



ChIP-seq: секвенирование, технологии и задачи биоинформатики

Юрий Львович Орлов



План лекции

ЦКП «Биоинформатика»

История секвенирования. «Геном человека»

Технологии высокопроизводительного секвенирования

Приложения и задачи секвенирования

Профиль ChIP-seq

Качество чтения фрагментов ДНК

Сайты связывания транскрипционных факторов в геноме, определяемые ChIP-seq

ВЫЧИСЛИТЕЛЬНЫЕ РЕСУРСЫ ССКЦ

Кластер **НКС-30T New !**

(hp BL2X220c)

Общее число
процессоров
Intel Xeon
E5450/E5540/
X5670

576 (2688 ядер);
InfiniBand, GE;

30 ТФлопс

Кластер **НКС-160**

(hp rx1620)

168 процессор.
Itanium 2,
1,6 ГГц;
InfiniBand,
Gigabit
Ethernet (GE);
> 1 ТФлопс

Кластер гибридной архитектуры

80 процессор.
CPU (X5670) –
480 ядер;

120 процессор.
GPU(Tesla M 2090) -
61440 ядер.

85,4ТФлопс

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Системное

Общематематическое
Прикладное (ППП)

СИСТЕМЫ ХРАНЕНИЯ ДАННЫХ (СХД)

Параллельная файловая
система IBRIX
для НКС-30T
32 Тбайт

СХД
для НКС-30T
36 Тбайт
(max-120 Тбайт)

СХД
для НКС-160
3,2 Тбайт

**Сервер
с общей памятью**
(hp DL580 G5)

4 процессора
(16 ядер)
Intel Xeon Quad
Core X7350,
2,93 ГГц;
256 Гбайт
общая память;
187,5 ГФлопс

GE

Сеть
ИВМиМГ

GE

Сеть
Internet ННЦ

Тренды науки в области геномики – точка зрения

Орлов Ю.Л., к.б.н.

Сингапур: 2006-2011,
5 лет в Биополисе,
в Геномном Институте Сингапура

Задачи геномики и
транскриптомики, компьютерной
биологии



Genome Institute
of Singapore

**Биомедицинские науки
Институты A-STAR**



Bioinformatics Institute (BII)

Bioprocessing Technology Institute (BTI)

Experimental Therapeutics Centre (ETC)

Genome Institute of Singapore (GIS)

Institute of Bioengineering and Nanotechnology (IBN)

Institute of Medical Biology (IMB)

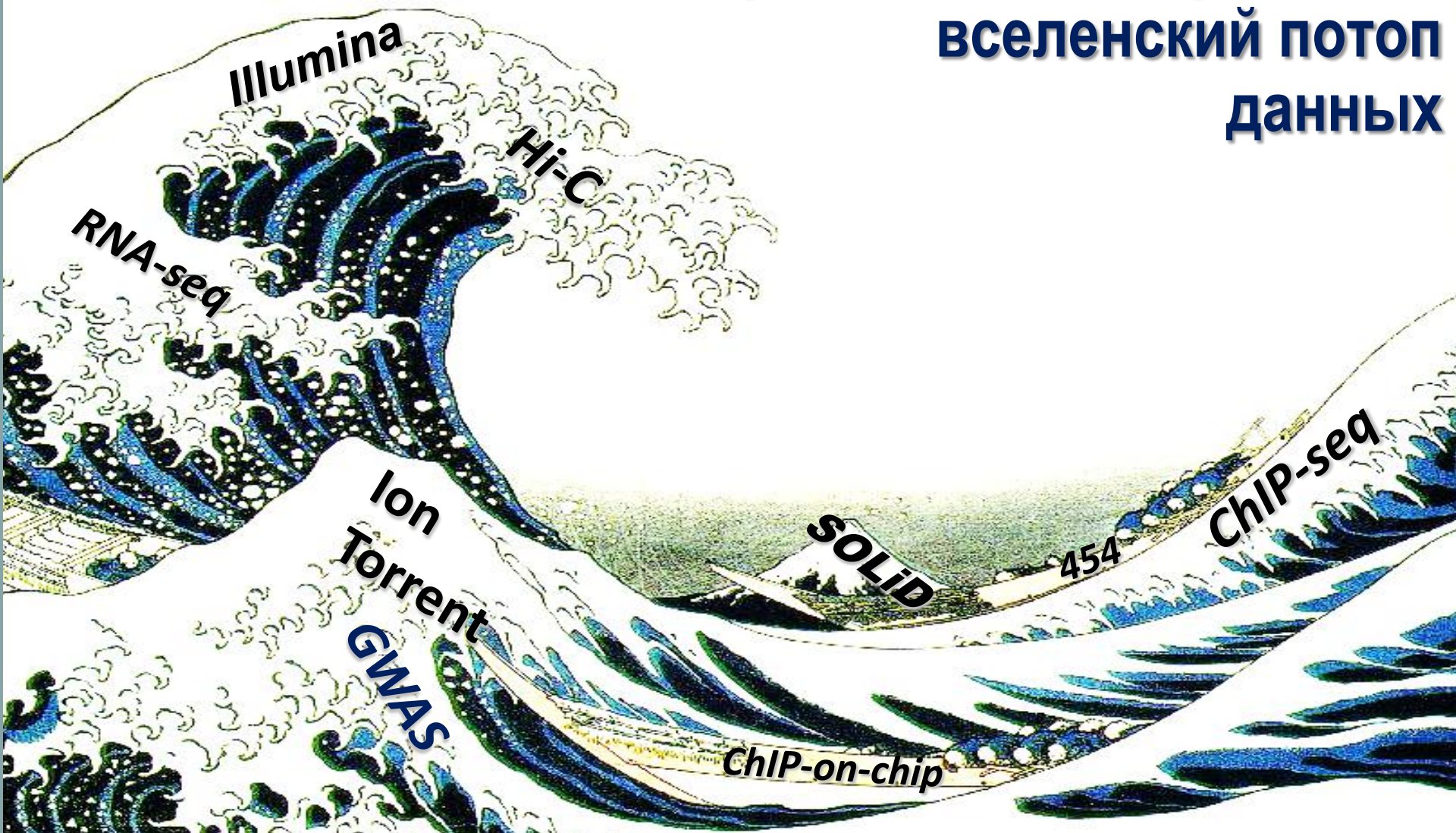
Institute of Molecular and Cell Biology (IMCB)

A*STAR - Duke-NUS

The Agency for Science, Technology and Research (A*STAR)



Геном человека и проекты секвенирования: вселенский потоп данных



Аббревиатуры, которые надо знать – «Чип» и «ChIP»



HUGO – Human Genome Organization

Организация генома человека

Проект «Геном человека»



Mirzabekov (USSR)

HUGO was conceived in 1988, at the first meeting on genome mapping and sequencing at Cold Spring Harbor. From a 42 scientists of 17 countries membership association, HUGO has increased its membership base to over 1,200 members, both established and aspiring of 69 countries after two decades.

1988

Cold Spring Harbor. April, 1988, Victor McKusick
(founding)

<http://www.hugo-international.org/>

PROFESSOR EDISON T LIU RE-ELECTED HUGO PRESIDENT



The Human Genome Organisation (HUGO), one of the oldest genome organisations, is pleased to announce the re-election of Professor Edison T. Liu as its President, effective 1 February 2010 by the HUGO Council, the governing body of the organisation.

Professor Liu is the Executive director of the Genome Institute of Singapore and has been a council member and President of HUGO since 2007. Since 2007, the President under the direction of the HUGO Council has made significant changes in the organisation. Membership rules were restructured, web and social networking for the organisation established. A new journal, the HUGO Journal has been initiated with Springer Publishing Group, and the symposia schedule enhanced. More recent collaborations with the WHO, the Archon X Prize, and the OECD have extended HUGO's participation in global policy making.

HUGO boss

Нынешний председатель HUGO –
Проф. Эдисон Лю

Немного истории..

Ежегодные конференции HUGO –
следующая, 2012 г. – в Австралии



HGM 2012, Sydney

Накопление данных шло многие годы параллельно с развитием технологий

Базы данных последовательностей ДНК

LANL (Los Alamos National Laboratory) Нью-Мексико, США

EMBL (European Molecular Biology Laboratory) Гейдельберг, Германия

1979 - прототип GenBank разрабатывался в LANL 1982-1992,

далее в NCBI (National Center for Biotechnology Information)

www.ncbi.nlm.nih.gov

1980 - EMBL Data Library (www.ebi.ac.uk)

1984 – DDBJ (www.ddbj.nig.ac.jp)

1983 - ГЕНЭКСПРЕСС и Программа "Геном человека" СССР
(А.А.Баев)

Сейчас эти центры образуют международный консорциум

International Nucleotide Sequence Database Collaboration

www.ncbi.nlm.nih.gov/collab

Описание баз данных (более тысячи мол.биол. баз данных в мире)

NAR (Nucleic Acid Res.) - ежегодный первый выпуск (N1, январь)



Margaret Dayhoff
(1925-1983)

1960-e Protein
Sequence Atlas

От анализа последовательностей ДНК

к секвенированию полных геномов, параллельно шло развитие компьютерных технологий,
баз и банков данных

Геном человека – основные вехи



2008

- Landmark Paper: [Mapping and sequencing of structural variation from eight human genomes](#),
• *Nature*, May 1, 2008

2004

- Landmark Paper: [Finishing the euchromatic sequence of the human genome](#), *Nature*, Oct. 2004
- [Human Gene Count Estimates Changed to 20,000 to 25,000](#), October 2004.

2003

- [Human Genome Project Declared Complete](#), April 2003 [Corresponding [Nature](#) and [Science](#)]

2002

- [Mouse Genome Sequencing Consortium publishes its draft sequence of mouse genome](#) in the December 5, 2002, issue of *Nature*.

2001

o**Publication of Initial Working Draft Sequence February 12, 2001**

Special issues of *Science* (Feb. 16, 2001) and *Nature* (Feb. 15, 2001) contain the working draft of the human genome sequence. human and several other genomes.

2000

- HGP leaders and President Clinton [announce](#) the completion of a "working draft" DNA sequence of the human genome.

1999

- [First Human Chromosome Completely Sequenced](#)! On December 1, researchers in the Human Genome Project announced the complete sequencing of the DNA making up human chromosome 22.

1998

[Celera Genomics](#) formed to sequence much of human genome in 3 years using HGP-generated resources.

1988

- [HUGO](#) founded by scientists to coordinate efforts internationally.

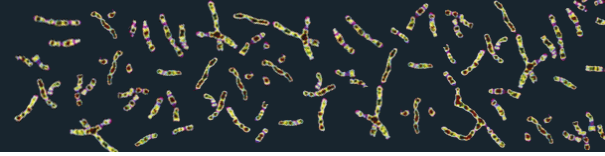
1986

- Following the Santa Fe conference, DOE OHER announces Human Genome Initiative. With \$5.3 million, pilot projects begin at DOE national laboratories to develop critical resources and technologies.

