Skoltech
Skolkovo Institute of Science and Technology

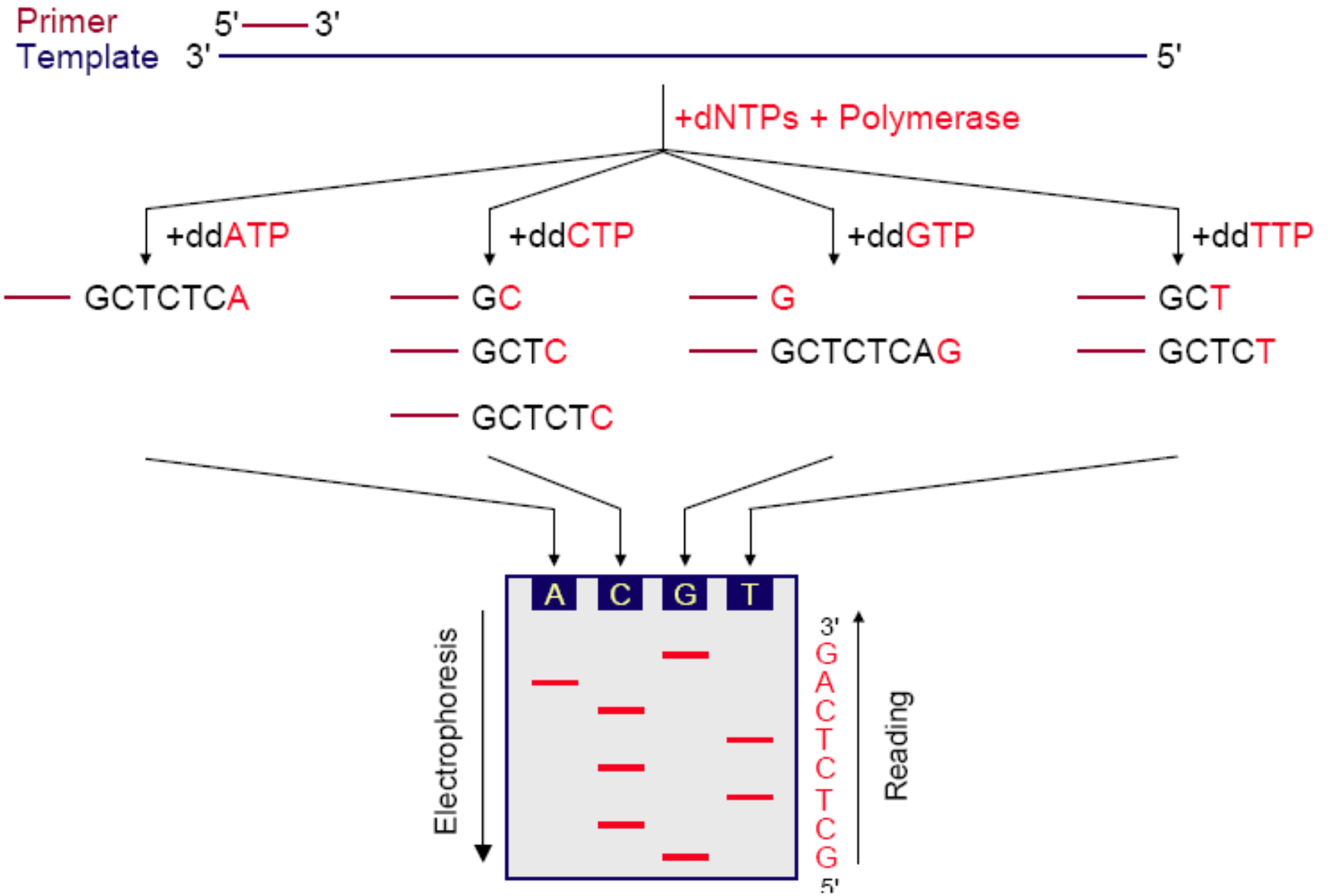
МЕТОДЫ СЕКВЕНИРОВАНИЯ НОВОГО ПОКОЛЕНИЯ

Дмитрий Первушин

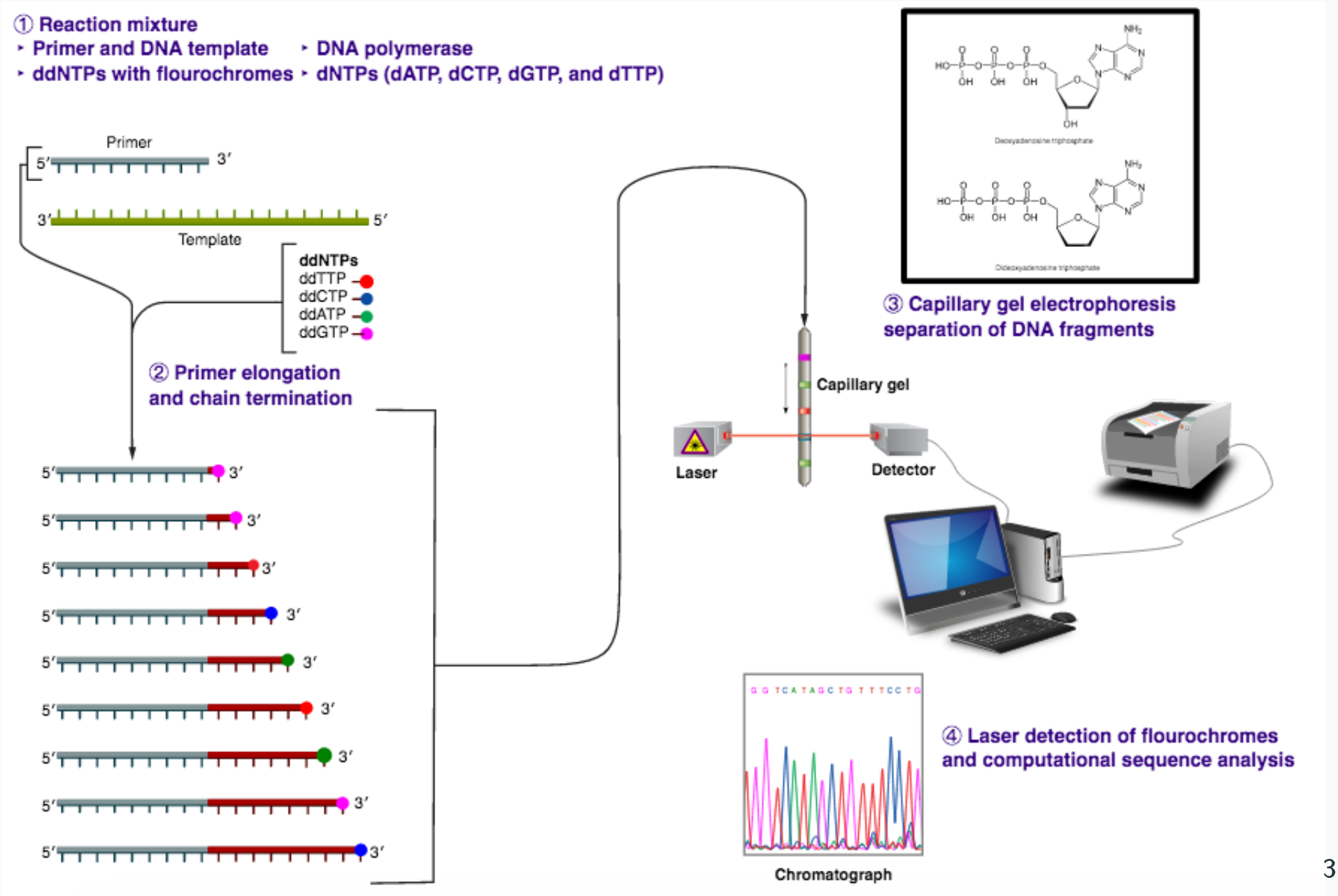
19 апреля 2017г.

Сколковский Институт Науки и Технологии
Московский государственный университет им. МВ Ломоносова
Государственный Университет - Высшая Школа Экономики
Centre Regulaciò Genómica, Parc Reserca Biomèdica Barcelona
Universitat Pompeu Fabra

Метод дидезоксинуклеотидов (Нобелевская премия 1980)



Секвенирование по Сангеру



Suggested Reading

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Coming of age: ten years of next-generation sequencing technologies

Sara Goodwin[1], John D. McPherson[2] and W. Richard McCombie[1]

Abstract | Since the completion of the human genome project in 2003, extraordinary progress has been made in genome sequencing technologies, which has led to a decreased cost per megabase and an increase in the number and diversity of sequenced genomes. An astonishing complexity of

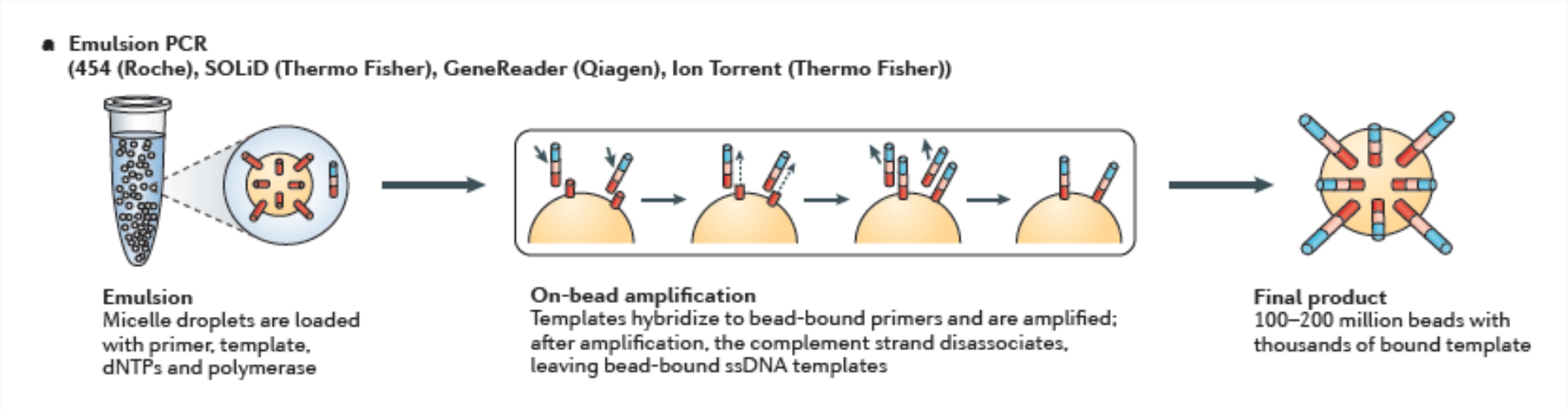
Массовое секвенирование коротких фрагментов

Секвенирование коротких ридов

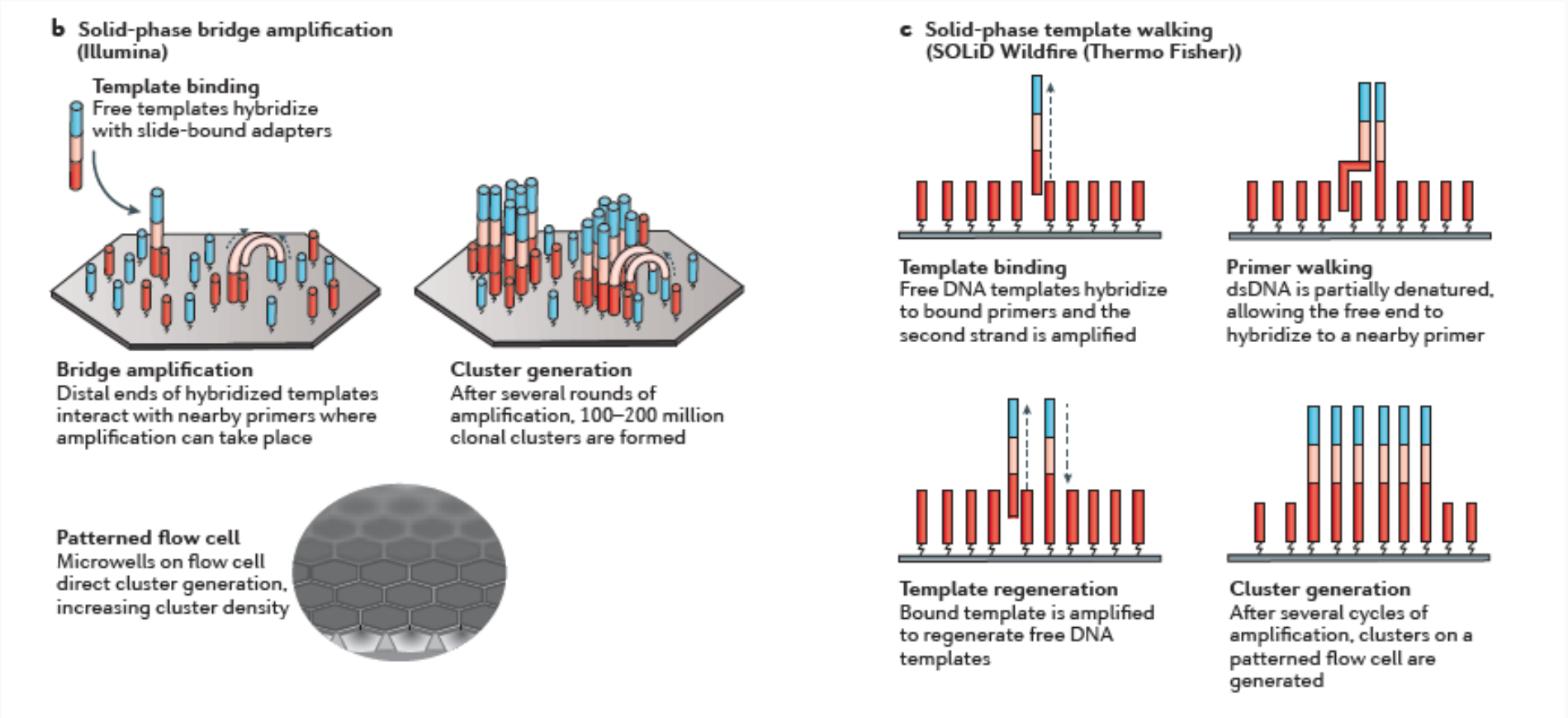
- Амплификация шаблона : цель – получить много тысяч идентичных молекул ДНК в одной микрообласти пространства для того, чтобы можно было регистрировать сигнал и отличать его от шумае
  - Эмульсионная ПЦР
  - Твердофазная генерация кластеров
  - Генерация кластеров в растворе
- Секвенирование : локальное секвенирование амплифицированных шаблонов
  - Секвенирование через лигирование (SBL)
  - Секвенирование через синтез (SBS)
    - Циклическая обратимая терминация (CRT)
    - Однонуклеотидное приращение (SNA)

