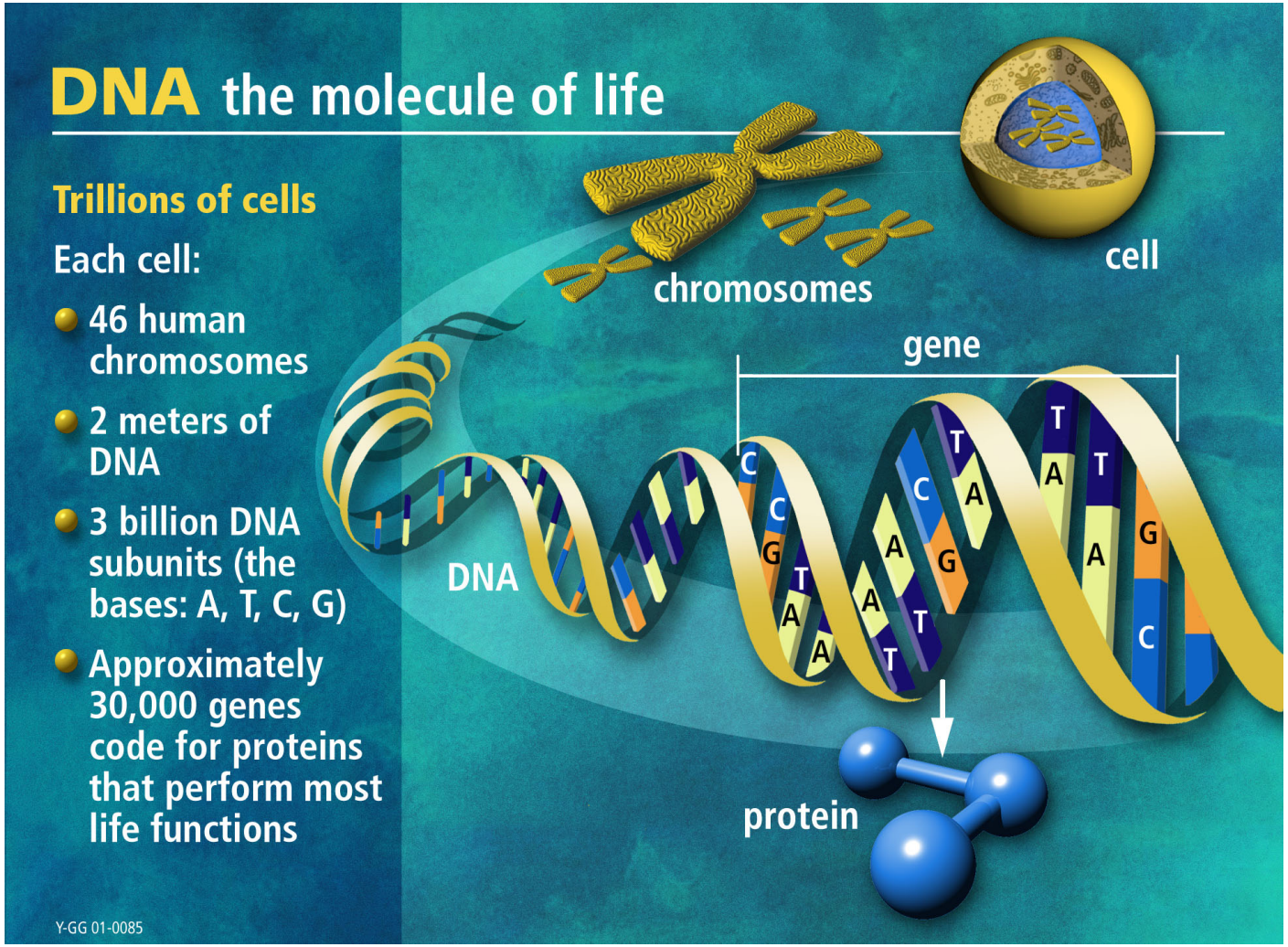
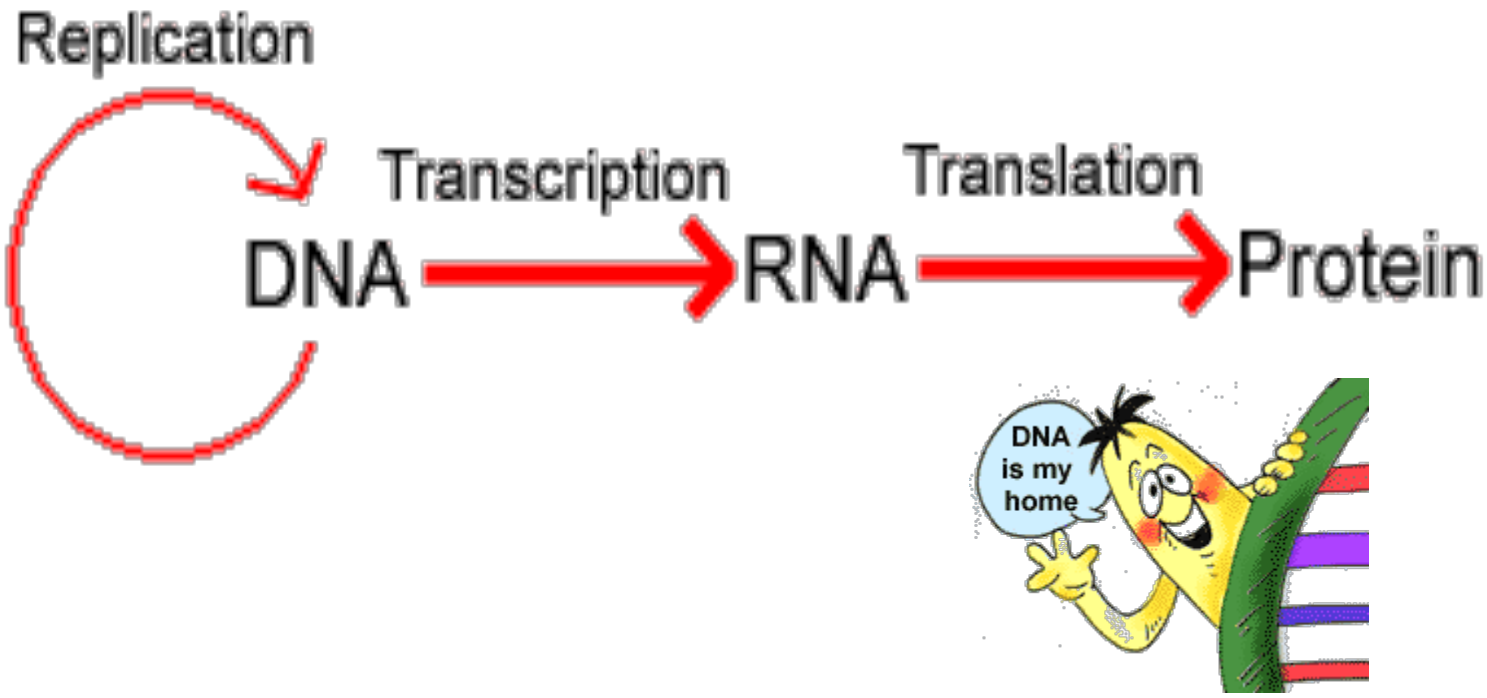
Stat 8750.04 – Autumn 2023

