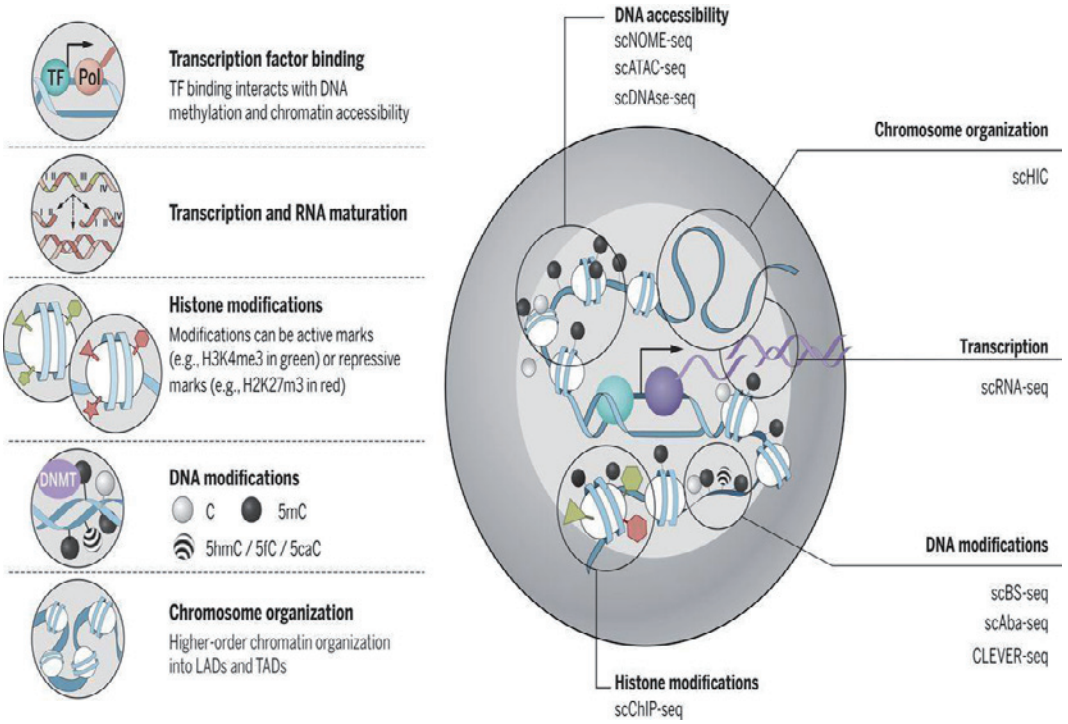




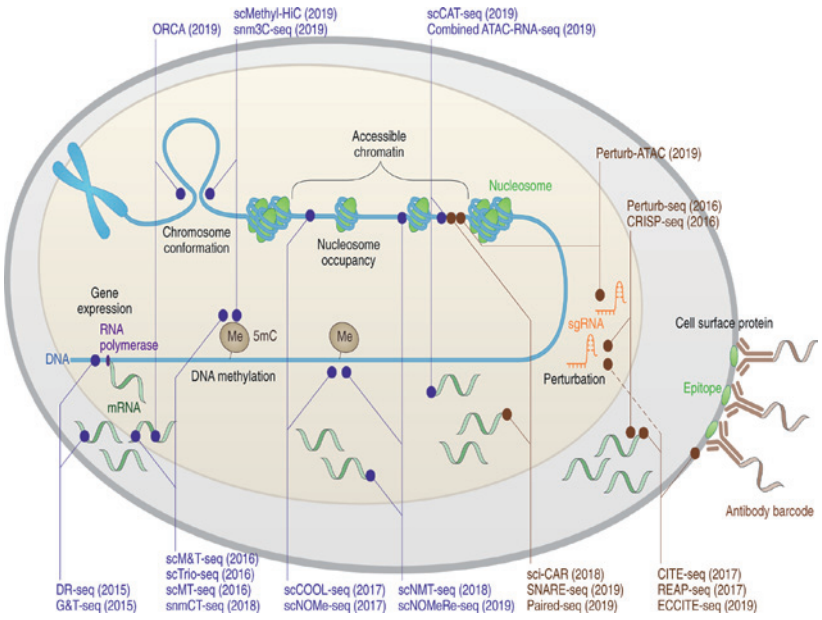
암조직 샘플

단일 세포 데이터 생성 및 분리

세포종별 데이터 분리 및 비교



NGS applications

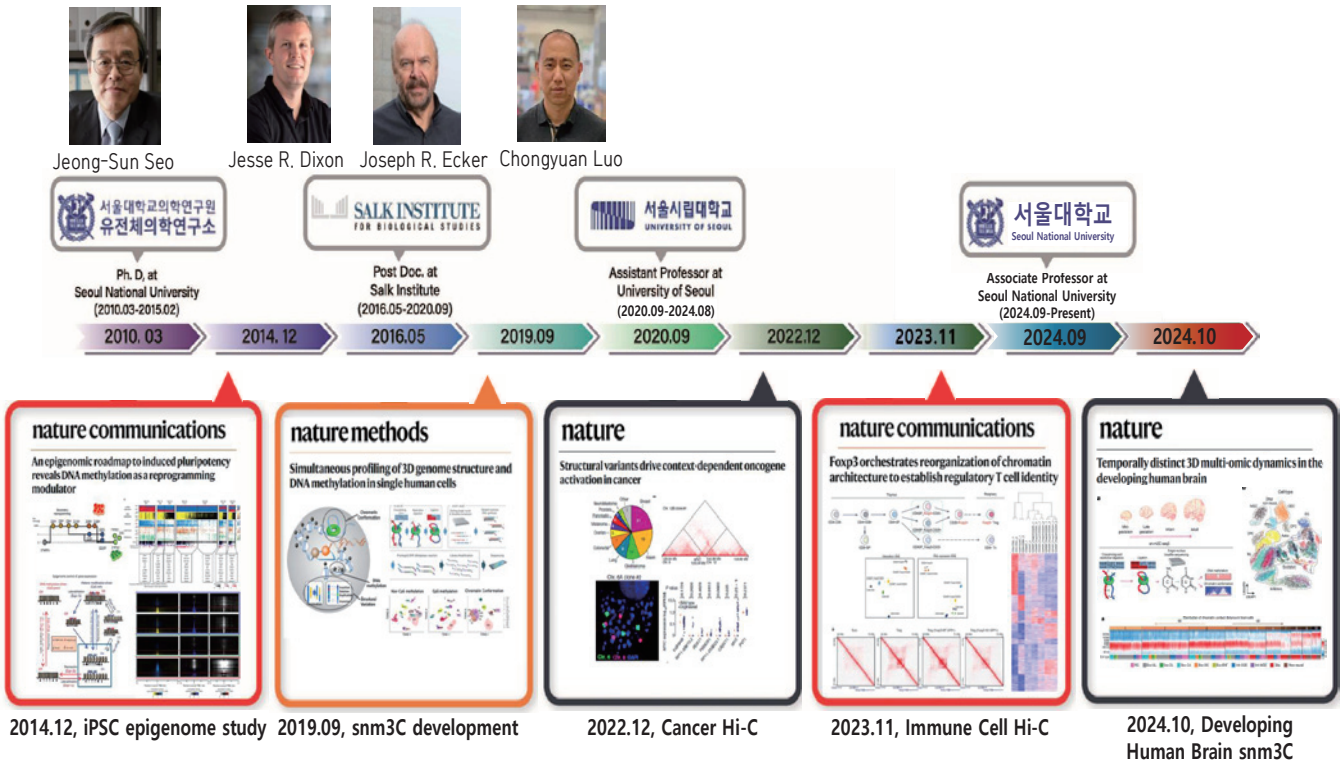


Single cell Multi-modal omics

Method of the Years

Method of the Year	Year
Methods for Modeling Development	2023
Long Read Sequencing	2022
Protein structure prediction	2021
Spatially resolved transcriptomics	2020
Single-cell multimodal-omics	2019
Imaging in freely behaving animals	2018
Organoids	2017
Epitranscriptome analysis	2016
Cryo-EM	2015
ligh-sheet fluorescence microscopy	2014
Single cell sequencing	2013
Targeted proteomics	2012
Genome engineering (TALEN)	2011
Optogenetics	2010
Induced pluripotency	2009
Super-resolution fluorescence microscopy	2008
Next-generation sequencing	2007

Major Career and Research Achievements

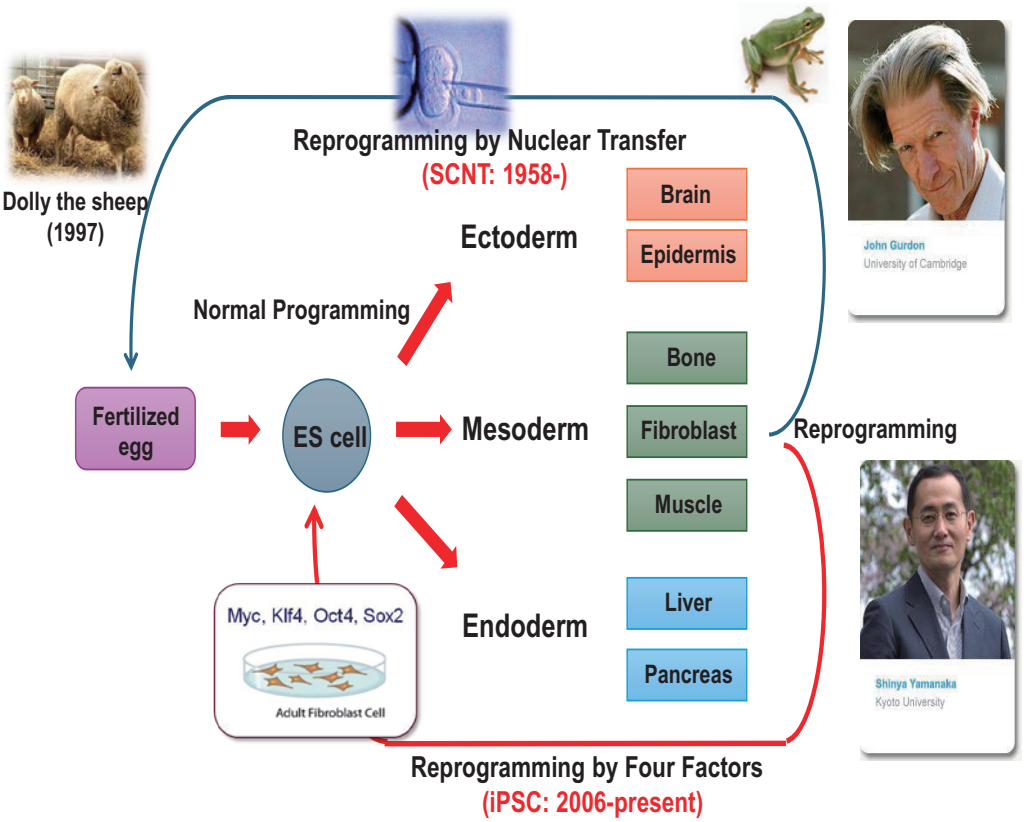


2014.12, iPSC epigenome study | 2019.09, snm3C development | 2022.12, Cancer Hi-C | 2023.11, Immune Cell Hi-C | 2024.10, Developing Human Brain snm3C

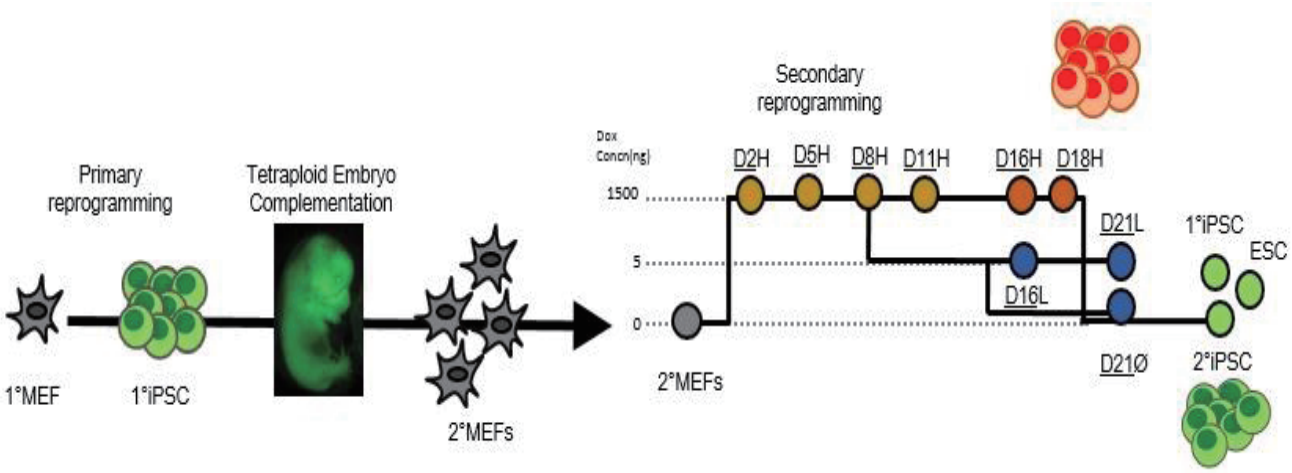
Epigenomic changes during iPSC reprogramming

used techniques:
- Bisulfite seq
- ChIP seq
  - H3K4me3
  - H3K27me3
  - H3K36me3)
- RNA seq

Normal development vs Reprogramming
: Cell differentiation vs Dedifferentiation



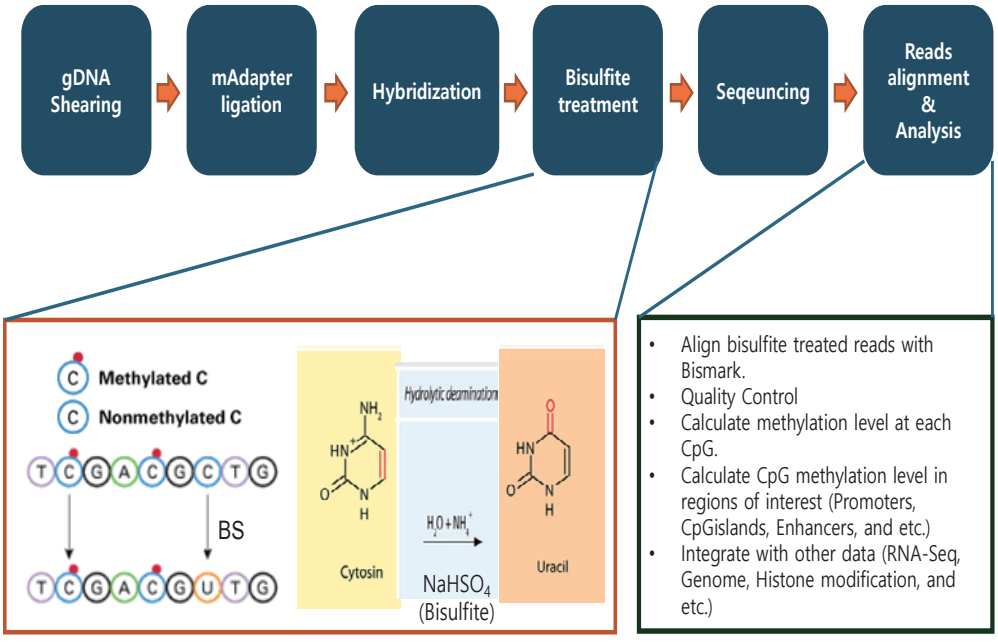
Experimental Design



Lee et al., *Nature Comm.* 2014

Whole Genome Bisulfite Sequencing (WGBS)

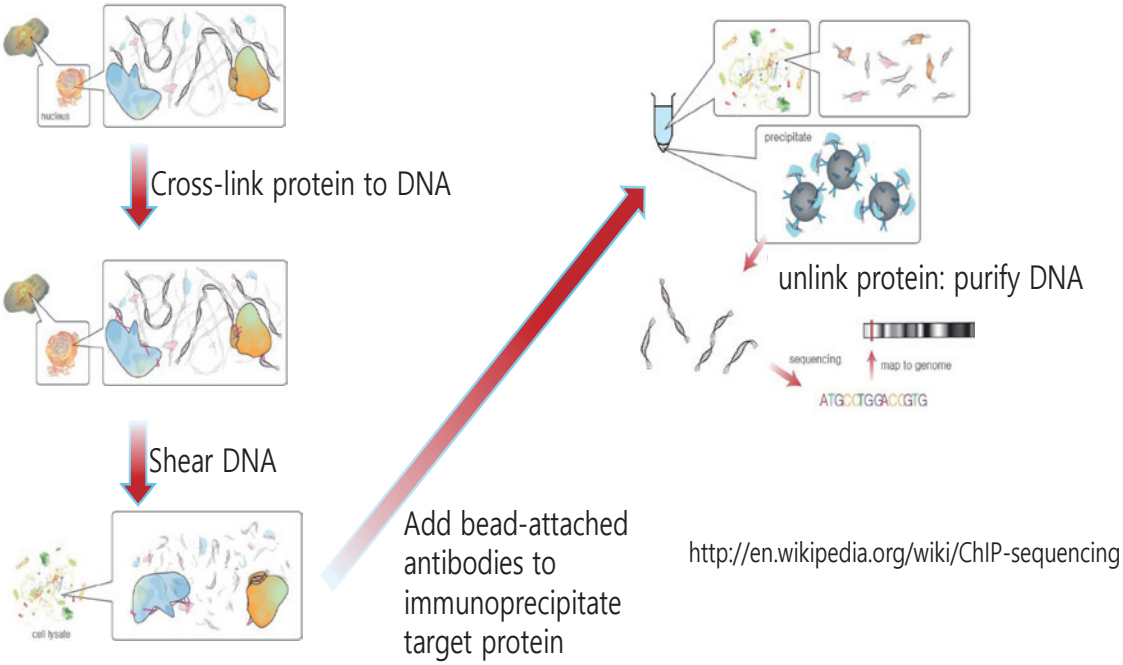
: for study methylome in whole-genome scale at base resolution



- Align bisulfite treated reads with Bismark.
- Quality Control
- Calculate methylation level at each CpG.
- Calculate CpG methylation level in regions of interest (Promoters, CpGislands, Enhancers, and etc.)
- Integrate with other data (RNA-Seq, Genome, Histone modification, and etc.)

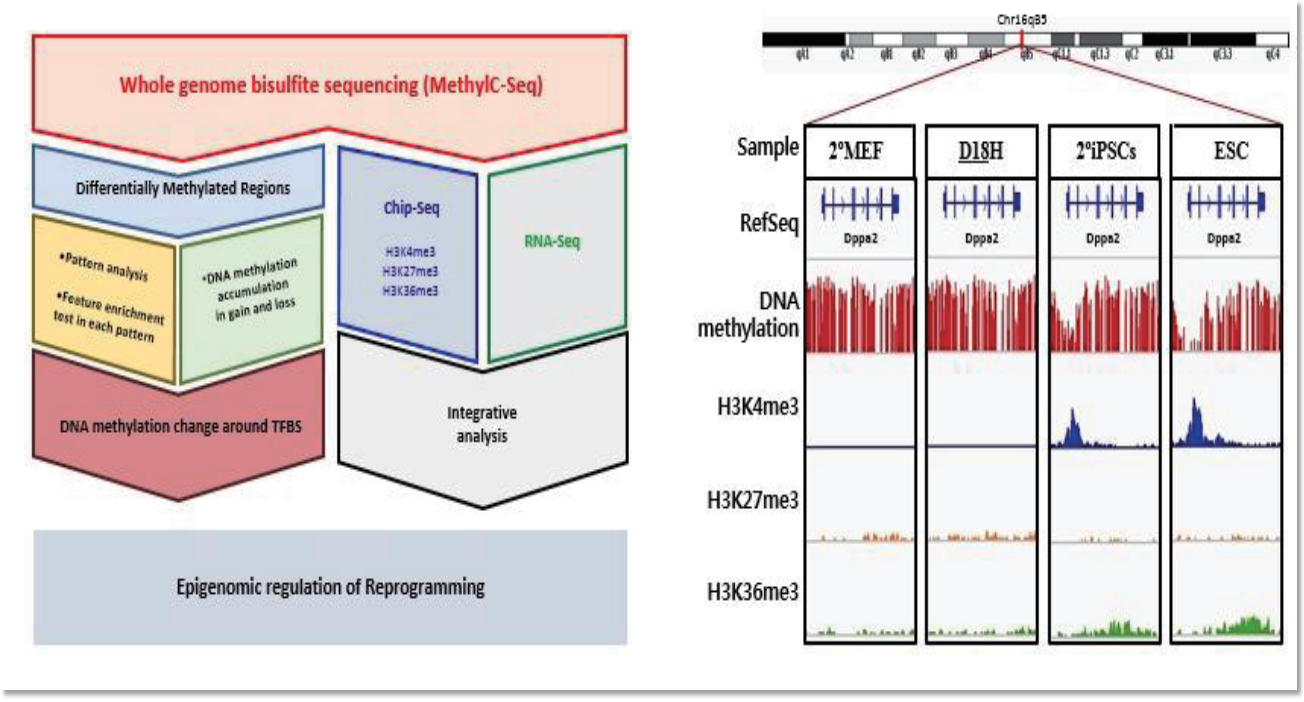
Chromatin immunoprecipitation sequencing

(Chip-Seq): for study histone marks (H3K4/K27/K36me3)



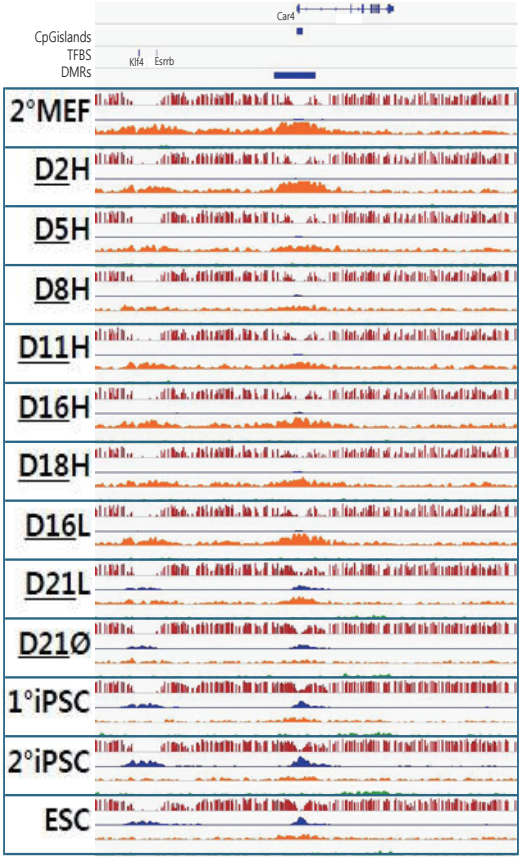
http://en.wikipedia.org/wiki/ChIP-sequencing

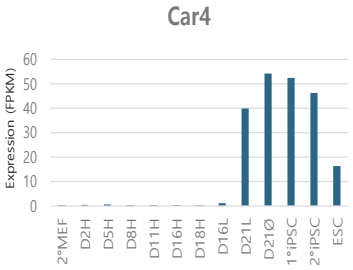
Overview of Data Analysis

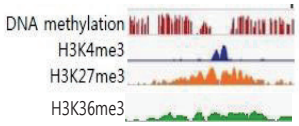


Lee et al., *Nature Comm.* 2014

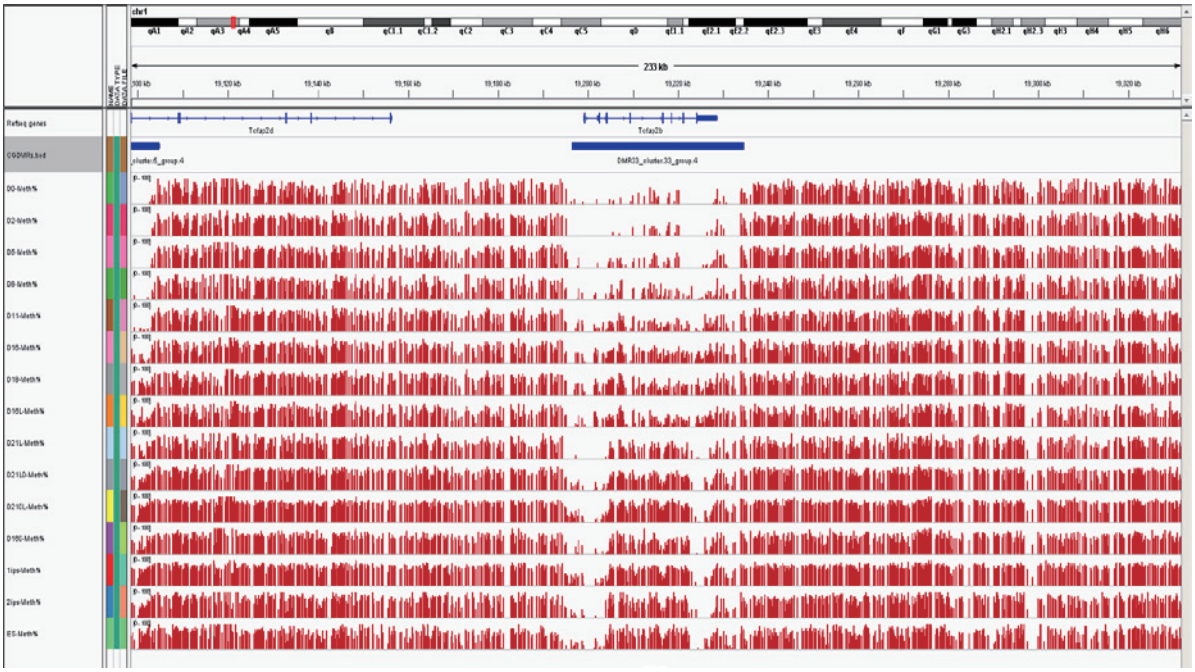
Visulized Data in Base-Resolution







Differentially methylated regions (DMRs)



Differentially methylated regions (DMRs)

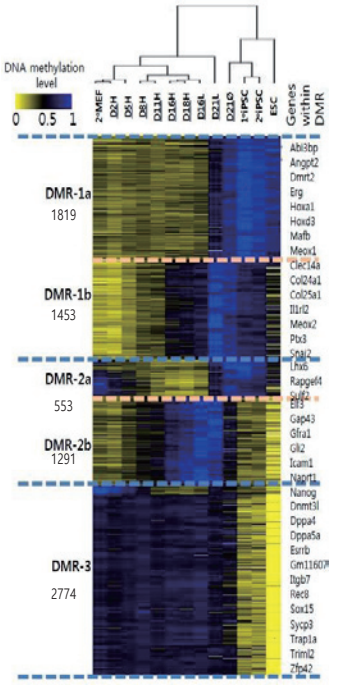
Make windows for 30 CpGs starting from every CpG(Sliding window)

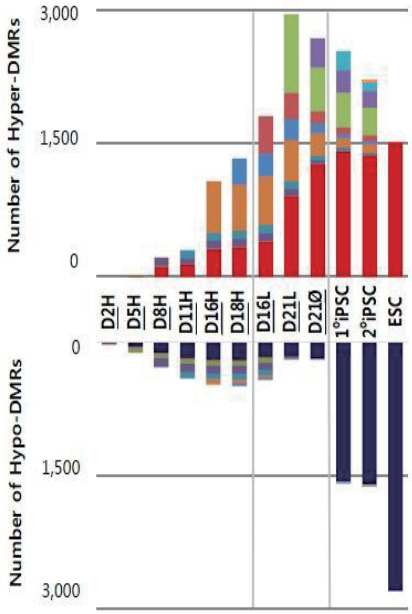
Filter windows:
Maximum-minium>=30%
Maximum/minimum>=4
Size of window<=6kb
All the samples should covered more than 5 in at least 15 CpGs/window

anova-test for every window, Benjamini-Hochberg FDR<0.01

Cluster all the differentially methylated windows if a region contains more than 50% of differentially methylated region: 7890(21,618,964bp)

Changes of DNA methylation during reprogramming





Lee et al., *Nature Comm.* 2014

Are there any specific factors that push DNA methylation change into a certain direction?