**Официальная
точка отсчета**



Первый геном человека – 2001 год – был «сборным» из ДНК разных анонимных пациентов



1000 KG

Individual genomes



The International 1000 Genomes Project was proposed by BGI in China, Sanger Institute in UK, and the 3G centers in USA in 2006, and later joined by groups in Germany and 3 major sequencer-producers. The project is designed to identify SNPs with lower MAF (Minor Allele Frequency) than those by HapMap, and other genomic variations such as CNVs (Copy Number Variations), by sequencing hundreds of samples from the 3 major populations.

~250 genomes to end of this year

phase III of the HapMap project (269 individuals)

Four populations were selected for inclusion in the HapMap: 30 adult-and-both-parents trios from Ibadan, Nigeria (YRI), 30 trios of U.S. residents of northern and western European ancestry (CEU), 44 unrelated individuals from Tokyo, Japan (JPT) and 45 unrelated Han Chinese individuals from Beijing, China (CHB).

Взрыв объема данных – от одного генома – к тысячам

Текущие проекты – 1000 индивидуальных геномов человека, HapMap, тысяча геномов растений, 10000 геномов микроорганизмов...

1000 Genomes: The first pilot intends to genotype 180 people of 3 major geographic groups at low coverage (2x). For the second pilot study, the genomes of two nuclear families (both parents and an adult child) are going to be sequenced with deep coverage (20x per genome). The third pilot study involves sequencing the coding regions (exons) of 1,000 genes in 1,000 people with deep coverage (20x).

The following populations will be included in the study: Yoruba in Ibadan, Nigeria; Japanese in Tokyo; Chinese in Beijing; Utah residents with ancestry from northern and western Europe; Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Toscani in Italy; Gujarati Indians in Houston; Chinese in metropolitan Denver; people of Mexican ancestry in Los Angeles; and people of African ancestry in the southwestern United States.

Один из лидеров секвенирования в мире –
Геномный Институт Пекина, в Шэньжэне, Китай



Prof Henry Yang Huanming
President & Professor
Beijing Genomics Institute (BGI)-
Shenzhen, China



Применение и возможности
секвенирования – растения,
Микроорганизмы,
Древняя ДНК

CHINESE

SEARCH

[Home](#) [About BGI](#) [Programs&Projects](#) [Technology&Platforms](#) [Collaboration&Service](#) [Education](#) [Careers](#)

Collaboration tastes sweet!

Cucumber genome
History of Silkworm Domestication
Building the sequence map of the human pan-genome
Faster human genome sequencing
Giant Panda Genome
The Ancient Human Genome
Our 'Other' Genome
Single base-resolution methylome of the silkworm reveals a sparse epigenomic map

1 2 3 4

Веб-сайт и проекты
Пекинского Геномного
Института

Для сравнения – Мощности секвенирования
MIT, Broad Institute USA >200 машин (США)
BGI ~100 машин (Китай) GIS ~20 (Сингапур)
ИЦиГ+ИХБФМ – 3 (Новосибирск)



Призы за технические достижения.. Независимый негосударственный фонд X-prize

The X Prize Foundation is a non-profit prize institute

In 1996, entrepreneur Peter Diamandis offered **a \$10 million prize** to the first privately financed team that could build and fly a three-passenger vehicle 100 kilometers into space twice within two weeks. The contest, later titled the Ansari X Prize for Suborbital Spaceflight, motivated 26 teams from seven nations to invest more than **\$100 million**.

Winner: Mojave Aerospace Ventures The prize was won on October 4, 2004, the 47th anniversary of the Sputnik 1 launch. (Brothers Ansari)



Role	Spaceplane
Manufacturer	Scaled Composites
Designed by	Burt Rutan
First flight	20 May 2003
Retired	4 October 2004
Status	Museum piece

* **The Ansari X Prize**, the original X Prize competition to create a private spaceflight vehicle

* **The Archon X Prize**, a genomics competition (100 human genomes)

* The Automotive X Prize

* The Wirefly X Prize Cup

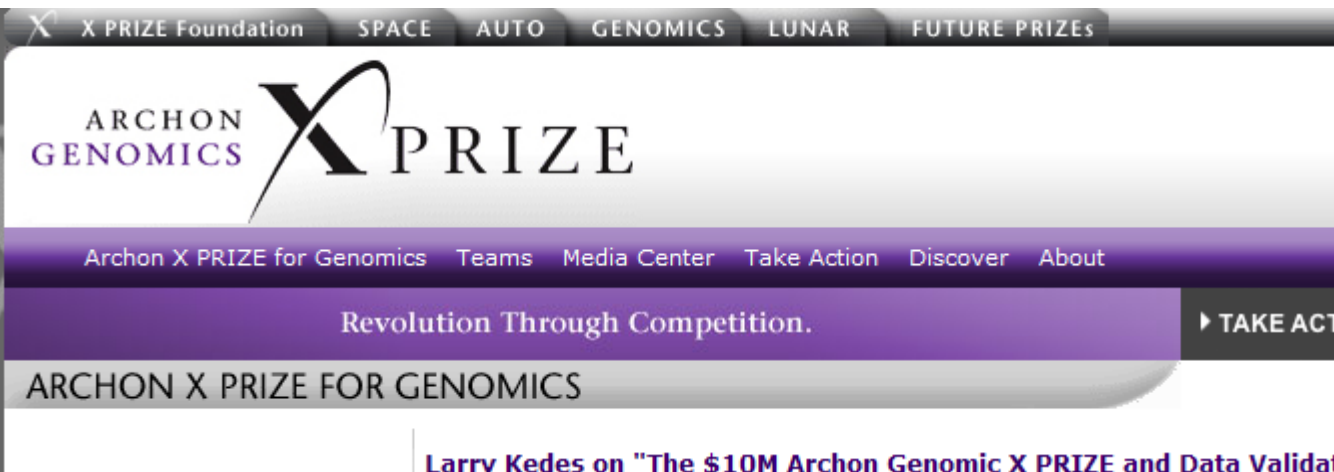
* The Google Lunar X Prize, a competition to put a robot on the moon

* The Ocean X Prize

The \$10 million was donated by Canadian geologist and philanthropist Stewart Blusson, who co-discovered the Ekati Diamond Mine. The name "Archon" is the name of Blusson's company

Примеры поддержки космических технологий





Larry Kedes on "The \$10M Archon Genomic X PRIZE and Data Validation"



Prof Larry Kedes
Professor Larry Kedes
Scientific Director and Senior Advisor, X PRIZE Foundation

Страница Фонда X-prize,
его председатель



1.1 The Device

The device and the processes it uses must be repeatable as determined by the Judges. A repeatable process in this context means that the device is capable of continuing to provide complete human genome sequences at the same or faster speed (see section 1.2), and with the same or better accuracy (see section 1.3) and coverage (see section 1.4), and at the same or lower cost (see section 1.5) as required for those genomes sequenced during the competition phase.

- There are no restrictions on types of devices or techniques.

1.2 Speed:

During each X PRIZE competition test, a TEAM must use its device to sequence within 10 days 100 human diploid genomes of 6 Gbp (6 giga base pairs, i.e., six billion pairs of DNA base molecules) each.

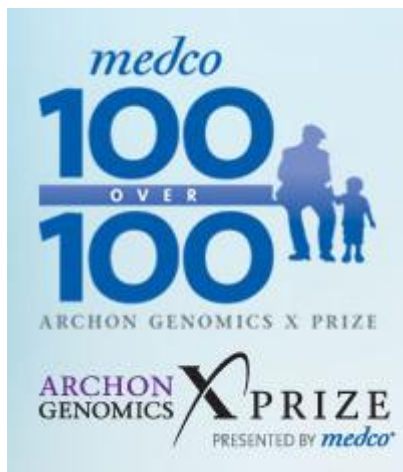
1.3 Accuracy:

The sequences of each of the 100 genomes, when presented to the Judges, must contain no more than one error in every 100,000 base pairs. The Judges will assess the accuracy based in part on sequences predetermined by several techniques. The predetermined information will be kept confidential from the TEAMS.

Условия
соревнования
: скорость,
качество
данных –
указаны (по
английски)

Приз 10 млн. \$ – секвенировать 1 геном за 1000 \$

Скачки и гонки технологий секвенирования



Продолжается соревнование компаний за разработку более дешевой, доступной технологии полного секвенирования индивидуального генома человека

Игроки в области геномного секвенирования («гонщики»). Перечислены международные компании, развивающие свои технологии



Llumina (Solexa), Sequenom, 454 Life Sciences, Pacific Biosciences, Complete Genomics, Intelligent Bio-Systems, Genome Corp., ION Torrent Systems, and Helicos Biosciences.

Технические аспекты и преимущества различных технологий:

**Короче последовательности ->
Дешевле технология -> меньше
надежность**

Большой объем коротких фрагментов (50-70bp)(SOLiD)
Longer sequence reads, 150bp (Технология Illumina - Solexa)
300 bp reads - Reliable reads (Технология Roche 454)
Long reads, up 1K-10K bp (Технология PacBio – Pacific Biocomputing)
Single molecule sequencing Light emission / magnetic / physical movement measurements

Общий тренд: Большой объем данных фрагментов ДНК ведет к снижению качества, зато уменьшает цену. Соответственно возникают вычислительные сложности обработки

**Overall cost of whole human genome sequencing
-> 10K\$ -> 6K\$ ->3 K\$**

Текущая стоимость секвенирования индивидуального генома человека уже ниже 10 тыс.долларов (Цель – 1 тыс.долларов)

**10K USD price is claimed (BGI),
3K\$ is minimal price per genome sequencing (but pilot project only)**

Представители компаний:



**Dr Mostafa
Ronaghi**
Illumina Inc



Dr Francisco M. De La Vega
**Life Technologies -
SOLiD™ System**

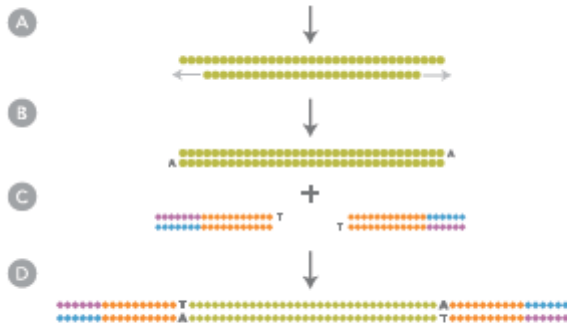


Dr Stephen Turner
PacBio

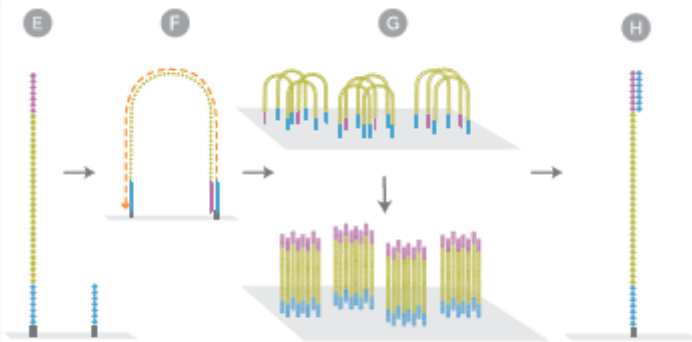


Dr Radoje Drmanac
Complete Genomics, Inc.