Some Random Notes on
Statistical Genomics and Bioinformatics

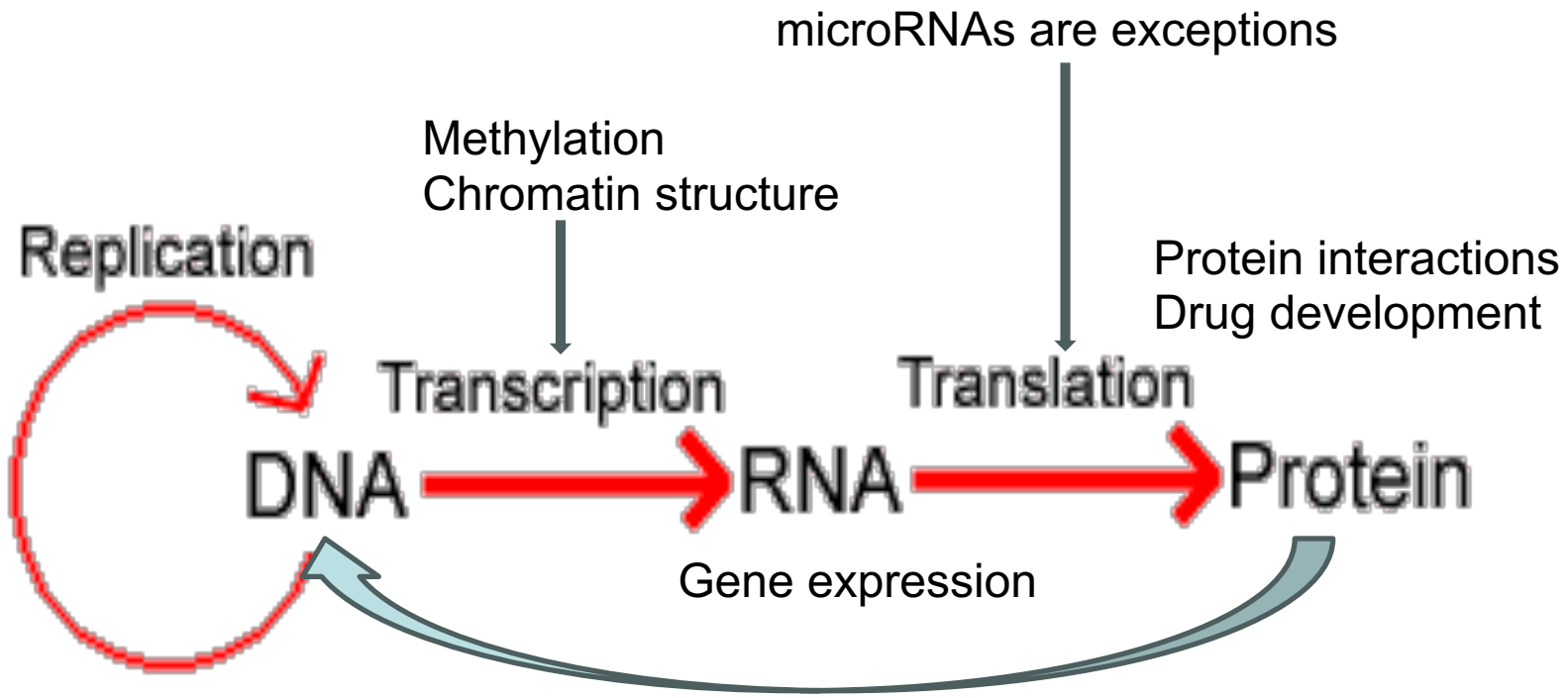
The Human Genome



The Central Dogma of Genetics (simplistic view – one-way street)



The Central Dogma

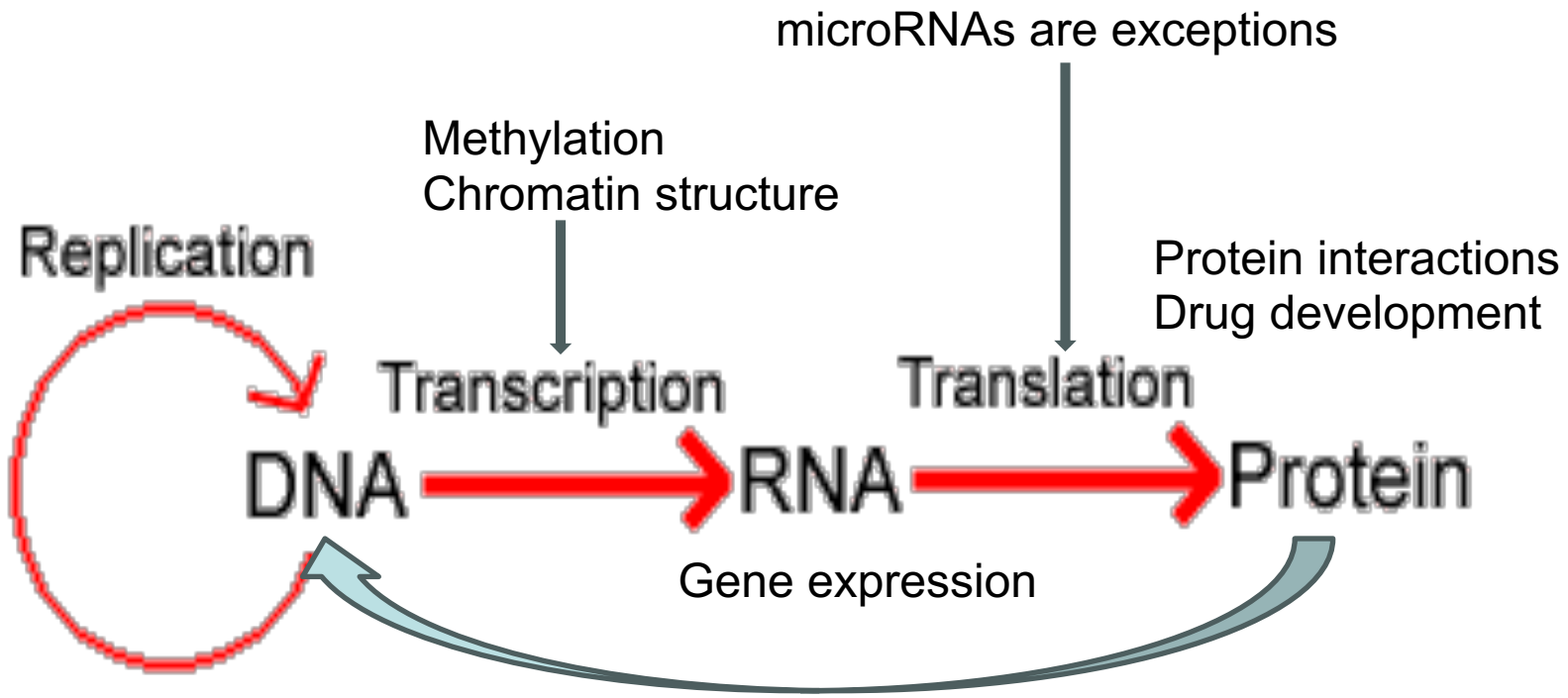


Traditional statistical genetics:
study of static state

Statistical Genetics/Genomics/Bioinformatics

- Studies of randomness in the genome (more traditional sense; the DNAs)
- Studies of gene expressions and protein networks
- Understanding (post) transcriptional regulations
- Investigate host-microbes/pathogen association (e.g. microbiome)
- Related/overlapped with bioinformatics (biodata mining; multi-omics)
- Interdisciplinary area in which probability modeling and statistical methods are used to
  - analyze genetic data
  - understand biological processes
  - aid medical researches