Эмульсионная ПЦР

- К ДНК пришивается адаптер
- К наночастицам пришиваются олигонуклеотиды, комплементарный адаптеру
- С каждой наночастицей связывается только одна молекула ДНК
- Раствор, содержащий реагенты ПЦР, превращается в обращенную эмульсию так, что каждая наночастица попадает в одну мицеллу
- В процессе ПЦР на наночастице образуется колония идентичных молекул
- Эмульсия обращается, частицы осаждаются

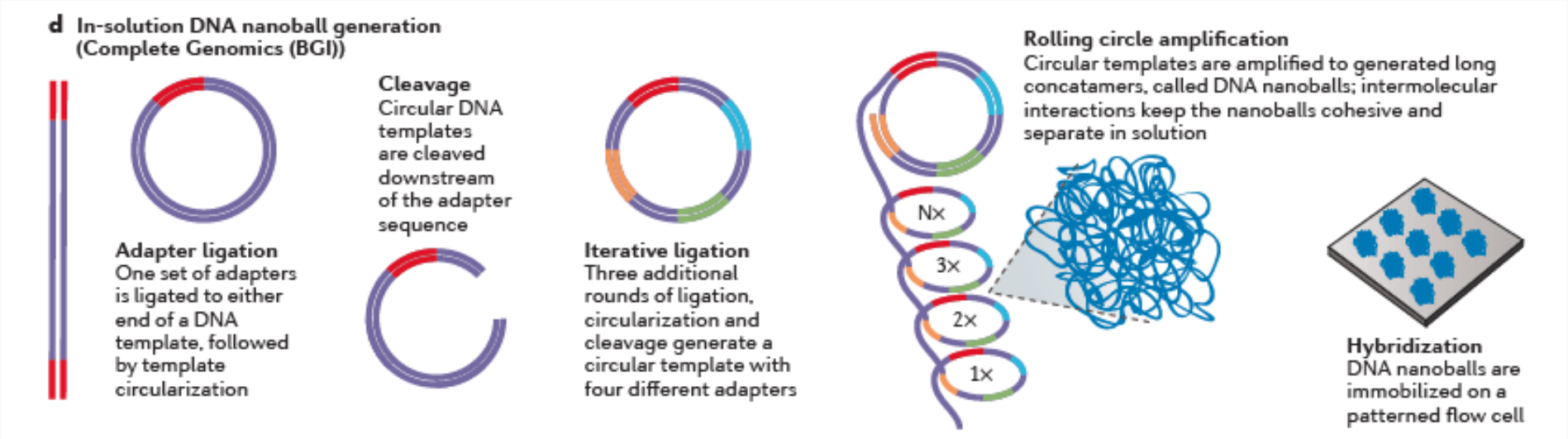


Твердофазная генерация кластеров

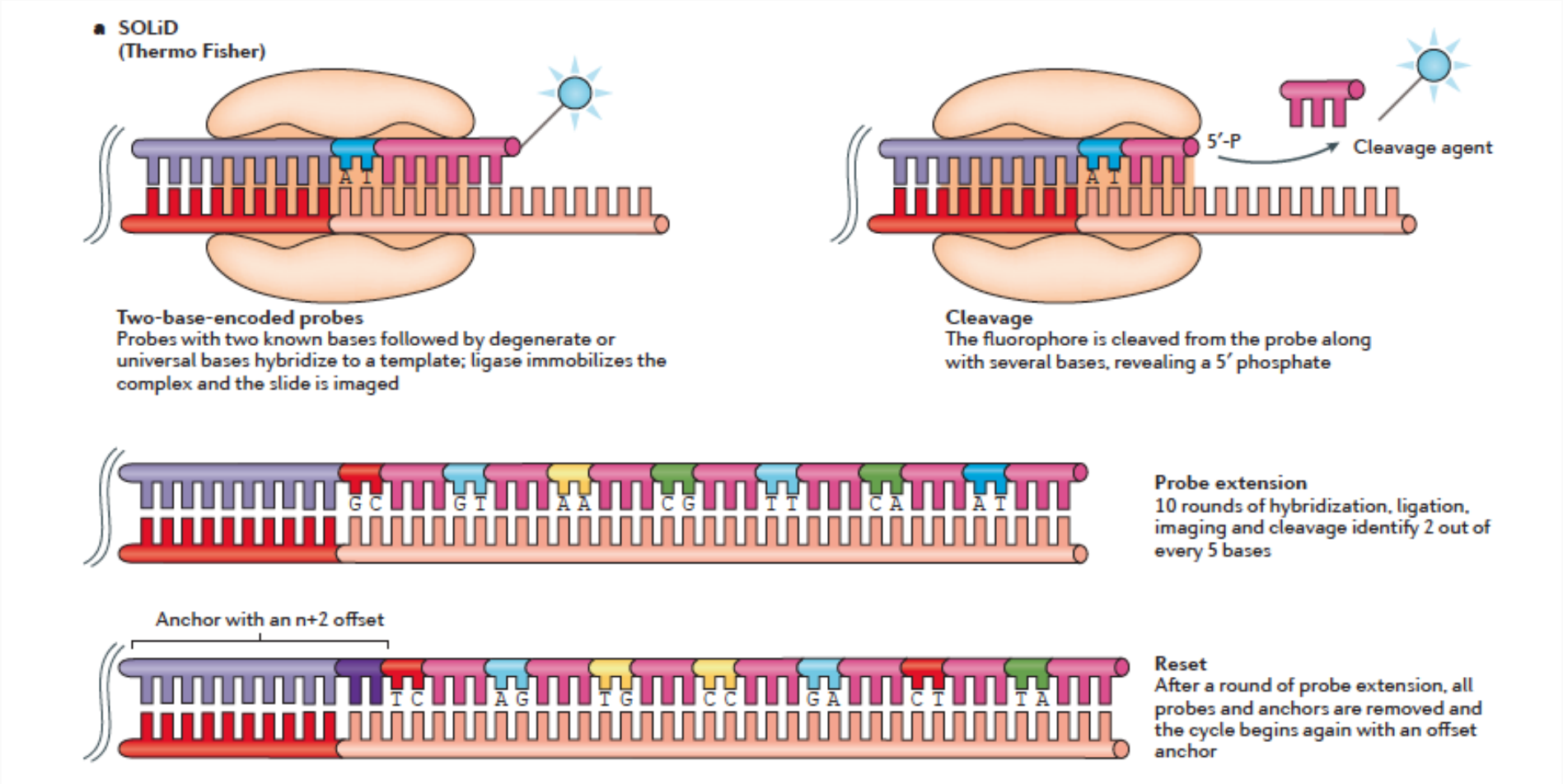


ДНК наношары (Пекинский Институт Биоинформатики)

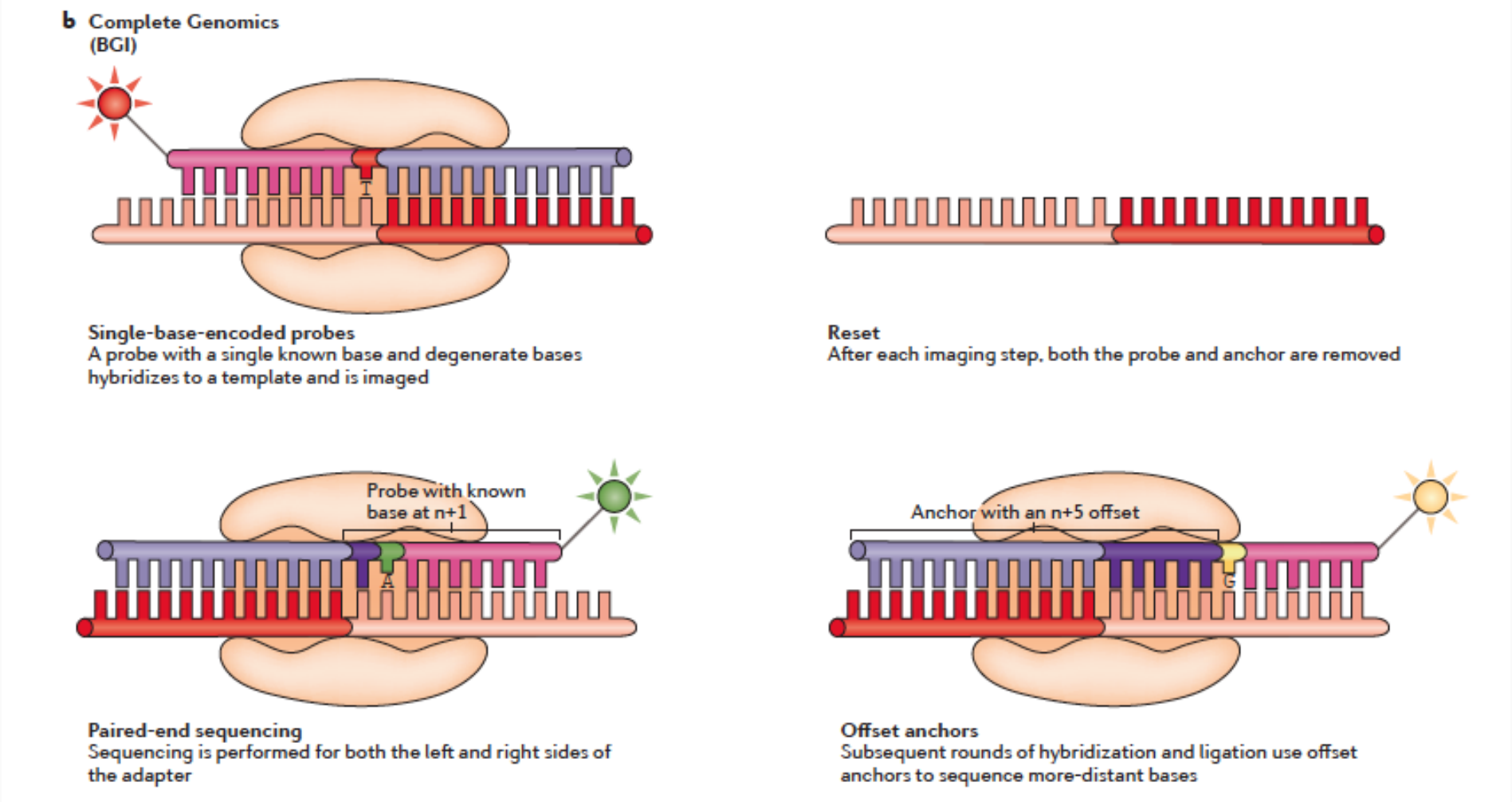
- К ДНК пришивается адаптер и молекула закольцовывается
- Затем ДНК разрезается в другом месте и процесс повторяется несколько раз
- В процессе амплификации катящегося кольца образуются длинные конкатамеры



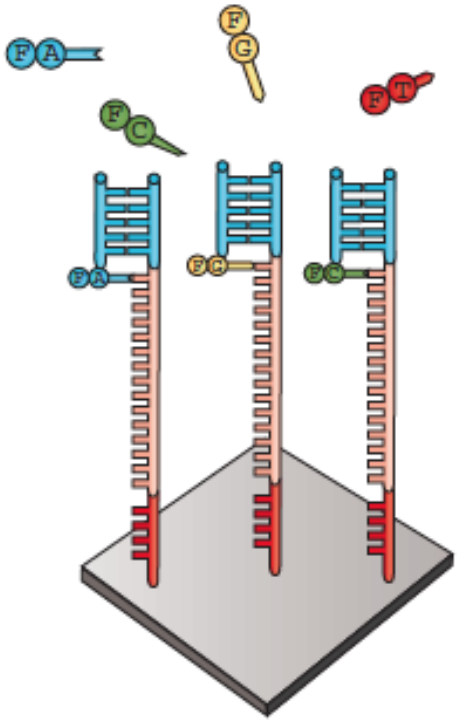
Секвенирование через лигирование (SOLiD)