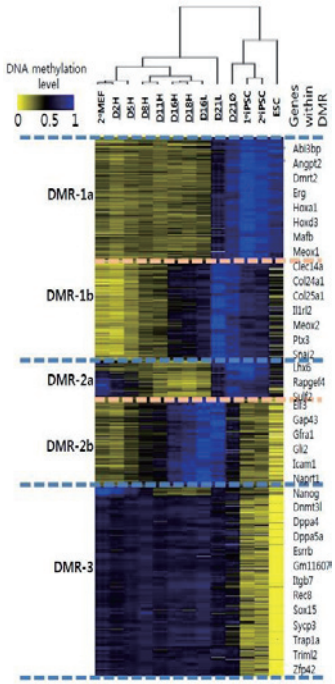


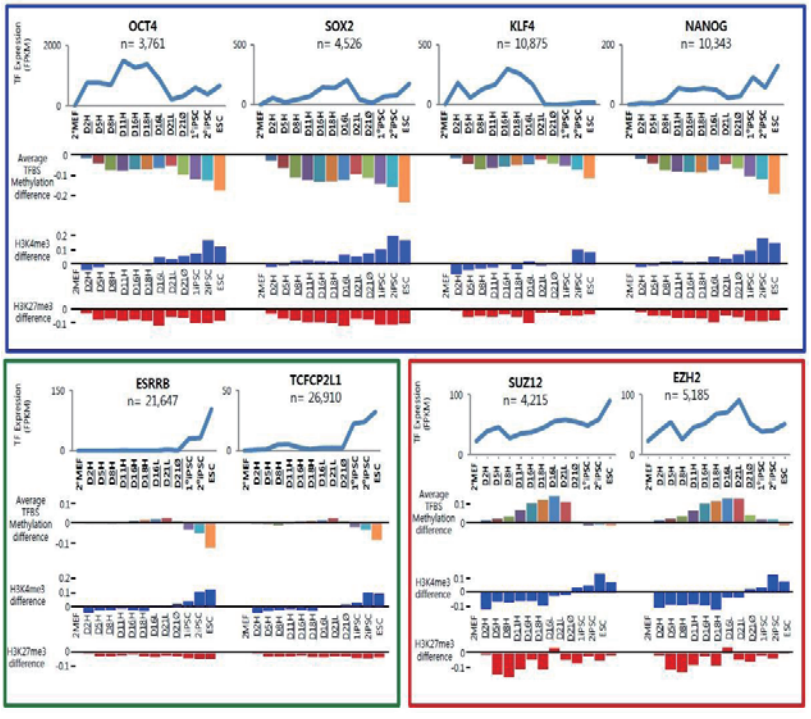
Table 1 Enrichment of Transcription factor binding sites in each DMR group

DMR Groups	Number	sequence-specific transcription factors													Transcription regulators			
		TET1	CTCF	Oct4	SOX2	NANOG	ESRRB	ZFX	KLF4	cMYC	nMYC	E2F1	TCFCP2L1	SMAD1	STAT3	p300	EZH2	SUZ12	RING1B
1a	1819	NE	NE	-	-	-	-	-	-	-	-	-	-	-	-	-	NE	NE	NE
1b	1453	NE	NE	-	-	-	-	-	-	-	-	-	-	-	-	-	++	++	+
2a	553	NE	NE	-	NE	NE	NE	NE	++	NE	NE	NE	NE	NE	NE	-	-	-	-
2b	1291	+	NE	NE	NE	NE	+	++	NE	NE	+	NE	+	-	NE	+	+++	+++	+++
3	2774	NE	NE	+++	+++	+++	++	++	++	+++	+++	+++	+	+++	+++	+++	-	-	-

Fold enrichment vs Total DMRs: - < 0.75X ≤ NE (Not enriched) ≤ 1.25X ≤ + < 1.5X ≤ ++ < 1.75X ≤ +++

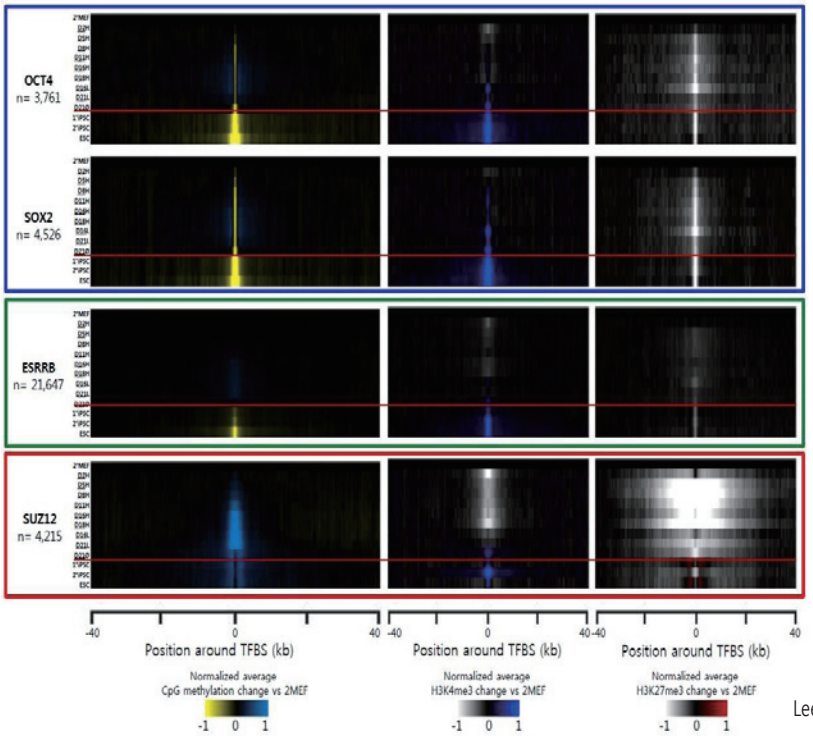
Lee et al., *Nature Comm.* 2014

What happens to each factor binding sites?



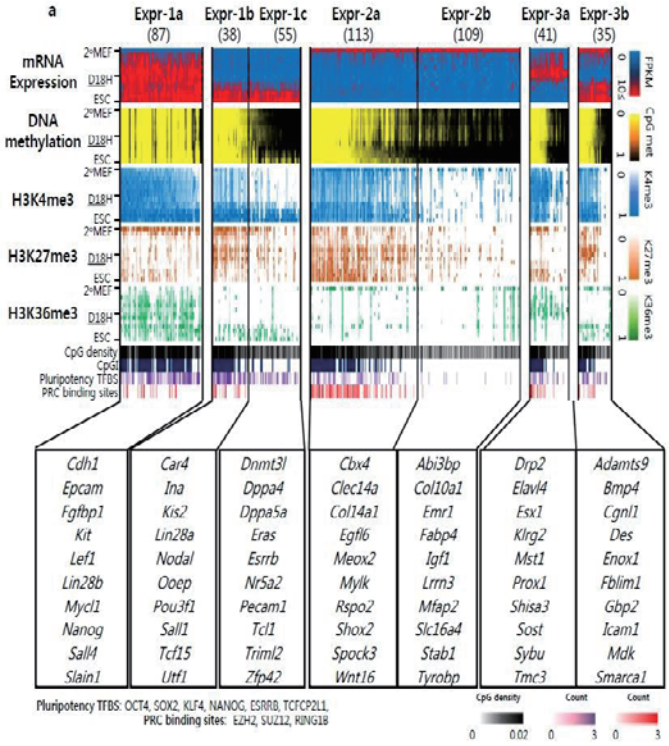
Lee et al., *Nature Comm.* 2014

Histone modification and DNA methylation change around transcription factor binding sites

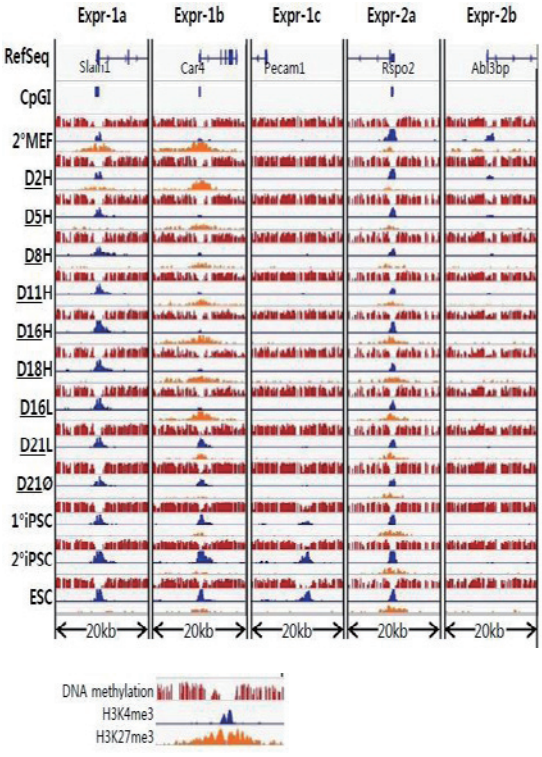


Lee et al., *Nature Comm.* 2014

Epigenomic change during iPSC reprogramming

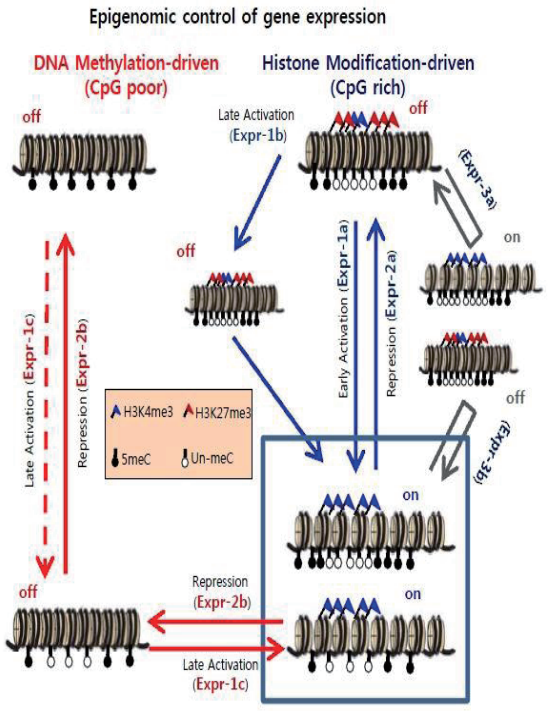


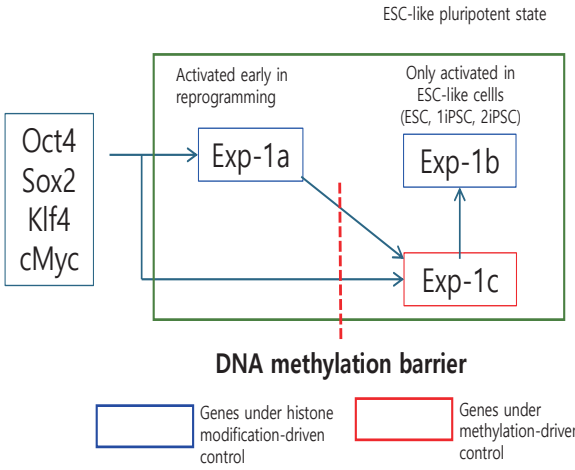
Lee et al., *Nature Comm.* 2014



Lee et al., *Nature Comm.* 2014

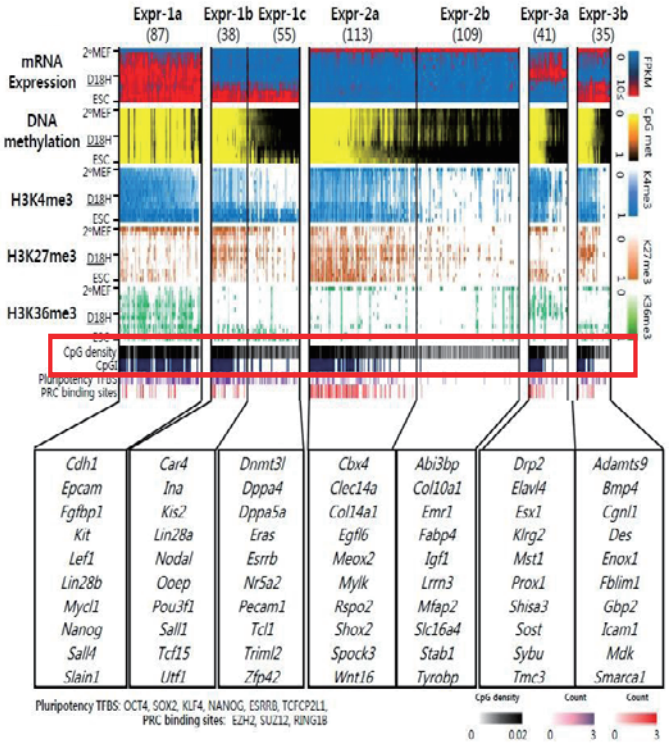
Model of gene expression control during reprogramming



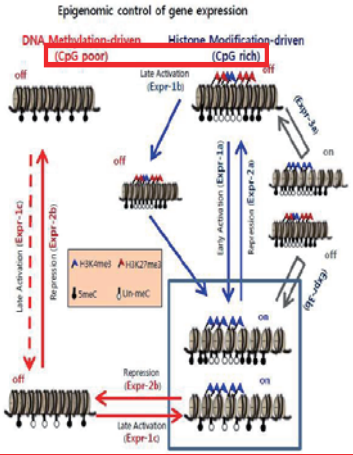


Lee et al., *Nature Comm.* 2014

Summary



Lee et al., *Nature Comm.* 2014

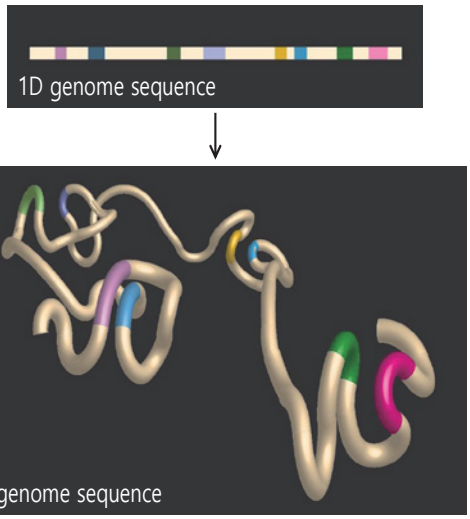


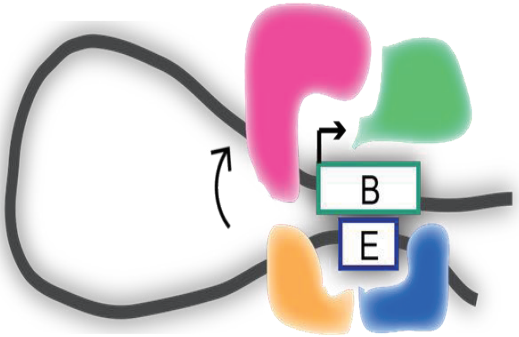
1. Gain of DNA methylation is easy, but the loss is restricted.
2. We found 2 groups of genes
   CpG poor promoter Group is mainly regulated by DNA methylation
   CpG rich promoter Group is mainly regulated by histone modification

Genes in **Group1** is hard to be activated when it's repressed

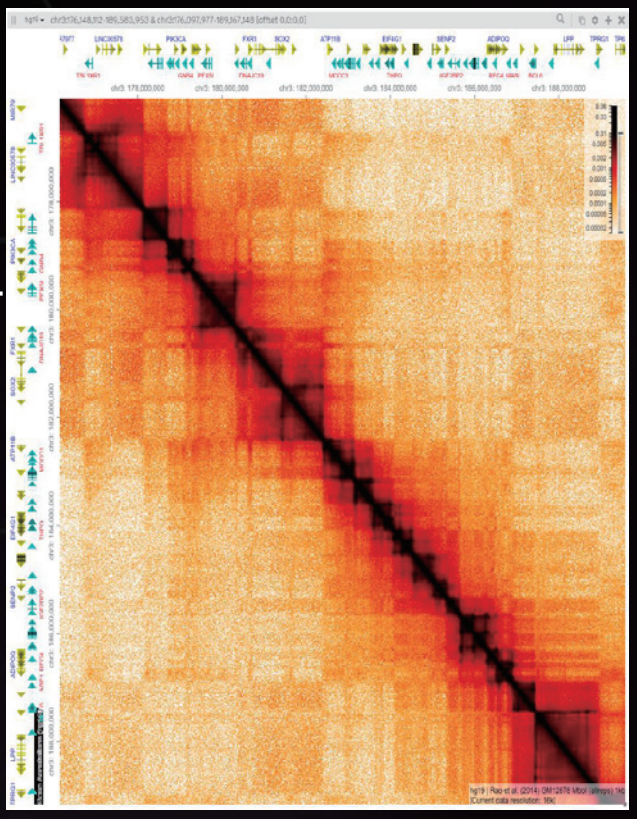
3D genome Organization
Chromatin Conformation

GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT
TTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAAGC
AGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGG
CGGAAAATGAGAAAATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCT
TTCGCCATCAACTAACGATTCTGTCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGC
TATGCTTGGCACGTTCGTCAAGGACTGGTTTAGATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATT
TTAAAAGAGCGTGGATTACTATCTGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCG

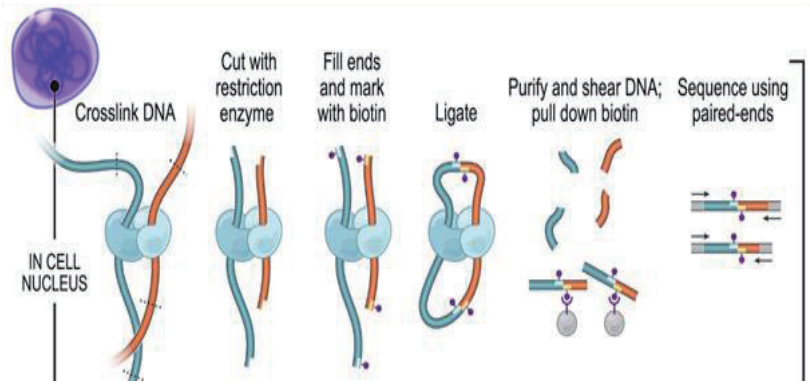


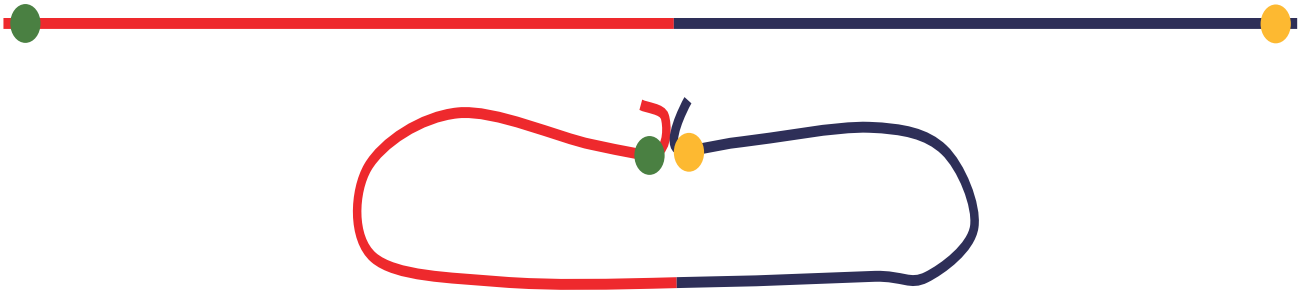


What is Hi-C?

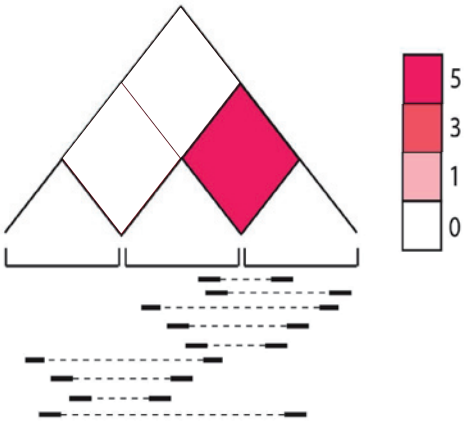


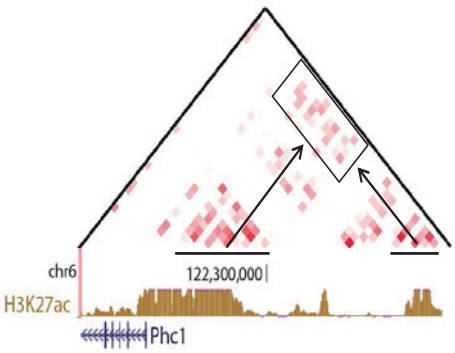
Hi-C experiment and basic analysis



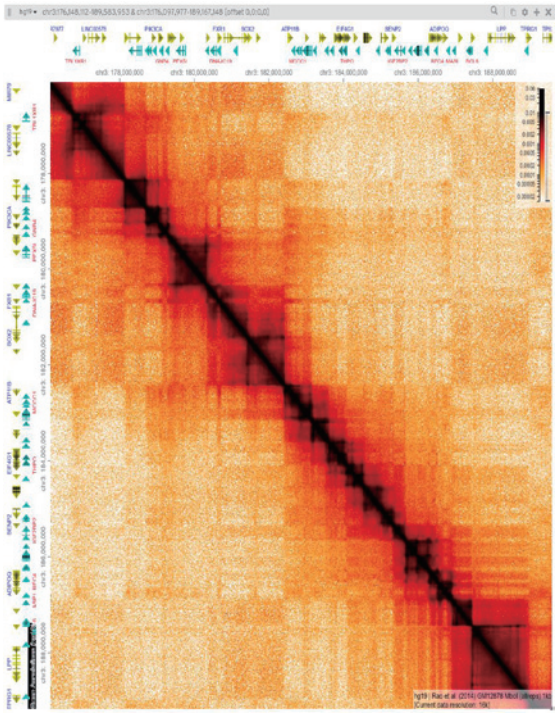


Visualizing Hi-C data

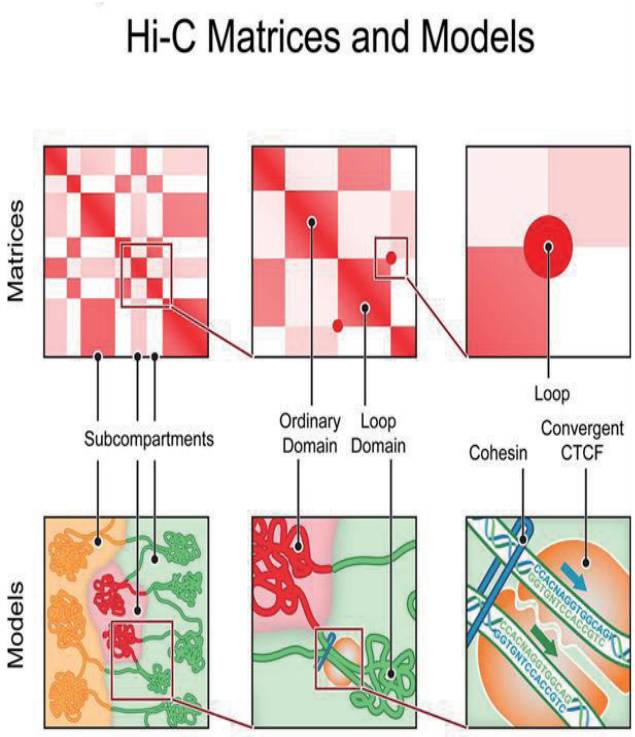




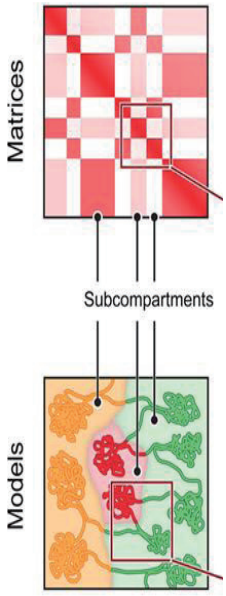
Hi-C experiment and basic analysis

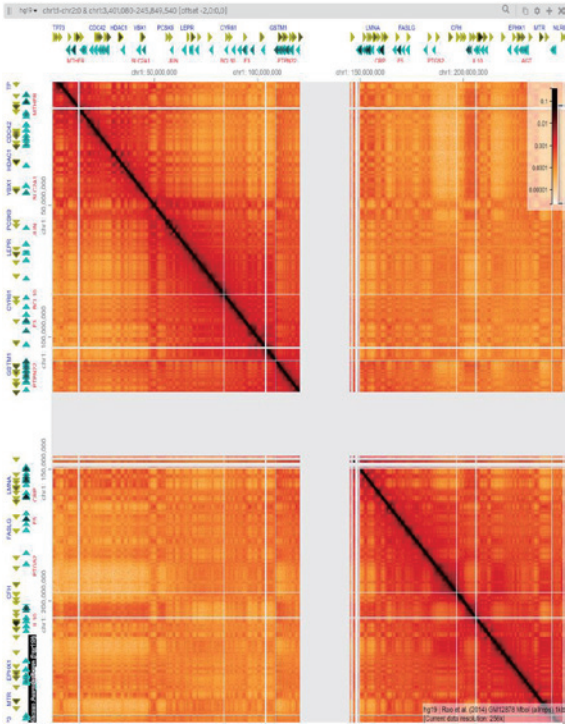


Hi-C experiment and basic analysis



Hi-C basic analysis-Compartment

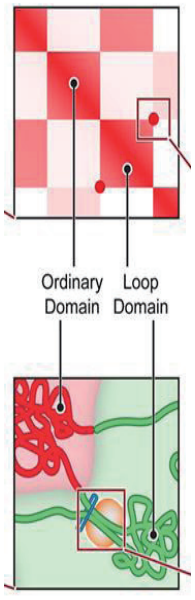


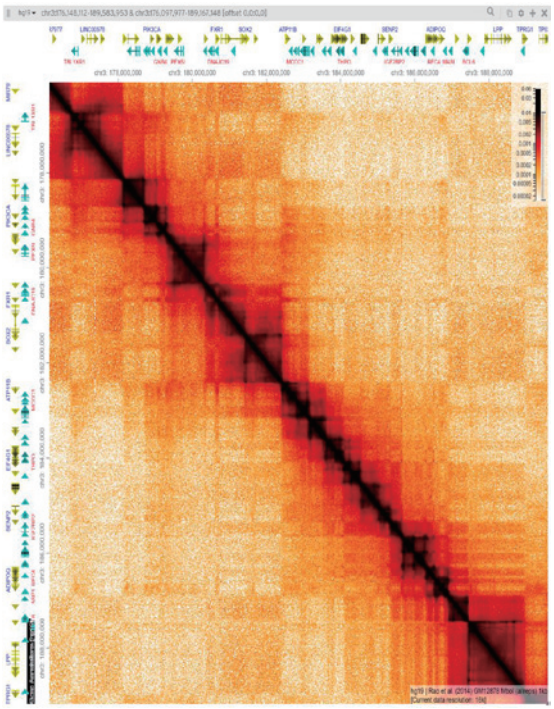


Compartment

- Long range interaction을 보았을 때 크게 2개의 pattern이 나타남. 이 두개의 패턴이 하나는 Euchromatin (Active region)을, 다른 하나는 Heterochromatin (Inactive region)을 나타냄.