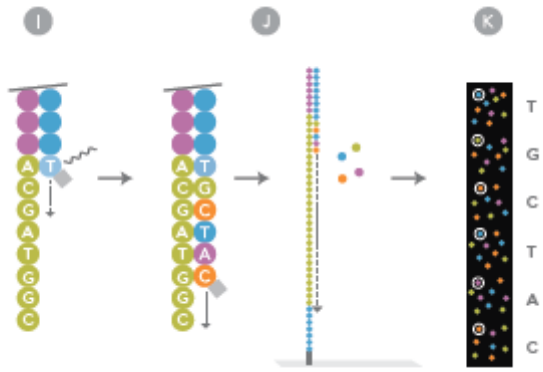
Illumina solexa sequencing



- A Fragment DNA
- B Repair ends/Add A overhang
- C Ligate adapters
- D Select ligated DNA



- E Attach DNA to flow cell
- F Perform bridge amplification
- G Generate clusters
- H Anneal sequencing primer



- I Extend first base, read, and deblock
- J Repeat step above to extend strand
- K Generate base calls

Пример технологии Illumina (SOLEXA – старое название до объединения компаний)

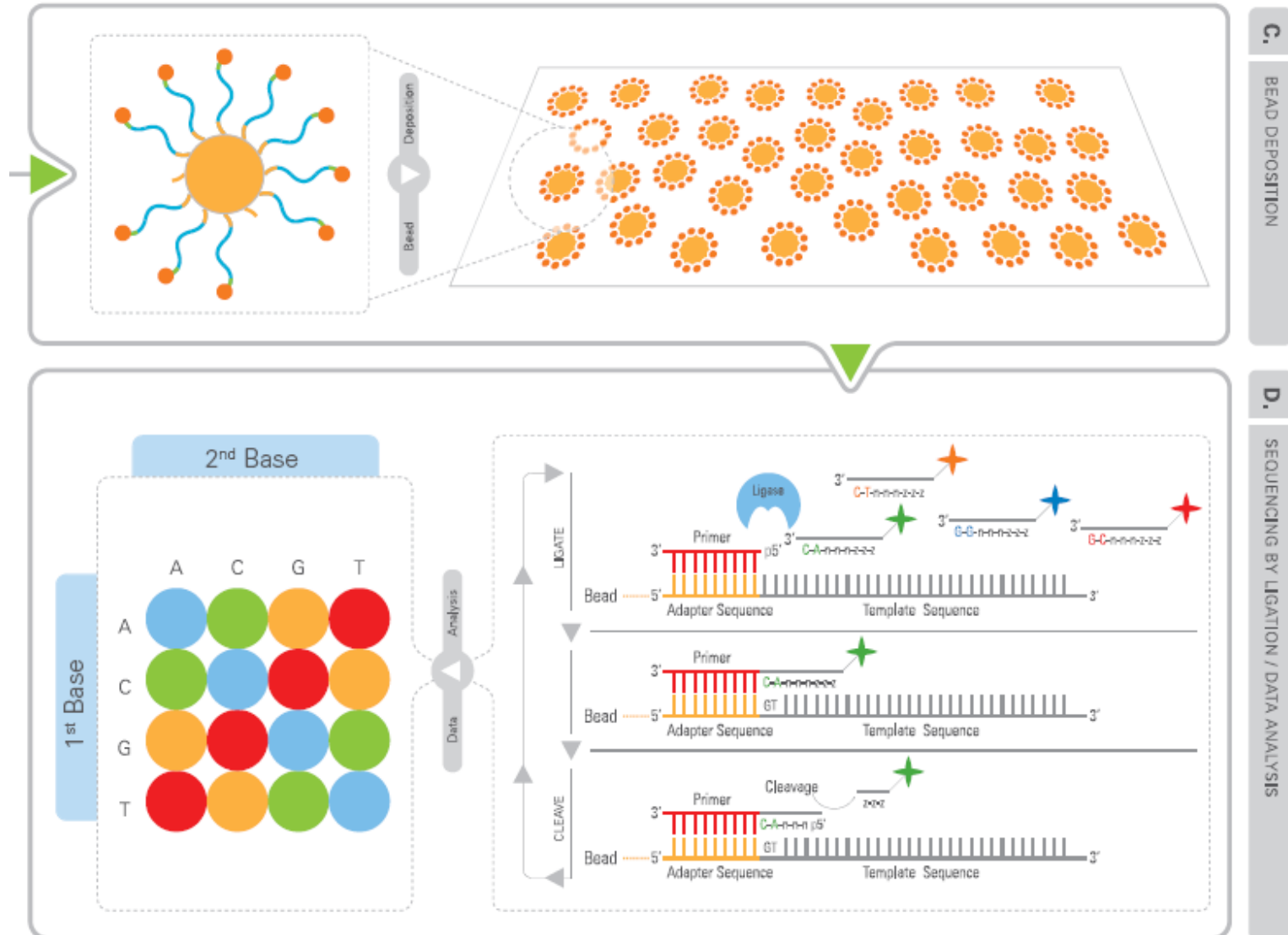
<http://www.illumina.com/>

Одна из наиболее распространенных платформ («рабочая лошадка» секвенирования).
Машины есть в Москве, в Новосибирске нет

SOLiD sequencing

Пример технологии SOLiD

Секвенирование идет по динуклеотидам которые представлены в цветовой кодировке (собственный формат данных)



Details available at the company web-site: <http://marketing.appliedbiosystems.com/>

Технологии высокопроизводительного секвенирования ДНК и возникающие статистические, алгоритмические и технические компьютерные задачи

Научные приложения:

- De novo genome sequencing
- Resequencing: analysis of polymorphisms
- TF binding
- Epigenomics
- Profiling (rearrangements, chromosome abnormalities)
- Transcriptomics
- Fishing for RNA Editing Sites
- “3D genome”**

Компьютерные задачи

Ассемблирование: Сборка длинной последовательности (мегабазы) из коротких (35-100 нуклеотидов)

Картирование. Поиск участков отличающихся от референсной последовательности (статистически)

Хромосомные участки отличающиеся по свойствам

Анализ всех транскриптов в геноме (~10000 генов X 50 транскриптов)

Редактирование РНК в процессе транскрипции (статистически)

Трехмерные контакты в хромосомах

Математически – много коротких последовательностей (25-75, до 100 нуклеотидов) многократно (10-100 раз) покрывающих геном (референсную последовательность (10Мб – 3Гб))



Профили ChIP-seq

Как это делается технически и
математически?

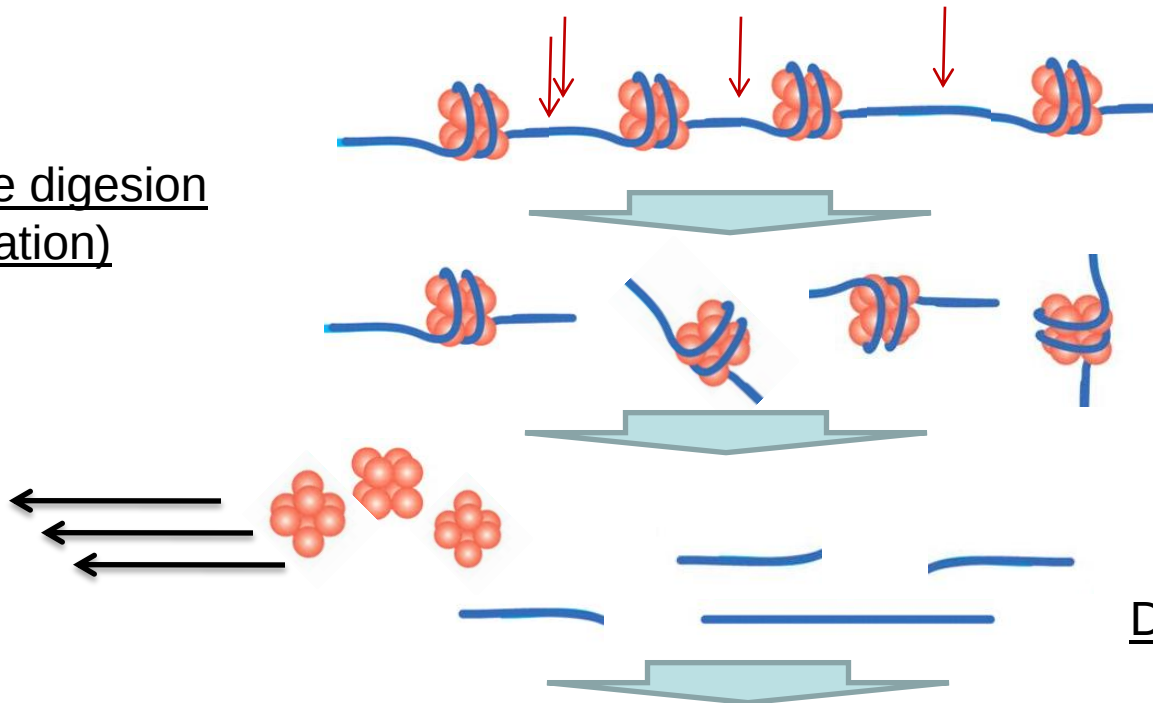
Аббревиатура

ChIP = Хроматин-иммунопреципитация

Общая схема геномного секвенирования (с иммунопреципитацией или без нее)



MNase digesion
(sonication)