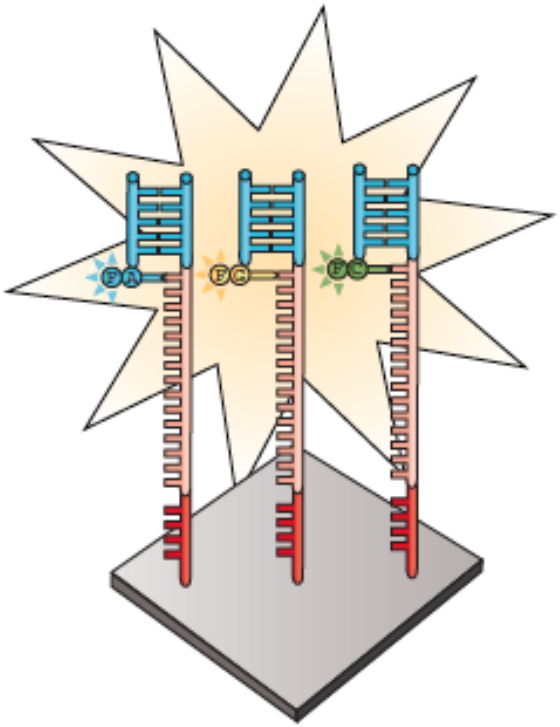
Секвенирование через лигирование (Complete Genomics)



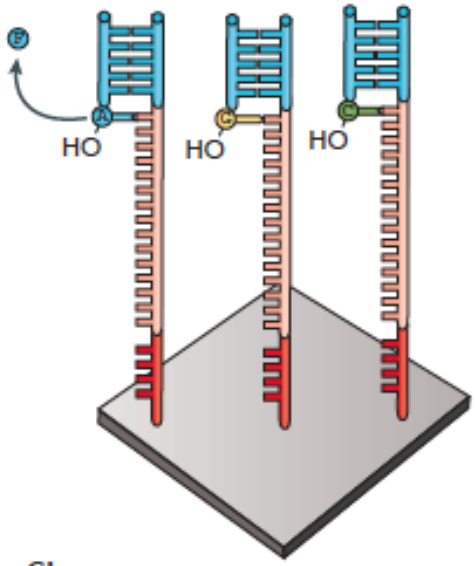
Секвенирование через синтез CRT (Illumina)



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

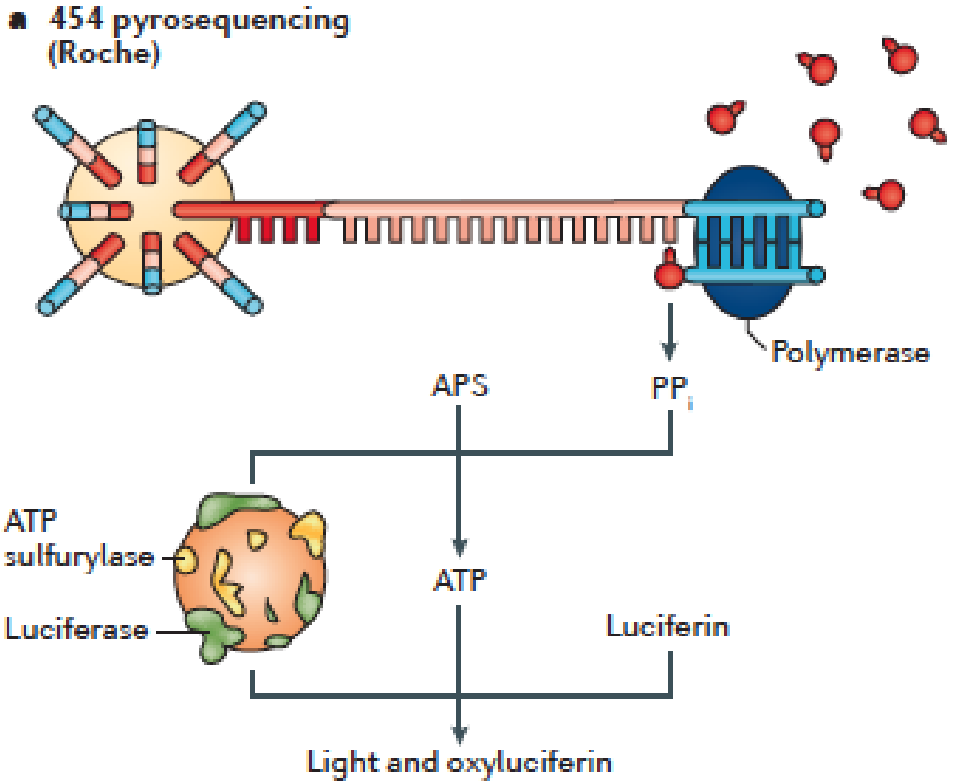


Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.



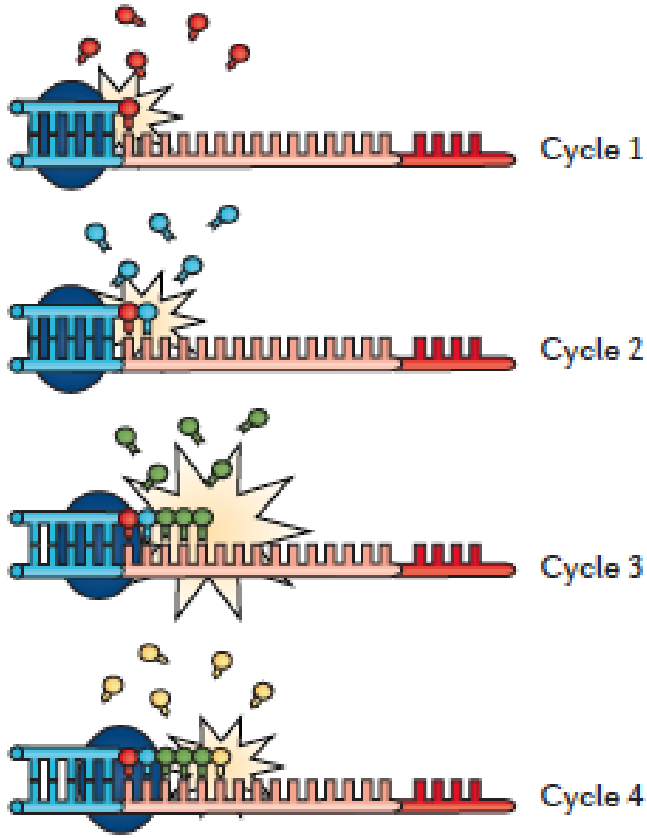
Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Секвенирование через синтез SNA (Roche 454)

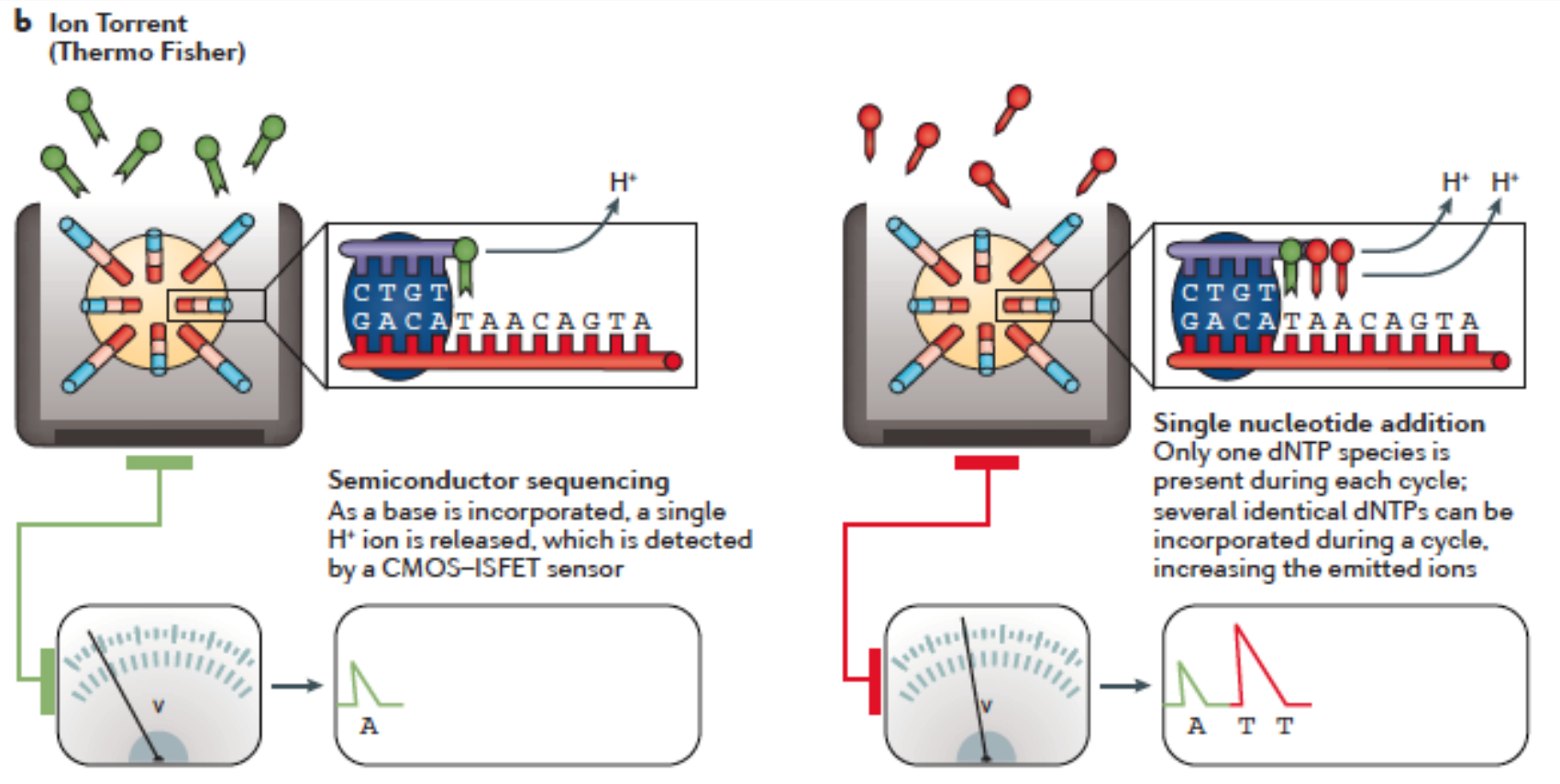


Pyrosequencing
As a base is incorporated, the release of an inorganic pyrophosphate triggers an enzyme cascade, resulting in light

Single nucleotide addition
Only one dNTP species is present during each cycle; multiple identical dNTPs can be incorporated during a cycle, increasing emitted light



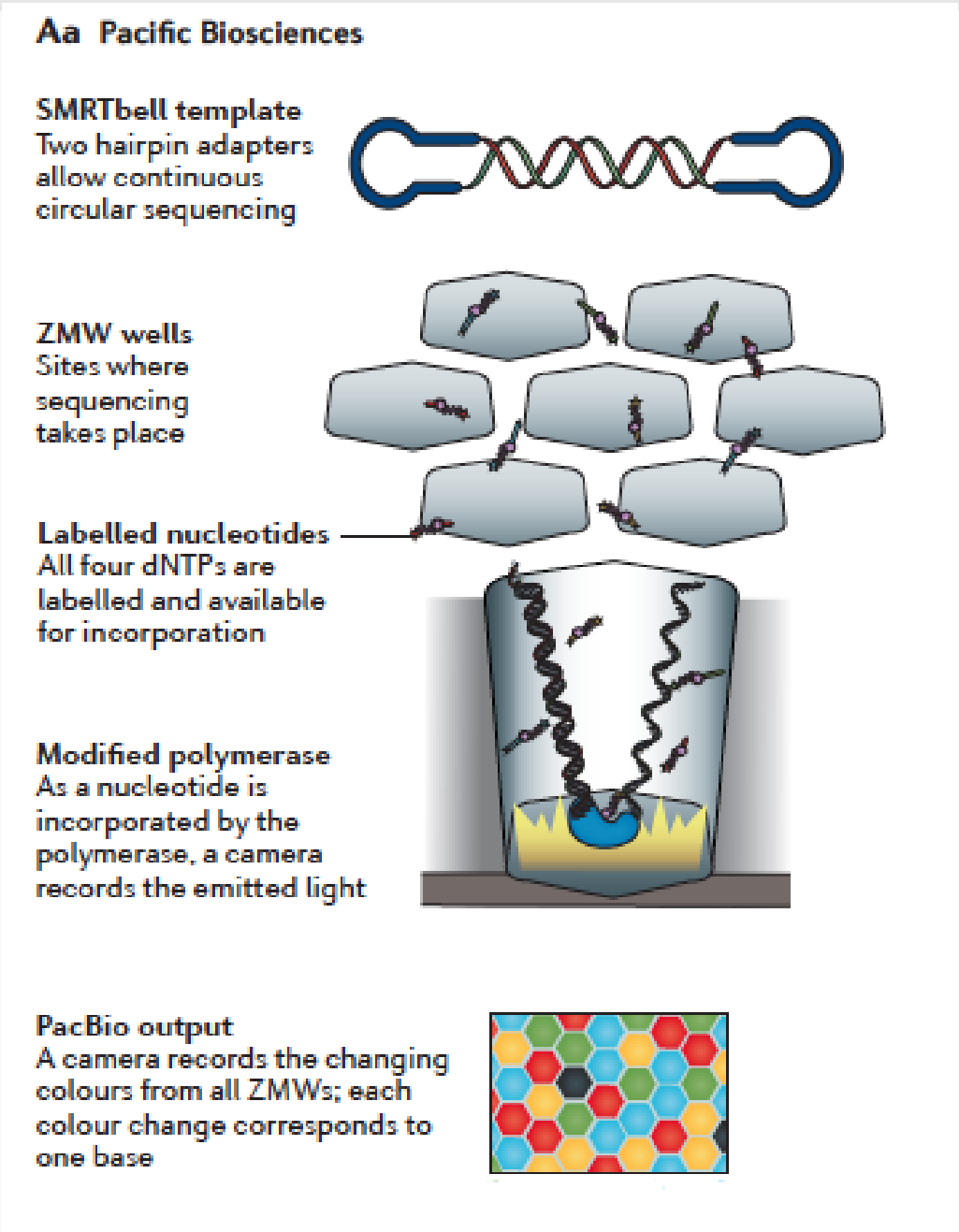
Секвенирование через синтез SNA (Ion Torrent)