The Central Dogma



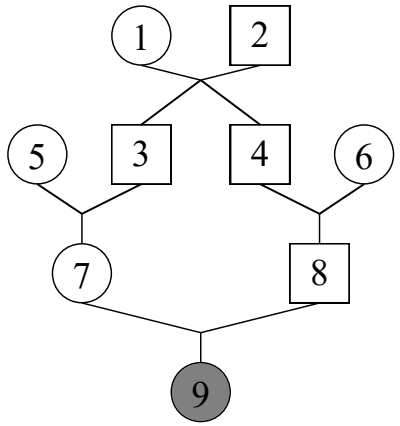
Traditional statistical genetics:
study of static state

Data for Genetic Association Study

- Phenotype (observable)
  - Binary (e.g. hypertensive status)
  - Quantitative (systolic/diastolic blood pressure)

- DNA data - marker genotypes (SNPs – single nucleotide polymorphisms)
  - A: major; a: minor allele with coding 0 (AA), 1 (Aa), 2 (aa)
  - Directly using ACGT

- Family relationships
  - Pedigree (graph)
  - Triplet identifier: id, fid, mid



Data Examples

- Individual-level data

```
1 1 0 0 1    1    A A    A A    A A    A A    A A
2 1 0 0 1    1    A C    A C    A C    A C    A C
3 1 0 0 2    1    A A    A A    A A    A A    A A
4 1 0 0 2    1    A C    A C    A C    A C    A C
```

- Summary statistics

Chr	SNP	bp	A1	A2	Freq	b	se	p
1	rs2286139	761732	C	T	0.1379	-0.0104056	0.00732416	0.155397
1	rs12562034	768448	A	G	0.10475	-0.00627592	0.00827054	0.447955
1	rs4970383	838555	A	C	0.247975	0.00946201	0.00587444	0.107243
1	rs1806509	853954	C	A	0.3912	0.0152744	0.00523012	0.00349507
1	rs13302982	861808	A	G	0.018025	-0.0180122	0.0189517	0.341895

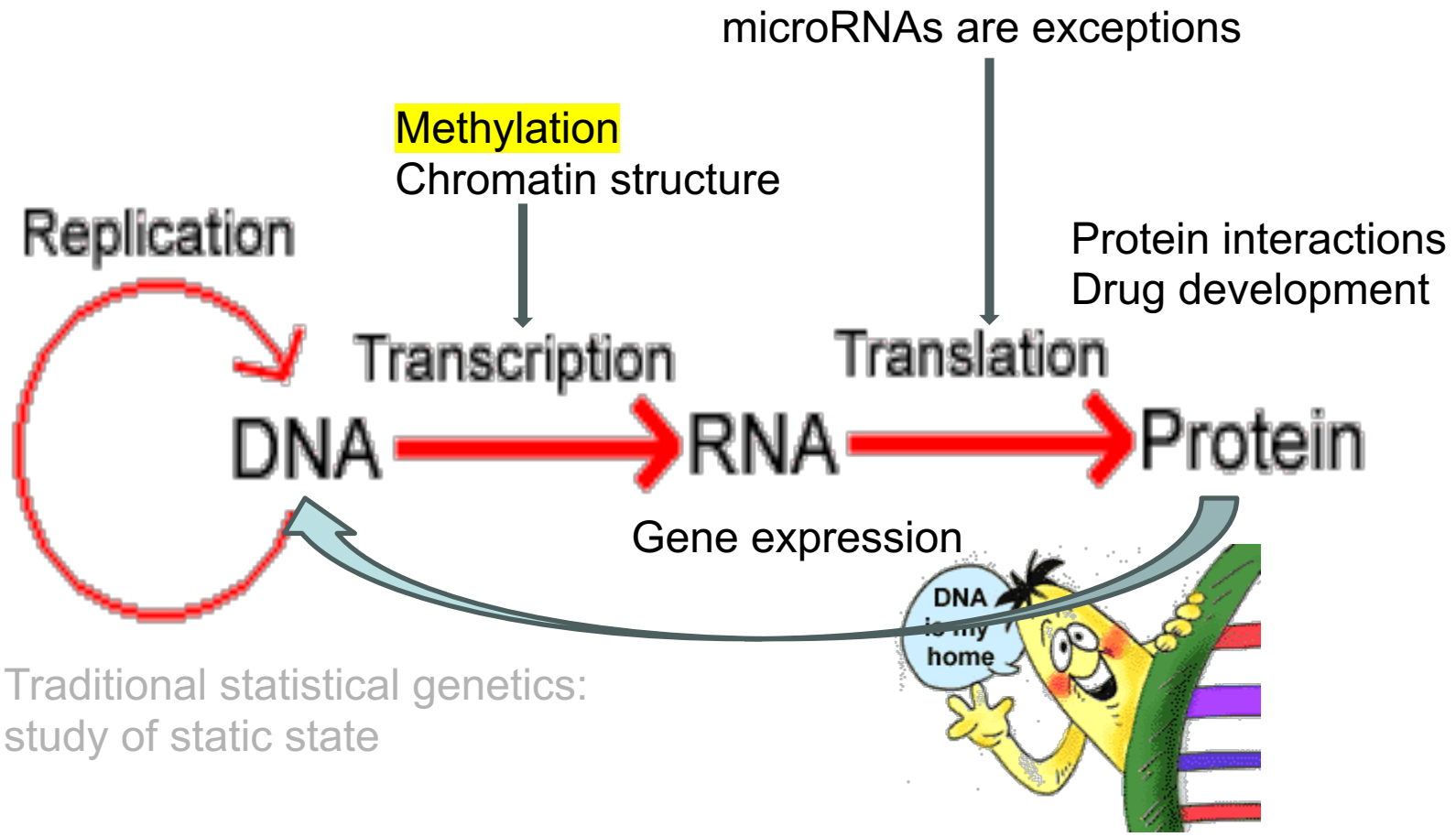
Research Areas (Stat Gene)

- Segregation analysis/Linkage analysis/association studies, including GXG and GXE (disease gene mapping)
  - Single trait/multi-trait; single variant/set-analysis;
  - Individual-level data/summary statistics
- Challenges
  - Missing data (Different platforms; imputation)
  - Non-independence (known relationships, cryptic relatedness)
  - Heterogeneity, population stratification, non-reproducibility
  - Non-linear relationships
  - Complex architecture — polygenic risk score
  - X chromosome

Commonly Used Statistical Methods

- Fisher exact/Chi-square tests for contingency tables
- Generalized linear model (LM and logistic regression)
- Mixed effects models (dependency)
- Dimension reduction /regularization methods
- Kernel-based methods
- Mixture modeling (varying coefficients)
- Multiple testing

The Central Dogma



Traditional statistical genetics:
study of static state

The role of Epigenetics

- A genome may have trillions of cells of different types, each carrying essentially the same genome in its nucleus.