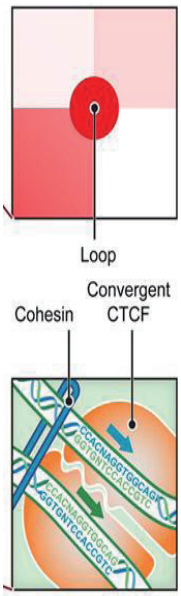
Hi-C basic analysis-TAD

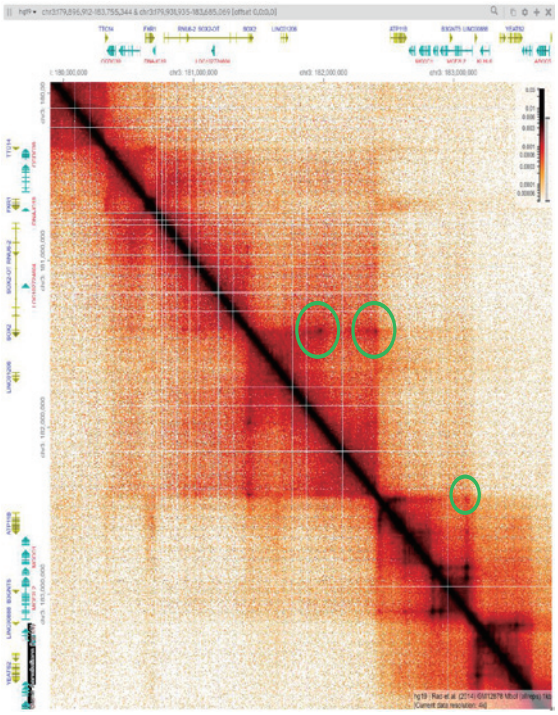




TAD (Topologically Associated Domain)
- Hi-C matrix를 zoom in 하면 short range에서 삼각형 구조가 발생함을 확인하였음. 또한 이러한 삼각형 구조들이 다른 삼각형들과는 isolate된 형태를 나타내고 있어 이를 topologically associated domain이라고 부름.

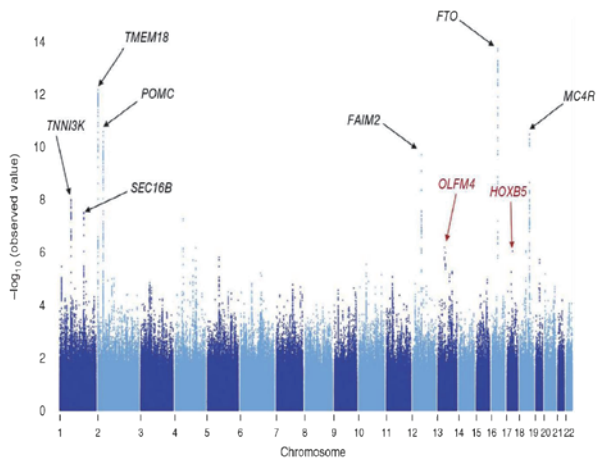
Hi-C basic analysis-Loop



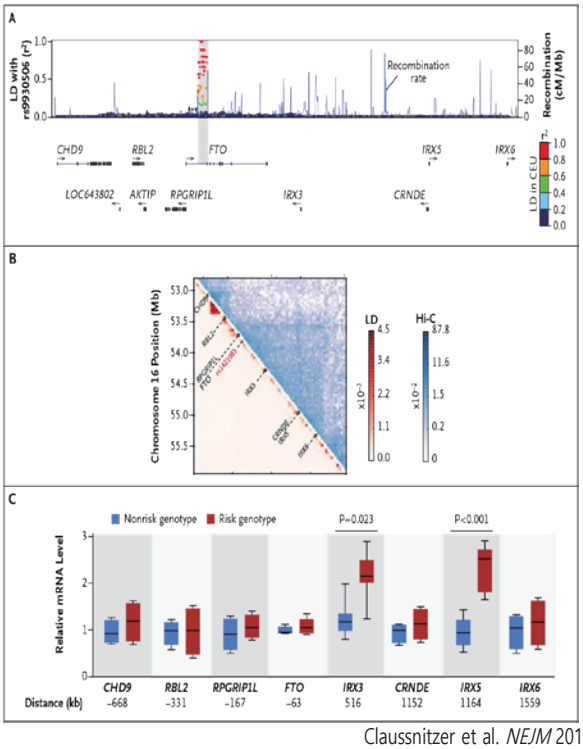


Loop
- 이러한 구조들 중 특히 하나의 점에서 강한 interaction을 나타내는 경우 이를 "loop이 형성되었다"고 함.

An example of using Hi-C to understand enhancer activity of non-coding variation

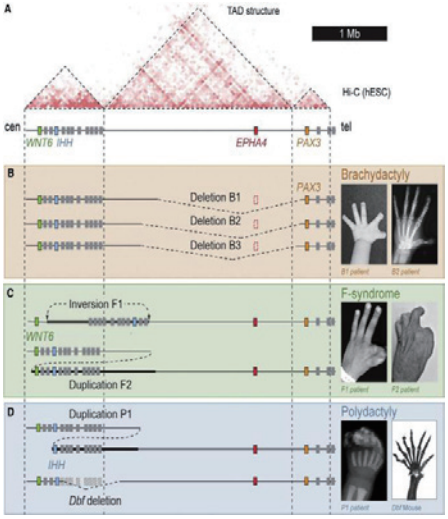


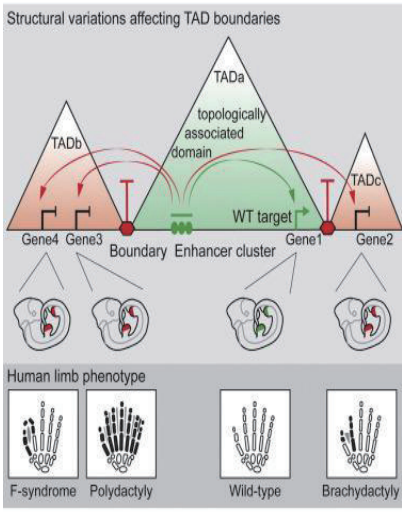
Bradfield et al., *Nature Genetics* 2012

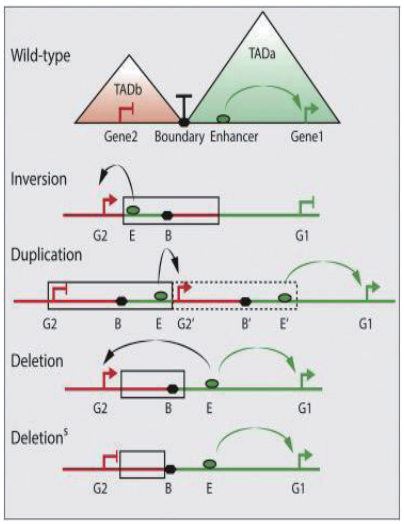


Claussnitzer et al. *NEJM* 2015

An example of using Hi-C to understand enhancer activity

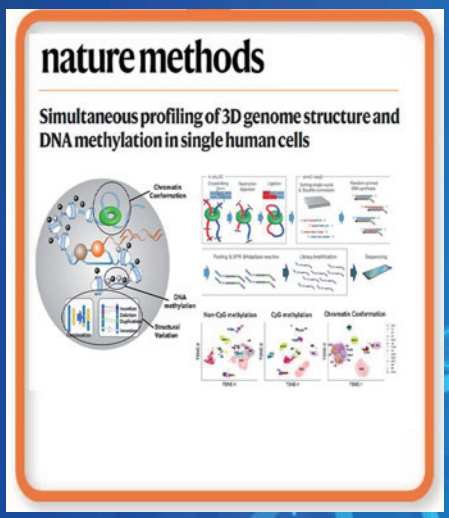


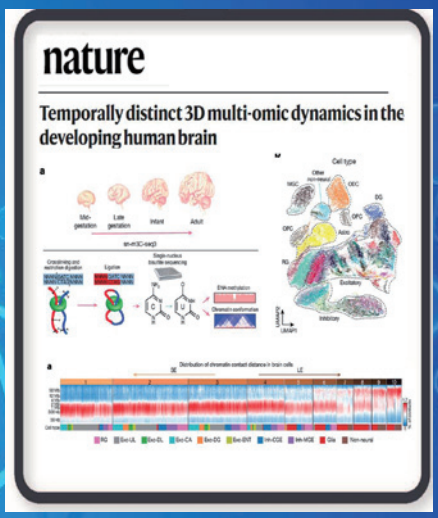




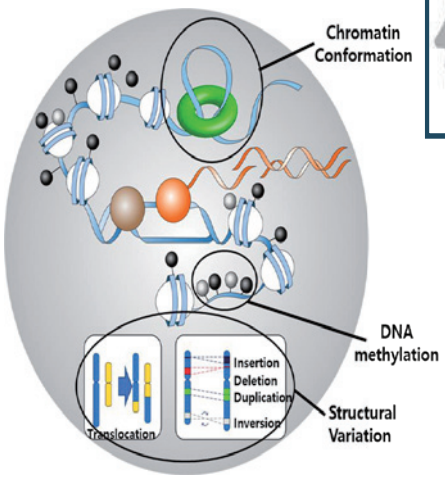
Darío G. Lupiáñez et al. *Cell* 2015

Development of Single nucleus methyl 3C and the Application

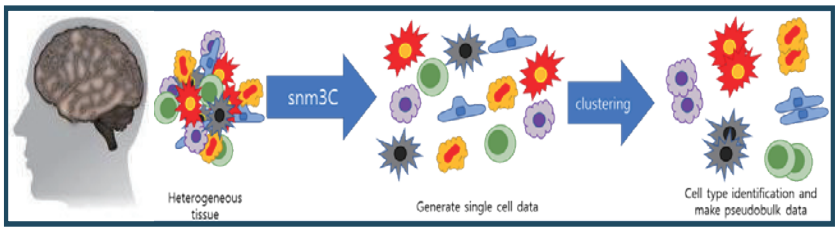


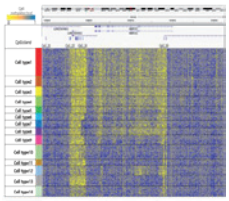


single nucleus methyl 3C (snm3C)

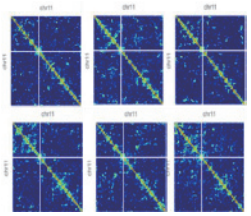


Single nucleus methyl-3C (sn-m3C)

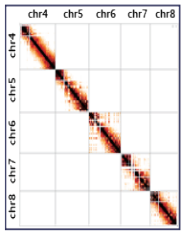




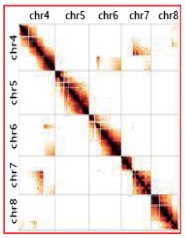
DNA methylation



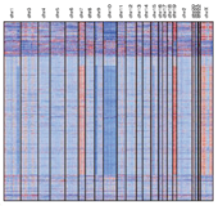
Chromatin organization



Hi-C matrix of a normal sample



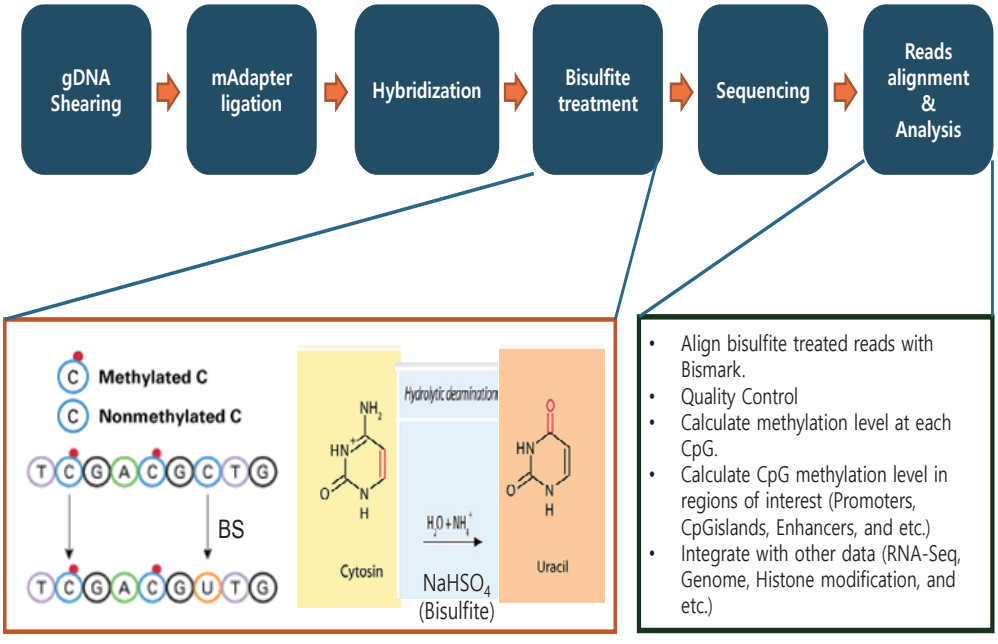
Hi-C matrix of a cancer sample with structural variants



Copy number variation

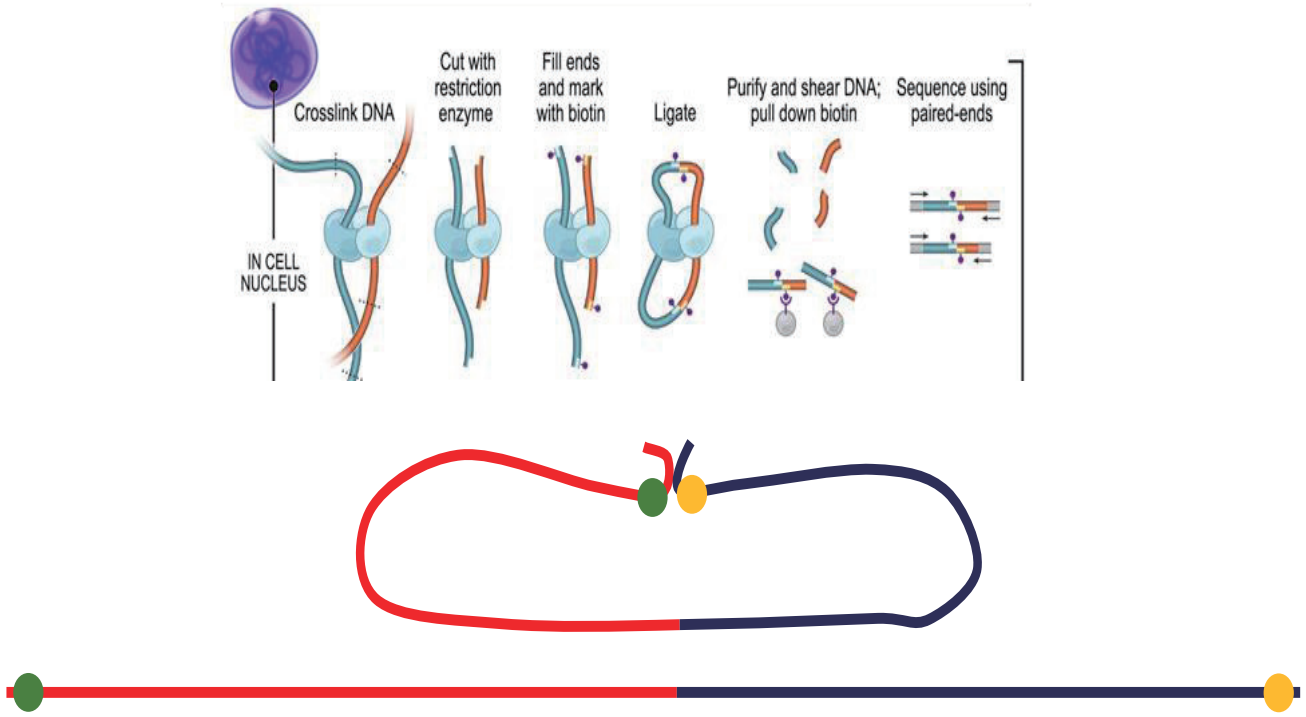
Lee et al., *Nature Methods* 2019

Whole Genome Bisulfite Sequencing (WGBS)

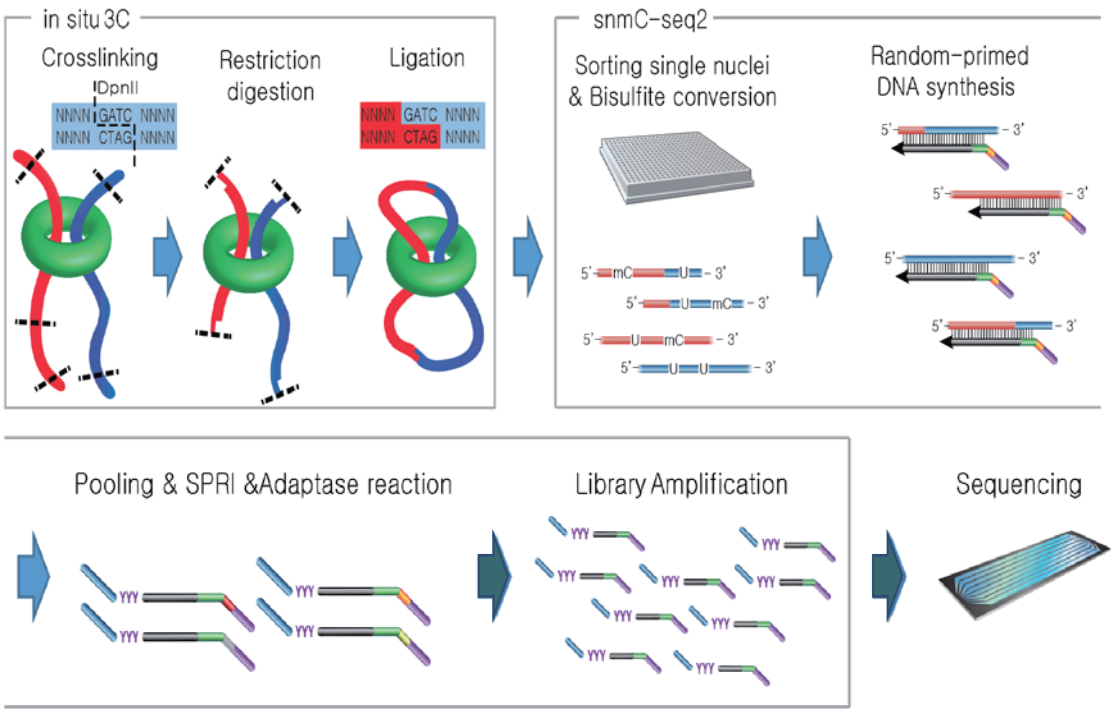


- Align bisulfite treated reads with Bismark.
- Quality Control
- Calculate methylation level at each CpG.
- Calculate CpG methylation level in regions of interest (Promoters, CpGislands, Enhancers, and etc.)
- Integrate with other data (RNA-Seq, Genome, Histone modification, and etc.)

3C (chromatin conformation capture)



Single nucleus methyl-3C Experiment

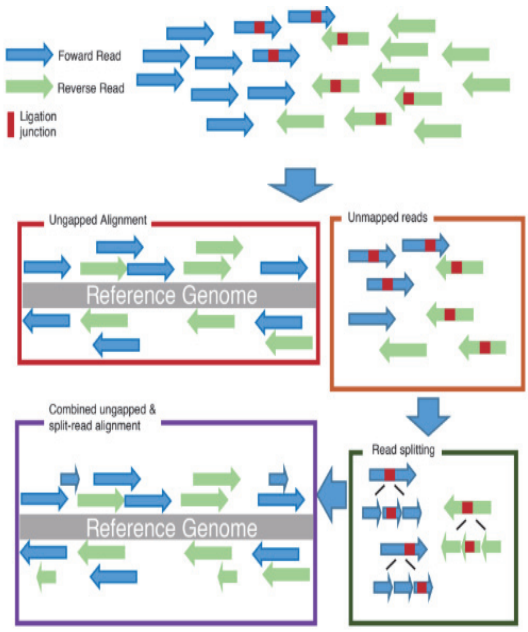


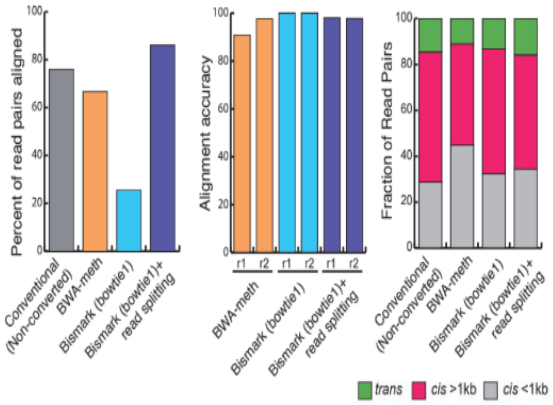
Lee et al., *Nature Methods*, (2019)

TAURUS-MH for methyl-3C data alignment

(Two-step Alignment with Unmapped Reads Using read Splitting for Methyl-HiC)

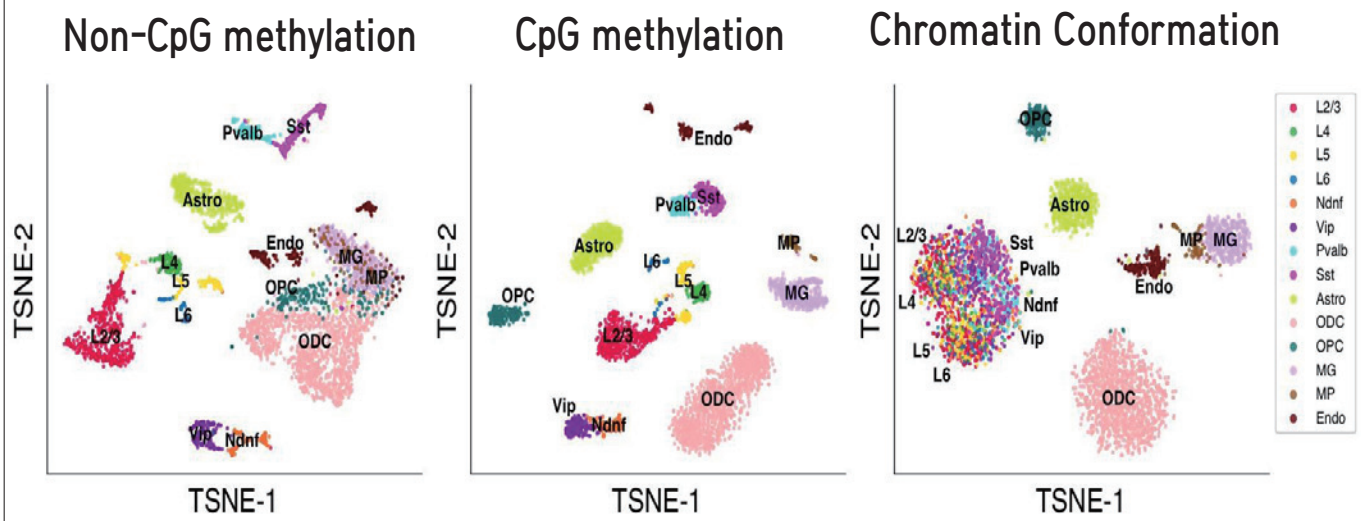
https://github.com/dixonlab/Taurus-MH





Lee et al., *Nature Methods,* (2019)

Human Brain Tissue snm3C



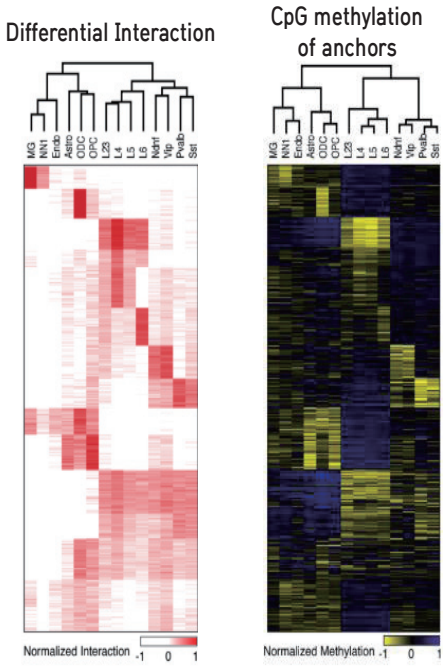
Total	L2/3	L4	L5	L6	Ndnf	Vip	Pvalb	Sst	Astro	ODC	OPC	MG	MP	Endo
4238	551	131	180	86	144	171	134	217	449	1245	203	422	100	205

Clustering is possible using each data, but cell type classification is possible only through methylation.