- комплементарный металло-оксидный полупроводник (CMOS)
- ион-чувствительный полевой транзистор (ISFET)

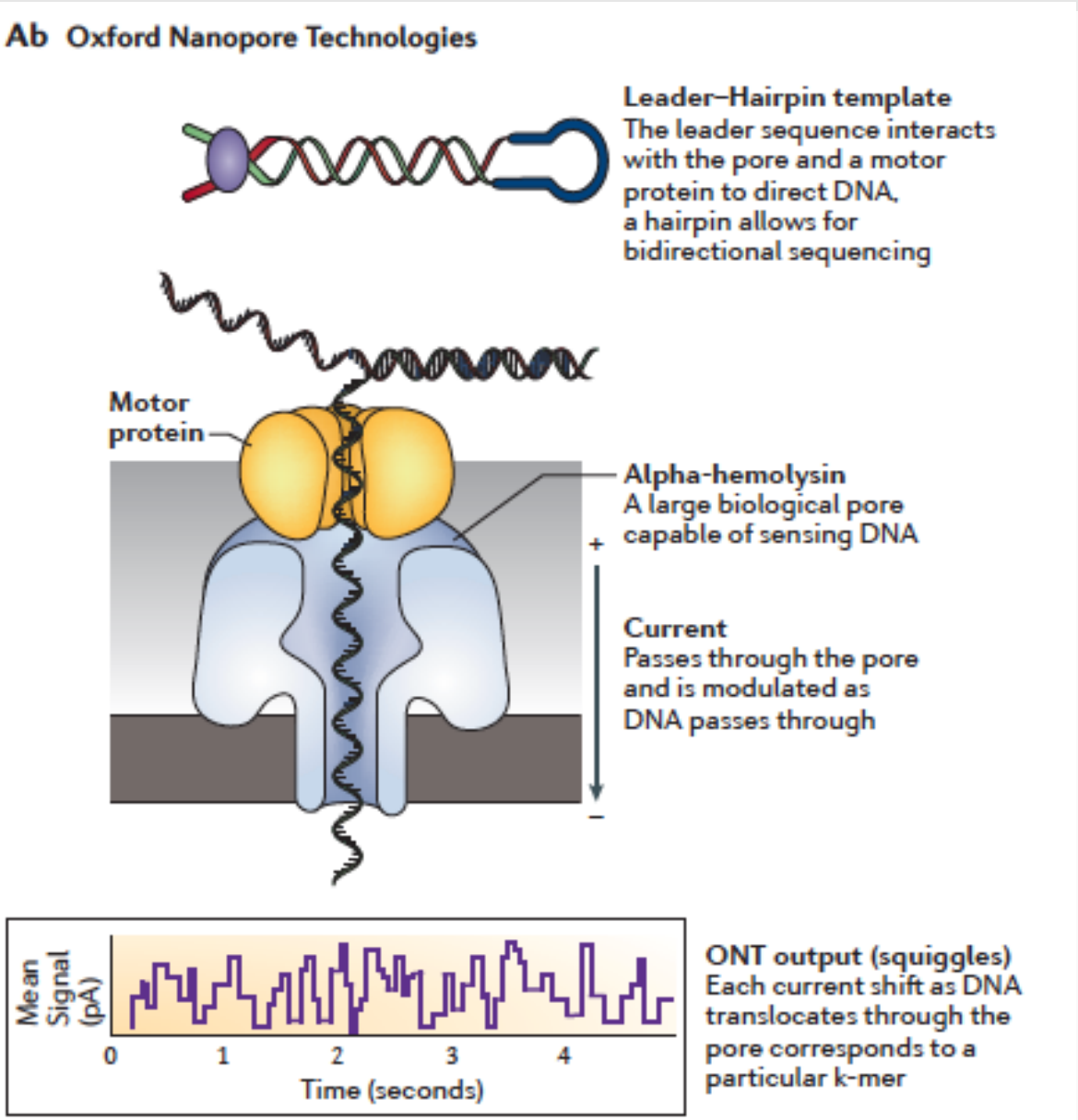
Секвенирование одиночных длинных ДНК

Секвенирование одиночных длинных ДНК (PacBio)



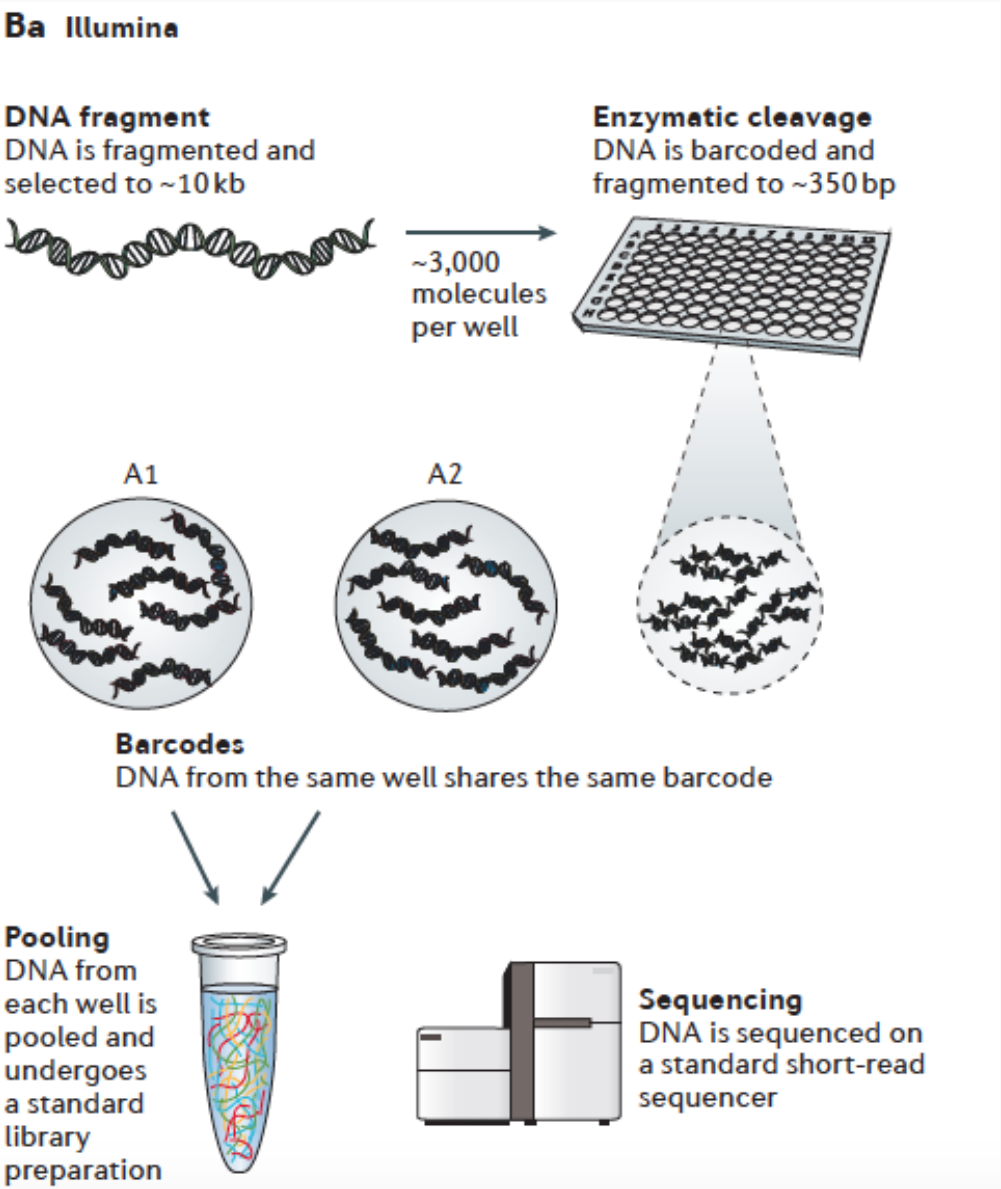
- Длинная ДНК лигируется к закольцованным адаптерам
- Модифицированная полимераза прикрепляется к дну микро-колодца ZMT (zero-mode waveguide)
- Полимераза отрезает флуорофор после включения нуклеотида
- Каждая молекула прочитывается многократно благодаря кольцевой структуре

Секвенирование одиночных длинных ДНК (Oxford Nanopore)



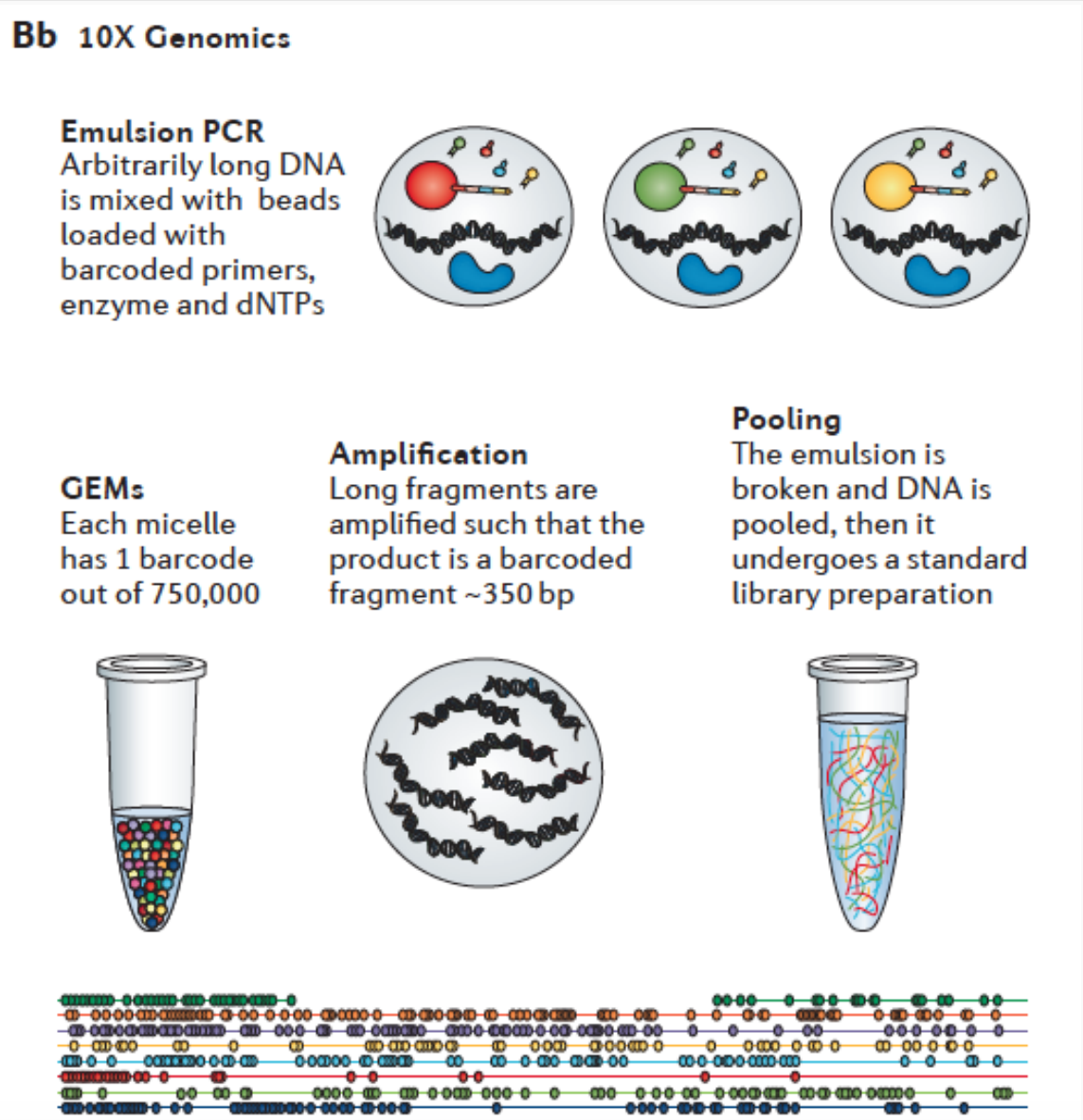
- Изменение в проводимости ионного канала зависят от того, какая последовательность находится в поре, что можно интерпретировать как k-мер

Синтетические длинные риды (Illumina Moleculo)