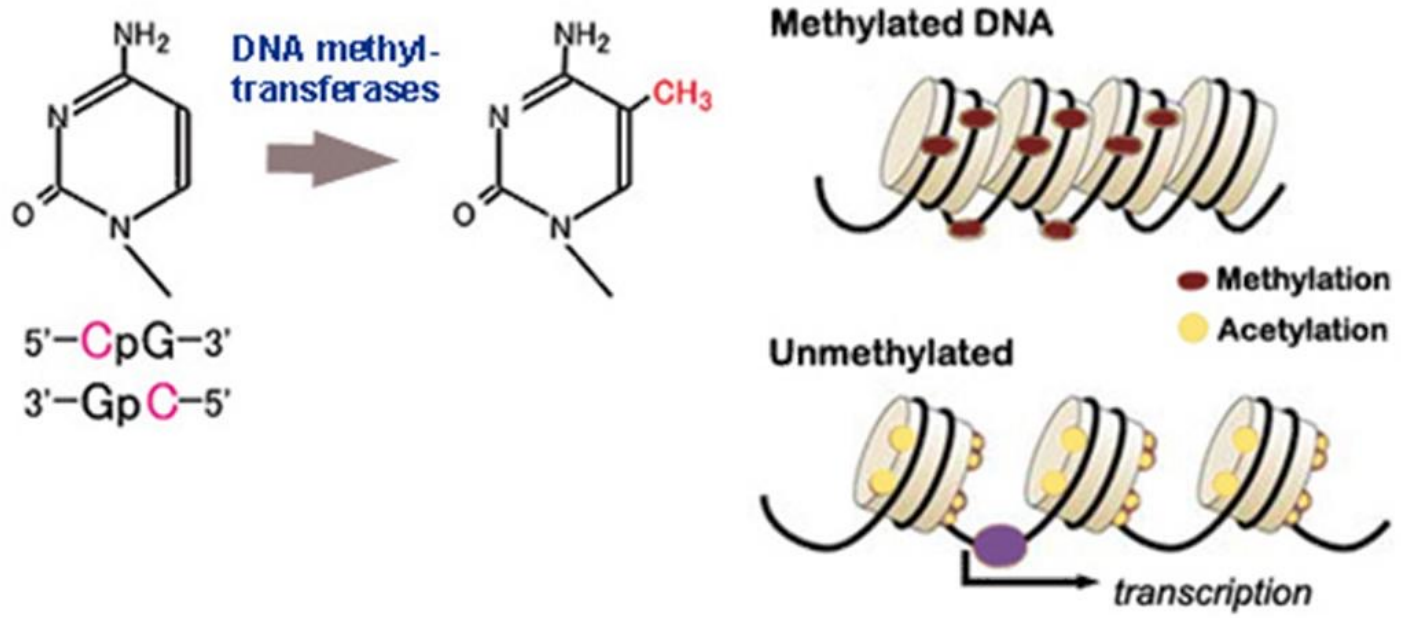
- The differences among different types of cells are determined by how and when different sets of genes are turned on or off.

- The epigenome (second set of genome) controls many of these changes to the genomic functions.
  - These epigenetic marks do not change the underlying DNA sequences.
  - Rather, they change the way that cells use the DNA's instructions.

DNA Methylation

- Methylation is the most well-known and best characterized epigenetic mark in eukaryotes
- First discovered epigenetic mark and remains the most studied.
- Involved in normal cell (embryonic) development – differentiation, genomic imprinting, lyonization and autoimmunity
- Aberrant DNA methylation, especially those occurring in the gene promoters, can lead to disease and malignancy through transcription repression
- DNA methylation plays a crucial role in the development of nearly all types of cancer

DNA Methylation and Disease - simplified cartoon guide



DNA Methylation Types:

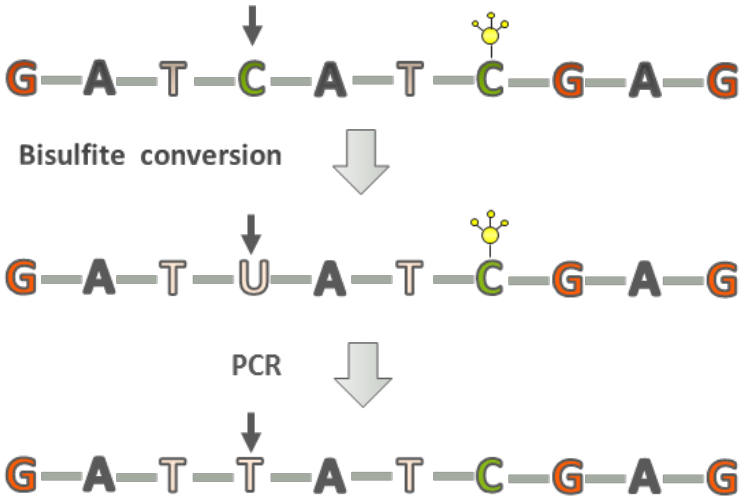
5-methylcytosine (5-mC)

5-hydroxymethylcytosine (5-hmC)

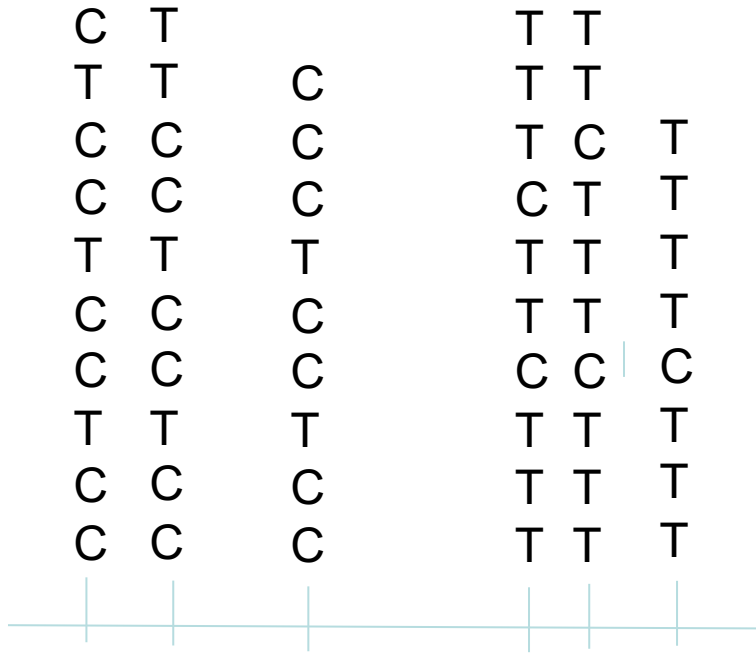
5-formylcytosine (5-fC)