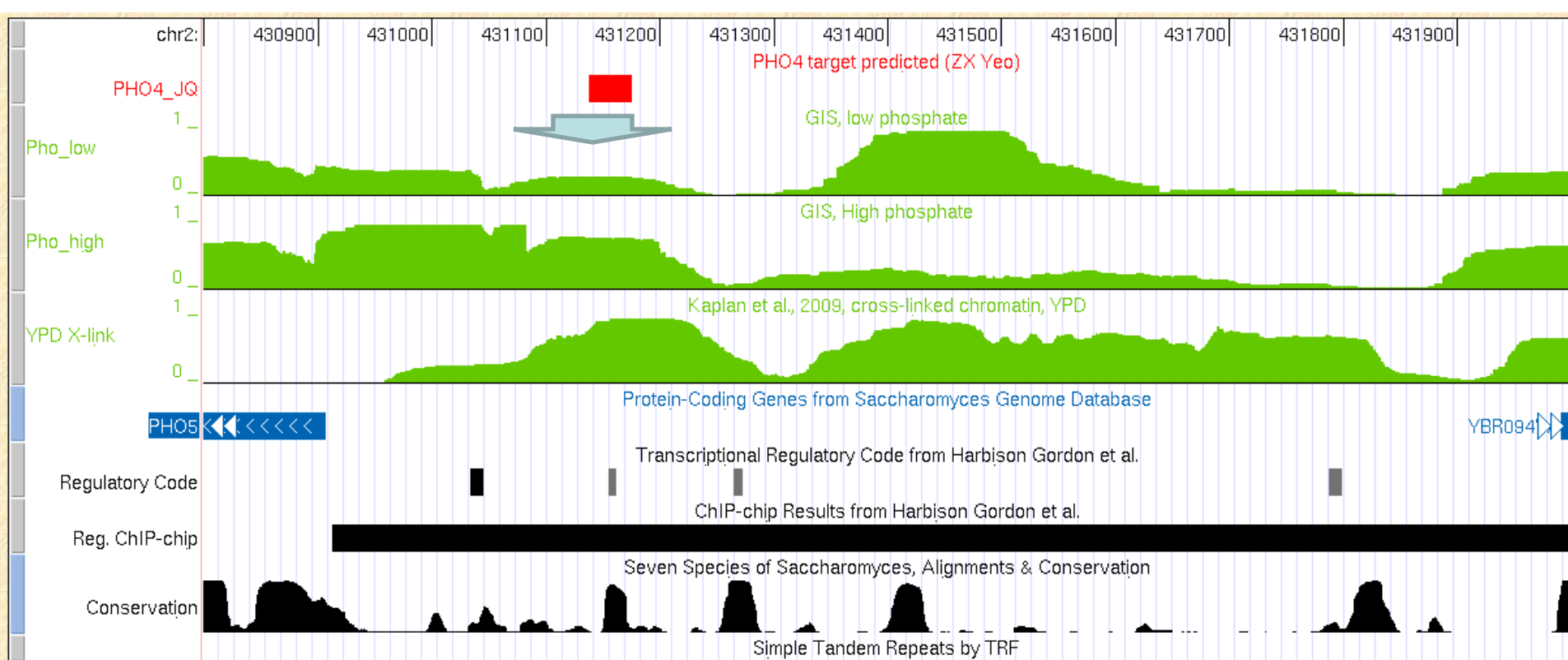


DNA extraction

High throughput sequencing, mapping to
reference genome

Фрагменты ДНК длиной по 150-300 нуклеотидов, но секвенируются (на машине) только начала (и концы) длиной 25, 35 или чуть более нуклеотидов

Геномный браузер – пример - Сравнение профилей нуклеосомной упаковки (здесь: геном дрожжей)



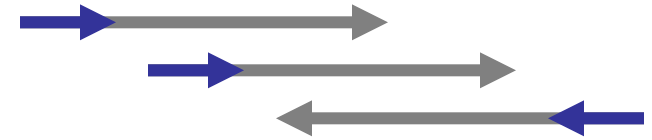
Nucleosome protecting *PHO4* binding sites is depleted in low phosphate condition

Красным помечен консервативный сайт связывания белка *PHO4* находящийся в промоторе гена *PHO5* дрожжей

Определение кластера (пика) сформированного парными или односторонними фрагментами (by paired or one-side ChIP reads) на хромосоме

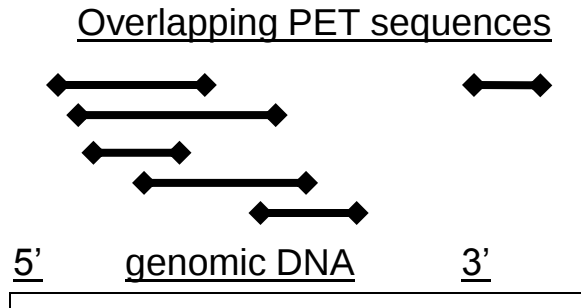


Pair of tags, possible variable length (<4Kb) of ChIP fragment



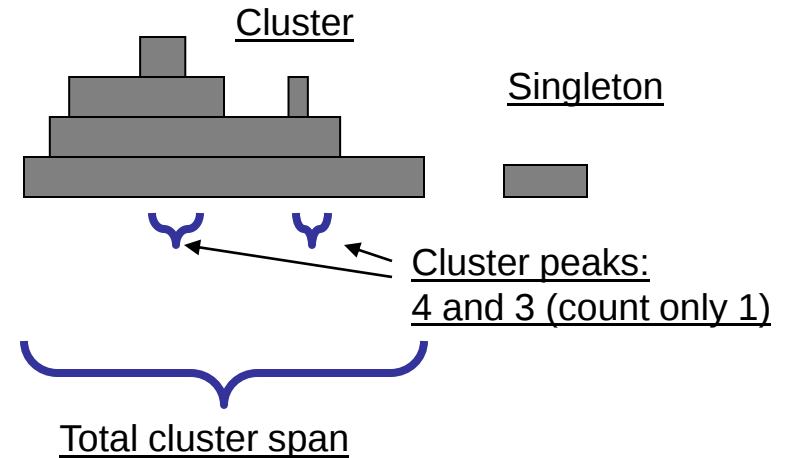
Single 25-35 nt tags extended to 200 nt, fixed length of ChIP fragment

cluster definitions



Cluster size = 5

Cluster member overlap = 4



Schematic example of sequence cluster, singleton, and cluster overlap. Cluster size is 5, and cluster member overlap is 4.

- Mapping - Картирование коротких фрагментов на геном (с несовпадениями)
- Peak calling – Определение пика (максимума и границ)
- Normalization to sequencing background – нормализация к контрольному секвенированию

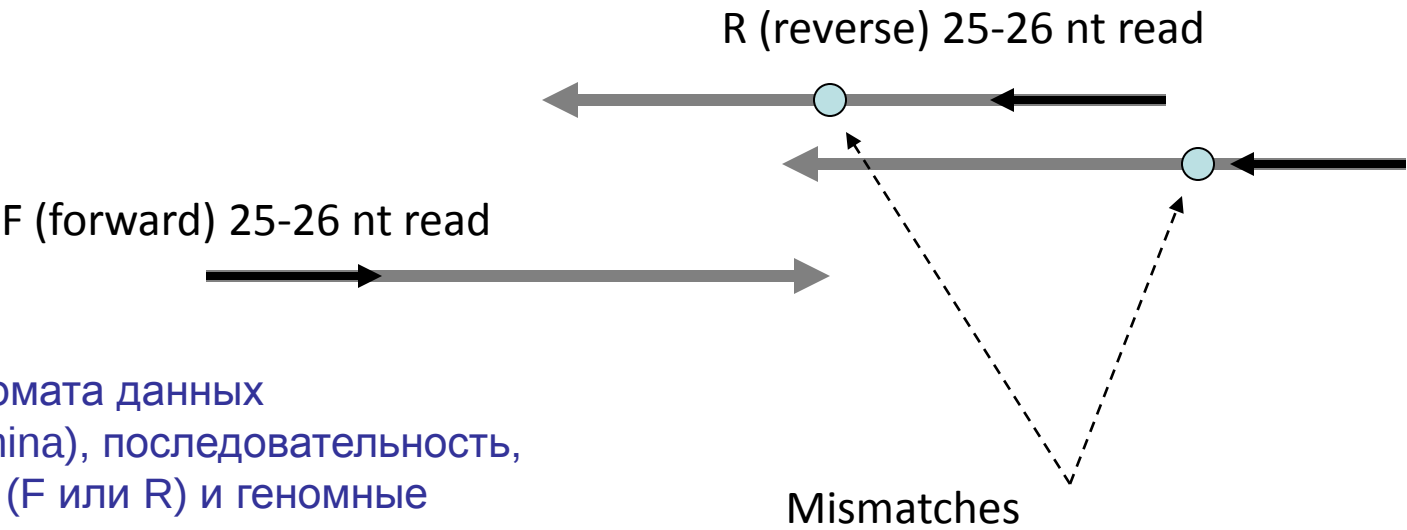
Задача картирования – найти (выровнять) короткие последовательности на геном

Mapping to the reference genome

Further validation of ChIP-seq reads:
bias in GC content, position mismatches,
ignored ‘N’ read letters in alignment

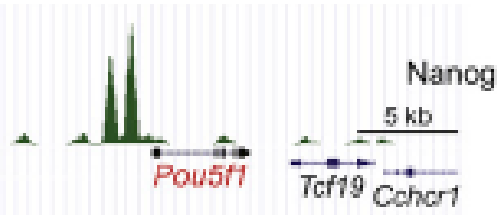
данные (после
картирования)

```
AAAAGCAACTTAGAGATTGCACCAC 12500 1 chr9:42004400 F AAAAGCAACTTAGAGATTGCACCAC 9359
GGTTTTATTATTATTTTAGGGGTTT 11453 1 chr13:116992396 F GGTTTTATTATTATTTATGGGTTT 9359
GTAATGTGTTTTTTGTGACTATTTT 11453 1 chr12:115188071 R AAAATAGACACAAAAAACACATTAC 9359
AAGATGCCAGAACTCTCAGCTCCTT 12500 1 chr7:135122171 F AAGATGCCAGAACTCTCAGCTCCTT 11453
TGCAAGCTTTCTTATTCCCTTTCAT 10406 1 chr6:128978691 R ATGTAAGGGAATAAGAAAGCTGGCA 9359
```