Lee et al. Nature Methods 2019

3D genome organization is strongly associated with DNA methylation

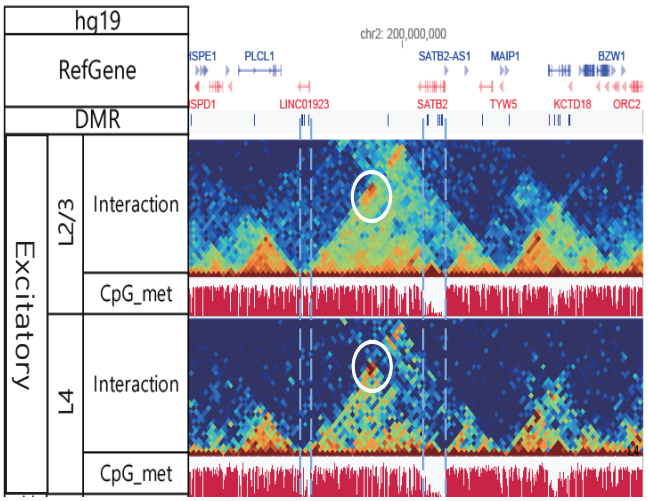
	Count	Overlap
Differential methylation	89,356	63,570 (71.14%)
Differential Interaction	36,559	31,669 (86.62%)

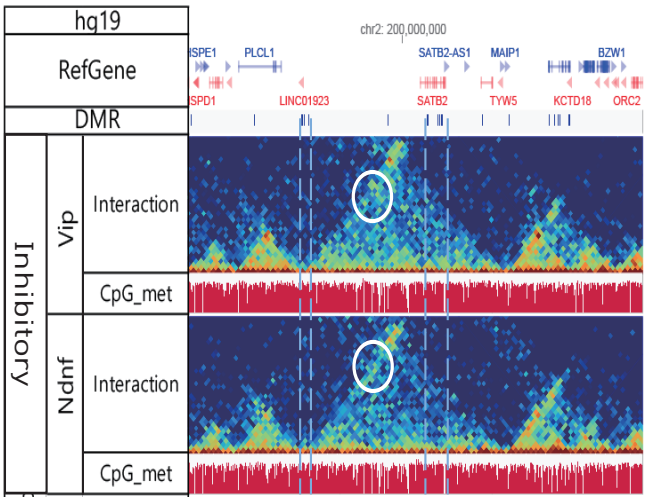


Lee et al., *Nature Methods,* (2019)

Cell type specific interaction and methylation

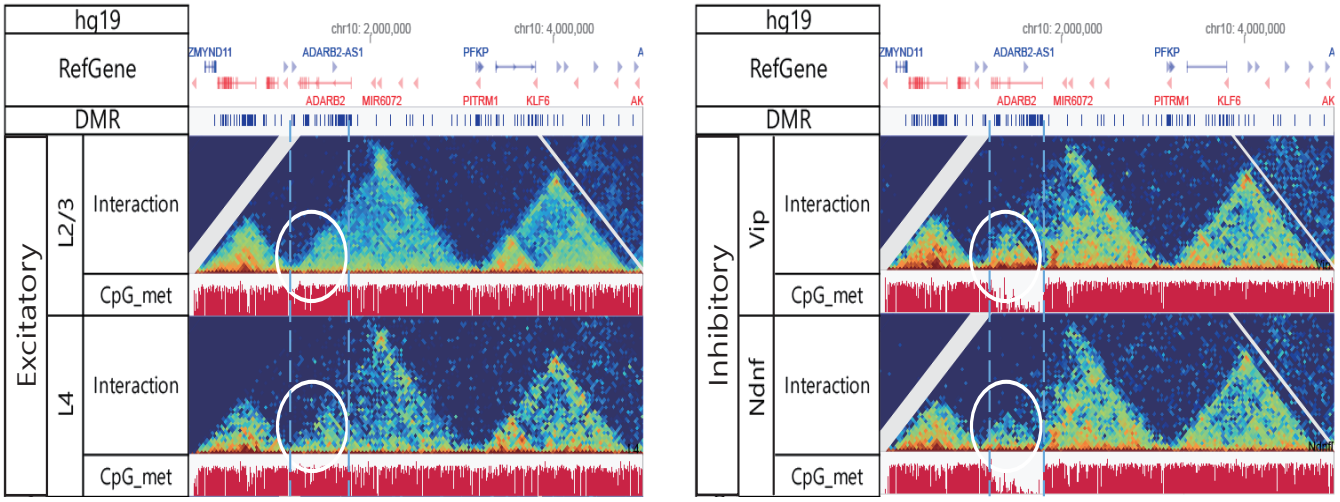
SATB2



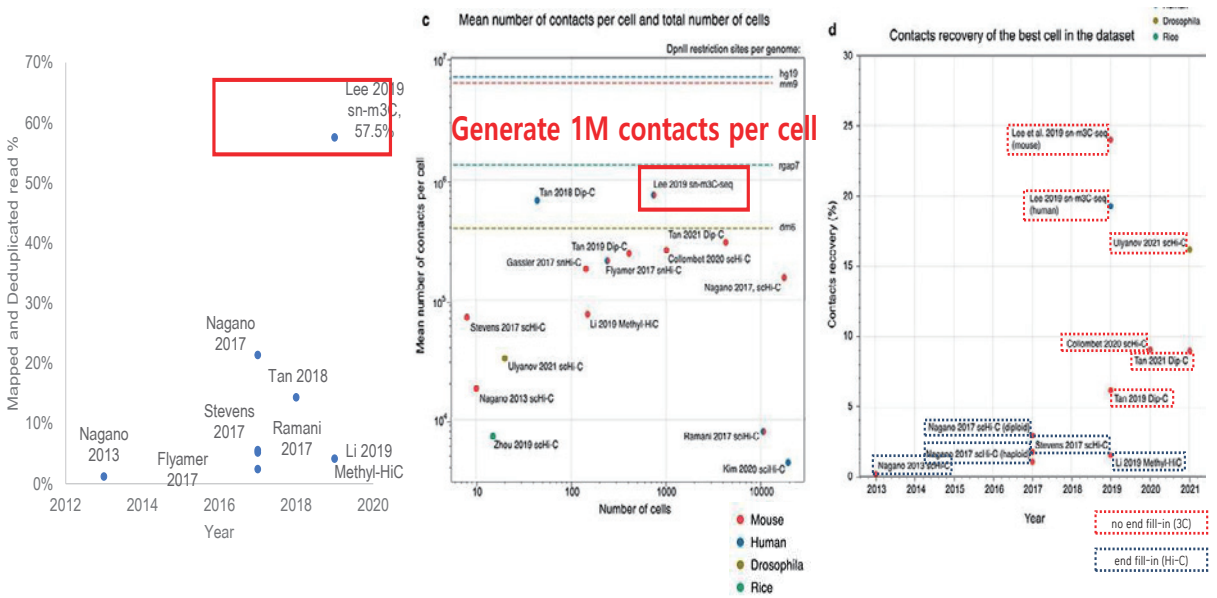


Cell type specific interaction and methylation

Adarb2



Comparison of single-cell Hi-C/3C methods



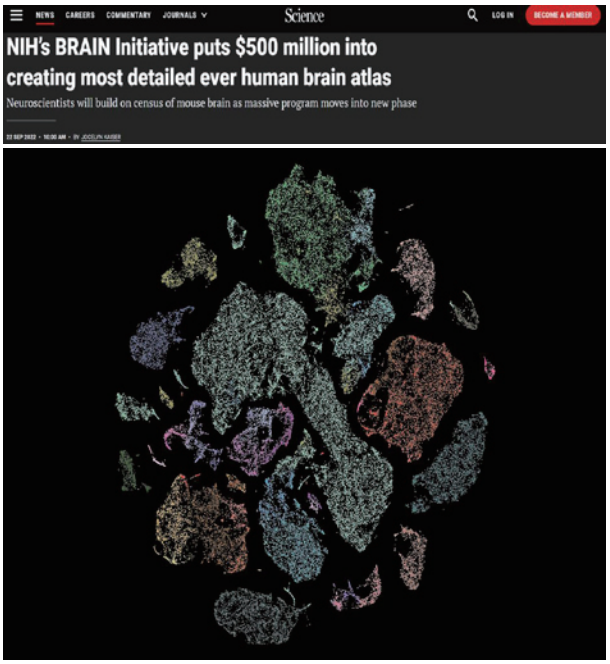
Galitsyna et al., *Briefings in Bioinformatics,* (2021)

BRAIN INITIATIVE

- U.S. government-led project announced on April 2, 2013
- Evaluated as a project comparable to the Human Genome Project (total amount of $3 billion)
- $2.4 billion invested in Phase 1 since 2014
- A total investment of $5 billion is planned by 2026.

- The main goal of Phase 1 is to develop new techniques.
- Phase 2 involves studying the brain using developed techniques.

BRAIN INITIATIVE uses snm3C



September 22, 2022

Salk Institute to lead $126 million effort to map the aging human brain

Largest grant in Salk history establishes new Center for Multiomic Human Brain Cell Atlas to detail the many individual cells that make up the human brain—their molecular features, where they are found and how they change with age

LA JOLLA—With a five-year, $126 million grant from the National Institutes of Health (NIH), a team led by Salk Institute scientists has launched a new Center for Multiomic Human Brain Cell Atlas. Part of the NIH's *Brain Research Through Advancing Innovative Neurotechnologies®* (BRAIN) Initiative, the project aims to describe the cells that make up the human brain in unprecedented molecular detail, classify brain cells into more precise subtypes, and pinpoint the location of each cell in the brain. What's more, the team will track how these features change from early to late life.

The goal is to better understand how neurotypical human brains work and age. The project will also establish a baseline against which scientists will be able to compare brains with neurological or psychiatric conditions such as Alzheimer's disease, autism, depression and traumatic brain injury.

From left: Margarita Behrens and Joseph Ecker
Click here for a high-resolution image.
Credit: Salk Institute

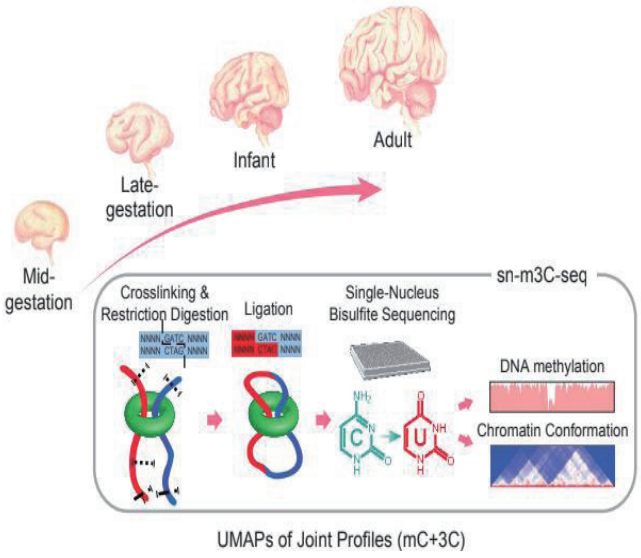
"The brain map we develop could help point disease researchers in the right direction—for example, we could say 'That's the region of the genome, in that specific subset of neurons, in that part of the brain, where a molecular event goes awry to cause that disease,'" says center leader Professor Joseph Ecker, director of the Genomic Analysis Laboratory at Salk and Howard Hughes Medical Institute investigator. "And ultimately this information might help us design gene therapies that target only the cell populations where the treatment is needed—delivering the right genes to the right place at the right time."

In addition to Ecker, the Center for Multiomic Human Brain Cell Atlas includes Margarita Behrens, research professor at Salk, Bing Ren at UC San Diego, Xiangmin Xu at UC Irvine, and Ting Wang at Washington University in St. Louis.

Salk will be awarded approximately $77 million of the center's funding, making it the largest single grant the Institute has received in its 62-year history.

Brain Initiative Cell Atlas Network (BICAN):
Characterization of human brain cell types in adult and developing brains

snm3C on Developing human brain



PFC

Age	Source	Cell Number
GW18	UCSF	1,199
GW20	UCSF	3,771
GW20	UCSF	2,942
GW23	UCSF	1,496
GW35	NBB	1,873
GW39	UCSF	1,904
4 mo	UCSF	2,228
7mo	UCSF	2,330
21 yr	NBB	7,377
29 yr	NBB	2,379
29 yr	NBB	917
31 yr	NBB	333
37 yr	NBB	942

Total: 29,691

HPC

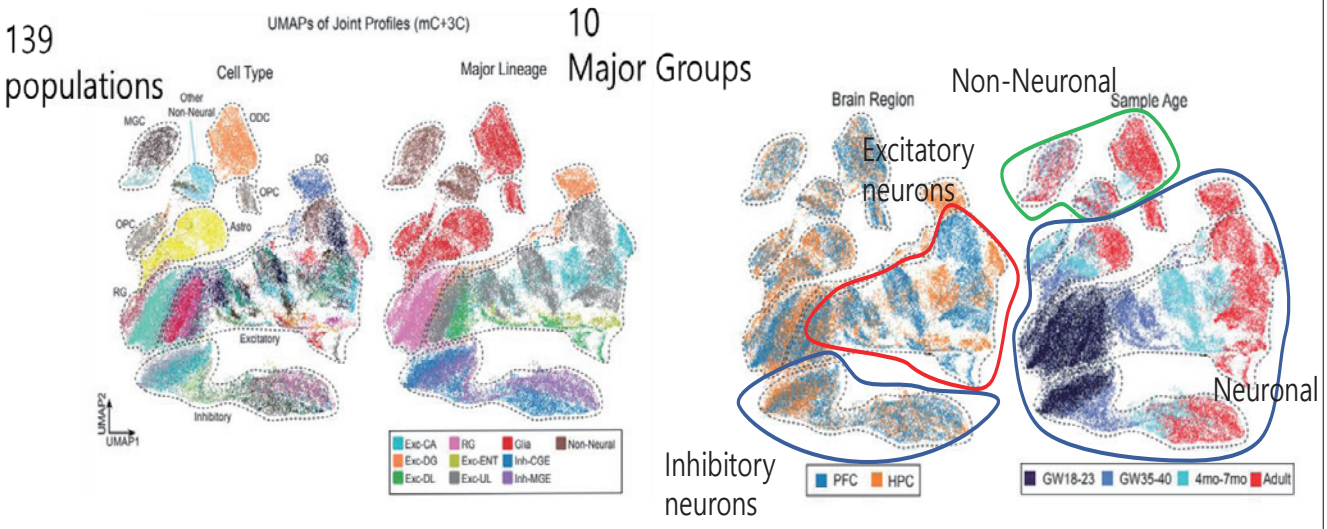
Age	Source	Cell Number
GW18	UCSF	4,432
GW20	UCSF	2,645
GW23	UCSF	1,359
GW35	NBB	2,627
GW39	UCSF	1,168
4 mo	UCSF	1,634
7mo	UCSF	3,353
29 yr	NBB	2,339
55 yr	NBB	3,815

Total: 23,372

Performed snm3C on Brain tissues (PFC and HPC) in different developmental stages.

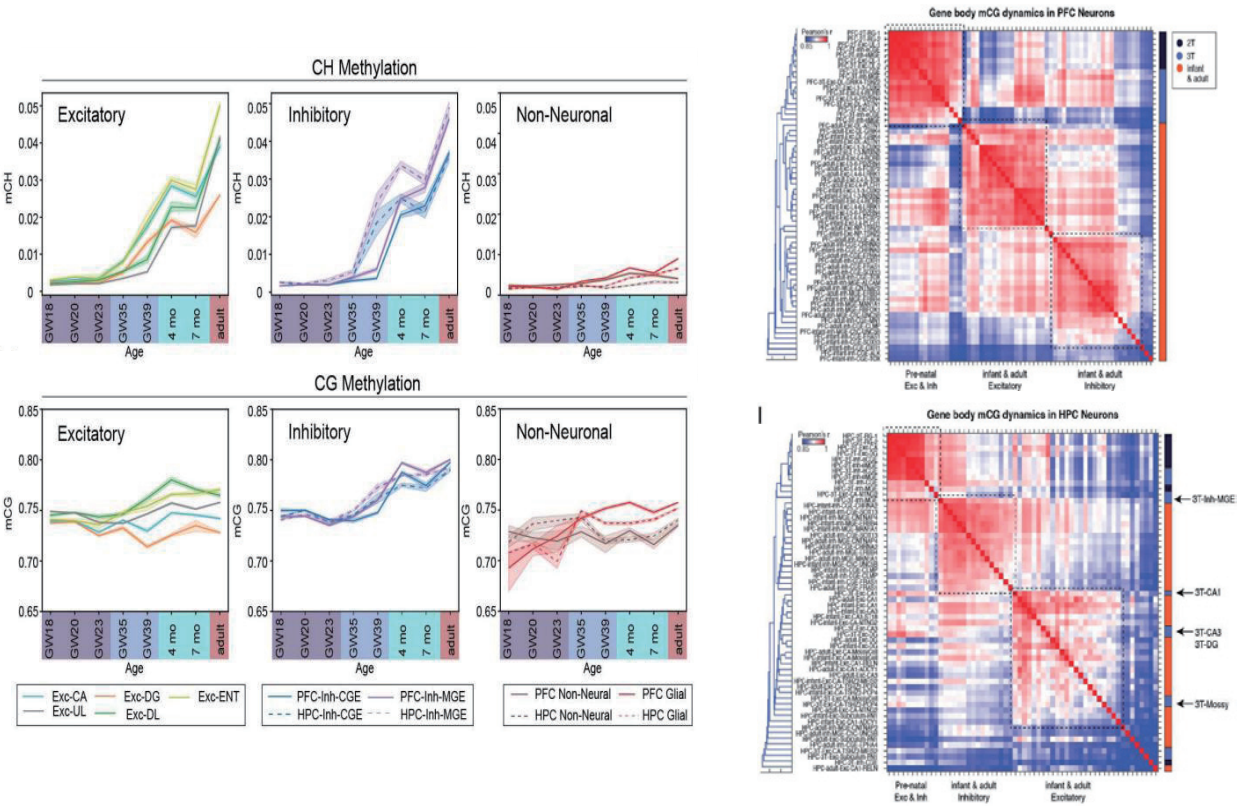
M. Heffel et al. Nature. 2024

Differences according to brain region and age



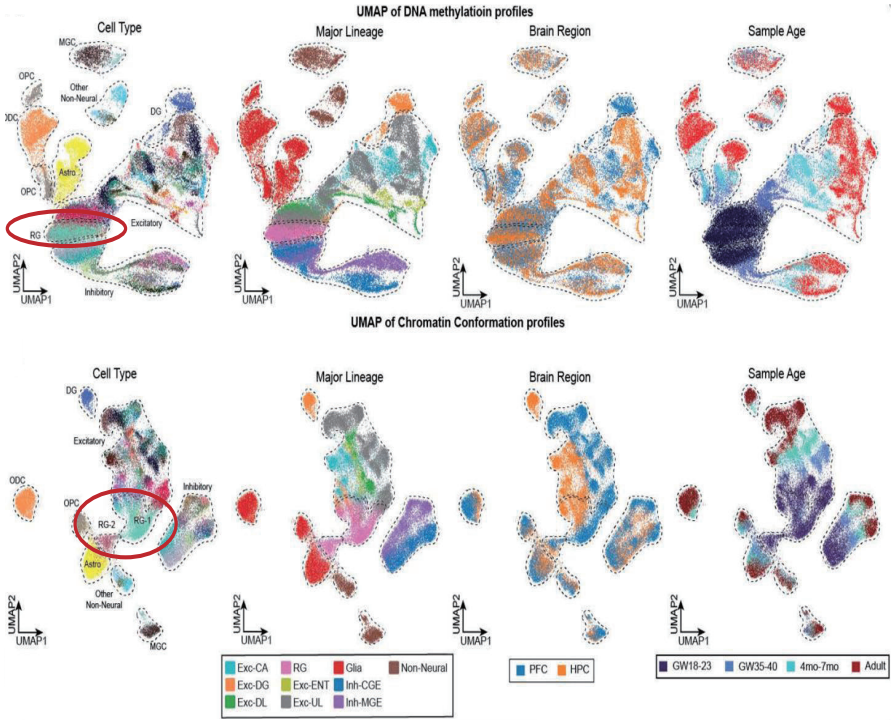
M. Heffel et al. Nature. 2024

Methylation remodeling in HPC precedes that in PFC



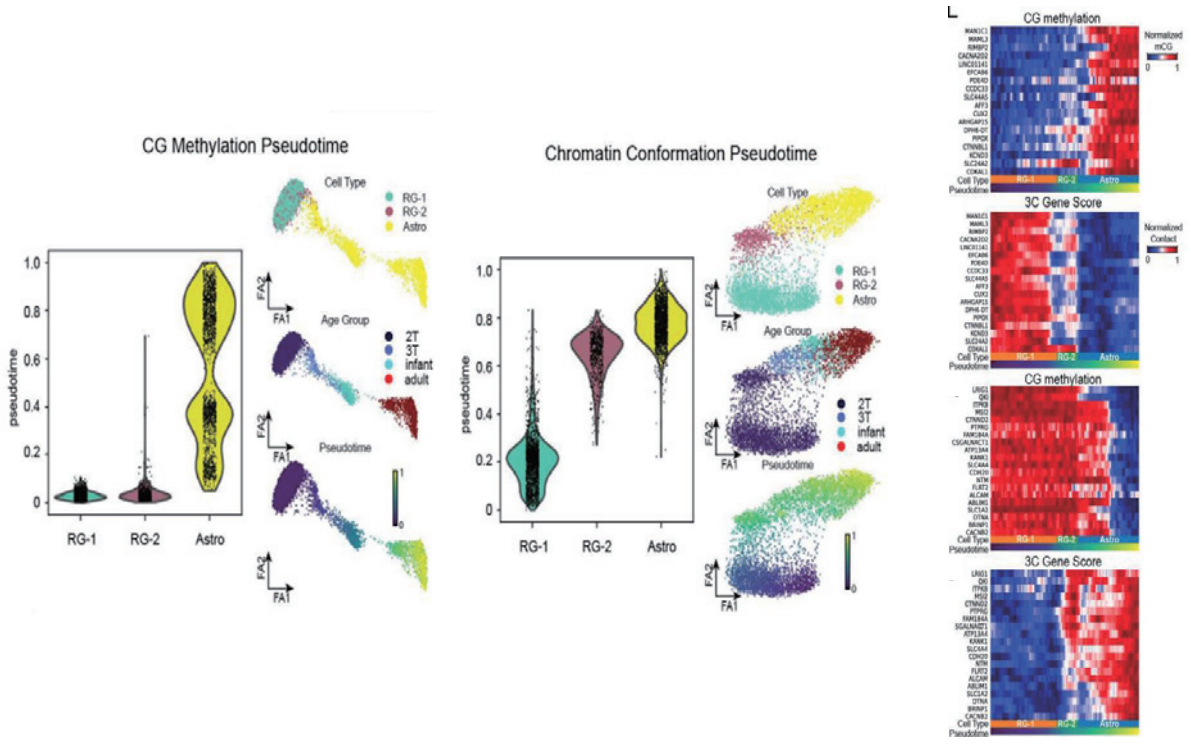
M. Heffel et al. Nature. 2024

Chromatin conformation dynamics precede the remodeling of DNA methylation during the differentiation