5-carboxylcytosine (5-caC)

BS-seq Data



- Each nucleotide resolution read is a binary variable
- Data at neighboring sites are more similar
- Missing data may exist



Sample Data

- Sequencing data (Binomial data)

chr	sites	g1c1	g1m1	g1c2	g1m2	g1c3	g1m3	g2c1	g2m1	g2c2	g2m2	g2c3	g2m3
chr21	9413763	6	3	9	6	7	2	8	5	13	9	10	10
chr21	9419355	10	7	19	14	10	8	14	14	9	6	8	8
chr21	9420237	4	3	7	7	7	6	6	5	8	6	5	4
chr21	10571455	26	21	12	9	13	5	23	22	14	14	13	13
chr21	10572570	3	1	12	12	15	7	3	2	9	7	3	3
chr21	10576274	5	3	5	5	5	4	5	4	5	5	6	6

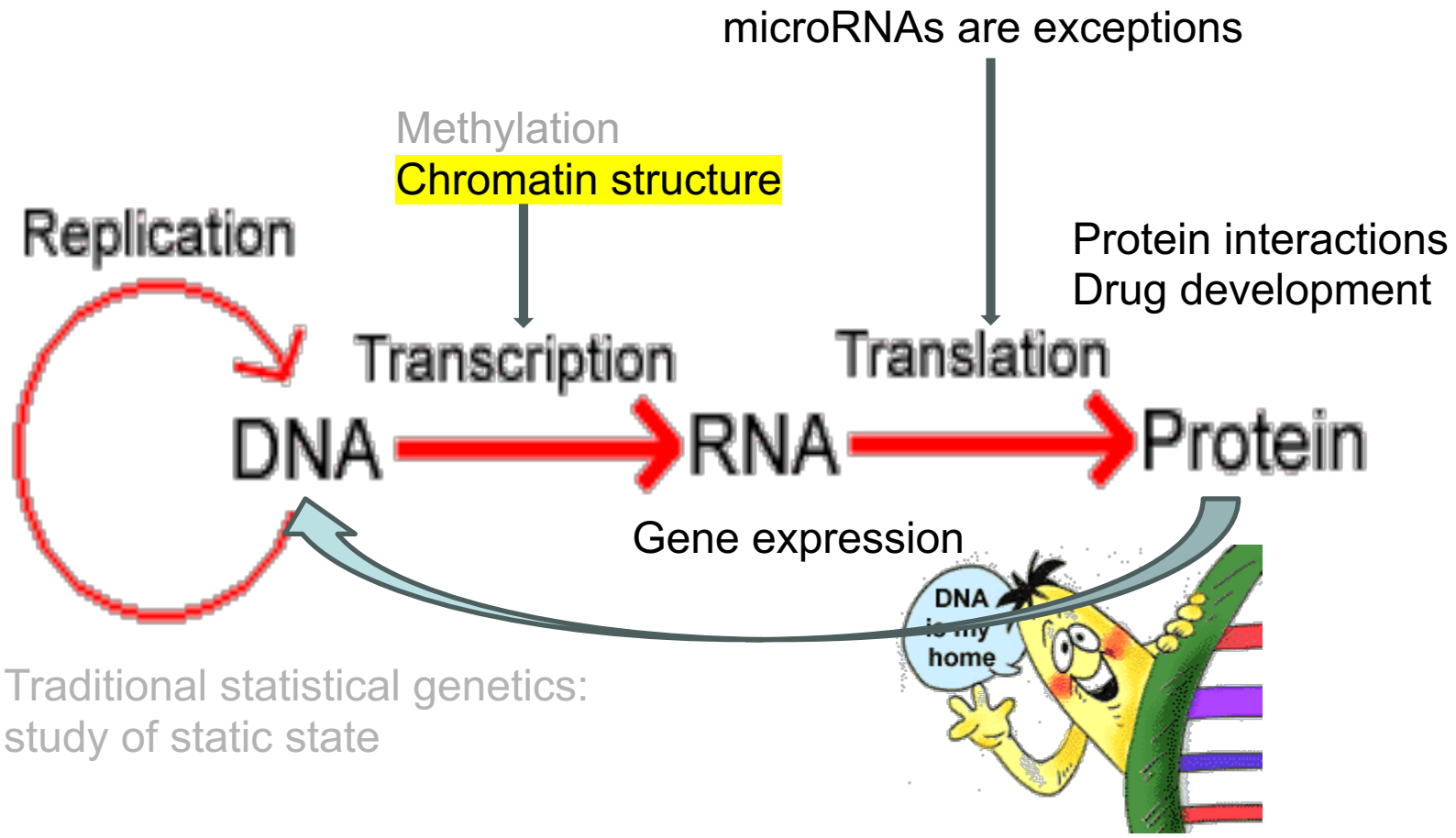
- Microarray data (beta values)

chr	sites	group1_1	group1_2	...	group1_12	group2_1	group2_2	...	group2_12
chr1	29407	0.2409	0.0321	...	0.0051	0.0594	0.3331	...	0.0988
chr1	29425	0.2921	0.0896	...	0.0299	0.6138	0.4514	...	0.1503
chr1	29435	0.2584	0.0241	...	0.0289	0.0002	0.0881	...	0.1354
chr1	1006781	0.9099	0.3094	...	0.2247	0.6834	0.5297	...	0.3761
chr1	1006818	0.9629	0.7839	...	0.9899	0.5172	0.6285	...	0.9832
chr1	1006915	0.0937	0.7839	...	0.9242	0.6673	0.1540	...	0.5892

Commonly Used Methods for BS-seq and Microarray

- Nucleotide resolution binomial data (BS-seq) or beta value (%methylation: Infinium methylation 450K, methylationEPIC)
  - Fisher's exact, logistic regression (BS-seq)
  - Beta-binomial distribution (BS-seq)
  - Smoothing (splines)
  - Empirical Bayes
  - Mixture modeling
  - Bayesian — informative priors

The Central Dogma



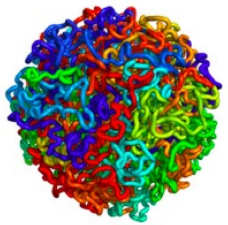
Traditional statistical genetics:
study of static state