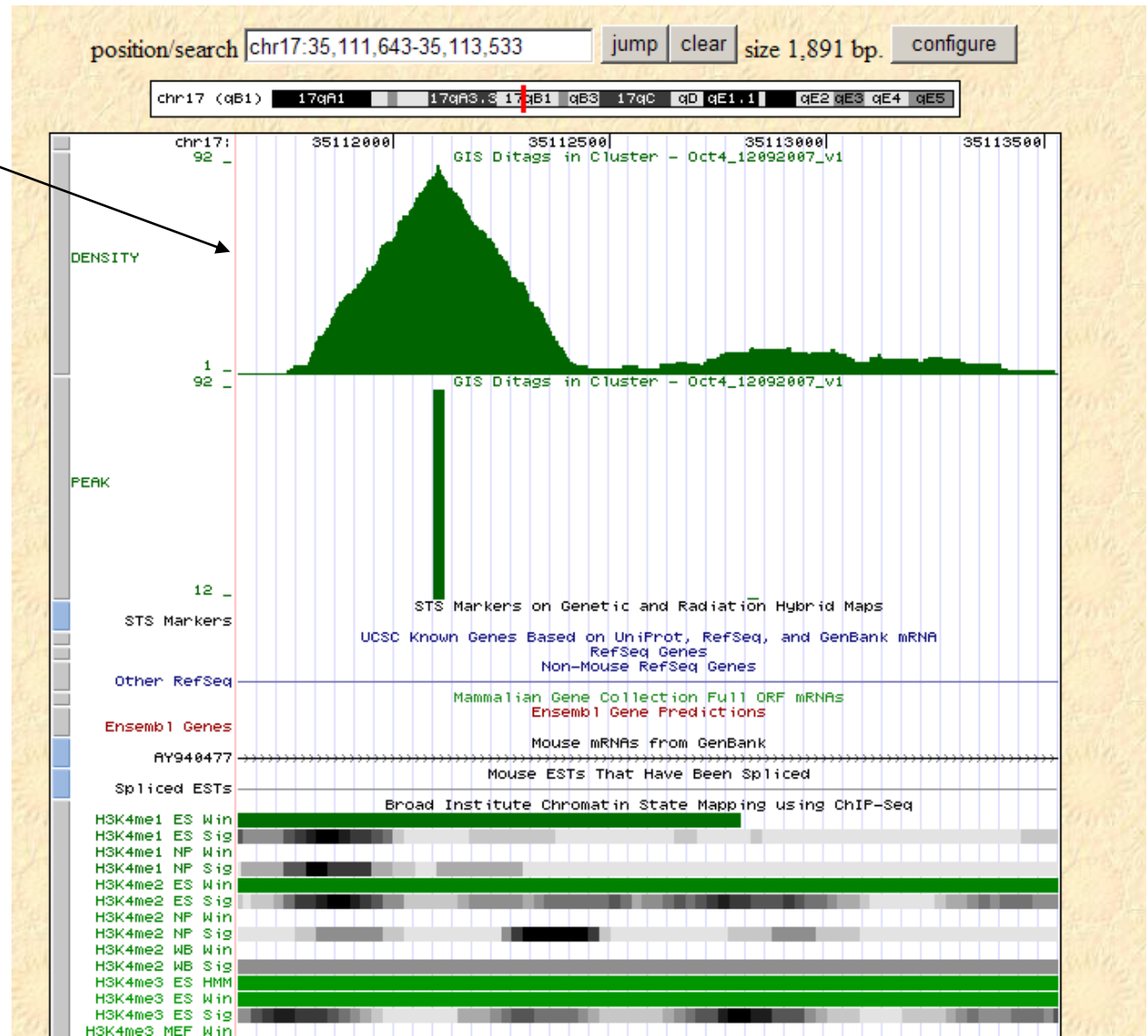


Пример формата данных
(Здесь: Illumina), последовательность,
ориентация (F или R) и геномные
координаты

Oct4 ChIP-seq cluster in Pou5f1 mouse promoter (link from T2G)



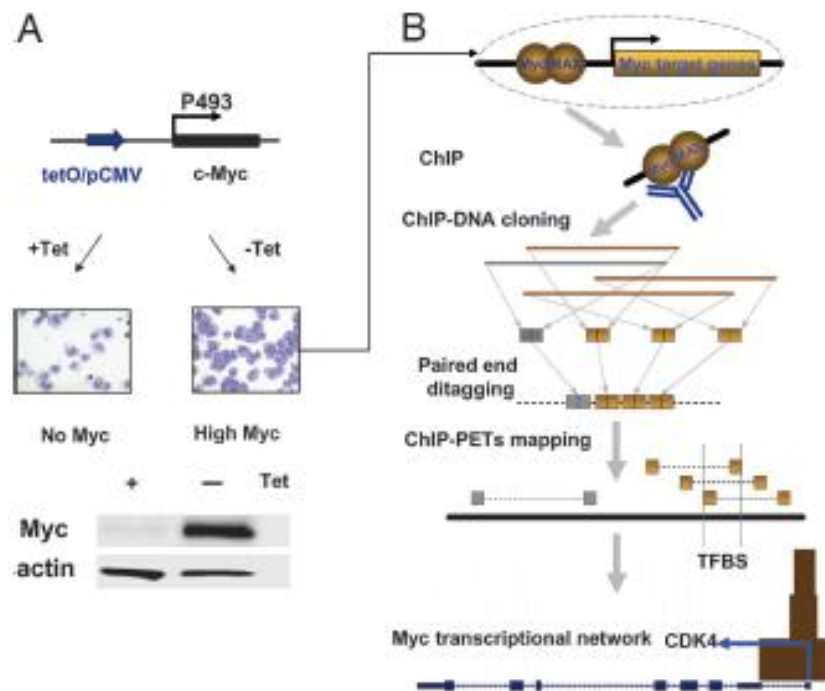
Четкий пик профиля (перекрывание коротких фрагментов)



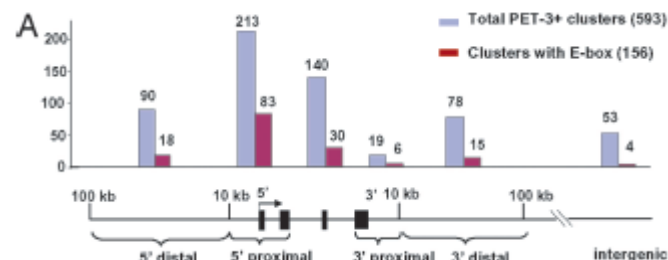
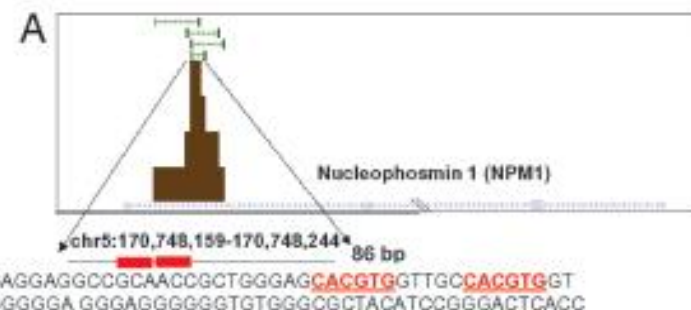
Пример четкого пика профиля ChIP-seq в хромосомных координатах

Пример анализа ТФ с-Мус

Экспериментальная схема от экстракции ДНК до картирования на геном



Определение пика профиля (ступенчатый вид)

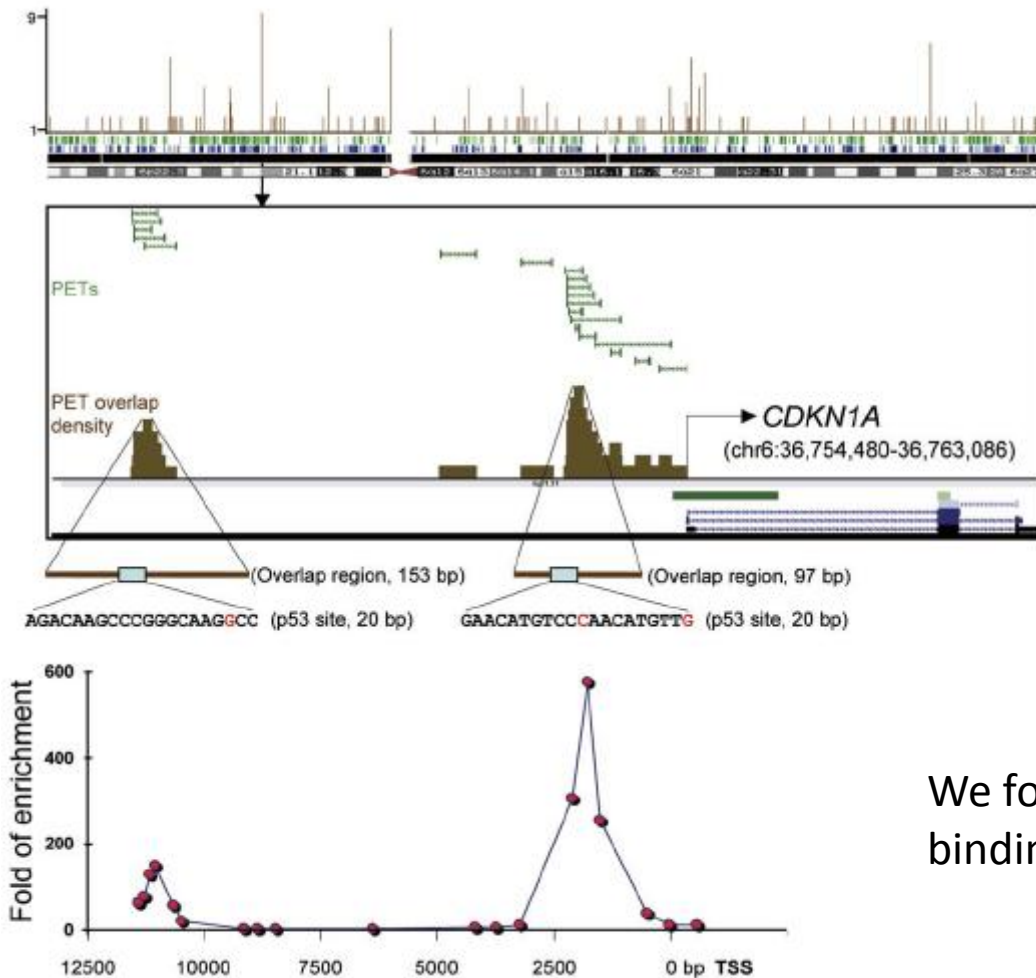


Частоты пиков (сайтов) по отношению к генам

Global mapping of c-Myc binding sites and target gene networks in human B cells

Karen I. Zeller*, XiaoDong Zhao†, Charlie W. H. Lee‡, Kuo Ping Chiu‡, Fei Yao‡, Jason T. Yustein*, Hong Sain Ooi‡, Yuriy L. Orlov‡, Atif Shahab‡, How Choong Yong‡, YuTao Fu§, Zhiping Weng§, Vladimir A. Kuznetsov‡, Wing-Kin Sung‡, Yijun Ruan‡, Chi V. Dang*¶, and Chia-Lin Wei†¶

*Department of Medicine and The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, The Johns Hopkins University School of Medicine, Baltimore, MD 21205; †Genome Institute of Singapore, Singapore 138672; ‡Bioinformatics Institute, Singapore 138671; and §Bioinformatics Program, Boston University, Boston, MA 02115



Validation of PET-Cluster-Identified p53 Binding Loci

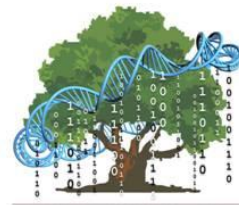
The whole-chromosome view of p53 ChIP-PETs mapping to chr6.

A genomic span of 23 kb that contains the CDKN1A gene and its 50 region is enlarged. CDKN1A was localized by two PET clusters; one contained 5 PETs, and the other contained 13 PETs. Both PET overlaps contained recognizable p53 binding motifs.

We found strong correlations between binding motifs and cluster size.

Большинство пиков содержат выраженные мотивы (консенсусные последовательности) сайтов связывания ТФ

Но это не соответствие один к одному!