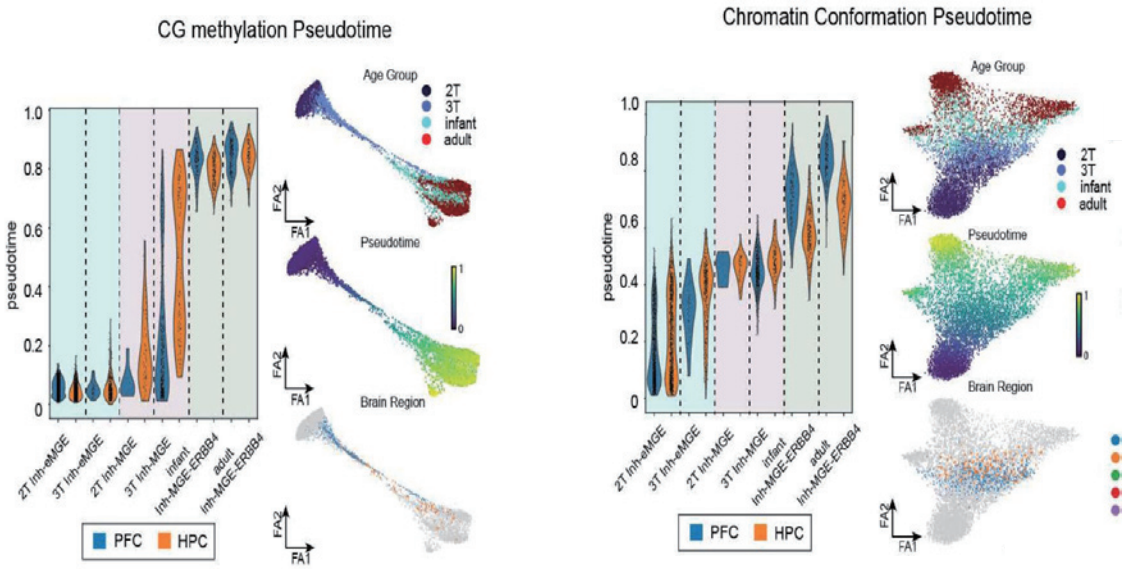
M. Heffel et al. Nature. 2024

Chromatin conformation dynamics precede the remodeling of DNA methylation during the differentiation



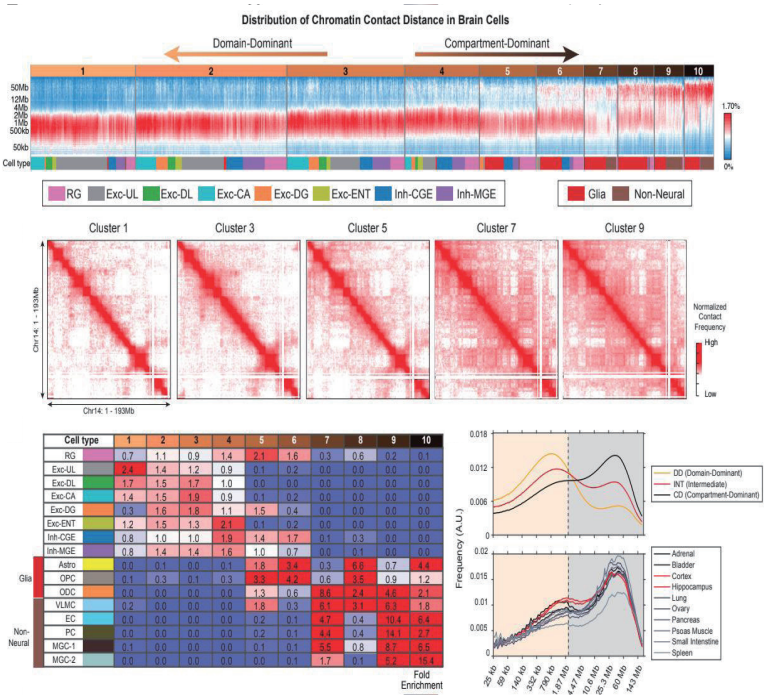
M. Heffel et al. Nature. 2024

Chromatin conformation dynamics precede the remodeling of DNA methylation during the differentiation



M. Heffel et al. Nature. 2024

a new layer of unique epigenomic regulation in neurons

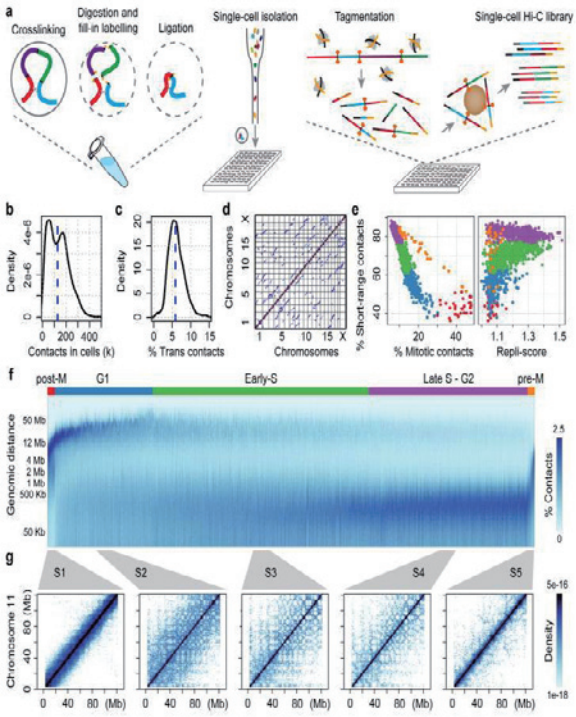


As in the research of other groups in 2017, we analyzed global interaction pattern. In fact, the analysis was started with the expectation that the cell cycle could be observed.

However, our observation indicates that this global interaction change does not appear to indicate the changes in the cell cycle (in fact, a lot of "S" phase is observed in adults) but can be used as a strong marker to distinguish neuronal and non-neuronal cells.

M. Heffel et al. Nature. 2024

Previous Single-cell Hi-C study (mono-omic) - Why do we have to do multi-modal omics?



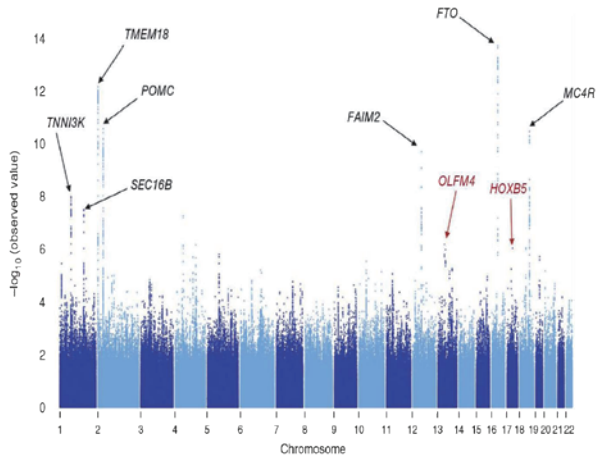
Nagano et al. Nature 2017

There are papers that performed Hi-C in single cell between 2013 and 2018.

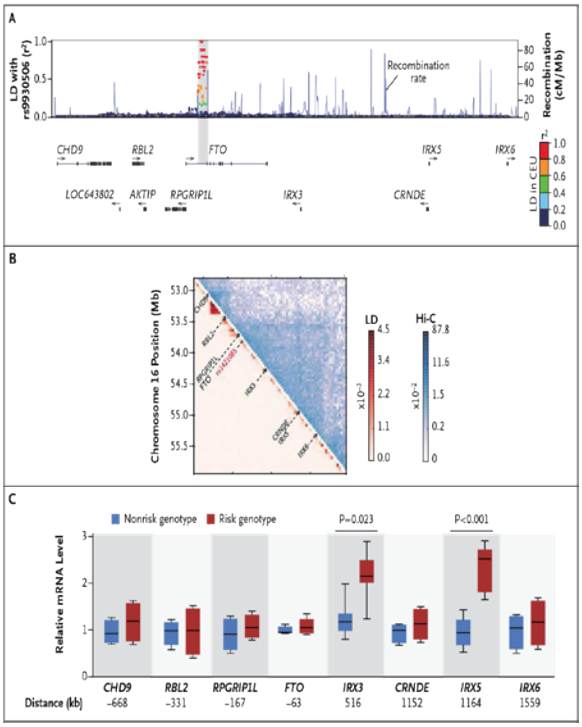
The biggest weakness of these papers was that because only chromatin conformation was profiled, the cell-types could not be divided, and even if the cell population was divided, the type could not be specified because there was no reference.

In the previous studies, the researchers found that the cells could be segregated into different groups using the interaction decay along the distance and claimed that this global trend reflects "cell cycle".

An example of using Hi-C to understand enhancer activity of non-coding variation

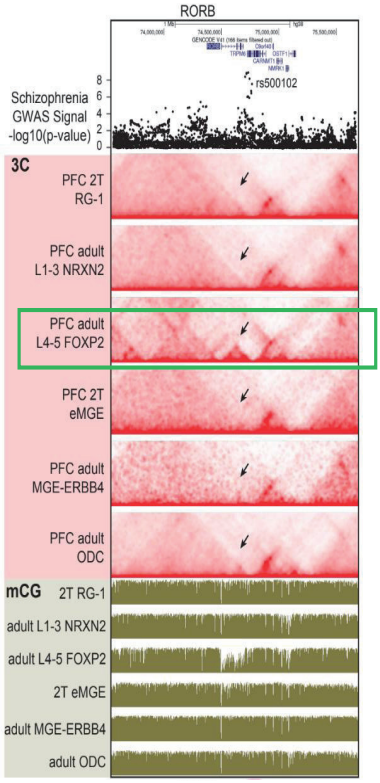


Bradfield et al., *Nature Genetics* 2012



Claussnitzer et al. *NEJM* 2015

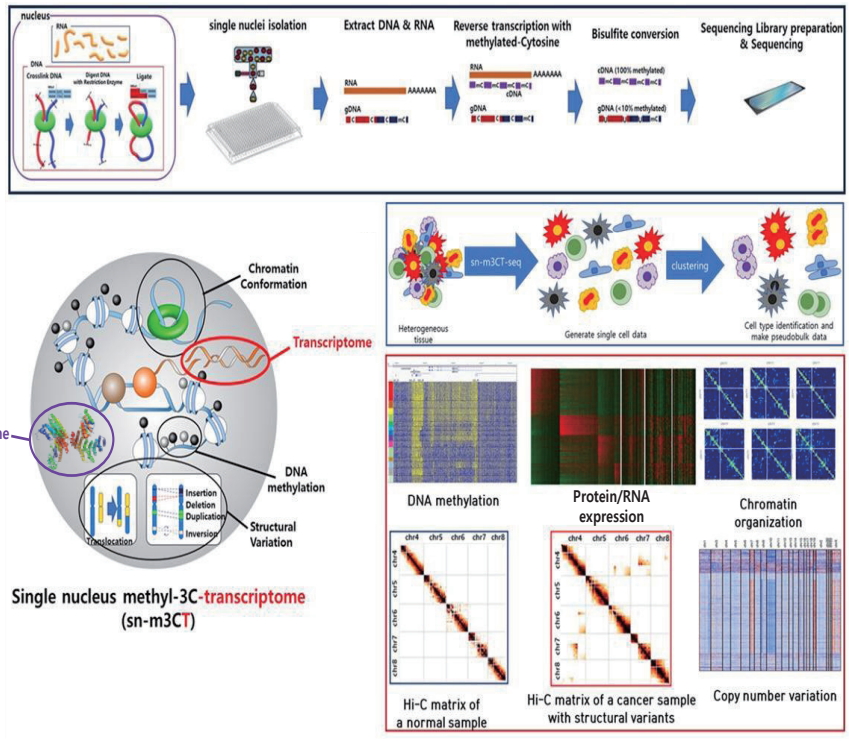
Identification of Cell type specific disease associated enhancer activity



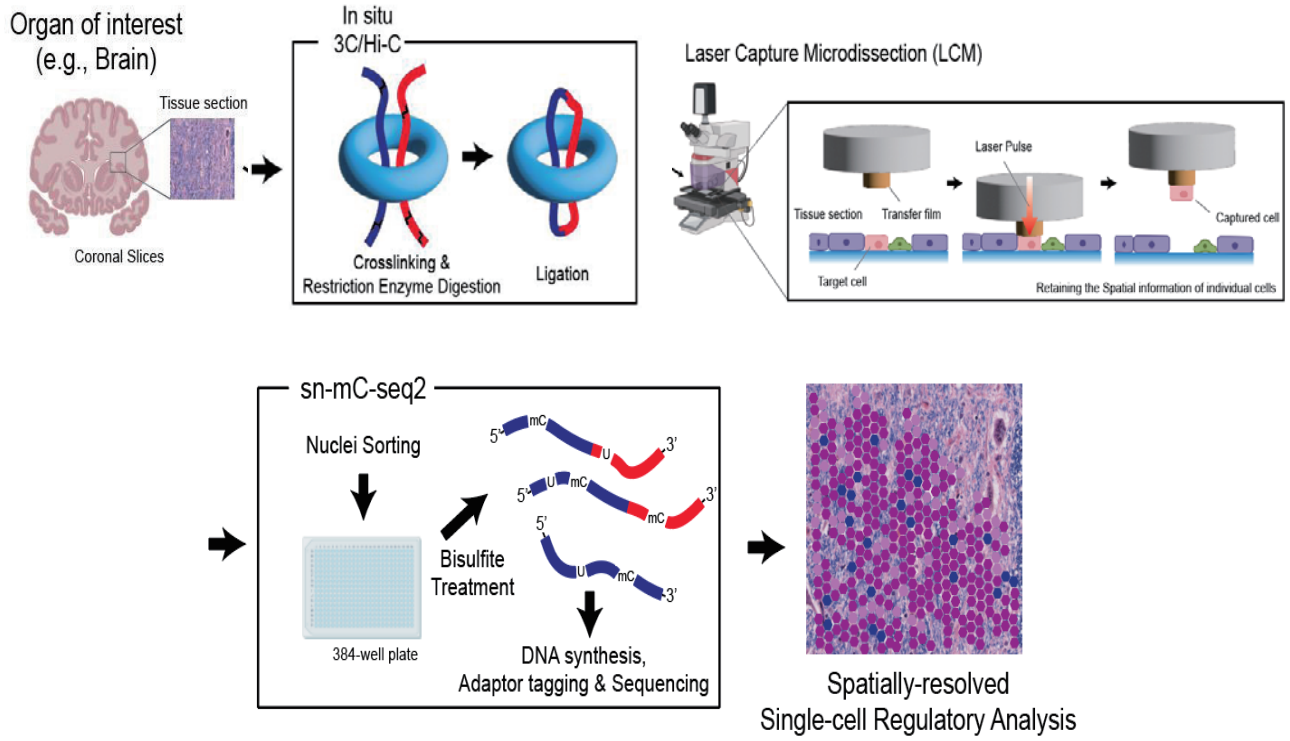
Schizophrenia high risk loci (TRPM6) obtained through GWAS shows a strong interaction only in a specific cell type (L4-5 Foxp2) with RORB.

M. Heffel et al. Nature. 2024

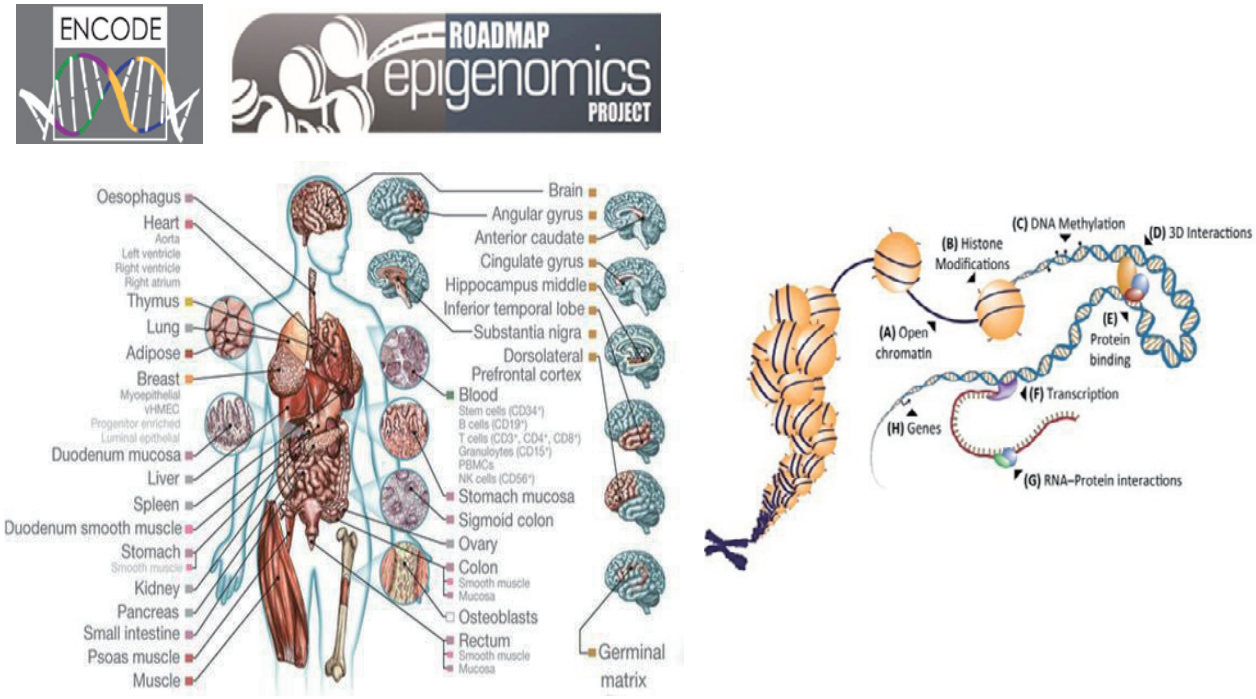
Ongoing and Future Research1: Expand single nucleus methyl 3C to other modalities



Ongoing and Future Research2: spatially resolved snm3C development



Ongoing and Future Research3: Human Cell Epigenome Atlas



Galaxy PROJECT

Introduction of Galaxy and NGS data

Galaxy PROJECT

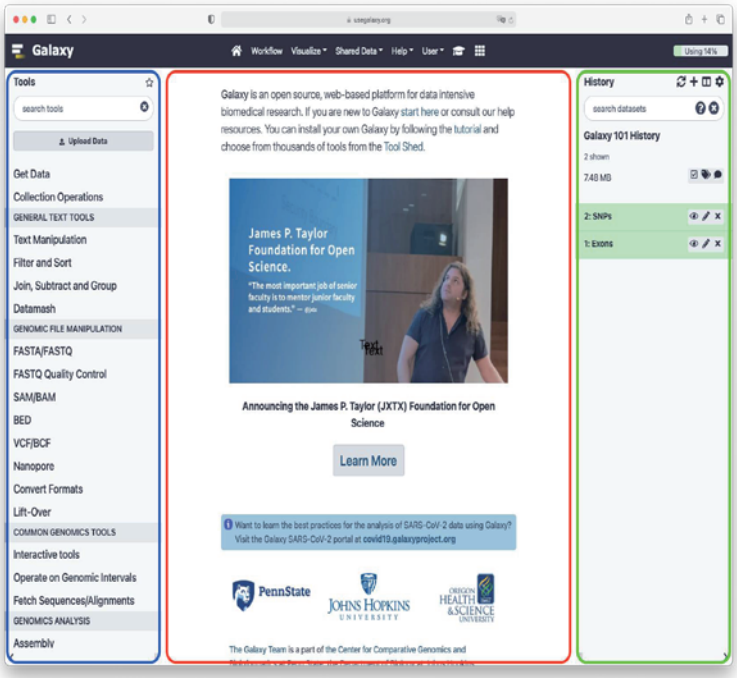
- **Data Analysis** platform
- Web-based
- **Easy** to use
- **Free** and Open Source

- Many tools (~9400Galaxy Tool Shed)
- Popular (>11.900 publications)
- Extensive tutorials available

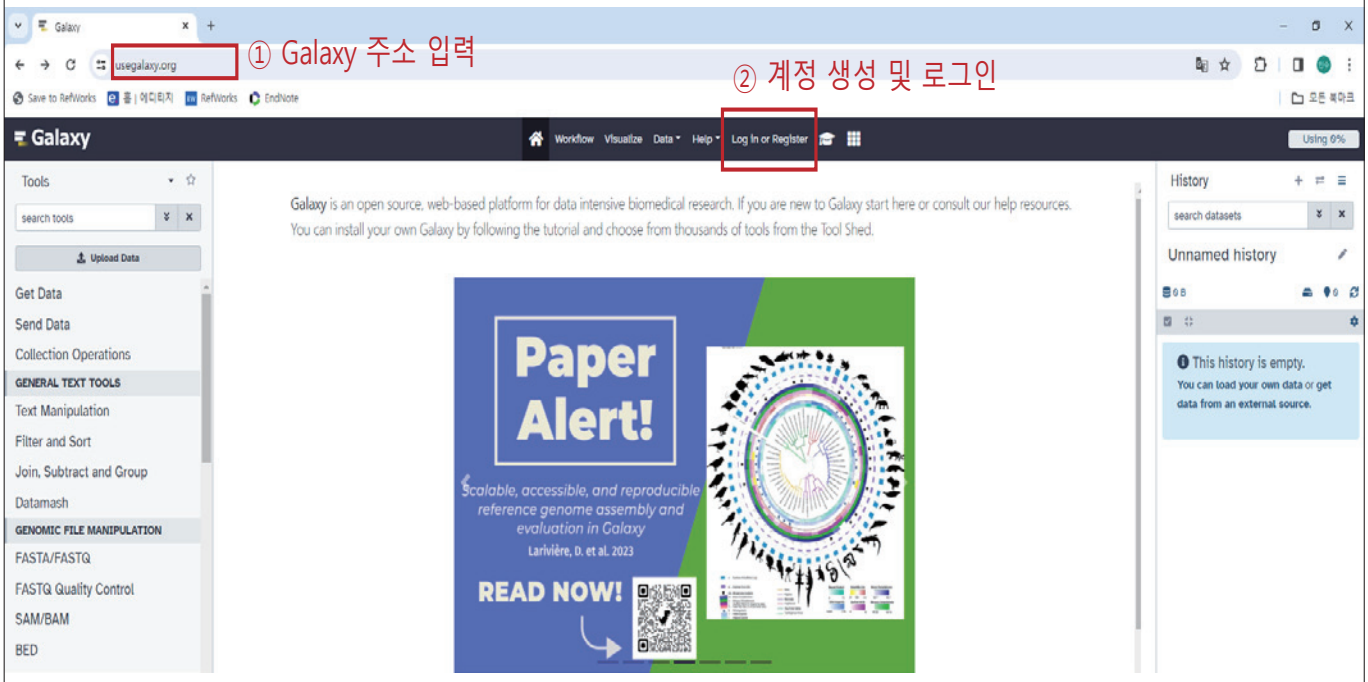
UseGalaxy.*: Galaxy Europe (UseGalaxy.eu), Galaxy Main (UseGalaxy.org), Galaxy France (UseGalaxy.fr), Galaxy Australia (UseGalaxy.org.au)

The Galaxy Interface

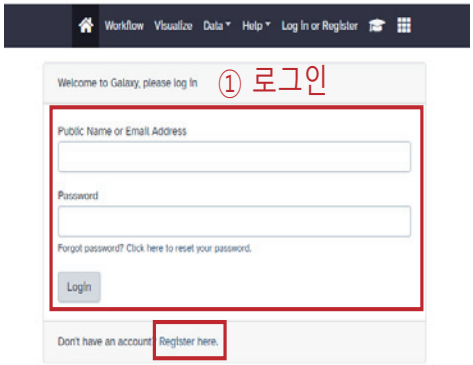
- Three main panels
  - **Left:** Available Tools
  - **Middle:** View your data and run tools
  - **Right:** Full record of your analysis **history**