Hierarchical Structure of a Genome

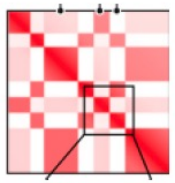
The 1-D genome

Length	Diameter	L reduction
$2.0 \times 10^6 \mu m$	$10.0 \mu m$	2.0×10^5

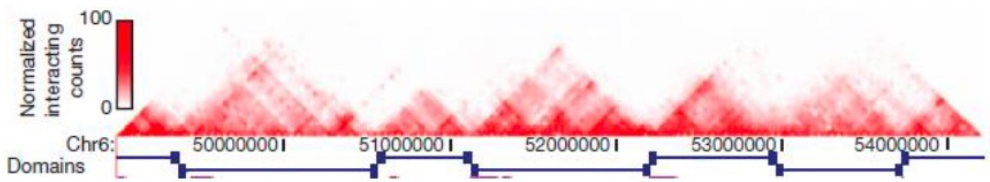
The 3-D genome



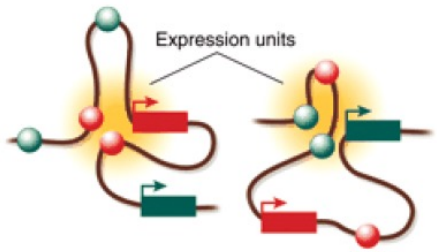
Compartments (A/B: active/inactive)



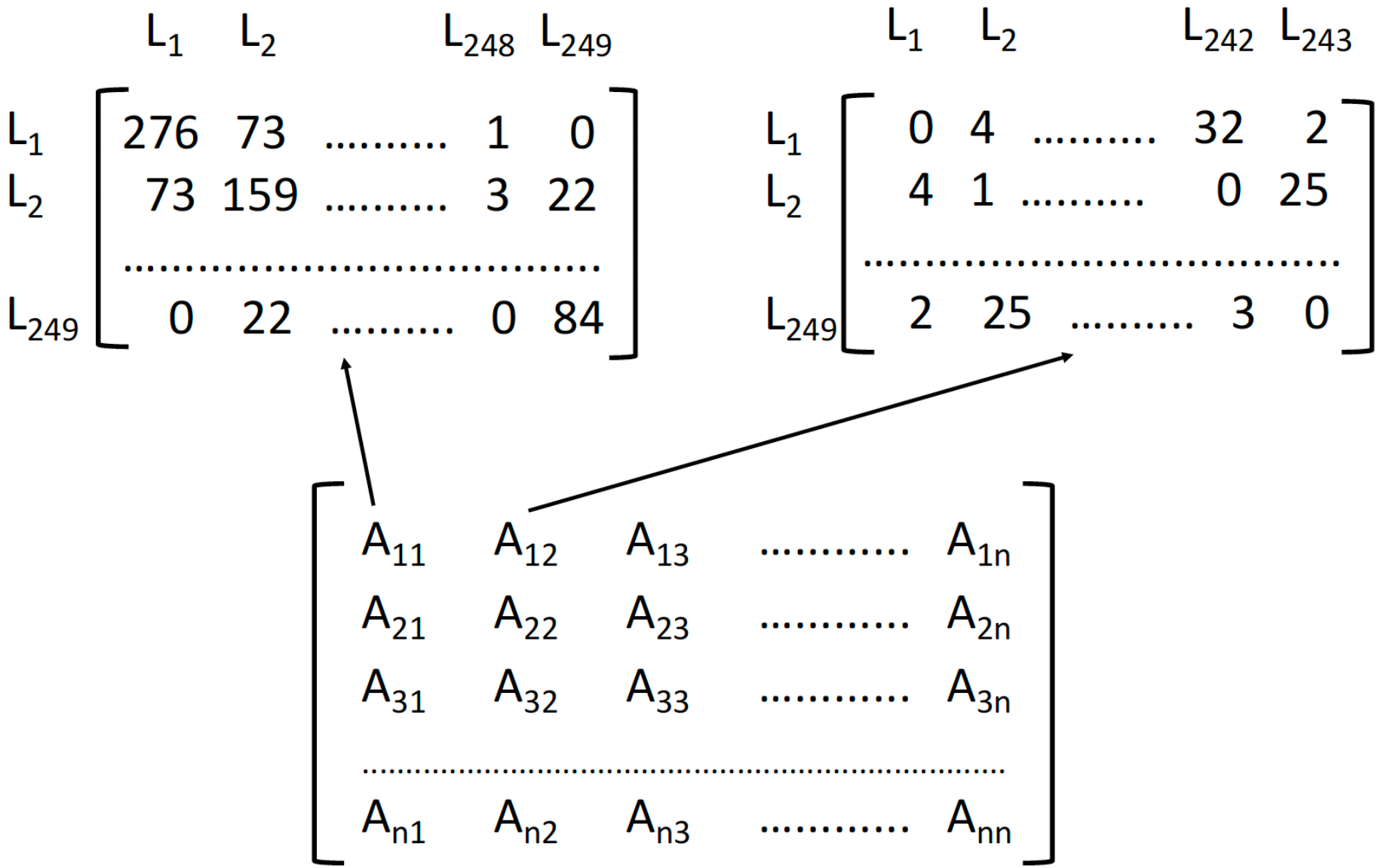
Topologically associated domains (TADs)



Long-range gene regulation (looping)