RNA-seq

(применение секвенирования)

(факультативно)

Ключевая обзорная статья

RNA-Seq: a revolutionary tool for transcriptomics

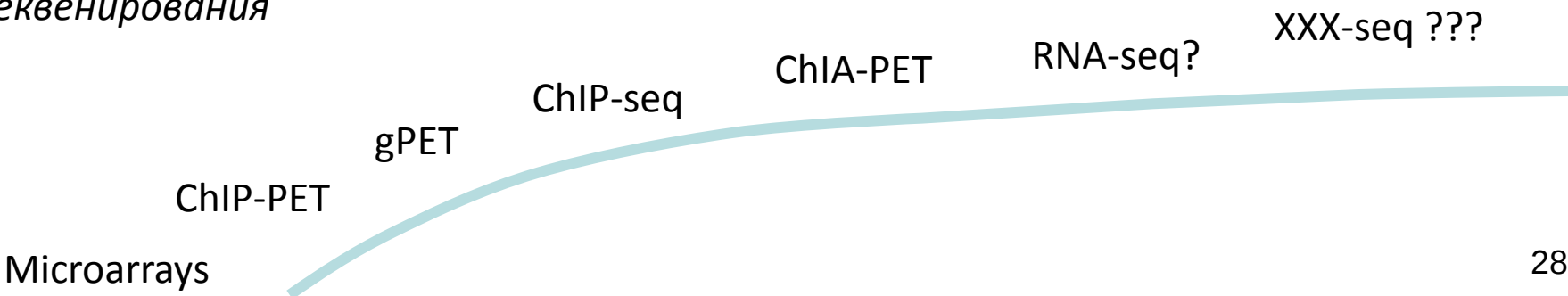
Zhong Wang, Mark Gerstein and Michael Snyder

NATURE REVIEWS | GENETICS
VOLUME 10 | JANUARY 2009

Особенности технологии – прямое определение количества транскриптов (не пробы, не микрочипы!)

- Новые задачи биоинформатики и статистических оценок экспрессии генов
- Возможность открытия новых транскриптов
- Проблемы картирования RNA-seq на геном (A.Mortazavi et. al.)
- Определение множественных «ридов» - выбор транскрипта, проблемы гомологичных генов

Развитие технологий и аббревиатуры секвенирования



Определение экспрессирующихся транскриптов в геноме (классификация)

Определение экспрессии генов через секвенирование (краткий обзор технологий)

Microarrays

Sequence-based

Expression
Exon array ...

cDNA

EST

SAGE - serial analysis of gene expression (3')

CAGE - cap analysis of gene expression (5')

MPSS - massively parallel signature

sequencing

(MPSS). A gene expression quantification method that determines 17–20-bp 'signatures' from the ends of a cDNA molecule using multiple cycles of enzymatic cleavage and ligation.

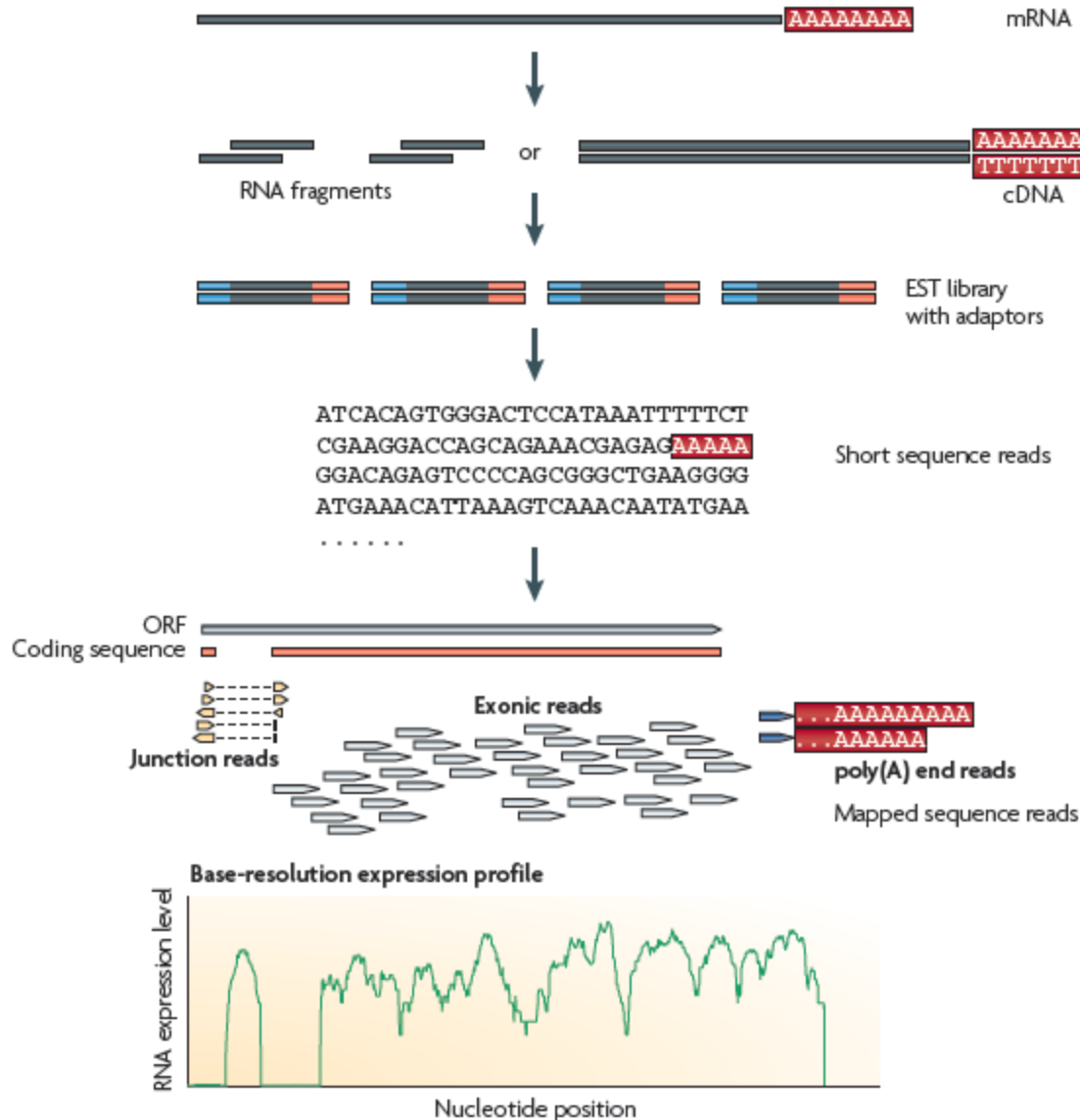
RNA-seq = RNA sequencing

RNA-seq library of cDNA fragments with adaptors attached to one or both ends

reads are of 30–400 bp – any (next-gen) technology
(может использоваться любая технология)

RNA-seq - экспрессия генов через секвенирование

Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing).

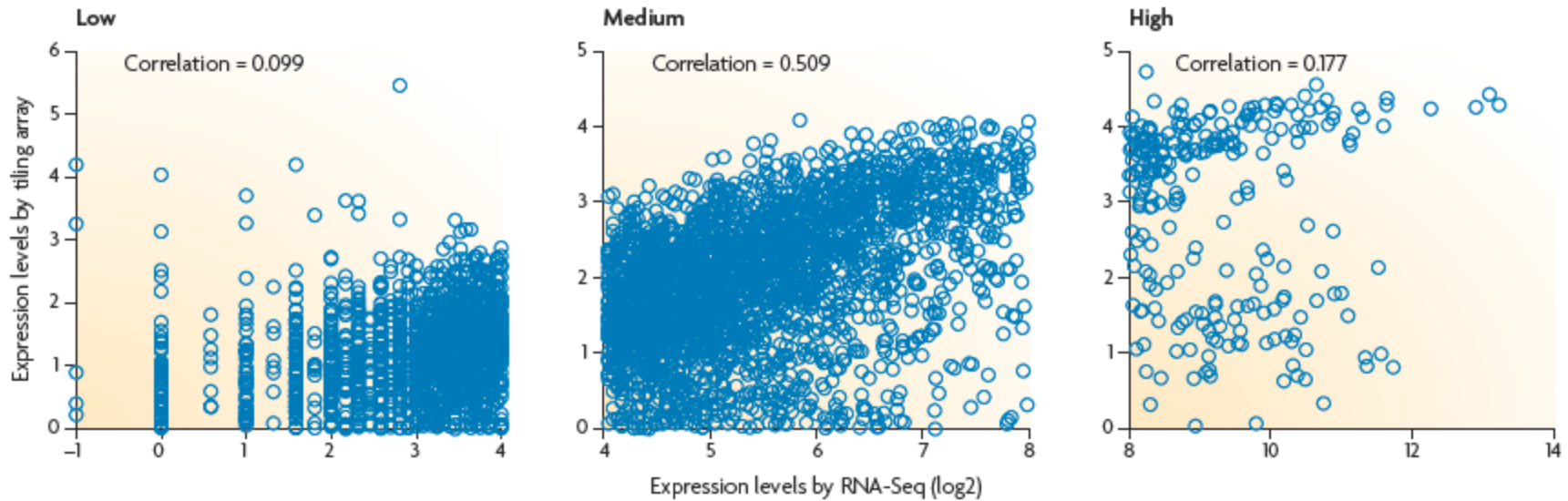


Any high-throughput sequencing technology can be used for RNA-Seq:
Illumina IG
Applied Biosystems SOLiD
Roche 454 Life Science
+ Helicos (perspective)

A typical RnA-seq experiment. Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. e - a base-resolution expression profile
 bottom - a yeast ORF with one intron is shown.

Comparison of RNA-seq with microarrays – where it is possible

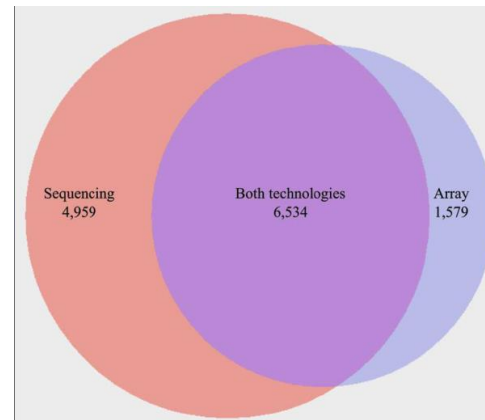
RNA-seq - экспрессия генов
через секвенирование



The two methods agree fairly well for genes with medium levels of expression (middle), but correlation is very low for genes with either low or high expression levels. (yeast)

Nagalakshmi U, et al. ..., Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008

(Факультативно) Показана корреляция экспрессии генов по RNA-seq с микрочипами –
Причем для низко-экспрессирующихся генов корреляция мала – микрочипы не работают



John C. Marioni et al
RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays
Genome Res. 2008

Технологии высокопроизводительного секвенирования ДНК - аббревиатуры

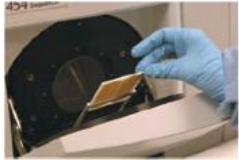
Capillary electrophoresis (Sanger)

Between 96 and 384 samples
(76 - 308 Kb/run)



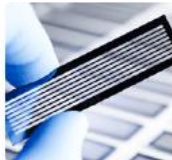
454

400,000 samples
(120Mb / run)