- ДНК фрагментируется до 8–10 килобаз и переносятся на подложку с микроколодцами
- Каждый фрагмент разрезается до ∼350 нт и к нему пришивается штрихкод, в каждом колодце свой
- ДНК с штрихкодами объединяются и секвенируются по обычному протоколу

Синтетические длинные риды 10X Genomics



- Фрагменты ДНК ∼100 килобаз помещаются в мицеллы, каждая мицелла содержит адаптер и уникальный штрихкод
- В каждой мицелле короткие фрагменты ДНК амплифицируются и помечаются штрихкодами, каждый из которых соответствует длинному фрагменту

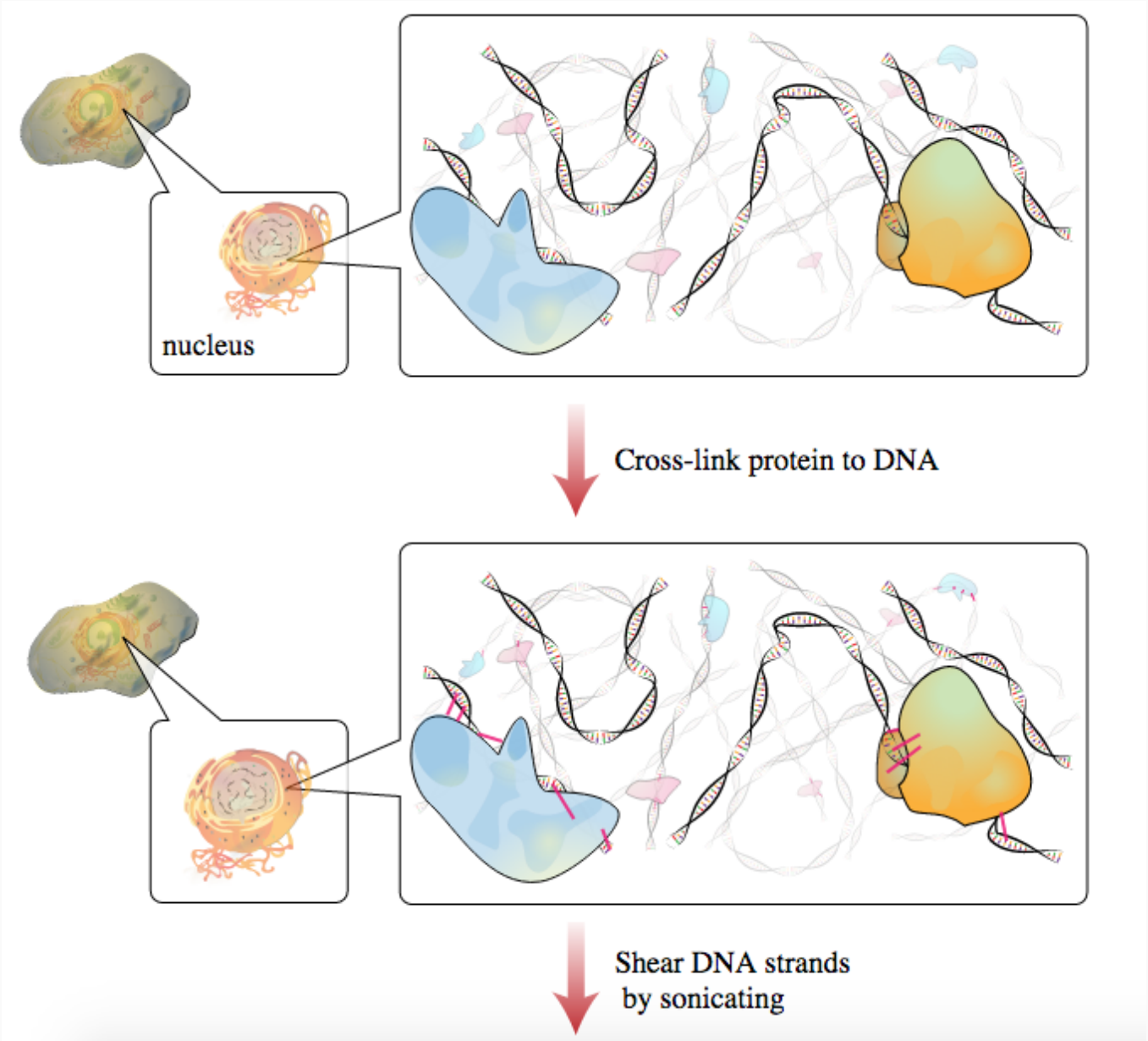
Приложения

Pull-down эксперименты : иммунопреципитация

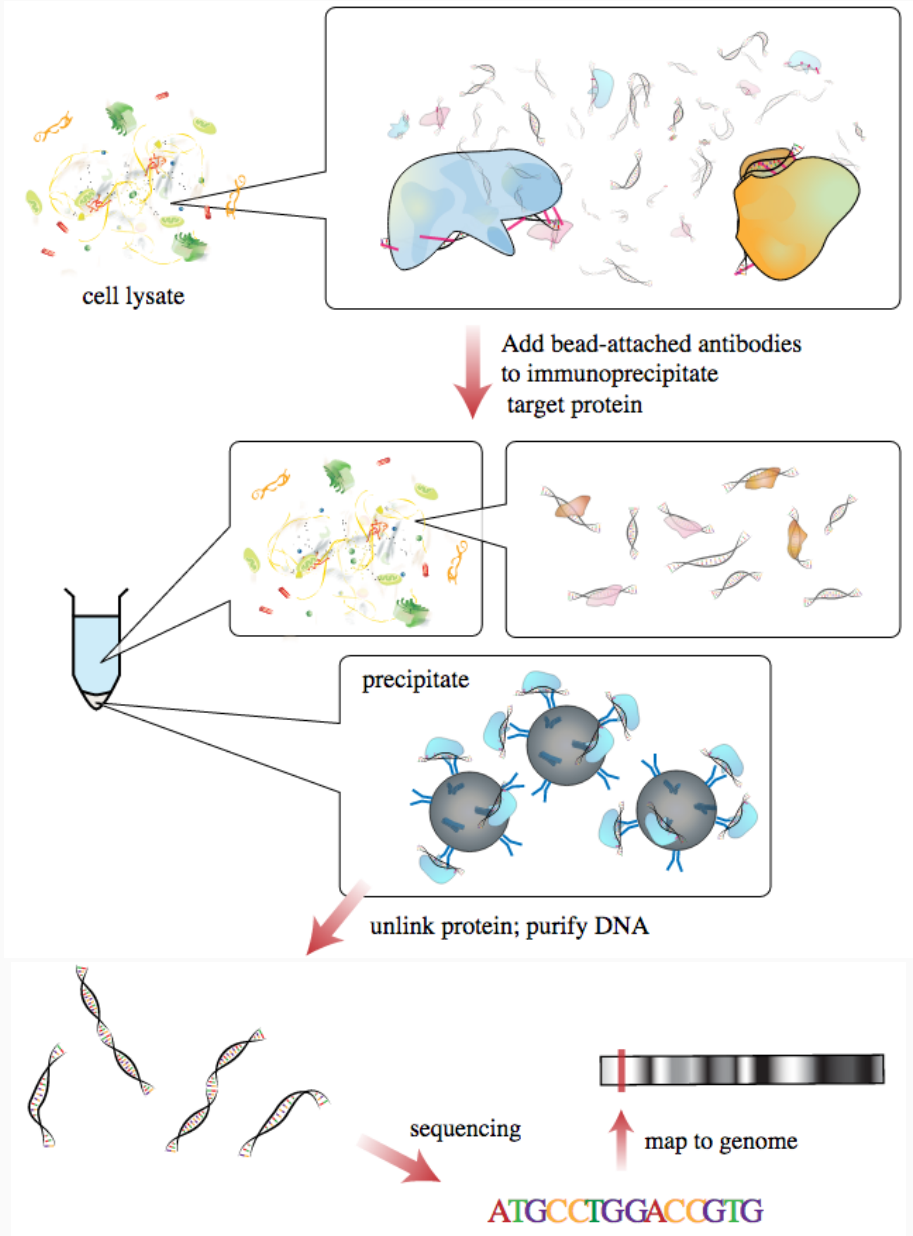
Иммунопреципитация (IP) – техника осаждения антигена путем специфическгого связывания с антителом

1. Иммунопреципитация отдельных белков (IP)
2. Иммунопреципитация белковых комплексов (Co-IP)
3. Иммунопреципитация хроматина (ChIP)
4. Иммунопреципитация рибопротеиновых комплексов (RIP, eCLIP)
5. Рибосомный профайлинг
6. Бромоуридиновый профайлинг (Bru-seq)
7. Бромоуридиновый пульсовый профайлинг

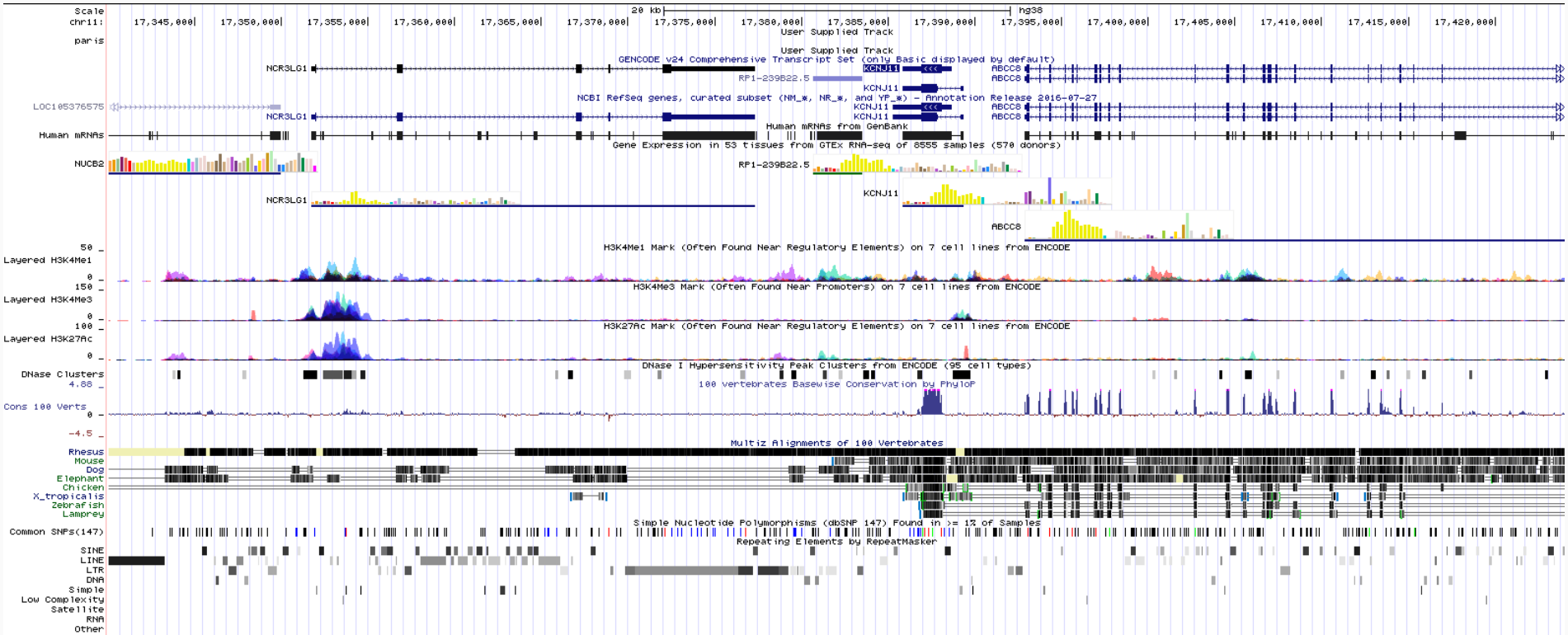
Иммунопреципитация (ChIP)



Иммунопреципитация (ChIP)



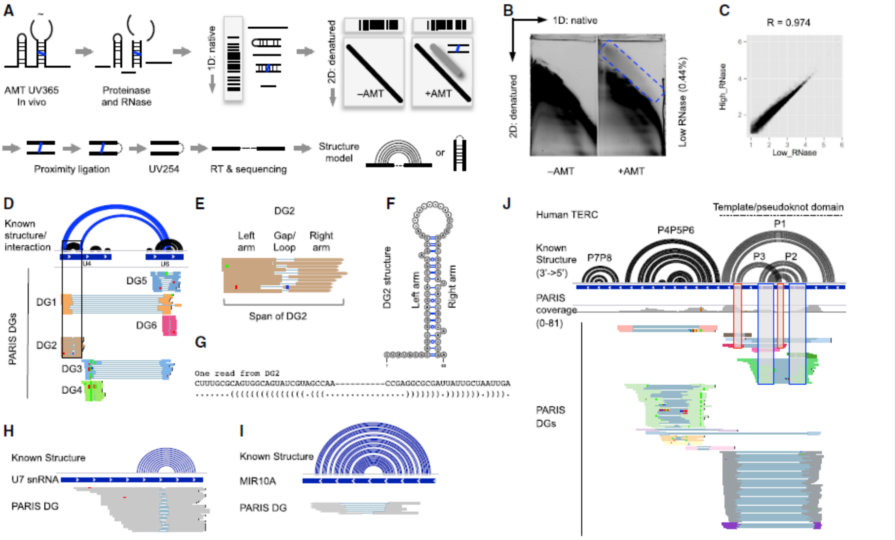
Пример


Scale
chr11:
17,345,000
17,350,000
17,355,000
17,360,000
17,365,000
17,370,000
17,375,000
17,380,000
17,385,000
17,390,000
17,395,000
17,400,000
17,405,000
17,410,000
17,415,000
17,420,000
20 kb
hg38
User Supplied Track
paris
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
NCR3LG1
KCNJ11
ABCC8
RP1-239B22.5
LOC105376575
NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*) - Annotation Release 2016-07-27
Human mRNAs
Human mRNAs from GenBank
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
NUCB2
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(147)
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other