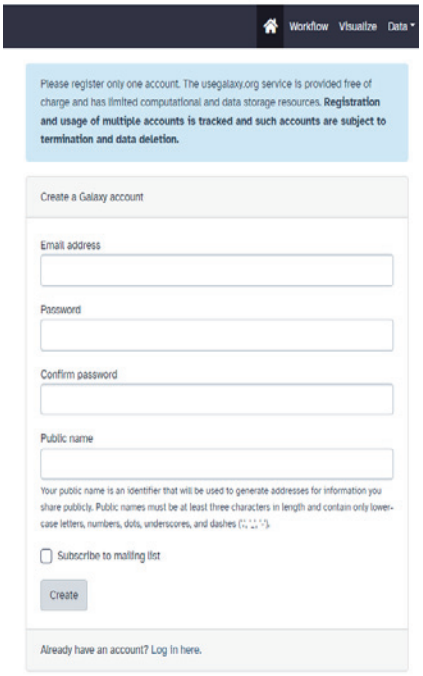
Start Galaxy



Start Galaxy



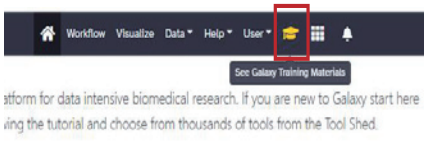
② 또는 계정 생성



학교 이메일 또는
사용중인 이메일을 통해
계정 생성 가능

Start Galaxy

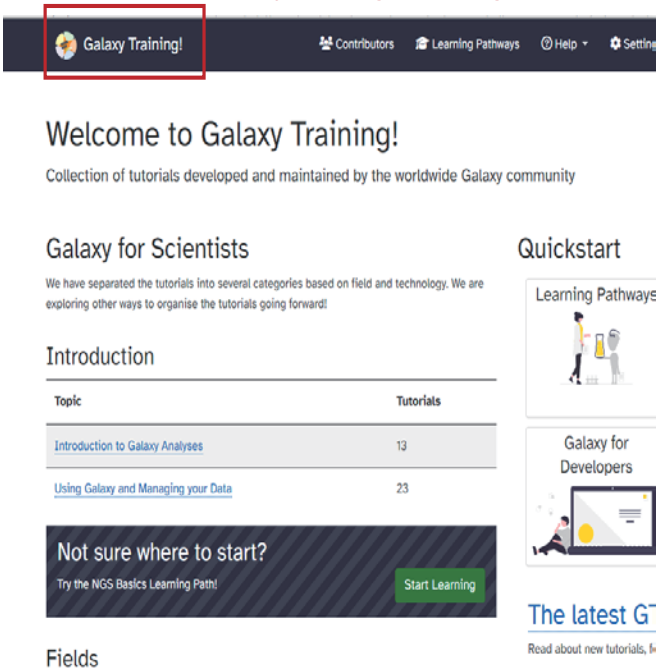
① Galaxy Training



Donate to the James P. Taylor Foundation for Open Science

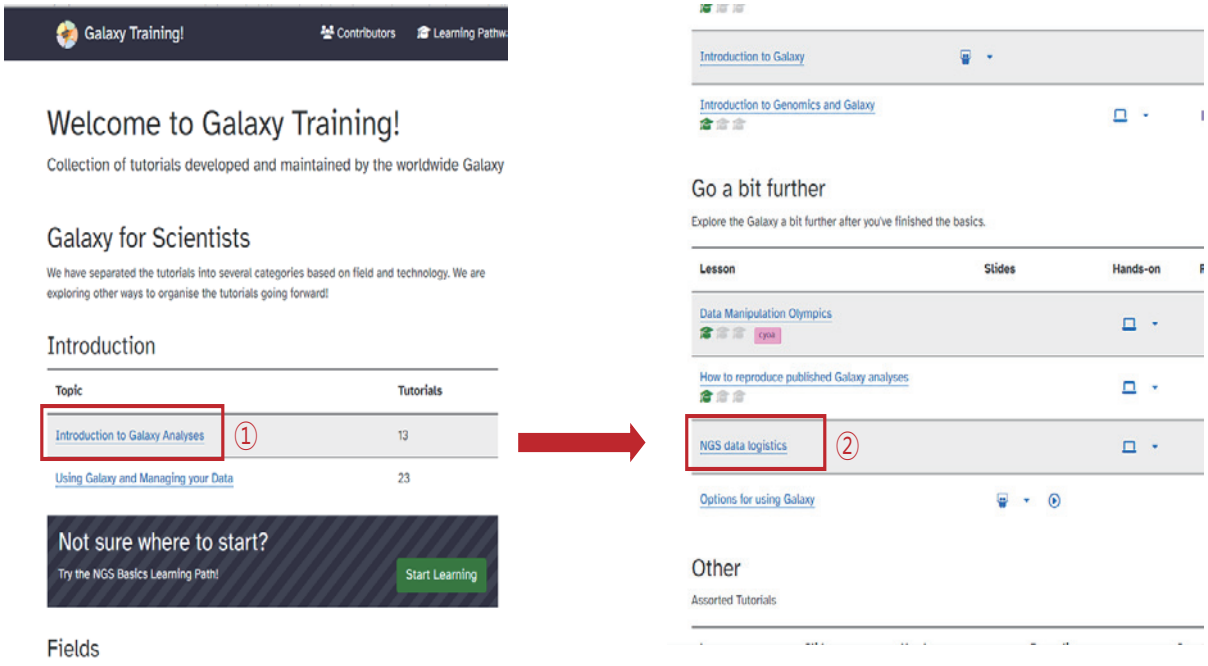
Learn More

Galaxy Training main page



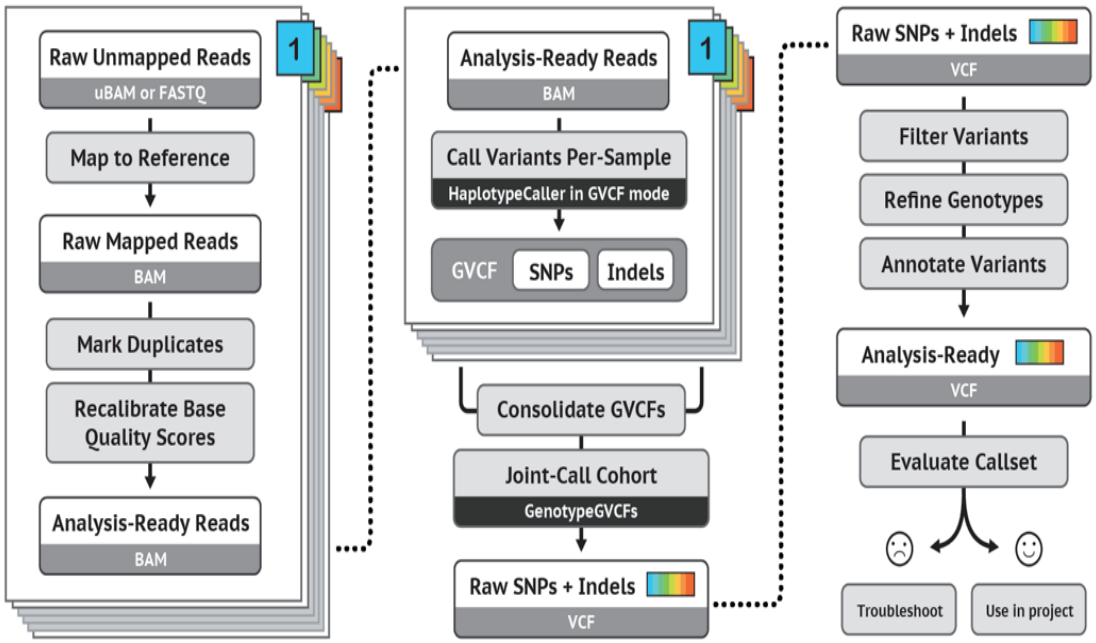
다양한 주제의 튜토리얼을 제공

Start Galaxy



Introduction to Galaxy Analyses -> NGS data logistics

Genome Sequencing data analysis pipeline



SNP : single nucleotide polymorphism
Indel : Insertion + deletion

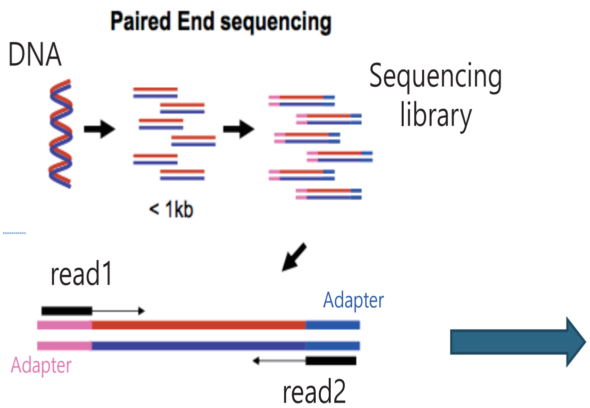
NGS data logistics

FASTA file

```
>chr21
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

- 보통 Reference genome을 넣는 파일
- sequence를 표현하는 가장 기본적인 포맷.
- DNA sequence 뿐만 아니라 Protein sequence도 저장한다.
- '>'로 주석을 표현하고 한줄당 50개의 sequence가 표현된다.

NGS data logistics



FastQ file

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCCAAGTAGCGTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE<,,CD,CEFC,@E9<FCFCF?9
@M02286:19:000000000-AA549:1:1101:15048:1299 1:N:0:23
CCTACGGGTGGCTGCAGTGAGGAATATTGGACAATGGTCGGAAGACTGATCCAGCCATGCCGCGTGCAG
+
ABC@CC77CFCEG;F9<F89<9--C,CE,--C-6C-,CE:++7:,CF<,CEF,CFGGD8FFCFCFEGCF
@M02286:19:000000000-AA549:1:1101:11116:1322 1:N:0:23
CCTACGGGAGGCAGCAGTAGGGAATCTTCGGCAATGGACGGAAGTCTGACCGAGCAACGCCGCGTGAGT
+
AAC<CCF+@@>CC,C9,F9C9@9-CFFFE@7@:+CC8-C@:7,@EFE,6CF:+8F7EFEEF@EGGGEEE
```

Line 1 : read의 ID와 optional information
Line 2 : read의 DNA 서열
Line 3 : additional information (중요x)
Line 4 : DNA 서열의 quality (ASCII code)

NGS data logistics

Base quality (phred quality score)

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

$$Q = -10 \log_{10} P$$

Q-Score는 염기를 호출할 때 발생할 수 있는 오류 가능성에 대한 수치.
-> 염기를 잘못 읽었을 확률 = P
-> $Q10 = -10 \log_{10} 10^{-1}$, $P = 0.1$

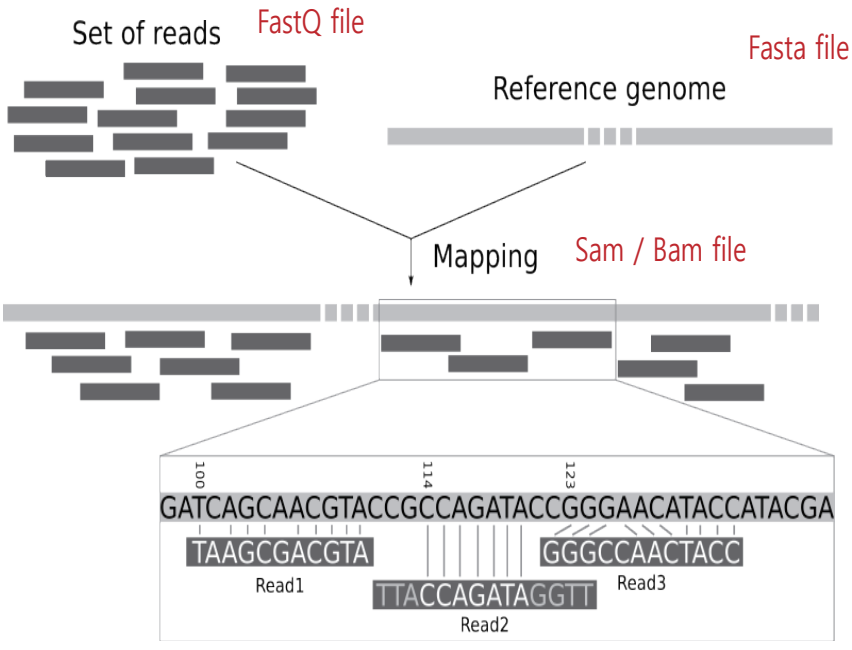
fastq파일 형식은 phred +33 또는 +64이며, 이 수치만큼 더해서 계산.
phred+33일때 quality score가 20이라면 53에 해당하는 '5'로 표기.

ASCII code

Ctrl	Dec	Hex	Char	Code	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
^@	0	00		NUL	32	20		64	40	@	96	60	`
^A	1	01		SOH	33	21	!	65	41	A	97	61	a
^B	2	02		STX	34	22	"	66	42	B	98	62	b
^C	3	03		ETX	35	23	#	67	43	C	99	63	c
^D	4	04		EOT	36	24	$	68	44	D	100	64	d
^E	5	05		ENQ	37	25	%	69	45	E	101	65	e
^F	6	06		ACK	38	26	&	70	46	F	102	66	f
^G	7	07		BEL	39	27	'	71	47	G	103	67	g
^H	8	08		BS	40	28	(	72	48	H	104	68	h
^I	9	09		HT	41	29	)	73	49	I	105	69	i
^J	10	0A		LF	42	2A	*	74	4A	J	106	6A	j
^K	11	0B		VT	43	2B	+	75	4B	K	107	6B	k
^L	12	0C		FF	44	2C	,	76	4C	L	108	6C	l
^M	13	0D		CR	45	2D	-	77	4D	M	109	6D	m
^N	14	0E		SO	46	2E	.	78	4E	N	110	6E	n
^O	15	0F		SI	47	2F	/	79	4F	O	111	6F	o
^P	16	10		DLE	48	30	0	80	50	P	112	70	p
^Q	17	11		DC1	49	31	1	81	51	Q	113	71	q
^R	18	12		DC2	50	32	2	82	52	R	114	72	r
^S	19	13		DC3	51	33	3	83	53	S	115	73	s
^T	20	14		DC4	52	34	4	84	54	T	116	74	t
^U	21	15		NAK	53	35	5	85	55	U	117	75	u
^V	22	16		SYN	54	36	6	86	56	V	118	76	v
^W	23	17		ETB	55	37	7	87	57	W	119	77	w
^X	24	18		CAN	56	38	8	88	58	X	120	78	x
^Y	25	19		EM	57	39	9	89	59	Y	121	79	y
^Z	26	1A		SUB	58	3A	:	90	5A	Z	122	7A	z
^[	27	1B		ESC	59	3B	;	91	5B	[	123	7B	{
^\	28	1C		FS	60	3C	<	92	5C	\	124	7C	|
^]	29	1D		GS	61	3D	=	93	5D	]	125	7D	}
^^	30	1E	▲	RS	62	3E	>	94	5E	^	126	7E	~
^_	31	1F	▼	US	63	3F	?	95	5F	_	127	7F	⌂

* ASCII code 127 has the code DEL. Under MS-DOS, this code has the same effect as ASCII 8 (BS). The DEL code can be generated by the CTRL + BKSP key.

NGS data logistics



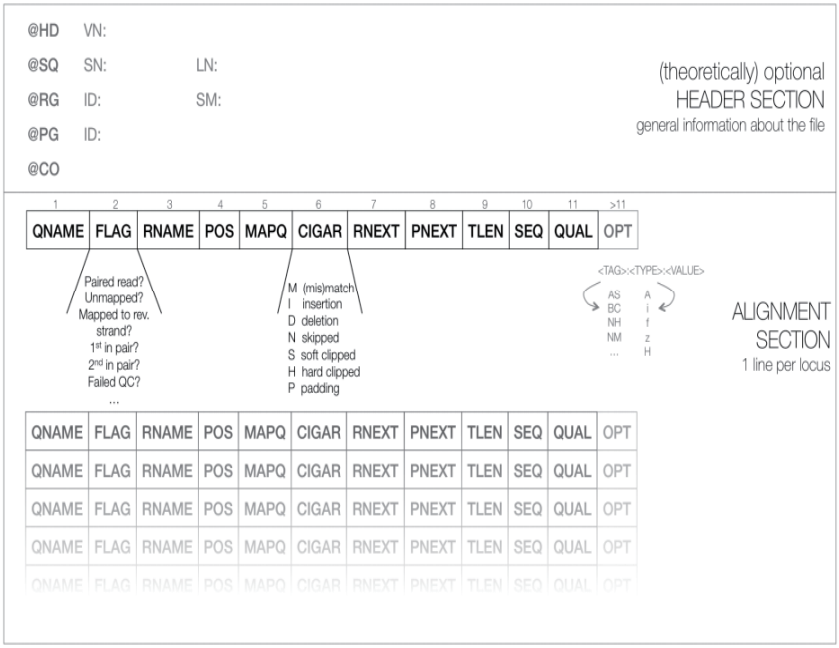
Fasta file

Ex)

>chr1
CAGTCGATCATCGCTAGCTA
ATCATATCGCCTTAGATATAG
CCGCGCTAATCGACTAGCA
:
>chr2
AATACGCTATTGATCAGTGC
ATAGCTAGCATTATACGTGC
GCGACTATAGTAGCAGCTAA
:

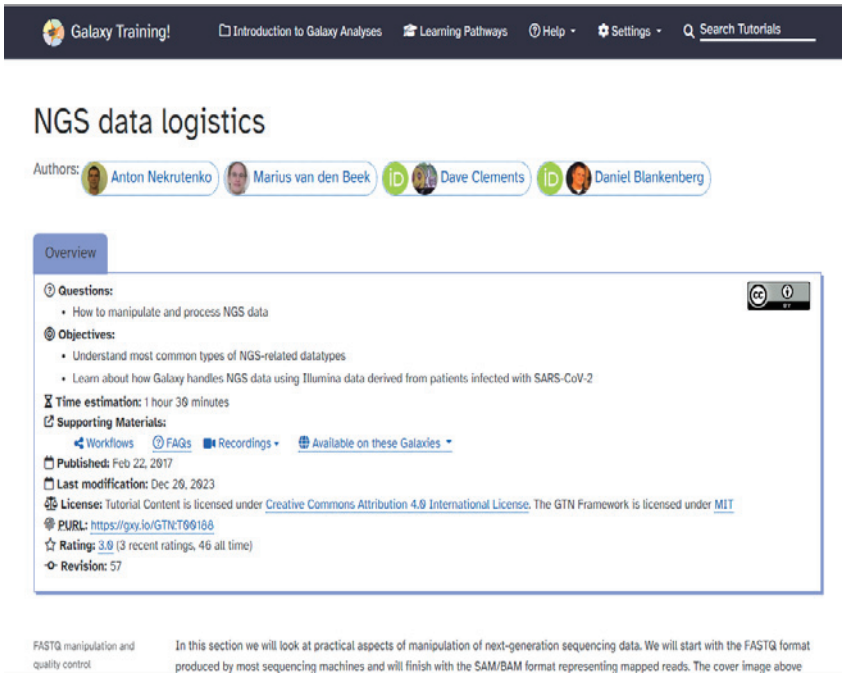
NGS data logistics

Sam / Bam file



QNAME	Read의 이름(ID)
FLAG	2진수로 표현된 mapping에 대한 정보
RNAME	Reference sequence의 이름 (chr1 등)
POS	Mapping된 위치
MAPQ	Mapping quality
CIGAR	Mapping된 read의 정보를 표현한 문자열
RNEXT	Paired-end sequencing일 때, read pair가 mapping된 reference sequence의 이름
PNEXT	Paired-end sequencing일 때, read pair가 mapping된 위치
TLEN	Paired-end sequencing일 때, 두 read의 왼쪽 끝에서 오른쪽 끝 까지의 거리
SEQ	Read의 DNA 염기서열
QUAL	ASCII code로 Phred quality score

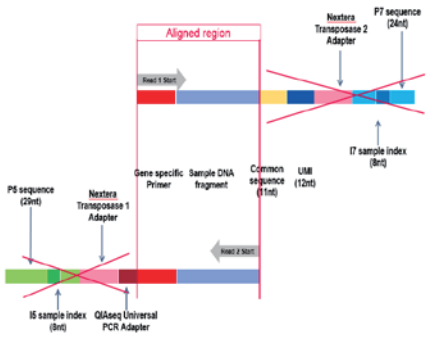
NGS data logistics

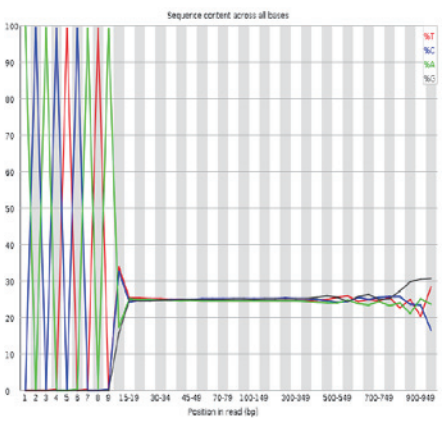


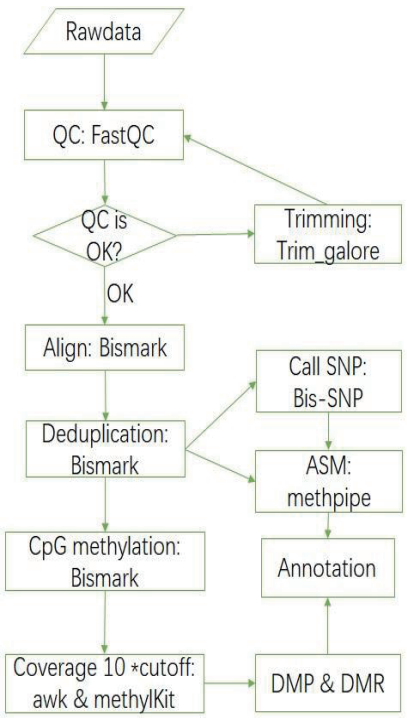
Introduction of WGBS analysis

Whole-genome methylation sequencing(WGBS) analysis

- **QC and Adapter Trimming**

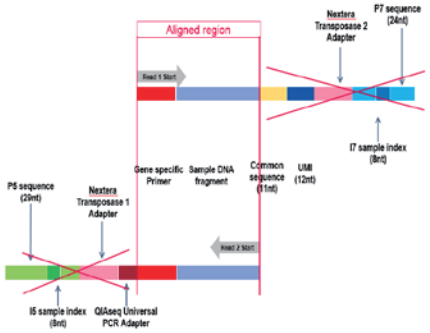


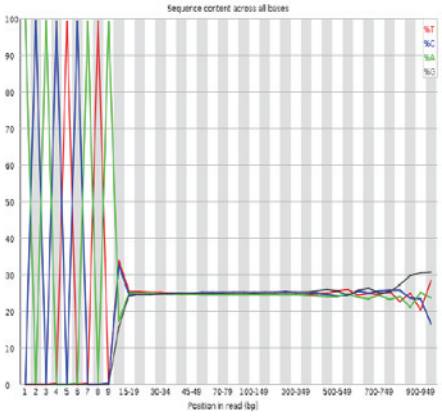


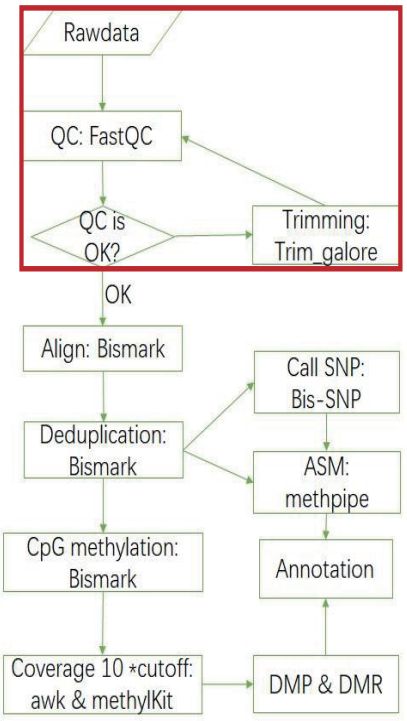


Whole-genome bisulfite sequencing(WGBS) analysis

- **QC and Adapter Trimming**

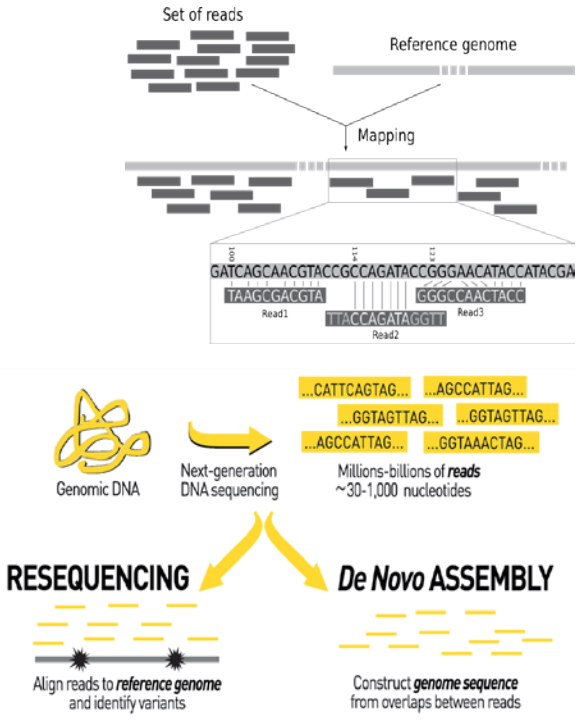


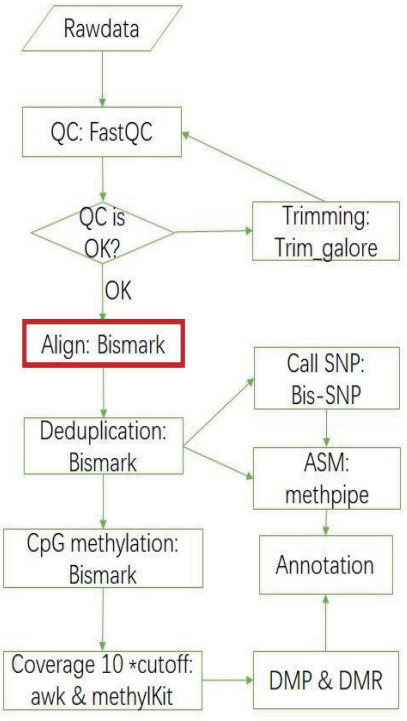




Whole-genome bisulfite sequencing(WGBS) analysis

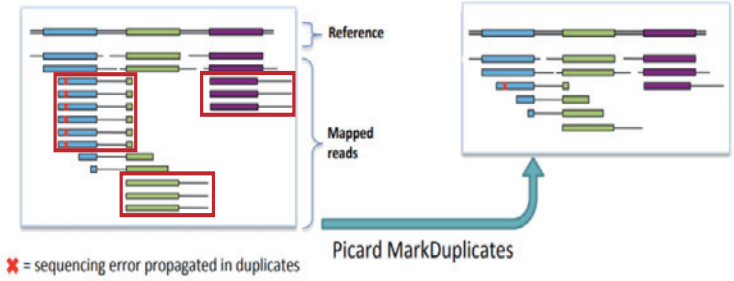
- **Alignment(mapping) and Reference genome**