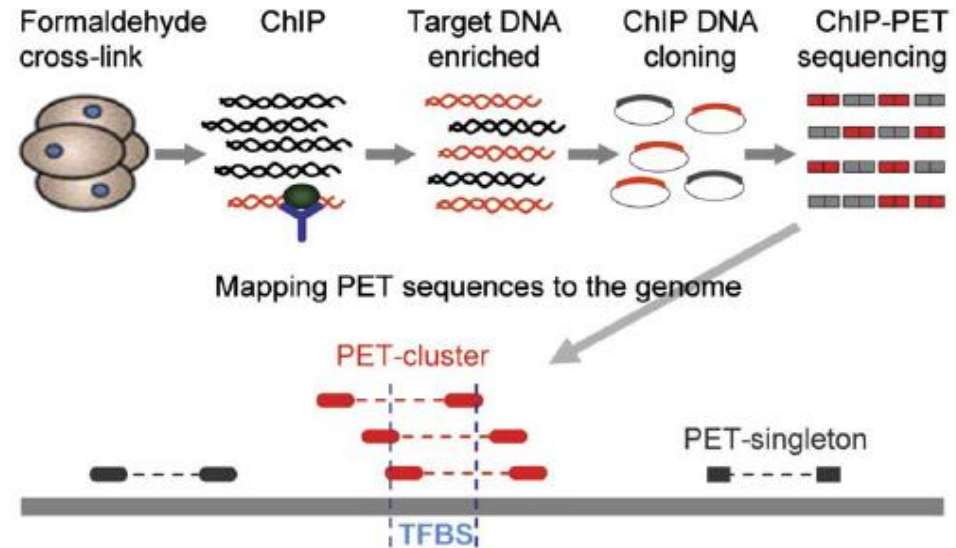
Solexa

32M samples
(2Gb / run)



Solid

40M samples
(2-6 Gb / run)



~~NGS~~ Next **Now** generation
sequencing

Wei CL et al. A global map of p53 transcription-factor binding sites in the human genome. *Cell*. 2006

**ChIP-PET (Chromatin
immunoprecipitation – Paired End
diTags)**

Микроэ́ррей (микрочипы,
микроматрицы)
Секвенирование 454 Roche
ChIP-PET
RNA-seq
ChIP-seq -> ChIA-PET, gPET

Анализ данных высокопроизводительного секвенирования ДНК и возникающие математические и алгоритмические задачи

Введение. Обзор методов высокопроизводительного секвенирования.

Ассемблирование геномов и метагеномов из коротких фрагментов:
нерешенные алгоритмические и статистические задачи

Картируемость геномов короткими фрагментами ДНК (mappability)

Определение сайтов связывания транскрипционных факторов (ТФ) в масштабе генома в профилях хроматин-иммунопреципитации (ChIP-seq)

Использование ChIP-seq данных (модификации хроматина, нуклеосомной упаковки) для предсказания сайтов связывания ТФ в геноме.

Связь локализации регуляторных сайтов относительно старта транскрипции и границ гена и дифференцировка экспрессии генов.

Анализ распределения сайтов связывания ТФ в геноме

**Рекомендуемый
ресурс:**



<http://seqanswers.com/>

Высокопроизводительное секвенирование – Next Gen Sequencing (NGS)

Genome Sequencer from Roche 454 Life Sciences (www.454.com), the Solexa Genome Analyzer from Illumina (www.illumina.com), the SOLiD System from Applied Biosystems (www.appliedbiosystems.com), the Heliscope from Helicos (www.helicos.com), and the commercialized Polonator (www.polonator.org).

Capillary electrophoresis (Sanger)

Between 96 and 384 samples
(76 - 308 Kb/run)



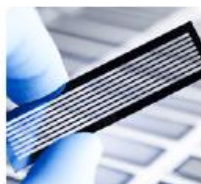
454

400,000 samples
(120Mb / run)



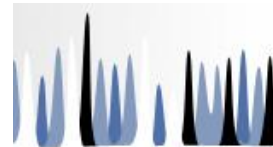
Solexa

32M samples
(2Gb / run)

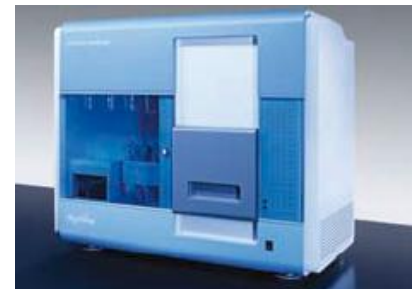


Solid

40M samples
(2-6 Gb / run)



SEQanswers
the next generation sequencing community



Ion Torrent

PacBio

CompleteGenomics

Картируемость (mappability) – «Уникома» - uniqueome - **КОНЦЕПЦИЯ**

Для каждой последовательности ДНК длиной XX нуклеотидов (обычно 18,20,25-35) определить число ее уникальных положений в референсном геноме (сверхдлинной ДНК)

/ тяжеловесная компьютерная задача /

И наоборот – для каждой позиции в хромосоме определить уникальность (число копий) фрагмента такой короткой длины

Существуют рассчитанные (добровольцами) карты для геномов человека, мыши, для фиксированных длин 35, 50 нуклеотидов (используются далеко не во всех программах)

Kohler et al (2010):

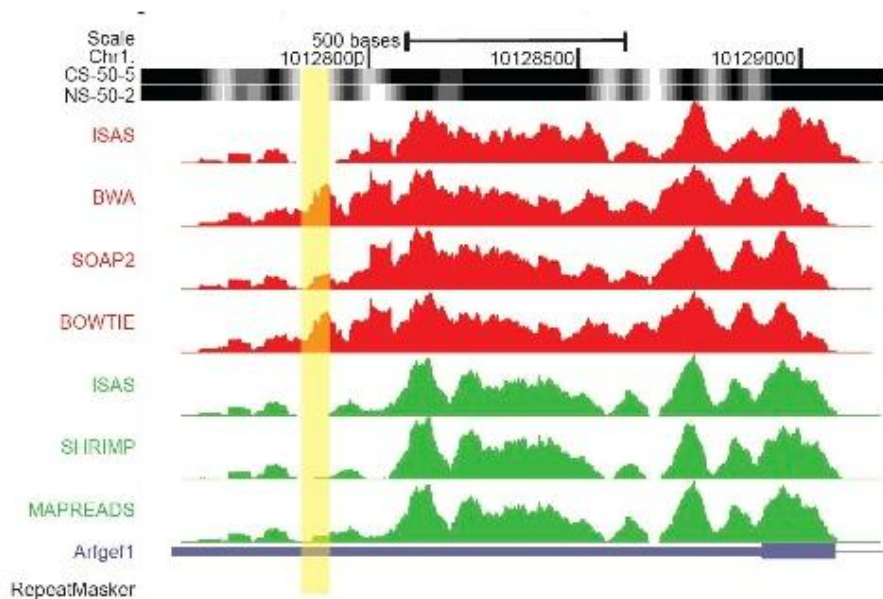


Figure 1. Color-space (CS-50-5) and nucleotide-space (NS-50-2) uniqueome plots visualized alongside RNAseq data. The same 50mer RNA-seq tags were aligned using several specialized short-read aligners in both nucleotide-space (red) and color-space (green). The yellow region highlights an area with no uniqueome coverage (confirmed by BLAT as a multi-mapping region), where tags have been falsely declared as “uniquely mapping” by the various aligners. No repetitive elements were detected by RepeatMasker. See Supplementary Methods for details.



Фрагменты (reads, «риды») ChIP-seq

Ошибки и артефакты

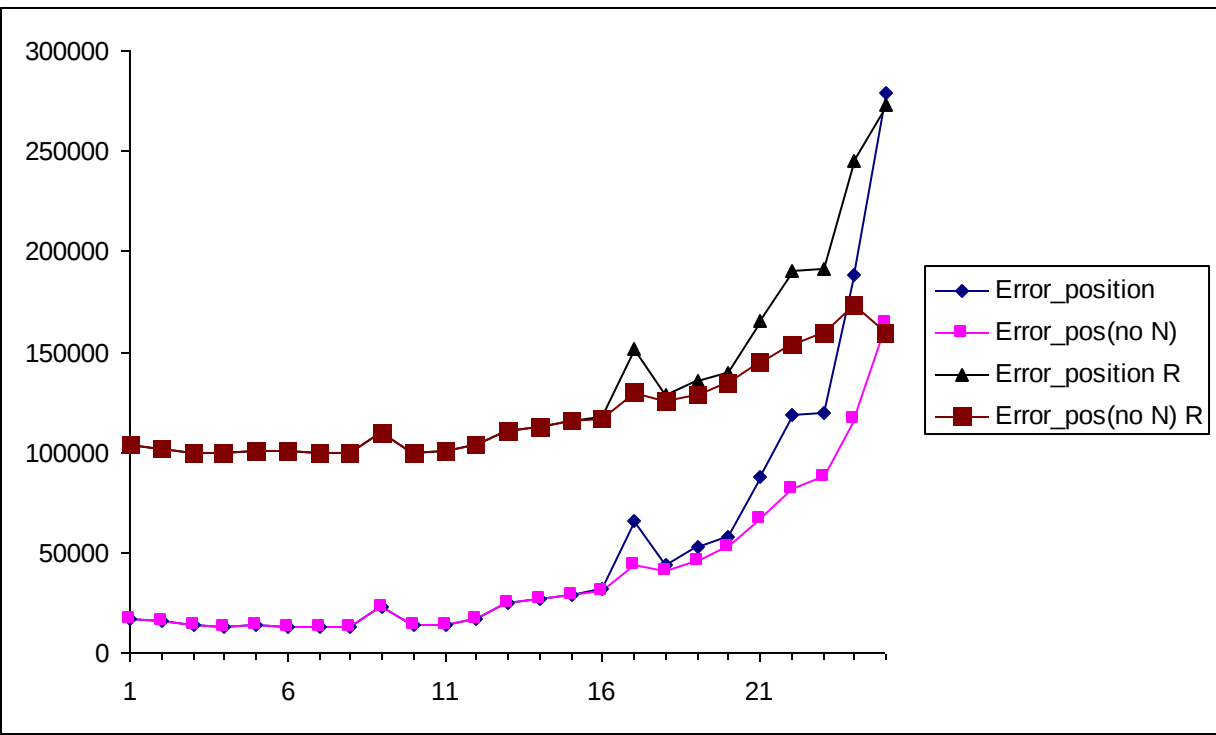
(факультативно)

Возможны ошибки (ошибки на концах - тогда надо картировать короче), либо политракты

Statistics of DNA sequence bias depends on technology.