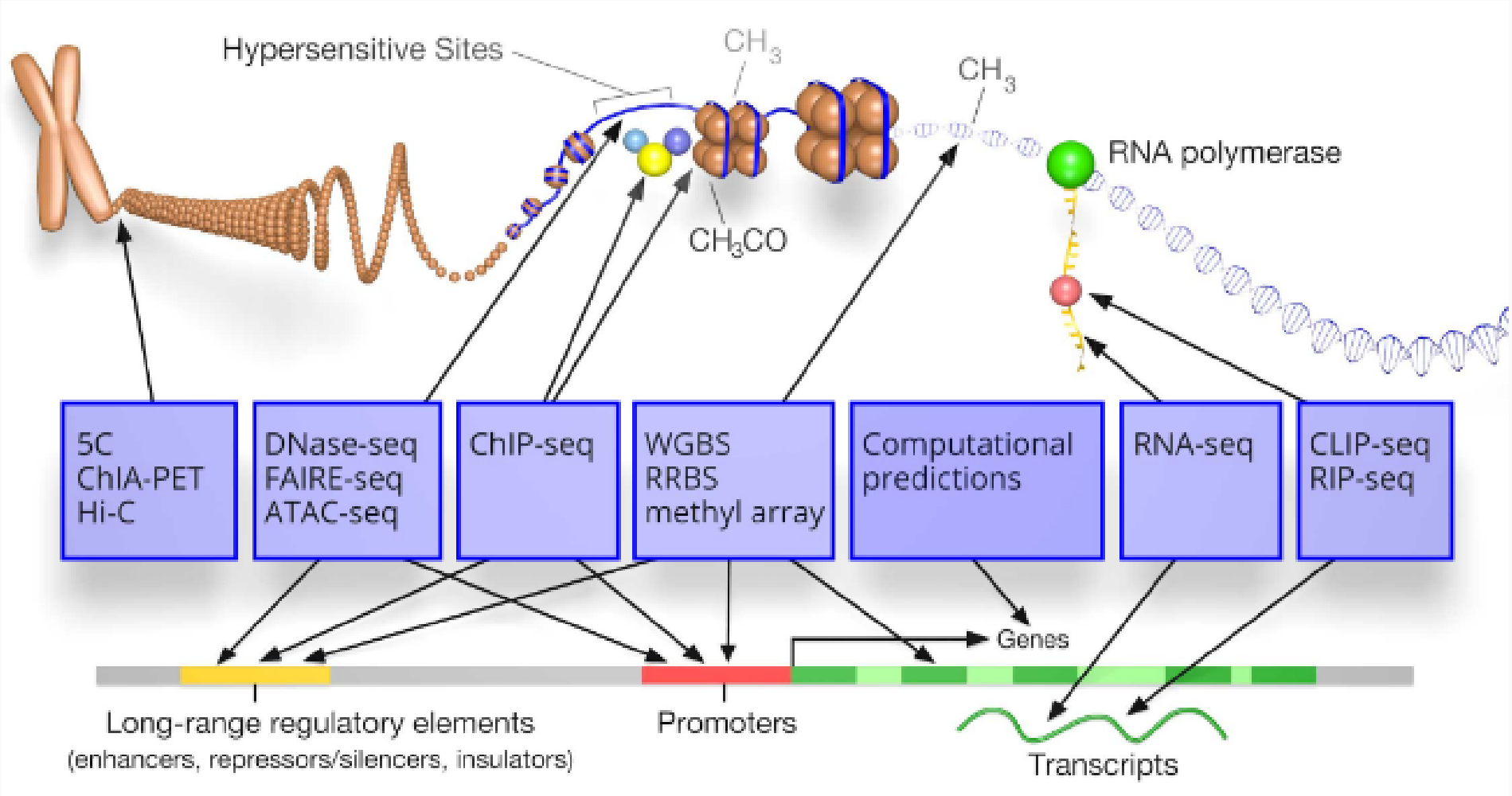
Further applications

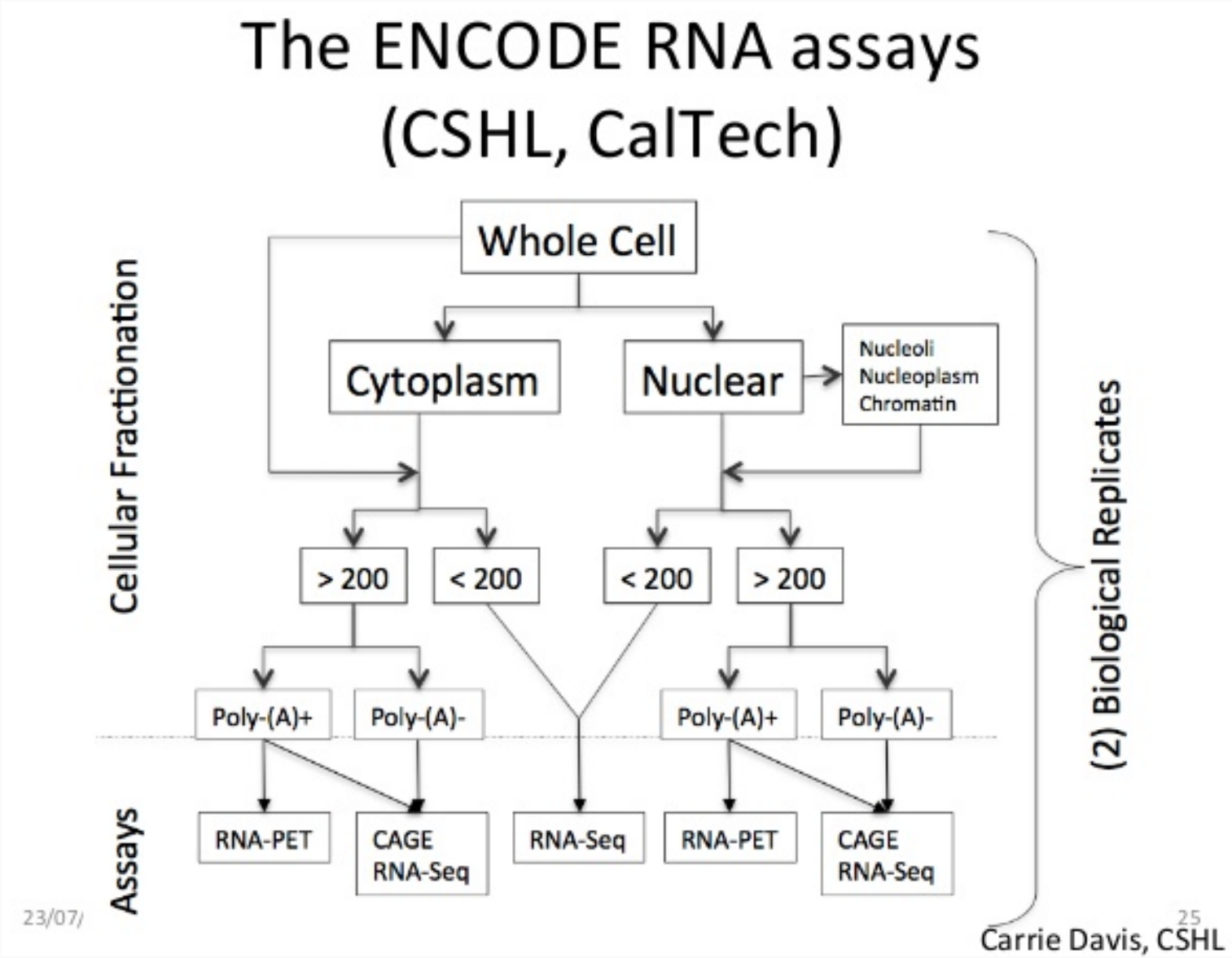
- Эксперименты обогащения/деплеции
  - РНКаза R для определения кольцевых РНК
  - Полиаденилированные РНК
  - Деплеция рРНК
- Определение доступности
  - Сайты гиперчувствительности ДНКазы
  - ATAC-seq
- Определение пространственной близости
  - Определение конформации хромосомы (3C)
  - Определение РНК дуплексов (PARIS)

Пример

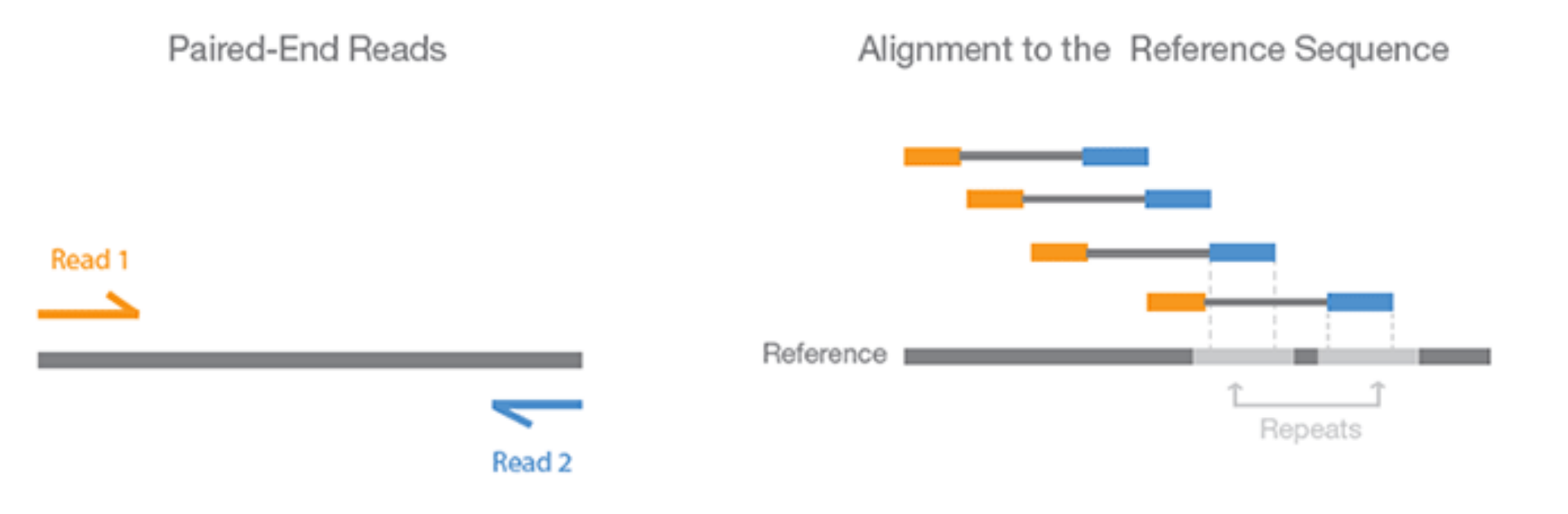

AMT UV365
In vivo
Proteinase
and RNase
1D: native
2D: denatured
–AMT
+AMT
Proximity ligation
UV254
RT & sequencing
Structure
model
or
1D: native
2D: denatured
Low RNase (0.44%)
R = 0.974
High_RNase
Low_RNase
Known
structure/
interaction
PARIS DGs
DG1
DG2
DG3
DG4
DG5
DG6
DG2
Left
arm
Gap/
Loop
Right
arm
Span of DG2
DG2 structure
Left arm
Right arm
One read from DG2
CUUUGCGCAGUGGCAGUAUCGUAGCCAA----------CCGAGGCGCGAUUAUUGCUAAUUGA
Human TERC
Template/pseudoknot domain
P4P5P6
P1
P7P8
P3
P2
Known
Structure
(3'->5')
PARIS
coverage
(0-81)
PARIS
DGs
Known Structure
U7 snRNA
PARIS DG
Known Structure
MIR10A
PARIS DG

ENCODE consortium (encodeproject.org)




The ENCODE RNA assays
(CSHL, CalTech)
Cellular Fractionation
Whole Cell
Cytoplasm
Nuclear
Nucleoli
Nucleoplasm
Chromatin
> 200
< 200
< 200
> 200
Poly-(A)+
Poly-(A)-
Poly-(A)+
Poly-(A)-
Assays
RNA-PET
CAGE
RNA-Seq
RNA-Seq
RNA-PET
CAGE
RNA-Seq
(2) Biological Replicates
23/07/
25
Carrie Davis, CSHL