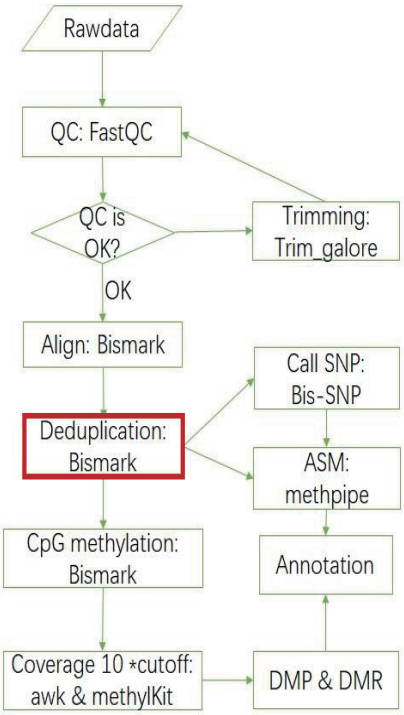




Whole-genome bisulfite sequencing(WGBS) analysis

- **Deduplication: remove PCR duplicates**

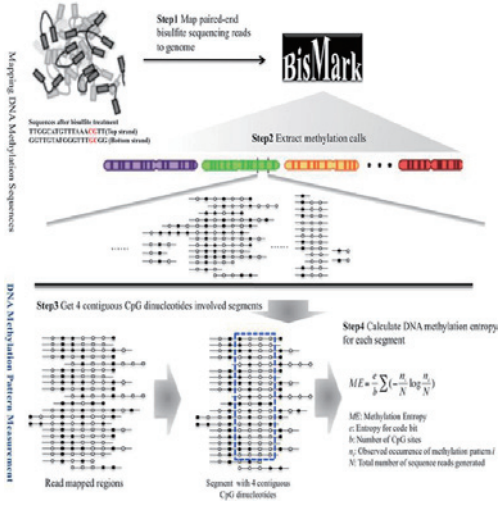


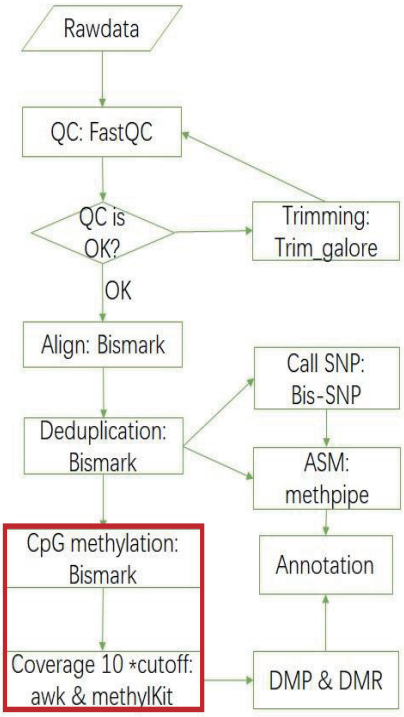


Whole-genome bisulfite sequencing(WGBS) analysis

- **Extract Methylation information of CpG context**

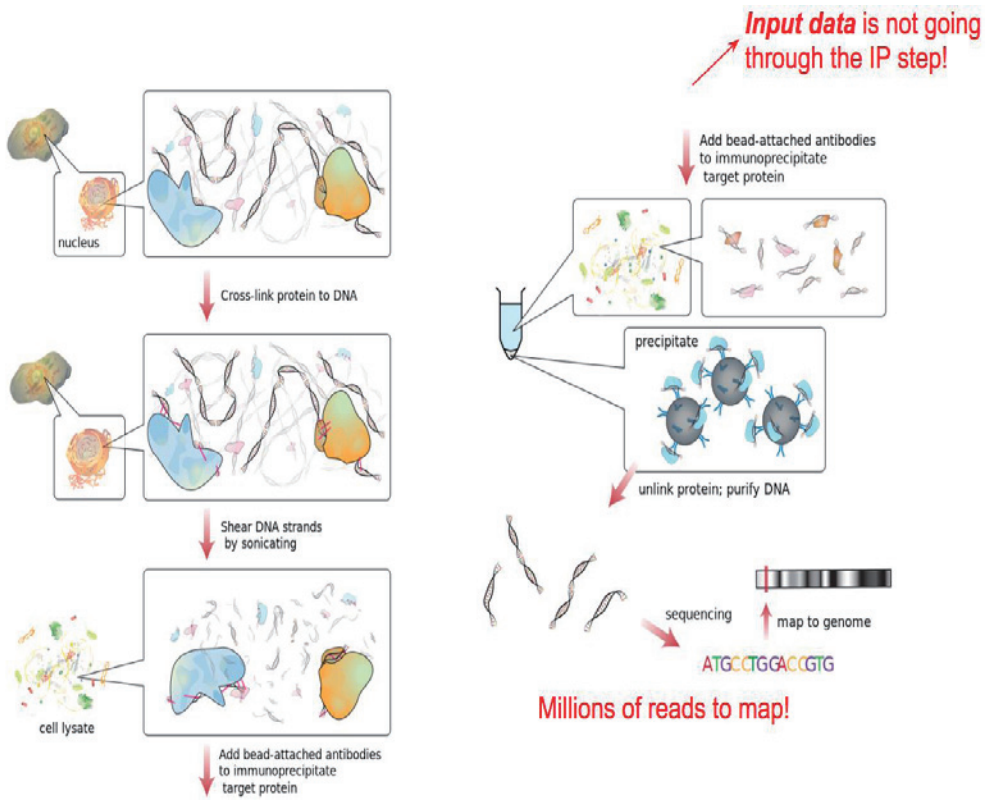
The bedGraph format allows display of continuous-valued data in track format. This display type is useful for probability scores and transcriptome data. This track type is similar to the wiggle (WIG) format, but unlike the wiggle format, data exported in the bedGraph format are preserved in their original state.



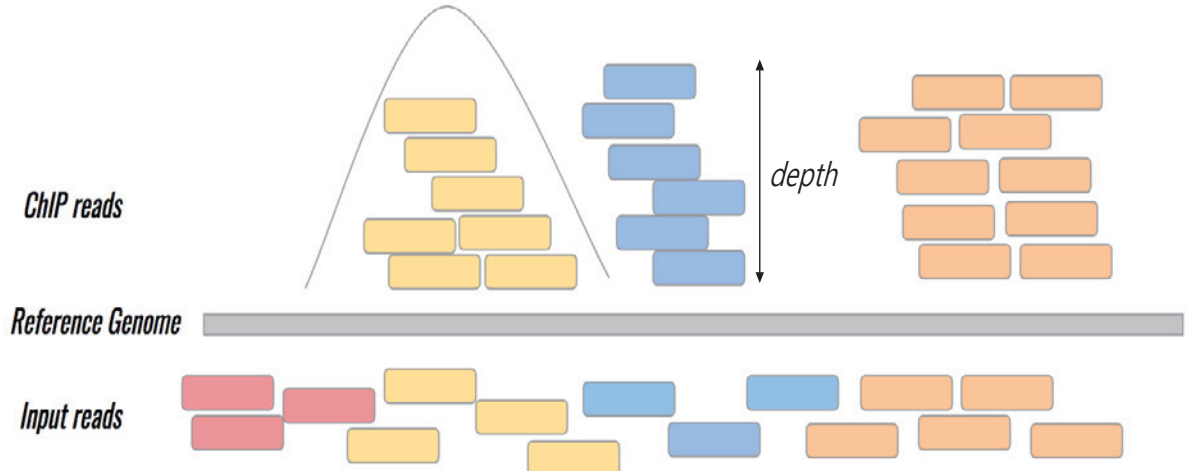


Introduction of ChIP-seq analysis

Methods of ChIP-seq

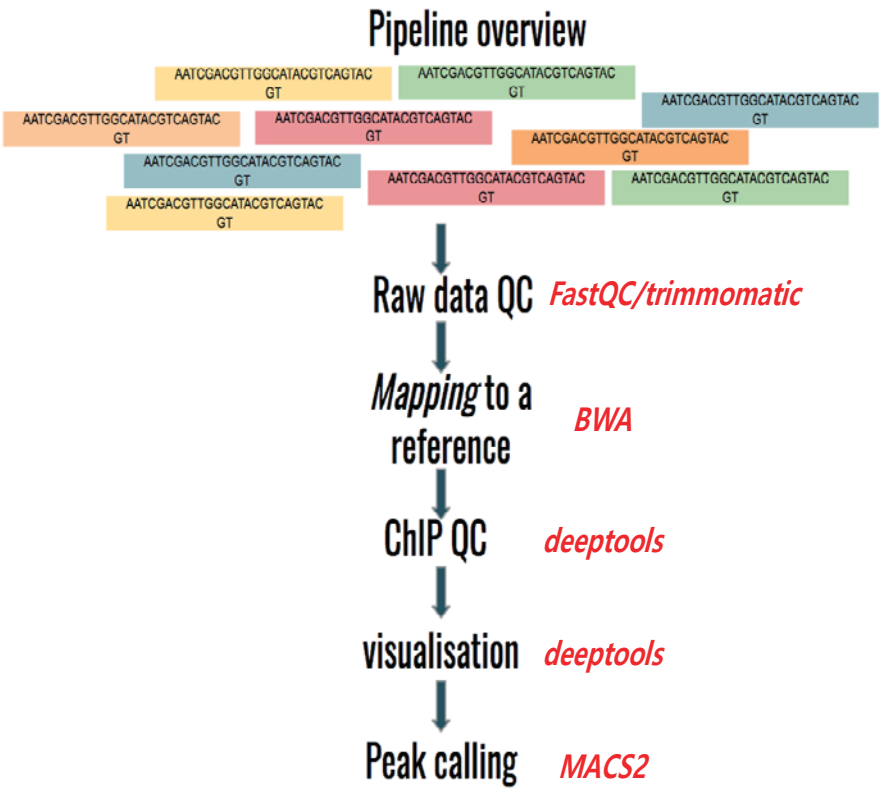


Basic understanding of ChIP-seq data



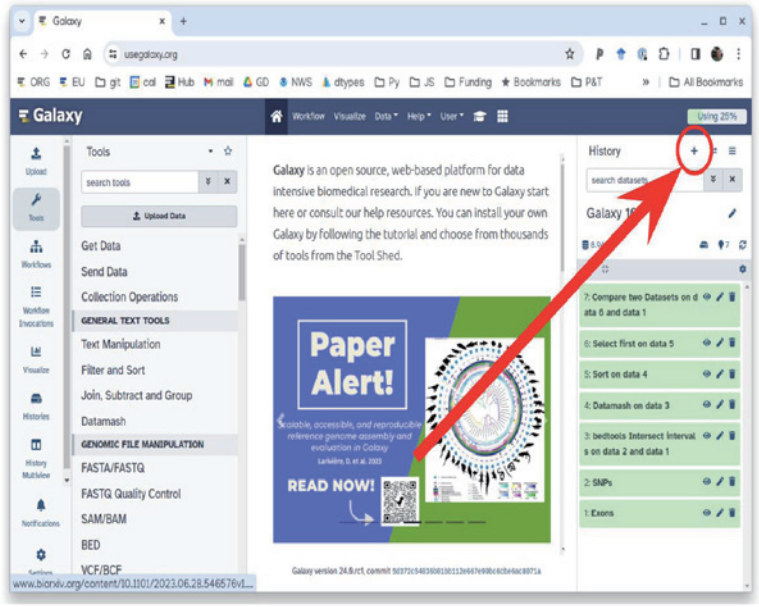
- Peak : enrichment of ChIP reads compared to input reads

Overview of the Data Processing



Tip: Creating a new history

Click the + icon at the top of the history panel:



Metadata for ChIP-seq

Table 1: Metadata for ChIP-seq experiments in this tutorial. SE: single-end.

Cellular state	Datatype	ChIP Ab	Replicate	SRA Accession	Library type	Read length	Stranded?	Data size (MB)
G1E	ChIP-seq	input	1	SRR507859	SE	36	No	35.8
G1E	ChIP-seq	input	2	SRR507860	SE	55	No	427.1
G1E	ChIP-seq	TAL1	1	SRR492444	SE	36	No	32.3
G1E	ChIP-seq	TAL1	2	SRR492445	SE	41	No	62.7
Megakaryocyte	ChIP-seq	input	1	SRR492453	SE	41	No	57.2
Megakaryocyte	ChIP-seq	input	2	SRR492454	SE	55	No	493.8
Megakaryocyte	ChIP-seq	TAL1	1	SRR549006	SE	55	No	340.3
Megakaryocyte	ChIP-seq	TAL1	2	SRR549007	SE	48	No	356.9

Quality control

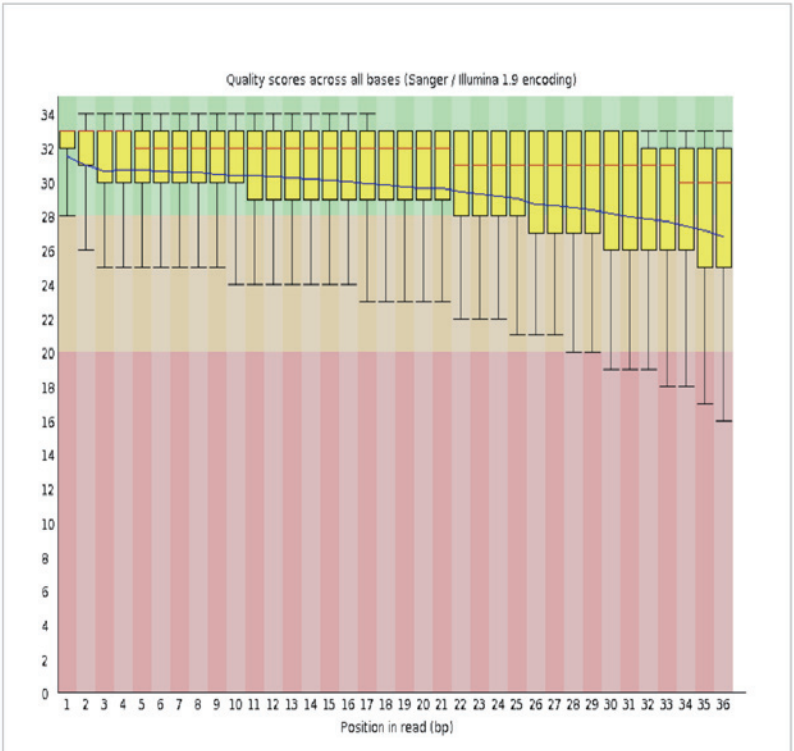
As for any NGS data analysis, ChIP-seq data must be quality controlled before being aligned to a reference genome. For more detailed information on NGS quality control, check out the tutorial here.

2. Import the ChIP-seq raw data (*.fastqsanger) from Zenodo.

Tip: Importing via links

- Copy the link location
- Click **Upload Data** at the top of the tool panel
- Select **Paste/Fetch Data**
- Paste the link(s) into the text field
- Press **Start**
- **Close** the window

4. **FastQC** (Galaxy version 0.72+galaxy1): Run FastQC on each FASTQ file to assess the quality of the raw data. An explanation of the results can be found on the FastQC web page.
    - "*Short read data from your current history*": The uploaded fastqsanger files.

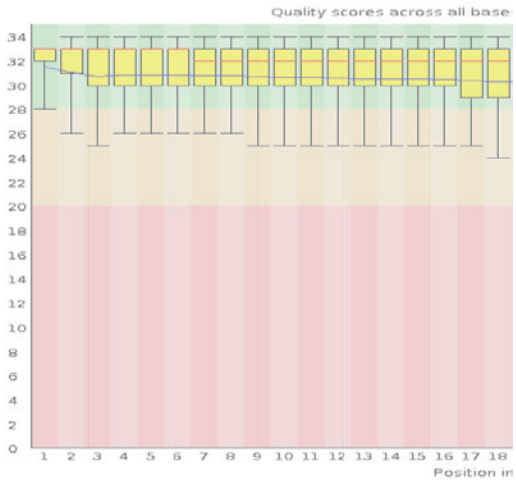


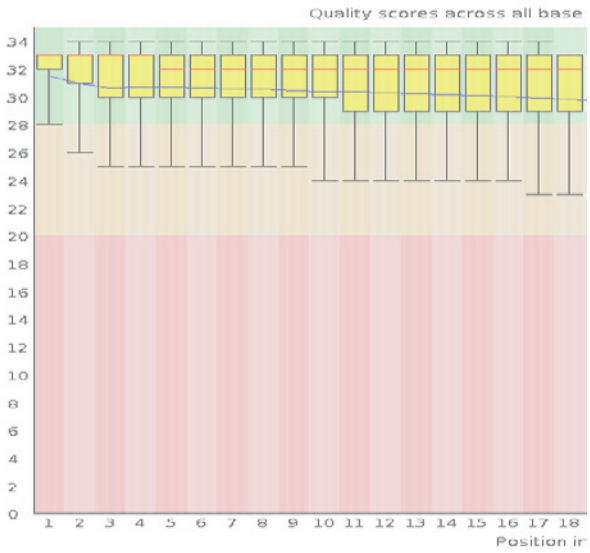
Trimming and clipping reads

It is often necessary to trim a sequenced read to remove bases sequenced with high uncertainty (*i.e.* low-quality bases). In addition, artificial adaptor sequences used in library preparation protocols need to be removed before attempting to align the reads to a reference genome.

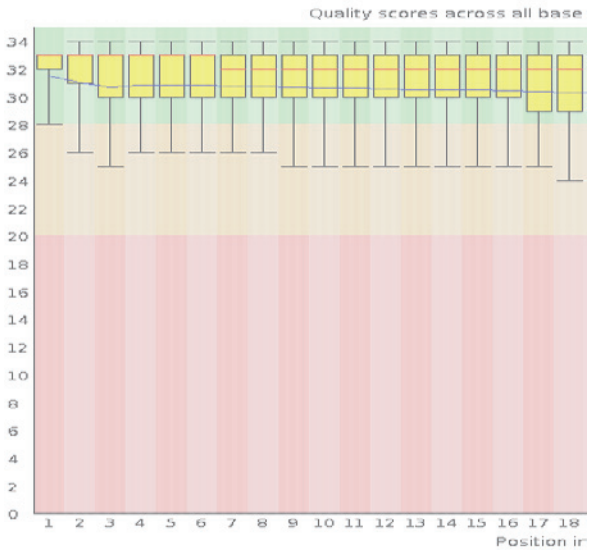
Hands-on: Trimming and clipping reads

1. **Trimmomatic** (Galaxy version 0.38.0): Run Trimmomatic to trim low-quality reads.
   - *"Single-end or paired-end reads?"*: `Single-end`
   - *"Input FASTQ file"*: Select all of the FASTQ files
   - *"Perform initial ILLUMINACLIP?"*: `No`
   - *"Select Trimmomatic operation to perform"*: `Sliding window trimming (SLIDINGWINDOW)`
   - *"Number of bases to average across"*: `4`
   - *"Average quality required"*: `20`





Before Trimming



After Trimming

Aligning reads to a reference genome

To determine where DNA fragments originated from in the genome, the sequenced reads must be aligned to a reference genome. This is equivalent to solving a jigsaw puzzle, but unfortunately, not all pieces are unique. In principle, you could do a BLAST analysis to figure out where the sequenced pieces fit best in the known genome. Aligning millions of short sequences this way, however, can take a couple of weeks. Nowadays, there are many read alignment programs for sequenced DNA, BWA being one of them. You can read more about the BWA algorithm and tool here.

Hands-on: Aligning reads to a reference genome

1. **BWA** (Galaxy version 0.7.17.4): Run BWA to map the trimmed/clipped reads to the mouse genome.
   - *"Will you select a reference genome..."*: `Use a built-in genome index`
   - *"Using reference genome"*: `Mouse (mus musculus) mm10`
   - *"Select input type"*: `Single fastq`
   - *"Select fastq dataset"*: Select all of the trimmed FASTQ files
2. Rename files to reflect the origin and contents.

> **Tip: Renaming a dataset**
>
> - Click on the **pencil icon** for the dataset to edit its attributes
> - In the central panel, change the **Name** field
> - Click the **Save** button

4. **Samtools idxstats** (Galaxy version 2.0.3): Run idxstats to get statistics of the BWA alignments.
   - *"BAM file"*: Select all of the mapped BAM files

deeptools for analysis and visualization

tool	type	input files	main output file(s)	application
tools/multiBamSummary	data integration	2 or more BAM	interval-based table of values	perform cross-sample analyses of read counts --> plotCorrelation, plotPCA
tools/multiBigwigSummary	data integration	2 or more bigWig	interval-based table of values	perform cross-sample analyses of genome-wide scores --> plotCorrelation, plotPCA
tools/plotCorrelation	visualization	bam/multiBigwigSummary output	clustered heatmap	visualize the Pearson/Spearman correlation
tools/plotPCA	visualization	bam/multiBigwigSummary output	2 PCA plots	visualize the principal component analysis
tools/plotFingerprint	QC	2 BAM	1 diagnostic plot	assess enrichment strength of a ChIP sample
tools/computeGCBias	QC	1 BAM	2 diagnostic plots	calculate the exp. and obs. GC distribution of reads
tools/correctGCBias	QC	1 BAM, output from computeGCbias	1 GC-corrected BAM	obtain a BAM file with reads distributed according to the genome's GC content
tools/bamCoverage	normalization	BAM	bedGraph or bigWig	obtain the normalized read coverage of a single BAM file
tools/bamCompare	normalization	2 BAM	bedGraph or bigWig	normalize 2 files to each other (e.g. log2ratio, difference)
tools/computeMatrix	data integration	1 or more bigWig, 1 or more BED	zipped file for plotHeatmap or plotProfile	compute the values needed for heatmaps and summary plots

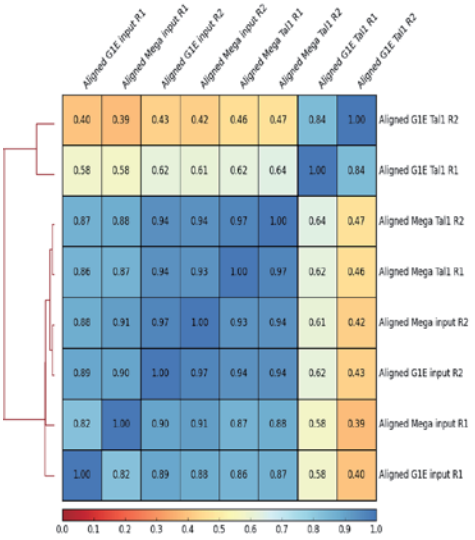
tools/estimateReadFiltering	information	1 or more BAM files	table of values	estimate the number of reads filtered from a BAM file or files
tools/alignmentSieve	QC	1 BAM file	1 filtered BAM or BEDPE file	filters a BAM file based on one or more criteria
tools/plotHeatmap	visualization	computeMatrix output	heatmap of read coverages	visualize the read coverages for genomic regions
tools/plotProfile	visualization	computeMatrix output	summary plot ("meta-profile")	visualize the average read coverages over a group of genomic regions
tools/plotCoverage	visualization	1 or more BAM	2 diagnostic plots	visualize the average read coverages over sampled genomic positions
tools/bamPEFragmentSize	information	1 BAM	text with paired-end fragment length	obtain the average fragment length from paired ends
tools/plotEnrichment	visualization	1 or more BAM and 1 or more BED/GTF	A diagnostic plot	plots the fraction of alignments overlapping the given features
tools/computeMatrixOperations	miscellaneous	1 or more BAM and 1 or more BED/GTF	A diagnostic plot	plots the fraction of alignments overlapping the given features

Assessing correlation between samples

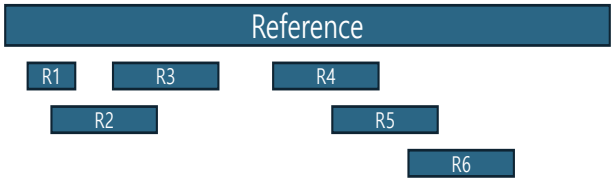
Correlation

- Assess the similarity between the replicates sequencing datasets.
- By using read counts for the different samples

1. **multiBamSummary** (Galaxy version 3.3.2.0.0): Run multiBamSummary to get read coverage of the alignments.
   - *"Sample order matters"*: `No`
   - *"Bam files"*: Select all of the aligned BAM files
   - *"Bin size in bp"*: `1000`
2. **plotCorrelation** (Galaxy version 3.3.2.0.0): Run plotCorrelation to visualize the results.
   - *"Matrix file from the multiBamSummary tool"*: Select the multiBamSummary output file
   - *"Correlation method"*: `Pearson`
   - *"Plotting type"*: `Heatmap`
   - *"Plot the correlation value"*: `Yes`
   - *"Skip zeros"*: `Yes`
   - *"Remove regions with very large counts"*: `Yes`

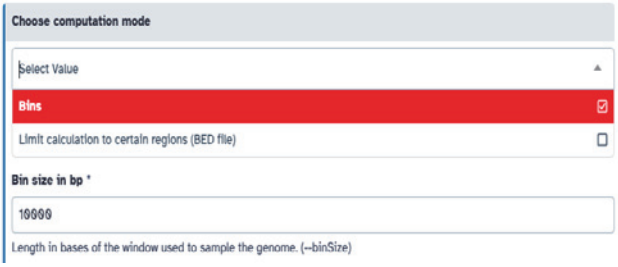


genomic bin vs bed-file



Genomic bin

- genomic region divided by specific length of bin (window)



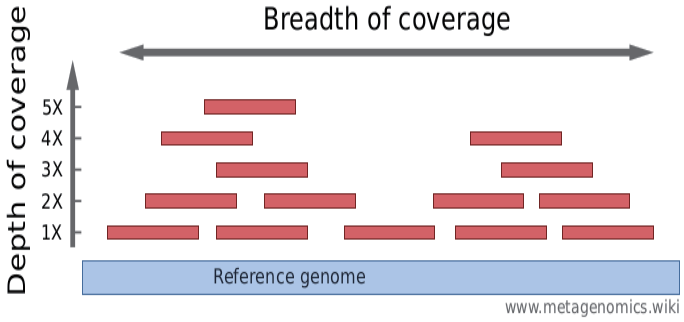
Bed (Browser Extensible Data) file

- a text file format used to store genomic regions as coordinates and associated annotations.
- at least 3 columns :

chromosome, chromStart, chromEnd

```
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    -    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    -    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    -    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    -    127480532    127481699    0,0,255
```

compute the read coverages for genomic region



Read coverage

- read count on a specific genomic region

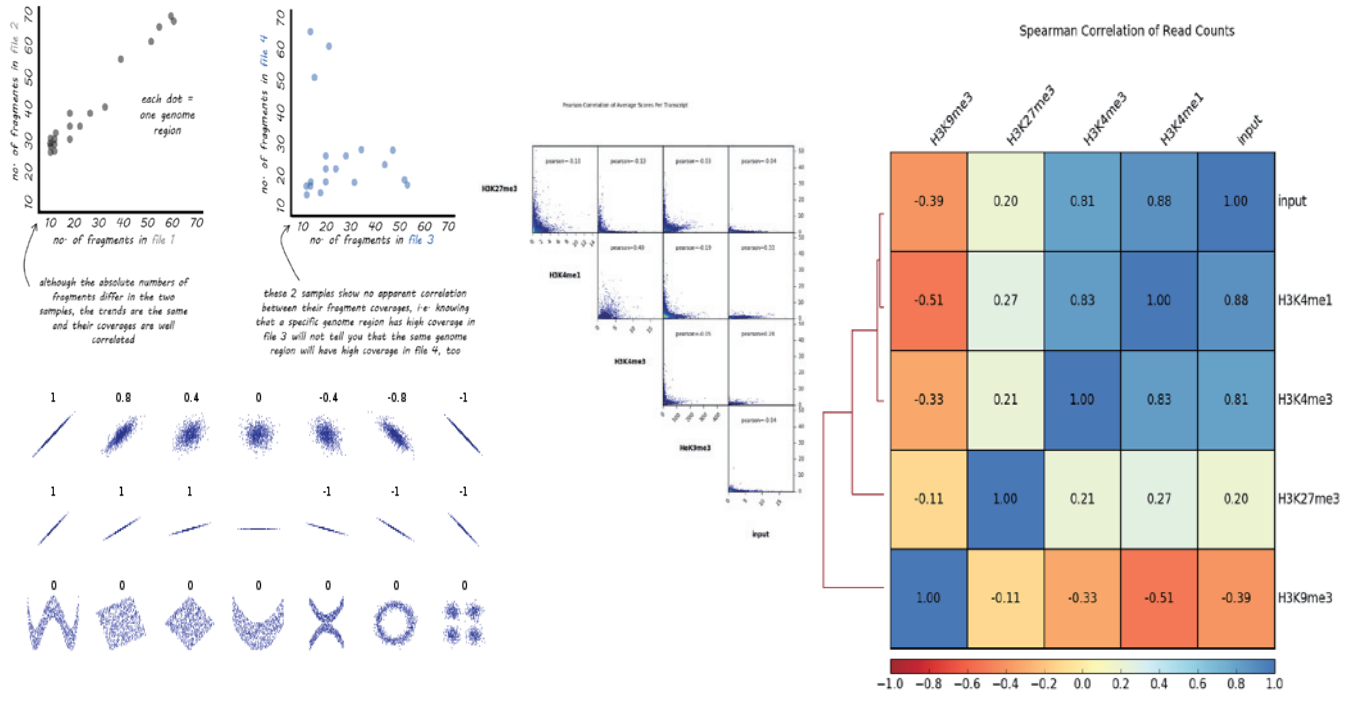
Depth

- per-base coverage, **the average number of times a base of a genome is sequenced**.
- the number of bases of all short reads that match a genome divided by the length of genome

Breadth

- the percentage of bases of a reference genome that are covered with a certain depth.
- ex) **90%** of a genome is covered at **1X** depth, **70%** of a genome is covered at **5X** depth.