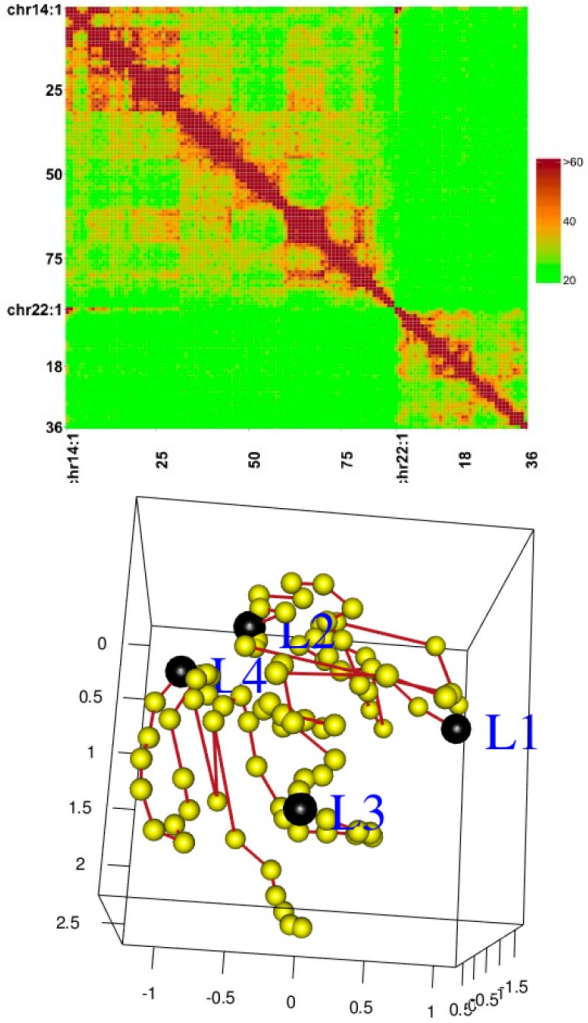
Data Structure



Data Visualization

Lymphoblastoid cell line - 1 Mb resolution

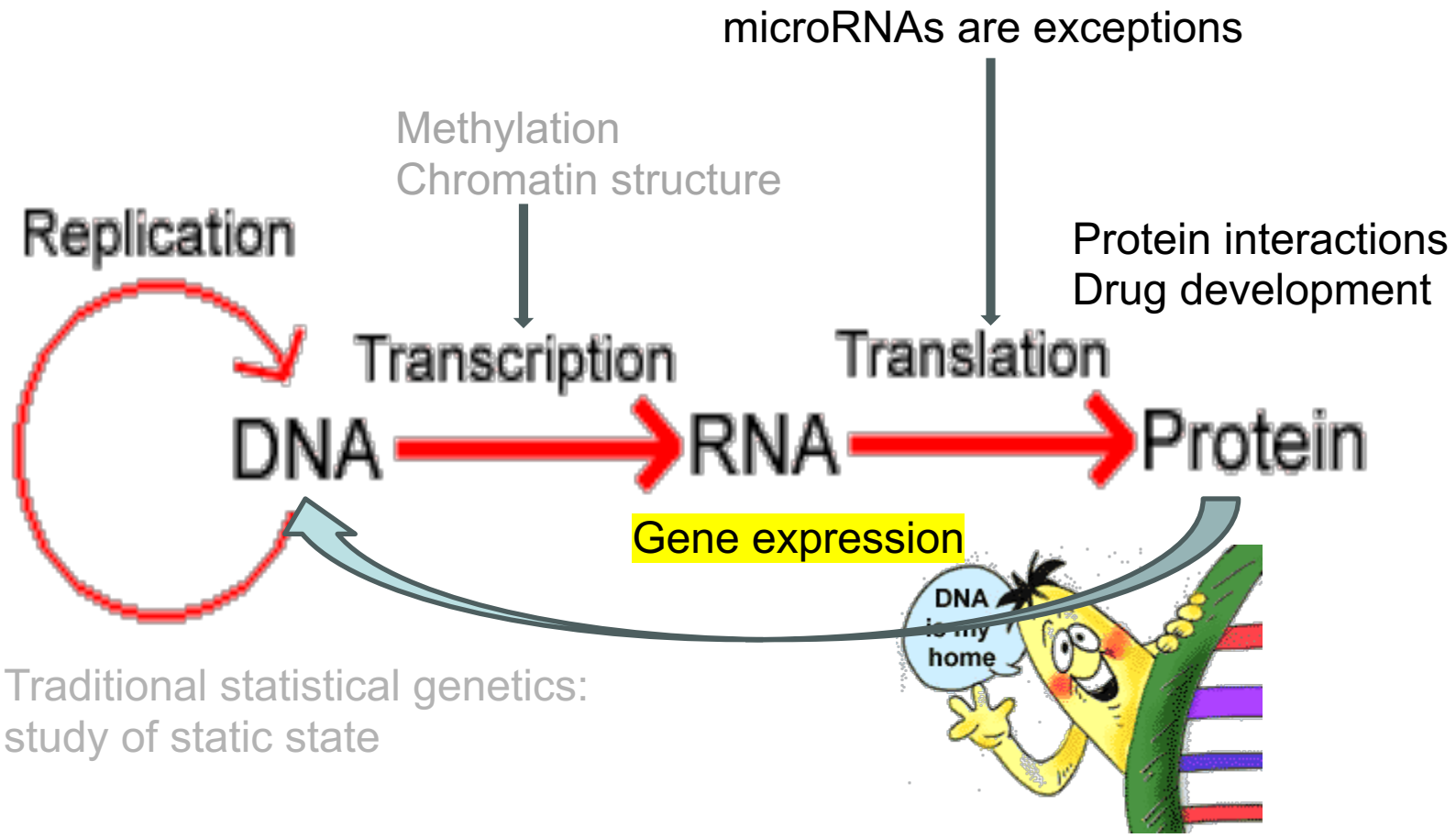
		chr14					chr22				
		l1	l2	...	l88	l89	l1	l2	...	l35	l36
chr14	l1	1079	657	...	0	1	990	218	...	7	1
	l2	657	1413	...	3	0	456	34	...	3	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	l88	0	3	...	733	130	0	1	...	0	2
	l89	1	0	...	130	444	1	1	...	0	4
chr22	l1	990	456	...	0	1	350	80	...	5	1
	l2	218	34	...	1	1	80	846	...	13	2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	l35	7	3	...	0	0	5	13	...	694	88
	l36	1	1	...	2	4	1	2	...	88	308



Challenges, Questions, and Methods

- Challenges in analyzing Hi-C data
  - Dependency, overdispersion, sparsity
- Scientific questions of interest (from statistical perspective)
  - 3D structure and variations
  - Significant interactions — peak detection
  - Clustering and subtype discoveries
- Commonly used methods/models
  - Zero-inflated and zero-truncated models
  - Negative binomial/Poisson-Gamma modeling
  - Random effects
  - Optimization, Bayesian, empirical Bayes