Парные прочтения и цепь-специфицеские прочтения

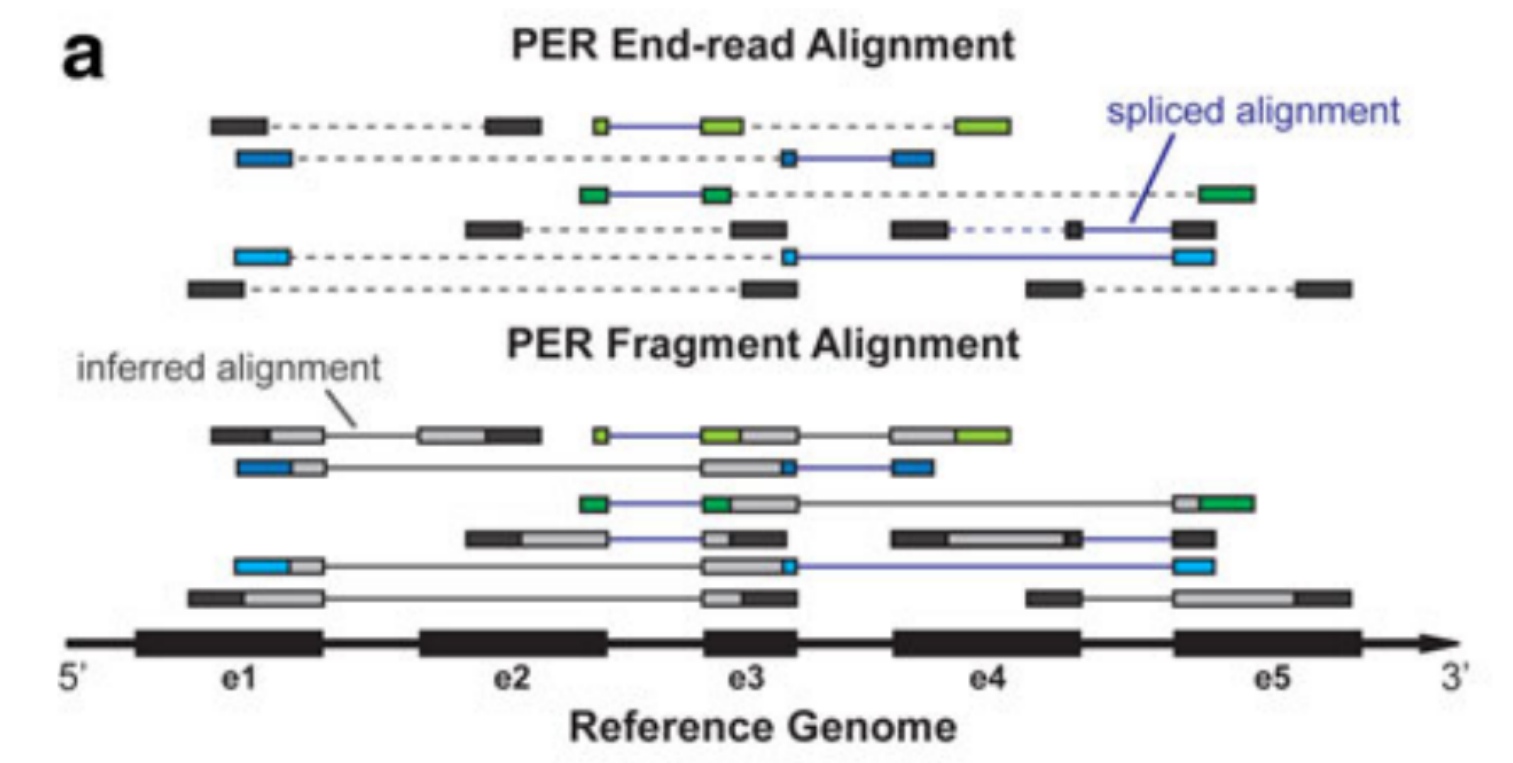


- Если секвенируем одноцепочечную РНК
  - Сначала синтезируем комплементарную цепь (RT)
  - Плюс-цепь синтезируем, но добавляем модифицированный нуклеотид (тиоуридин)
  - Двухцепочечный дуплекс лигируем с адаптером
  - Расщепляем тиоуридин

Data Analysis

- Картирование : имеется большое число коротких прочтений (10^8) и референсный геном
  - Высокая скорость
  - Небольшое число ошибок
  - Неоднозначные картирования
  - Разрывные/нелинейные картирования
- Квантификация : количественная оценка представленности каждого геномного участка
- Peak Calling : локализованный количественный сигнал
- Деконволюция : нужно восстановить представленности и структуру перекрывающихся транскриптов, основываясь на локальной информации о коротких ридах

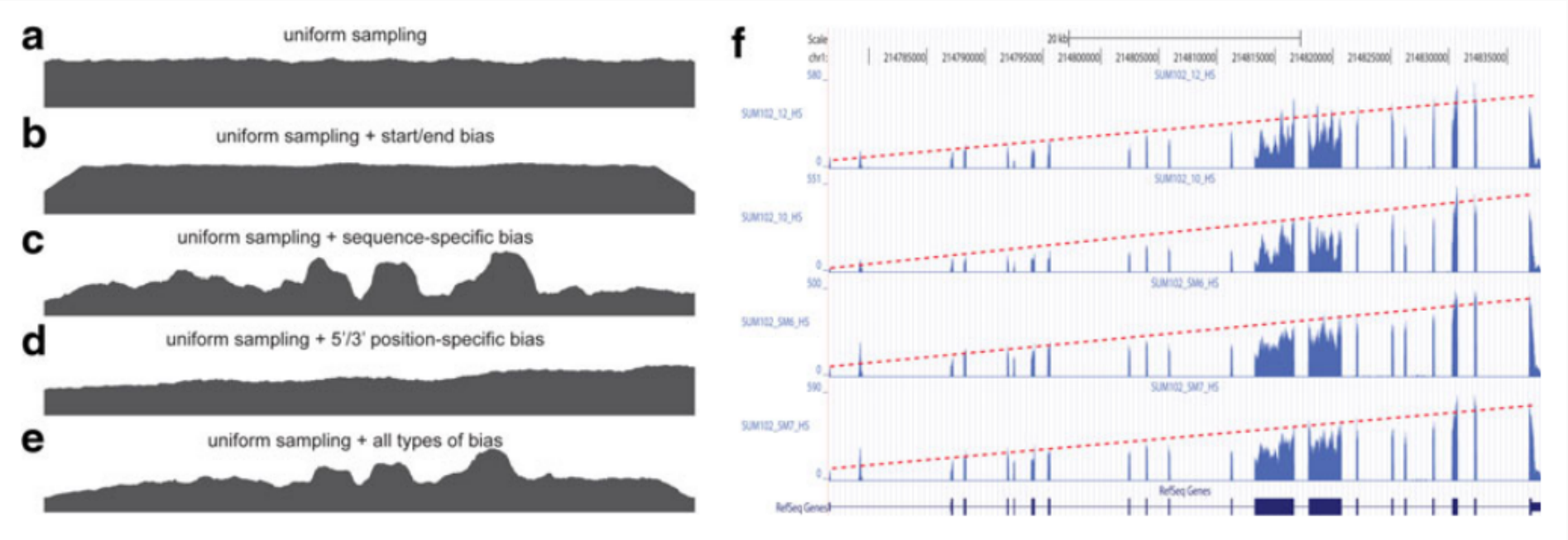
Деконволюция*



- Для данной модели транскрипта, какова вероятность получить наблюдаемое распределение ридов?
- Имея наблюдаемое распределение ридов, какая структура транскриптов максимизирует постериорную вероятность видеть то, что мы видим?
- Какая структура транскриптов максимизирует функцию потока в графе?

* Huang et al, JCB 2013

Неравномерное распределение ридов*

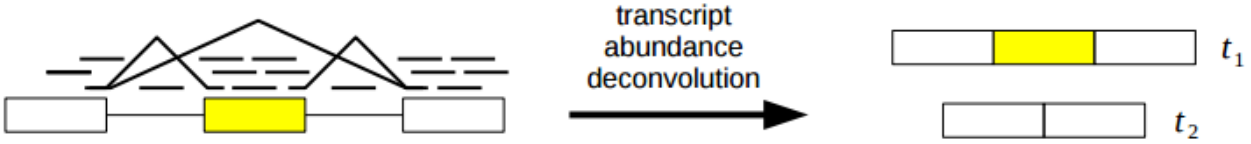


- 3'-смещение из-за прайминга обратной транскрипции
- ГЦ смещение – риды с высоким ГЦ-составом оказываются перепредставлены
- Эффект лаборатории
- Артефакты штрихкодов в новых технологиях Илюмины

* **Huang et al, JCB 2013**

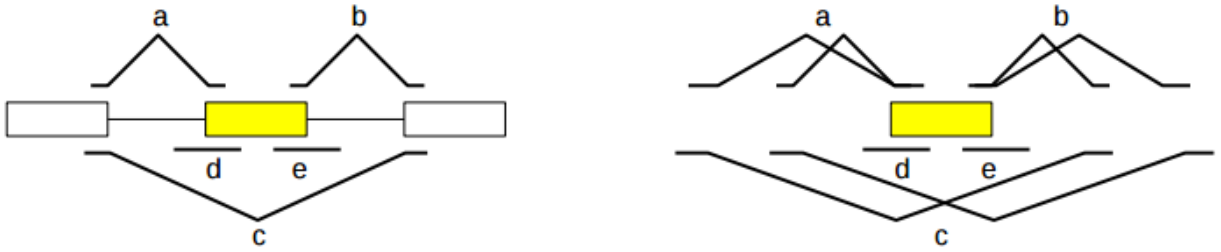
Количественные оценки сплайсинга

1. Transcript-centric



$$psi = \frac{t_1}{t_1 + t_2}$$

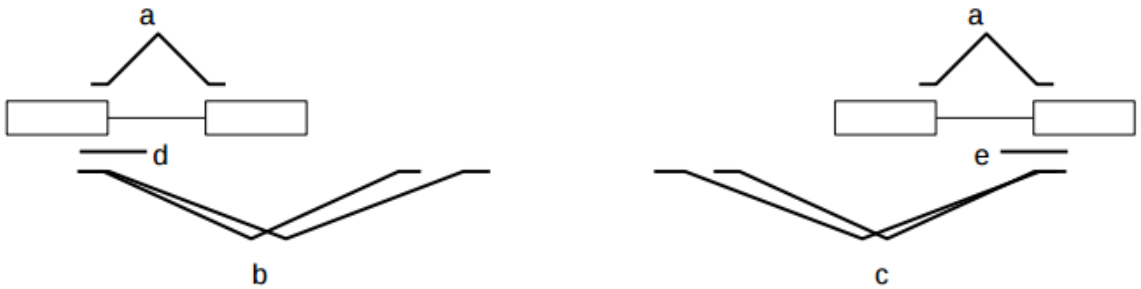
2a. Exon-centric



$$psi = \frac{a+b}{a+b+2c}$$

$$cosi = \frac{a+b+2c}{a+b+2c+d+e}$$

2b. Intron-centric (can be generalized to arbitrary splicing graphs)