Correlation



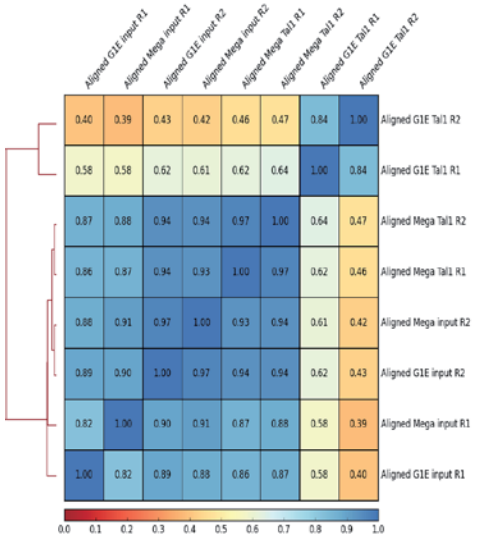
deeptools plotCorrelation

Assessing correlation between samples

Correlation

- Assess the similarity between the replicates sequencing datasets.
- By using read counts for the different samples

1. **multiBamSummary** (Galaxy version 3.3.2.0.0): Run multiBamSummary to get read coverage of the alignments.
   - *"Sample order matters"*: `No`
   - *"Bam files"*: Select all of the aligned BAM files
   - *"Bin size in bp"*: `1000`
2. **plotCorrelation** (Galaxy version 3.3.2.0.0): Run plotCorrelation to visualize the results.
   - *"Matrix file from the multiBamSummary tool"*: Select the multiBamSummary output file
   - *"Correlation method"*: `Pearson`
   - *"Plotting type"*: `Heatmap`
   - *"Plot the correlation value"*: `Yes`
   - *"Skip zeros"*: `Yes`
   - *"Remove regions with very large counts"*: `Yes`

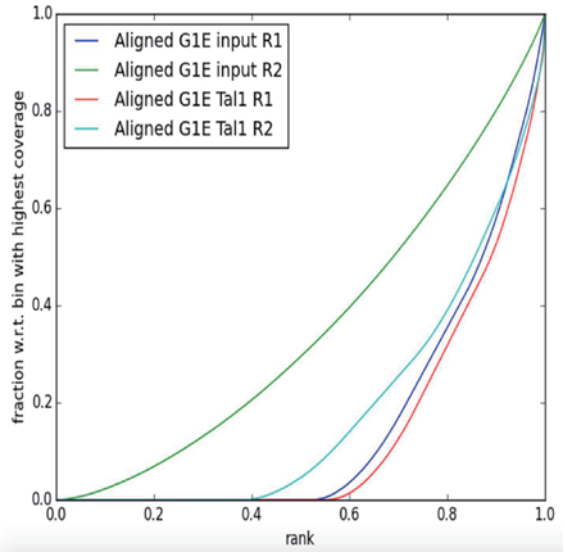


IP strength

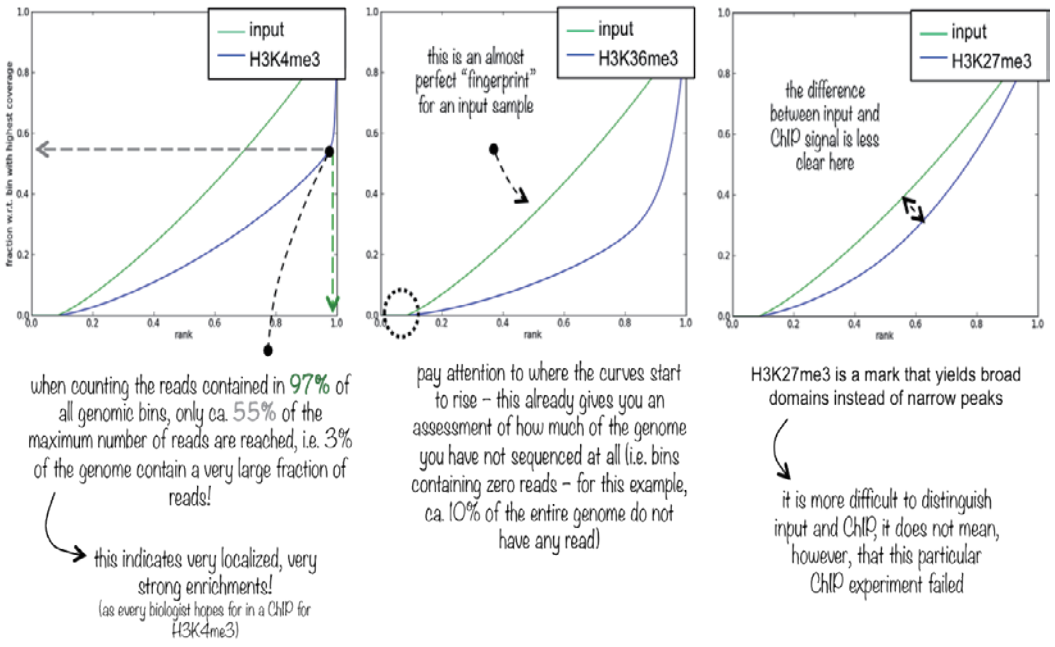
ImmunoPrecipitation strength

- evaluate the quality of the immunoprecipitation step in the ChIP-seq protocol.
- By compare input bam and ChIP bam

1. **plotFingerprint** (Galaxy version 3.3.2.0.0): Run plotFingerprint to assess ChIP signal strength.
   - *"Bam files"*: Select all of the aligned BAM files for the G1E cell type
   - *"Show advanced options"*: `yes`
   - *"Bin size in bases"*: `100`
   - *"Skip zeros"*: `Yes`



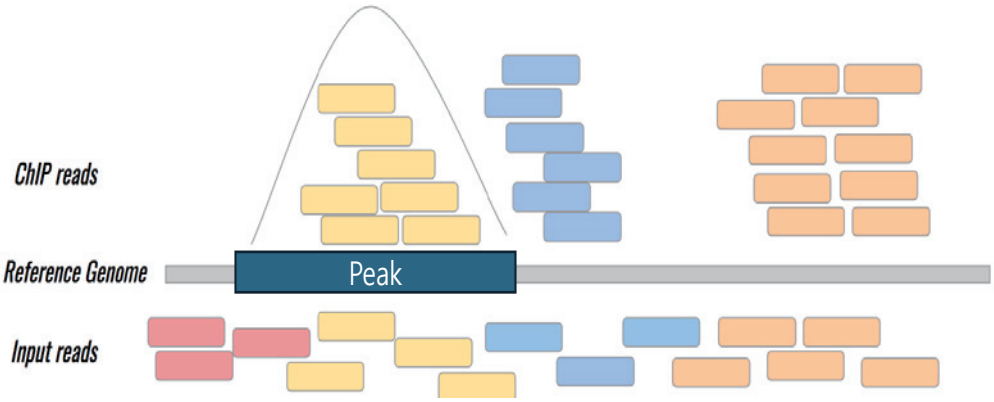
IP strength



Peakcall

Peakcall

- **Peak :** regions of protein occupancy
- Determined from pileups of sequenced reads across the genome that correspond to where protein binds

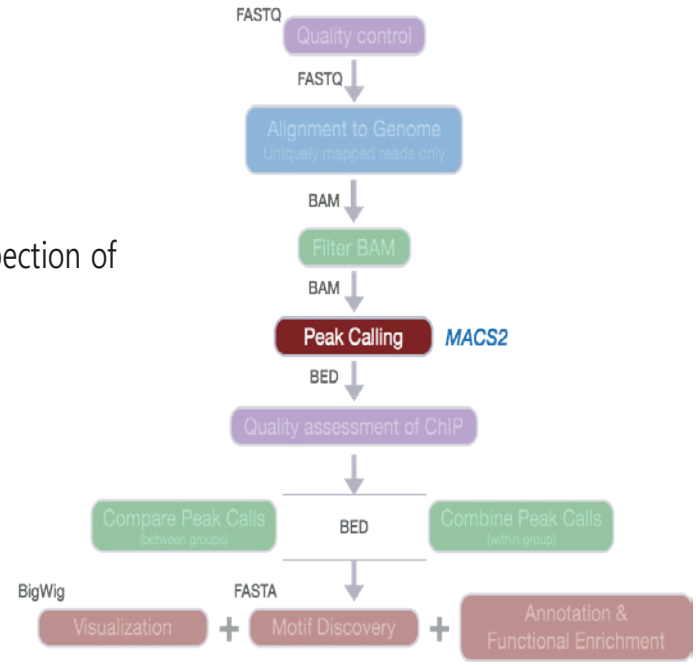


MACS2

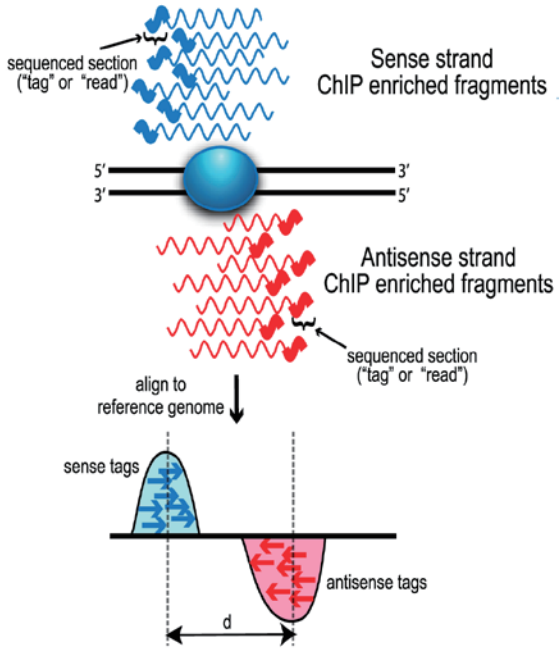
Model-based Analysis of ChIP-seq

Tasks:

- Identify regions of protein occupancy
- Generate bedgraph files for visual inspection of the data on a genome browser



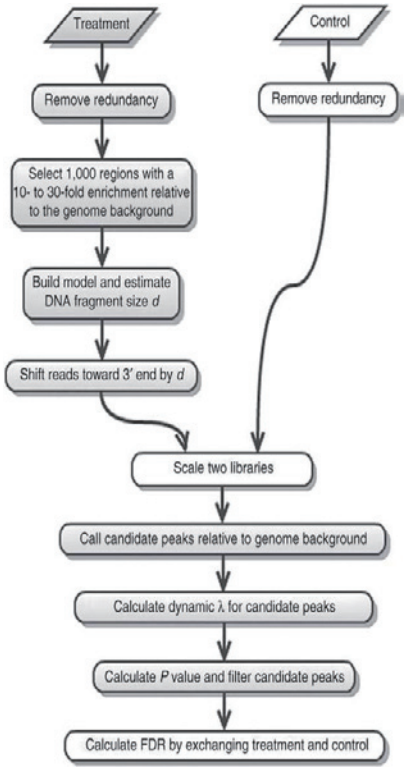
Sequencing data of ChIP-seq



Strand asymmetry with read densities on the +/- strand, centered around the binding site.

- The 5′ ends of the selected fragments will form groups on the positive- and negative-strand.
- The distributions of these groups are then assessed using statistical measures and compared against background (input or mock IP samples) to determine if the site of enrichment is likely to be a real binding site.

MACS2 algorithm



Removing redundancy

dealing with **duplicate reads** at the exact same location

- reads with the same coordination and the same strand

The Bad kind of duplicates

- If initial starting material Is low this can lead to overamplification of this material before sequencing.
- PCR, repeat sequence
- Masking these regions prior to analysis can help remove this problem

The Good kind of duplicates

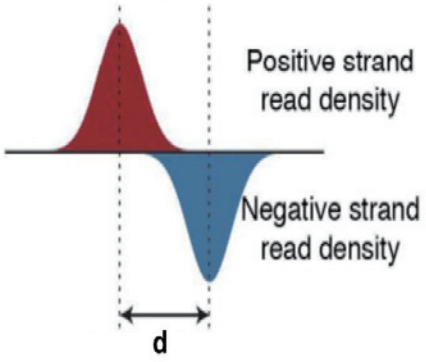
- only sequencing a small part of the genome.
- Increase if the depth of coverage is excessive or if the protein only binds to few sites.
- If there are a good proportion of biological duplicates, removal can lead to an underestimation of the ChIP signal

Modeling the shift size

- The read density around a true binding site should show a **bimodal enrichment pattern** (or paired peaks)

Build model

- scan the whole dataset searching for highly significant enriched rigions
- find regions with **reads more than fold-enriched relative to a random read genome distribution**
- randomly samples 1,000 of high-quality peaks, separates their +/- strand reads, and aligns them by the midpoint between their centers
- d : **distance between the modes the two peaks in the alignment**
- shifts all the reads by **d/2 toward the 3' end** to the most likely protein-DNA interaction sites.

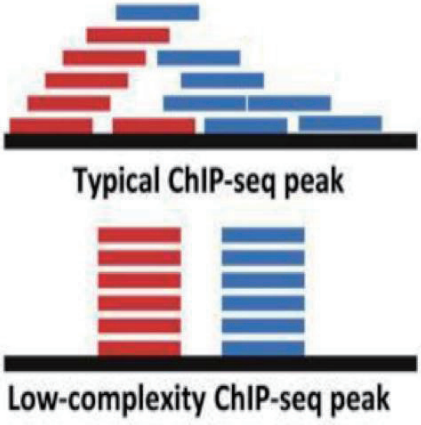


Scaling libraries

- Sequence depth differs between input and treatment samples
- Linearly scales the **total control read count to be the same as the total ChIP read count**.

Effective genome length

- λ is the **expected number of reads in that window**
- To calculate λBG from read count, MACS2 requires the **effective genome size** or the size of the genome that is mappable.
- Effective genome size to **correct for the loss of true signals in low-mappable regions**
- **Mappability** is related to the uniqueness of the k-mers at a particular position the genome.
- **Low complexity and repetitive regions** have low uniqueness which means low mappability.



Landt et al, Genome Res. 2012

Peak detection

- After MACS2 shifts every read by d/2, it then slides across the genome using a window size of 2d to find **candidate peaks**.
- The read distribution along the genome can be modeled by a Poisson distribution.
- Poisson is a one parameter model, where the parameter λ is the **expected number of reads in that window**

$$P_\lambda(X=k) = \frac{\lambda^k}{k! * e^{-\lambda}}$$

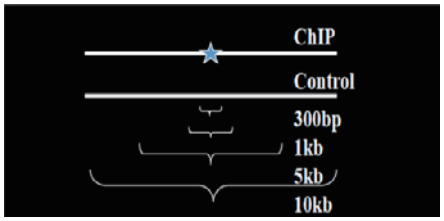
λ = mean = expectated value = variance

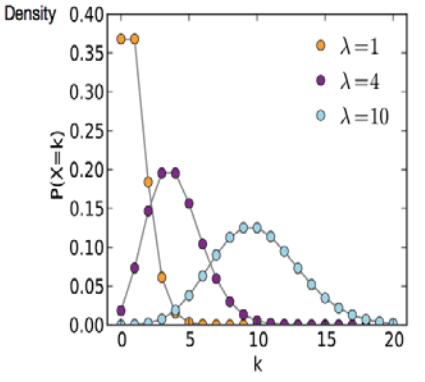
$$\lambda = \frac{\text{total number of events (k)}}{\text{number of units (n) in the data}}$$

$$= \frac{\text{Read length (nt)} * \text{Total read number}}{\text{Effective genome length (nt)}}$$

λlocal

- **dynamic parameter**
- taking the maximum value across window sizes
- λlocal = max(λBG, λ1k, λ5k, λ10k).
- robust against **occasional low read counts at small local regions**.





http://en.wikipedia.org/wiki/Poisson_distribution

Peak detection

- A region is considered to have a significant read enrichment if the p-value < 10e-5
- Overlapping enriched peaks are merged, and each read position is extend 'd' bases from its center.
- Summit : The location in the peak with the highest fragment pileup, precise binding location
- The fold enrichment : the ratio between the ChIP-seq read count and λlocal

Determining TAL1 binding sites

Now that BWA has aligned the reads to the genome, we will use the tool MACS2 to identify regions of TAL1 occupancy, which are called "peaks". Peaks are determined from pileups of sequenced reads across the genome that correspond to where TAL1 binds.

MACS2 will perform two tasks:

1. Identify regions of TAL1 occupancy (peaks).
2. Generate bedGraph files for visual inspection of the data on a genome browser.

More information about MACS2 can be found in Zhang *et al.* 2008.

1. **MACS2 callpeak** (Galaxy version 2.1.1.20160309.6): Run MACS2 callpeak with the aligned read files from the previous step as Treatment (TAL1) and Control (input).
   - *"Are you pooling Treatment Files?"*: `Yes`
   - *"ChIP-Seq Treatment File"*: Select all of the replicate ChIP-Seq treatment aligned BAM files for one cell type
   - *"Do you have a Control File?"*: `Yes`
   - *"Are you pooling Control Files?"*: `Yes`
   - *"ChIP-Seq Control File"*: Select replicate ChIP-Seq control aligned BAM files for the same cell type
   - *"Format of Input Files"*: `Single-end BAM`
   - *"Effective genome size"*: `M. musculus`
   - *"Additional Outputs"*: Select `Peaks as tabular file (compatible wih MultiQC)`, `Peak summits`, `Scores in bedGraph files (--bdg)`
2. Rename files to reflect the origin and contents.
3. Repeat for the other cell type.

Inspection of peaks and aligned data

It is critical to visualize NGS data on a genome browser after alignment to evaluate the "goodness" of the analysis. Evaluation criteria will differ for various NGS experiment types, but for ChIP-seq data we want to ensure reads from a Treatment/IP sample are enriched at peaks and do not localize non-specifically (like the control/input condition).

Inspection of peaks and aligned data with IGV

1. Open IGV on your local computer.
2. Click on each narrow peaks result file from the MACS2 computations on "display with IGV" -> "local Mouse mm10"
3. For more information about IGV see here

Identifying unique and common TAL1 peaks between stages

We have processed ChIP-seq data from two stages of hematopoiesis and have lists of TAL1-occupied sites (peaks) in both cellular states. The next analysis step is to identify TAL1 peaks that are *shared* between the two cellular states and peaks that are *specific* to either cellular state.

1. **bedtools Intersect intervals** (Galaxy version 2.29.0): Run bedtools Intersect intervals to find peaks that exist both in G1E and megakaryocytes.
   - *"File A to intersect with B"*: Select the TAL1 G1E narrow peaks BED file
   - *"File B to intersect with A"*: Select the TAL1 Megakaryocytes narrow peaks BED file
   - Running this tool with the default settings will return overlapping peaks of both files.
2. **bedtools Intersect intervals** (Galaxy version 2.29.0): Run bedtools Intersect intervals to find peaks that exist only in G1E.
   - *"File A to intersect with B"*: Select the TAL1 G1E narrow peaks BED file
   - *"File B to intersect with A"*: Select the TAL1 Megakaryocytes narrow peaks BED file
   - *"Report only those alignments that **do not** overlap the BED file"*: Yes
3. **bedtools Intersect intervals** (Galaxy version 2.29.0): Run bedtools Intersect intervals to find peaks that exist only in megakaryocytes.
   - *"File A to intersect with B"*: Select the TAL1 Megakaryocytes narrow peaks BED file
   - *"File B to intersect with A"*: Select the TAL1 G1E narrow peaks BED file
   - *"Report only those alignments that **do not** overlap the BED file"*: Yes
4. Rename files to reflect the origin and contents.

Generating Input normalized coverage files

We will generate Input normalized coverage (bigWig) files for the ChIP samples, using the bamCompare tool from deepTools2. bamCompare provides multiple options to compare the two files (*e.g.* log2ratio, subtraction). We will use log2 ratio of the ChIP samples over Input.

1. **bamCompare** (Galaxy version 3.3.2.0.0): Run bamCompare to get the log2 read ratios between treatment and control samples.
    - *"First BAM/CRAM file (*e.g.* treated sample)"*: Select the Megakaryocyte TAL1 aligned BAM file for replicate 1 (R1)
    - *"Second BAM/CRAM file (*e.g.* control sample)"*: Select the Megakaryocyte input aligned BAM file for replicate 1 (R1)
    - *"How to compare the two files"*: `Compute log2 of the number of reads`
2. Repeat this step for all treatment and control samples:
    - Megakaryocyte TAL1 aligned BAM R2 and Megakaryocyte input aligned BAM R2
    - G1E TAL1 aligned BAM R1 and G1E input aligned BAM R1
    - G1E TAL1 aligned BAM R2 and G1E input aligned BAM R2
3. Rename files to reflect the origin and contents.

Plot the signal on the peaks between samples

Plotting your region of interest will involve using two tools from the deepTools suite:

- computeMatrix: Computes the signal on given regions, using the bigwig coverage files from different samples.
- plotHeatmap: Plots heatMap of the signals using the computeMatrix output.

Optionally, you can use plotProfile to create a profile plot using to computeMatrix output.

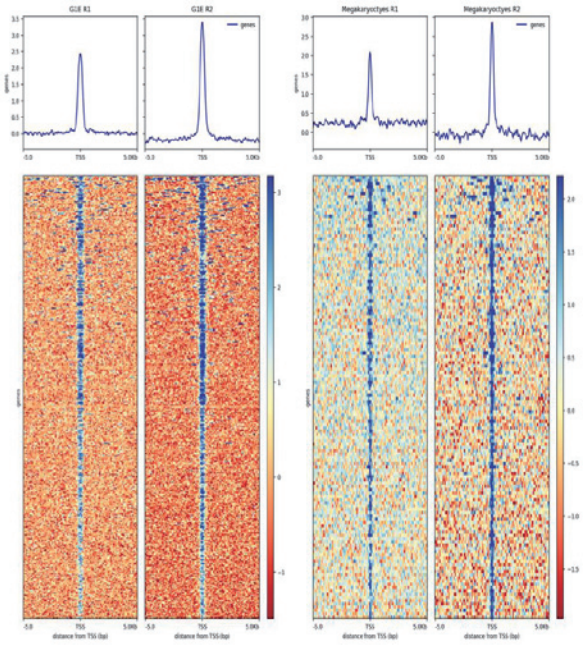
Hands-on: Calculating signal matrix on MACS2 output

1. **computeMatrix** (Galaxy version 3.3.2.0.0): Run computeMatrix to prepare data for plotting a heatmap of TAL1 peaks.
    - *Select Regions* > *"Regions to plot"*: Select the MACS2 narrow peaks files for G1E cells (TAL1 over Input)
    - *"Score file"*: Select the bigWig files for the G1E cells (log2 ratios from bamCompare)
    - *"computeMatrix has two main output options"*: `reference-point`
    - *"The Reference point for plotting"*: `center of region`
    - *"Distance upstream of the start site of the regions defined in the region file"*: `5000`
    - *"Distance downstream of the end site of the given regions"*: `5000`
    - *"Show advanced options"*: `Yes`
    - *"Convert missing values ot zero"*: `Yes`
    - *"Skip zeros"*: `Yes`
2. Repeat for Megakaryoctes.

Hands-on: Plotting a heatmap of TAL1 peaks

1. **plotHeatmap** (Galaxy version 3.3.2.0.1): Run plotHeatmap to create a heatmap for score distributions across TAL1 peak genomic regions in each cell type.
    - *"Matrix file from the computeMatrix tool"*: Select the computeMatrix output for G1E cells
    - *"Show advanced options"*: Yes
    - *"Labels for the samples (each bigwig) plotted"*: Enter sample labels in the order you added them in computeMatrix, separated by spaces.

2. Repeat for Megakaryocytes.



THANK YOU

감사합니다