The Central Dogma



Traditional statistical genetics:
study of static state

Data for Gene Expression Analysis

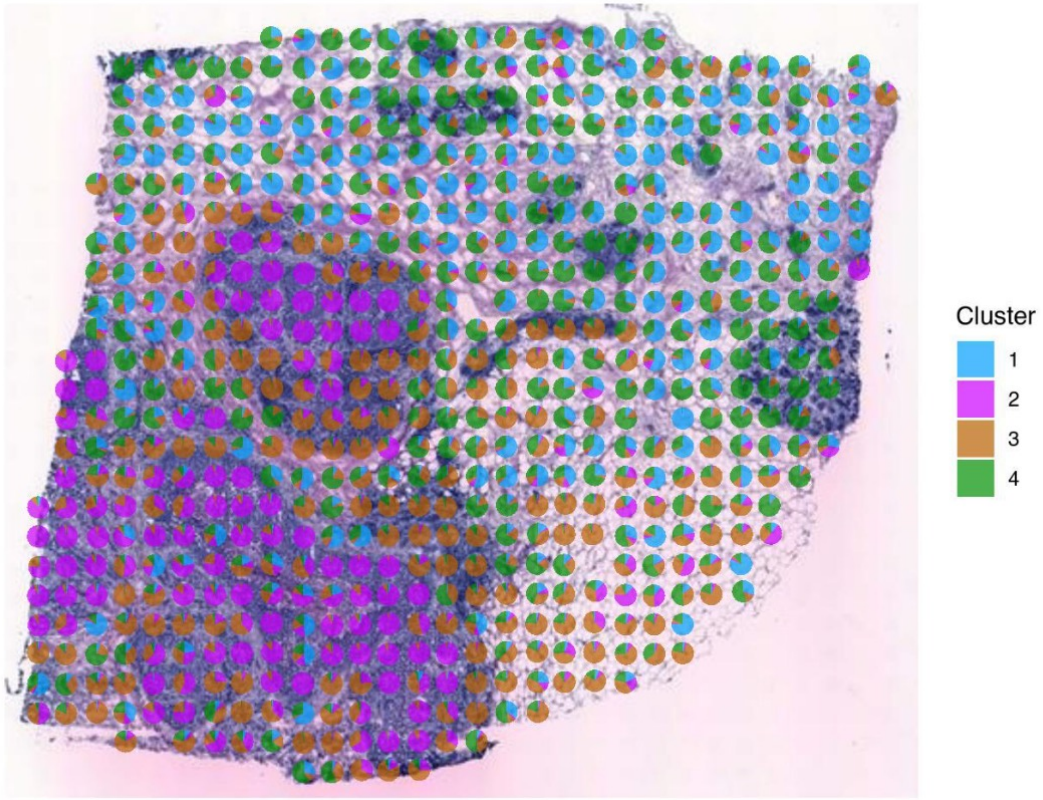
- Bulk RNA-seq data

	Gene.ID	Gene.Name	SRR975551	SRR975552	SRR975553	SRR975554
1	ENSG00000000003	TSPAN6	6617	1352	1492	3390
2	ENSG00000000005	TNMD	69	1	20	23
3	ENSG00000000419	DPM1	2798	714	510	1140
4	ENSG00000000457	SCYL3	486	629	398	239
5	ENSG00000000460	C1orf112	466	342	73	227
6	ENSG00000000938	FGR	75	95	158	107

Spatial Transcriptomics Data

	3x34	3x30	3x31	3x32	3x33
ACTB	2.511	2.116	2.910	3.792	2.432
CD74	1.744	2.323	1.666	3.061	0.000
CFL1	0.000	2.116	1.666	1.909	2.432
CST3	3.009	2.664	0.000	3.061	2.925
ERBB2	1.744	3.461	0.000	3.473	3.584

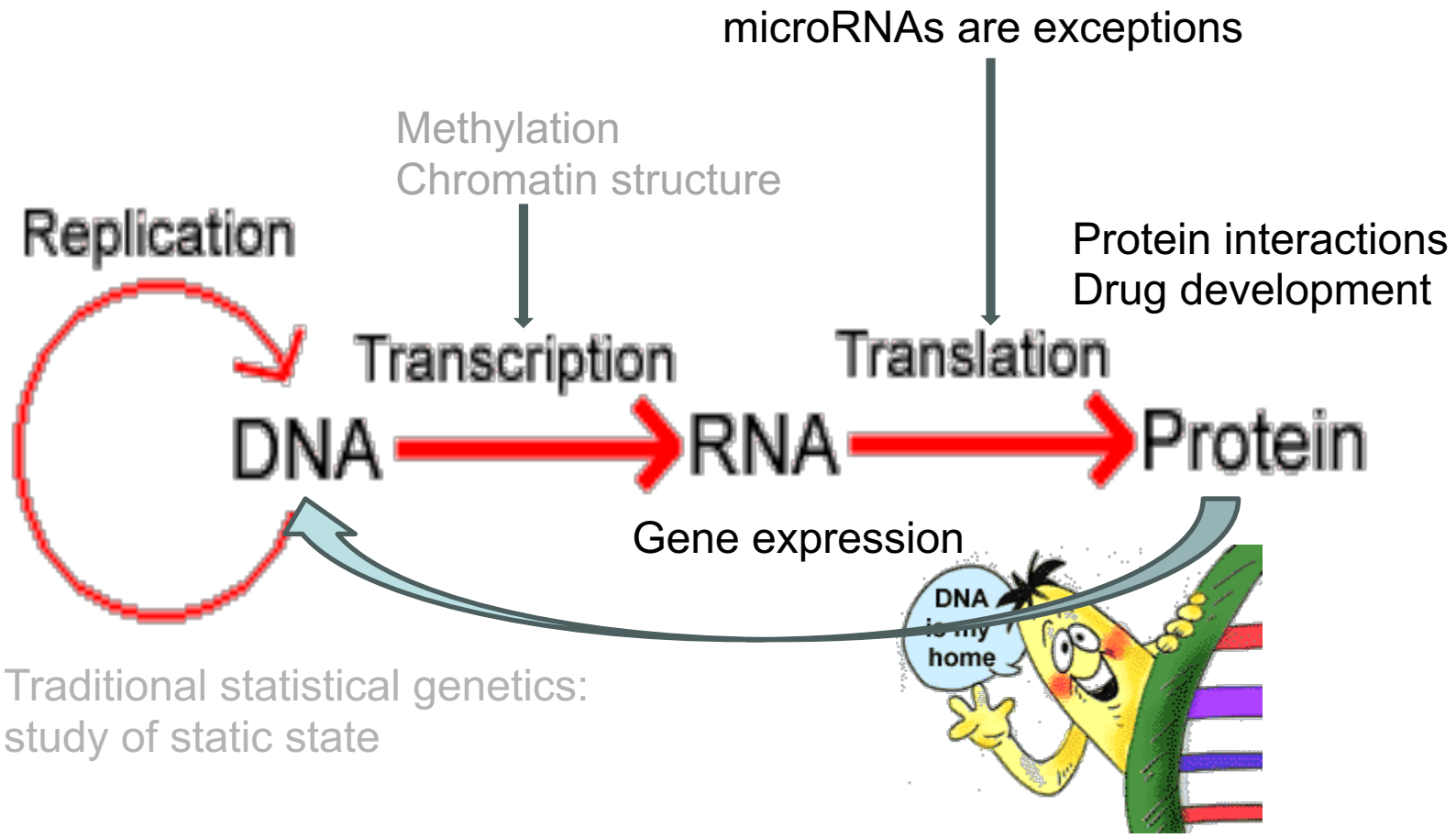
	x	y
3x29	29.36387	422.9227
3x30	29.05259	437.0339
3x31	29.10447	453.1166
3x32	29.20823	468.3692



Challenges, Questions, and Methods

- Challenges in analyzing transcriptomics data
  - Dependency, overdispersion, sparsity (scRNA-seq)
- Scientific questions of interest (from statistical perspective)
  - Differential expression
  - Clustering (single cells, and ST)
  - Cell decomposition (ST)
- Commonly used methods/models
  - Normalization/Preprocessing (quantile normalization)
  - Zero-inflated Negative binomial/Poisson-Gamma modeling
  - Bayesian, empirical Bayes
  - Multiple testing

The Central Dogma



Traditional statistical genetics:
study of static state

Metagenomics,16S rRNA, and Shotgun Sequencing

- Metagenomics is the study of genes from multiple genomes altogether.
- Applied to samples collected from the environment without the need for isolation and lab cultivation of individual species.
- 16S rRNA studies revel a profile of diversity in a natural sample, and that a vast majority of microbial biodiversity had been missed by cultivation-based methods.
- Shotgun sequencing can consider the entire samples (all genes) but can only construct Operational Taxonomic Units (OTUs).

Data and Some Research Questions

- Metagenomic count matrix – row samples, column OTUs (taxa)
- Compositional in nature
- Sample may be placed in a phylogenetic tree (16S rRNA)
- Dissimilarity measure between samples or taxa (OTUs) — UniFrac distance
- Study microbiome communities
- Relate similarity in composition of microbes to similarity in a trait (like genetic association studies)

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$