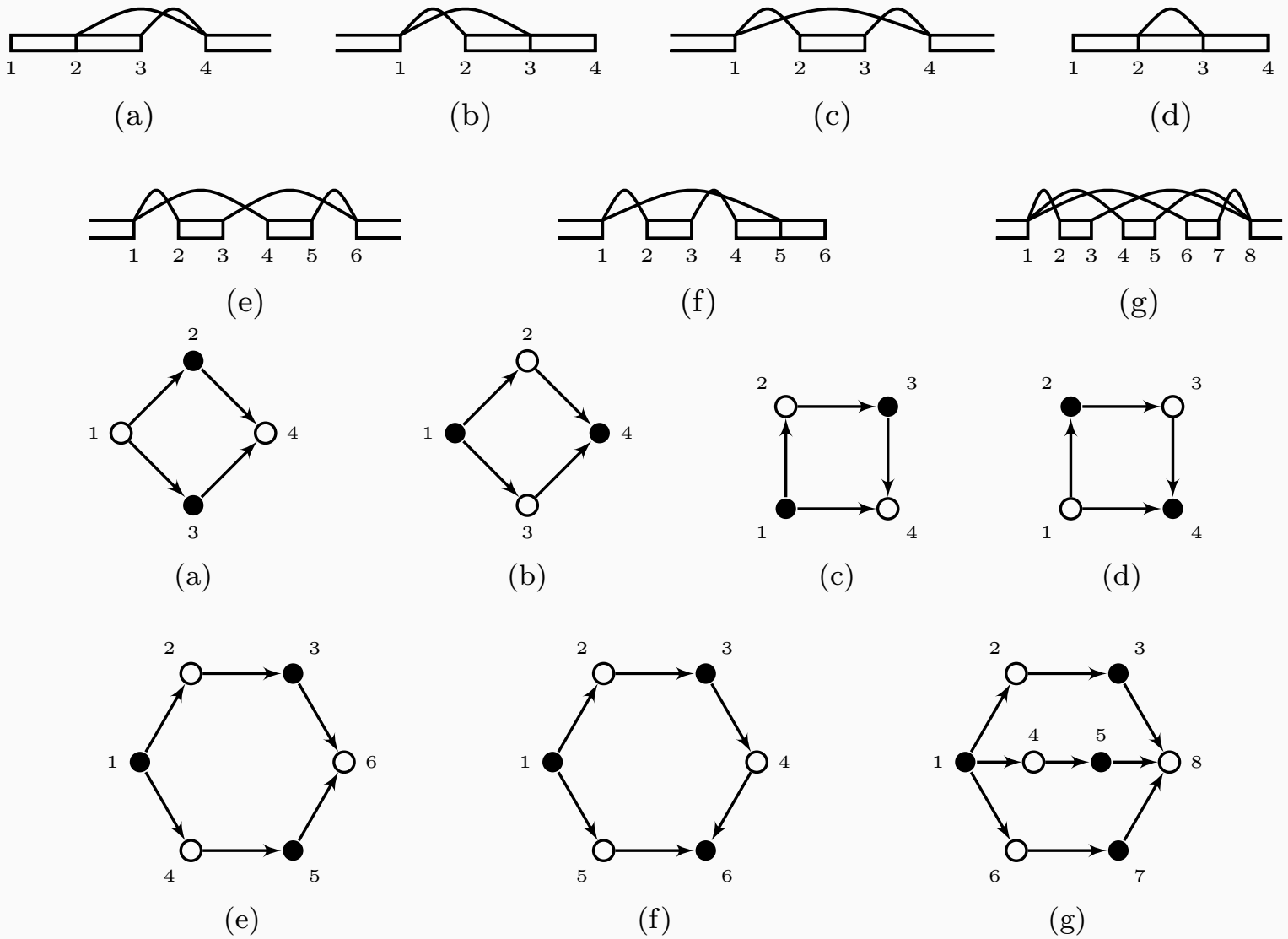
$$psi5 = \frac{a}{a+b}$$

$$psi3 = \frac{a}{a+c}$$

$$cosi5 = \frac{a+b}{a+b+d}$$

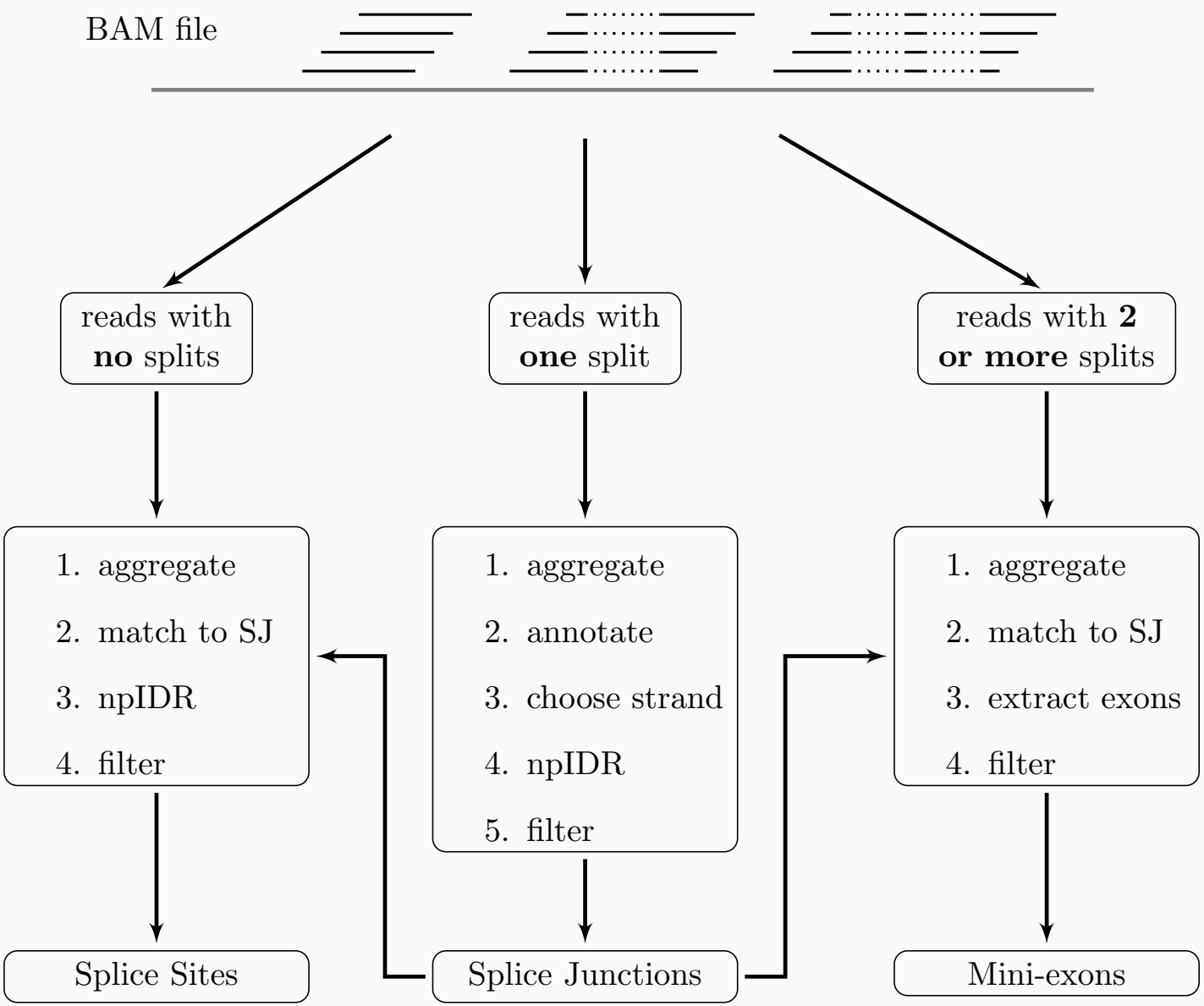
$$cosi3 = \frac{a+c}{a+c+e}$$

Граф сплайсинга



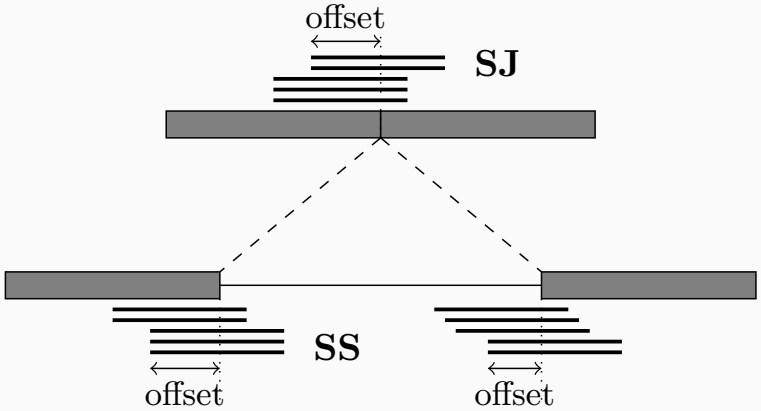
Ψ is a binomial (multinomial) variable defined for a group of vertex-independent paths

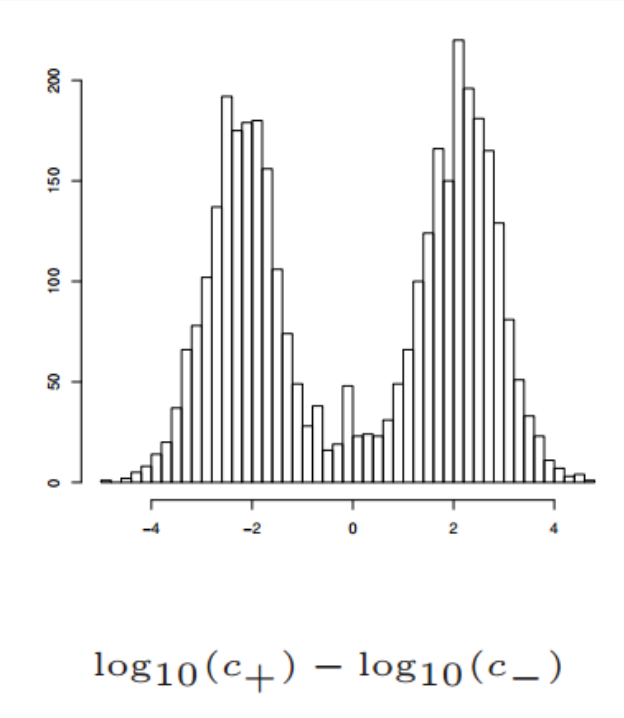
Пайплайн IPSA

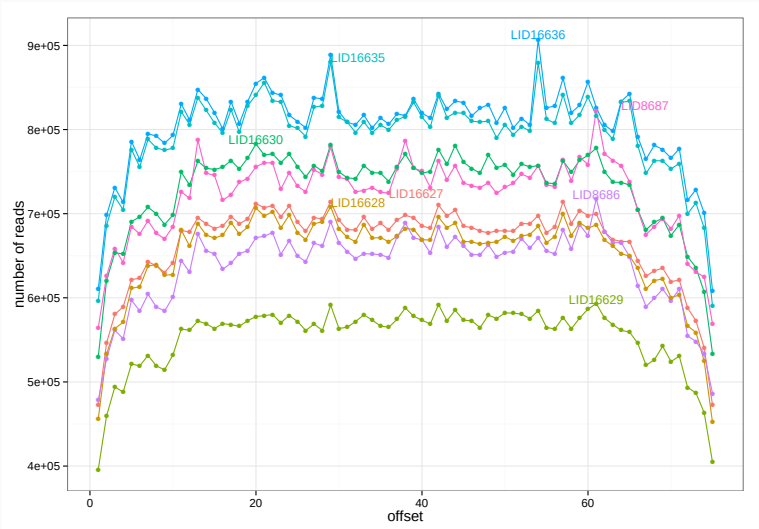


Offset

Offset is the position of the split within the read sequence





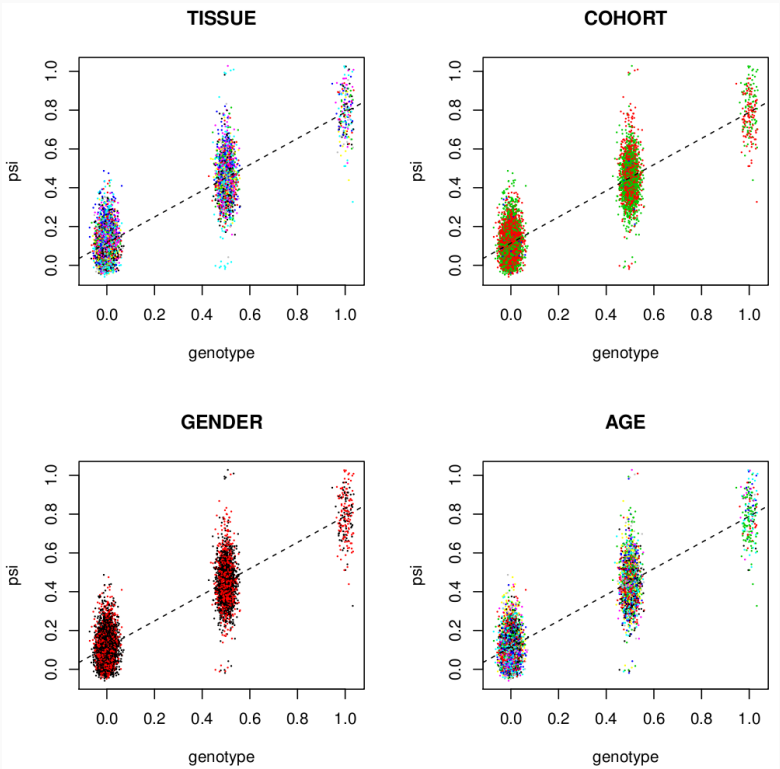


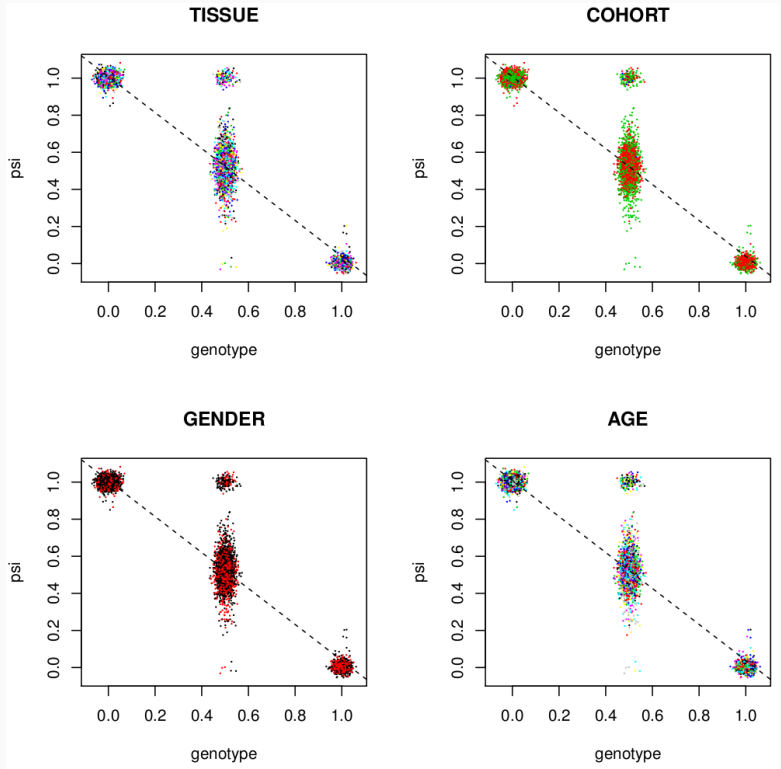
GTEx: splicing QTLs

$$\Psi = b_0 + b_1 * genotype, \quad 0 \leq \Psi \leq 1, \quad 0 \leq \text{genotype} \leq 1$$

$$\Delta\Psi = 0.1 \Leftrightarrow b_1 = 0.1$$

$$t = (|b_1| - 0.1)/SE(b_1)$$





Challenge: find sQTLs in unballanced design

Спасибо за внимание