Fraction of errors <??? (max 10%)

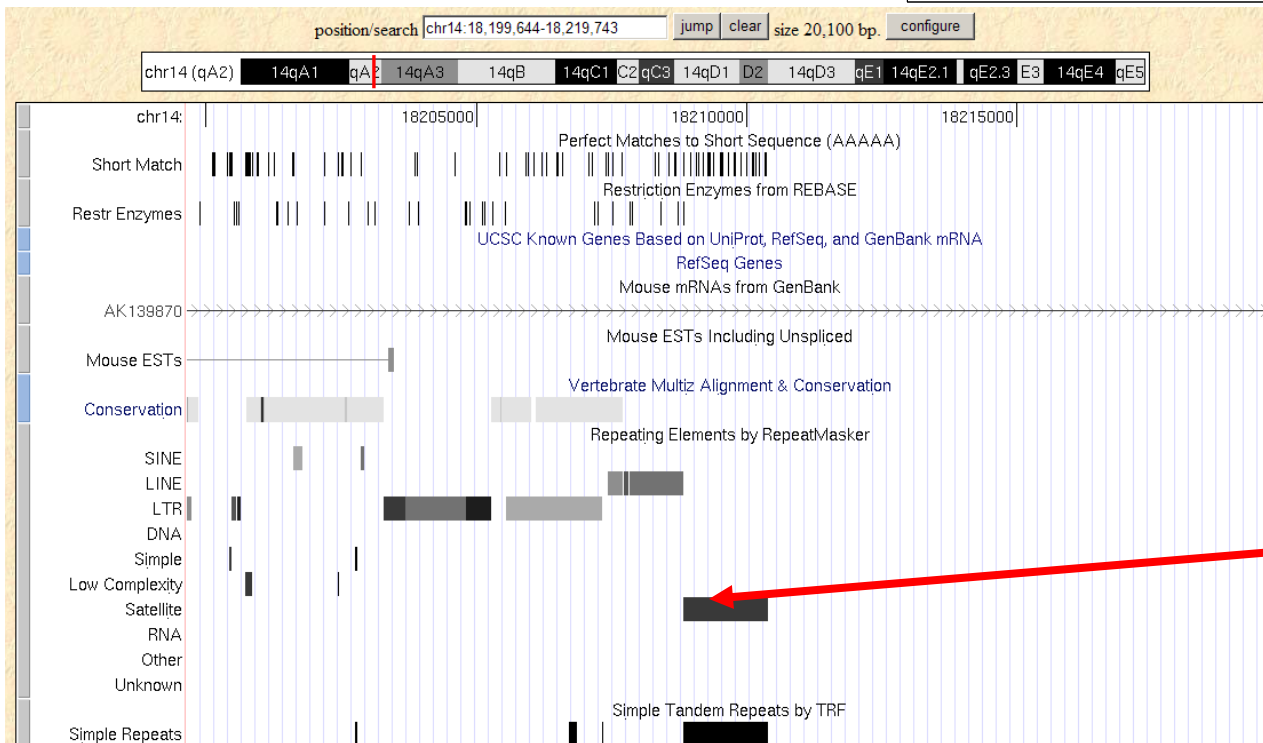
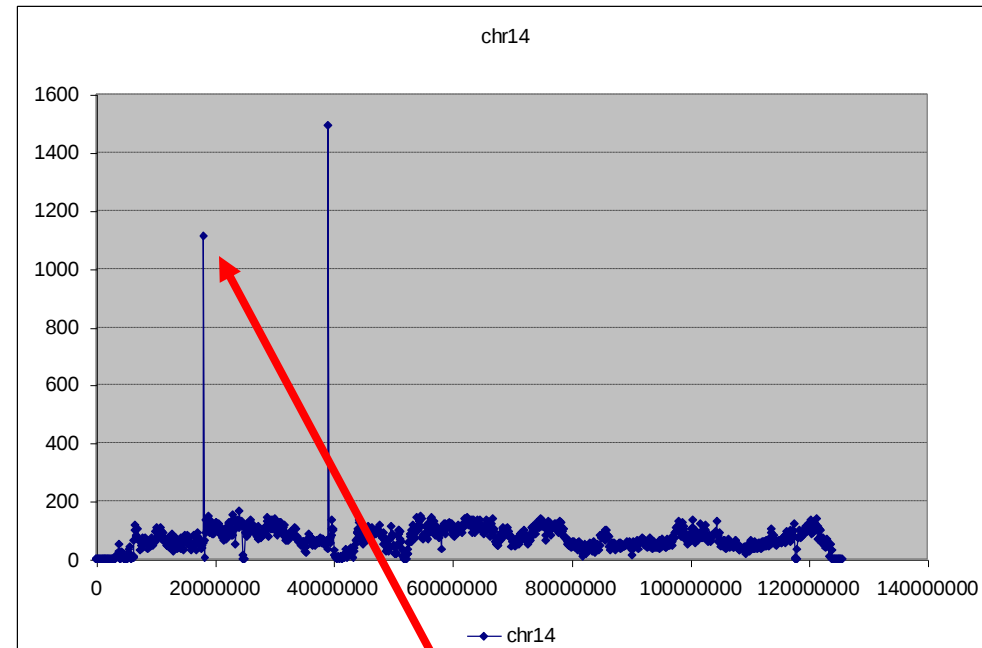
AAAAGCAACTTAGAGATTGCACCAC 12500 1 chr9:42004400 F AAAAGCAACTTAGAGATTGCACCAC 9359
GGTTTTATTATTATTTTAGGGGTTT 11453 1 chr13:116992396 F GGTTTTATTATTATTTATGGGTTT 9359
GTAATGTGTTTTTTGTGACTATTTT 11453 1 chr12:115188071 R AAAATAGACACAAAAAACACATTAC 9359
AAGATGCCAGAACTCTCAGCTCCTT 12500 1 chr7:135122171 F AAGATGCCAGAACTCTCAGCTCCTT 11453
TGCAAGCTTTCTTATCCCTTTCAT 10406 1 chr6:128978691 R ATGTAAGGGAATAAGAAAGCTGGCA 9359



Short (25-35 nt) sequence reads have bias relative to genomic DNA. Thus, first step is correct mapping of such short sequence tags to the reference genome (done)

Ложные пики (артефакты)

Artificial peaks
(technological bias in
sequencing)

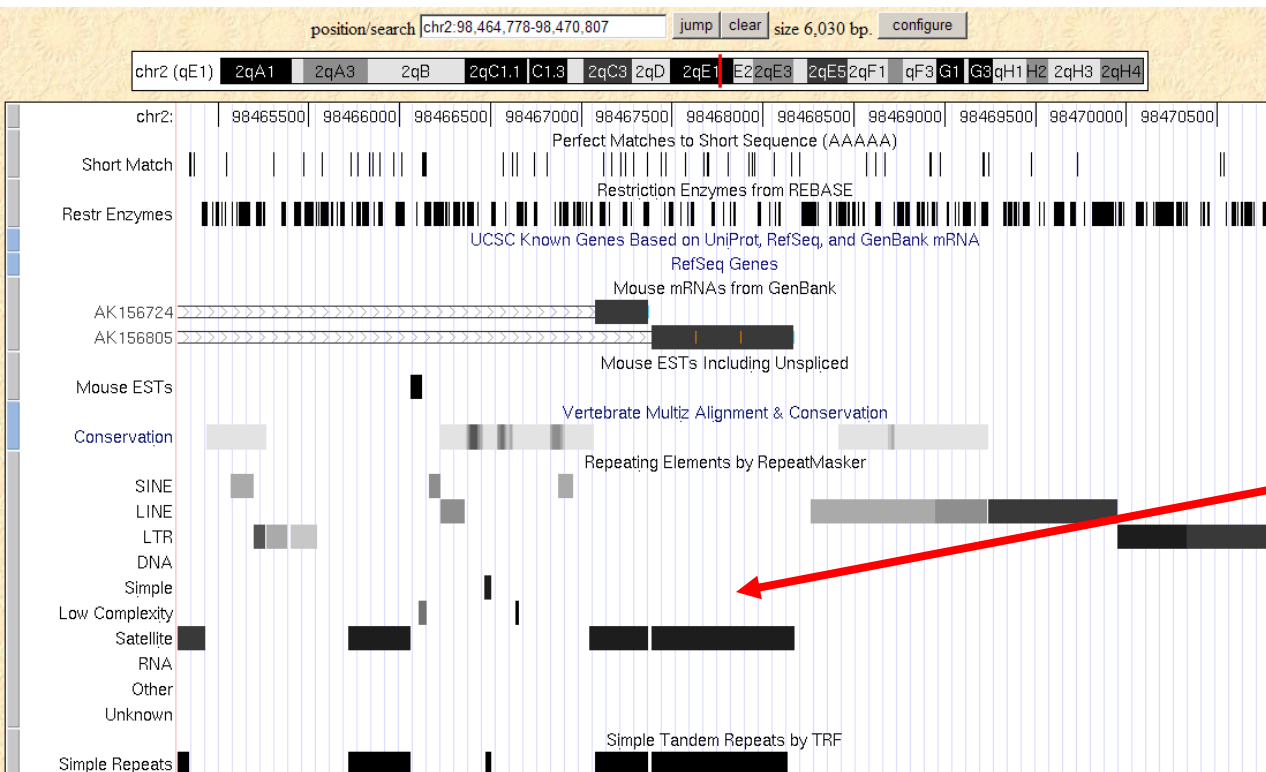
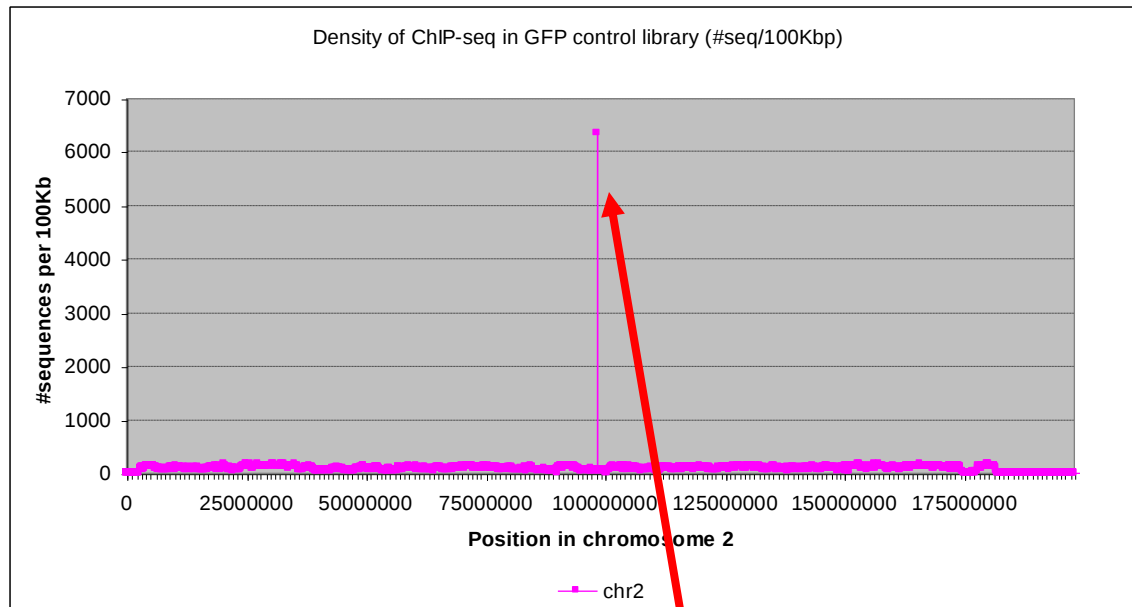


Huge cluster

No genes (no
mRNA, no EST)

Satellite repeat

Another pre-filtering example:



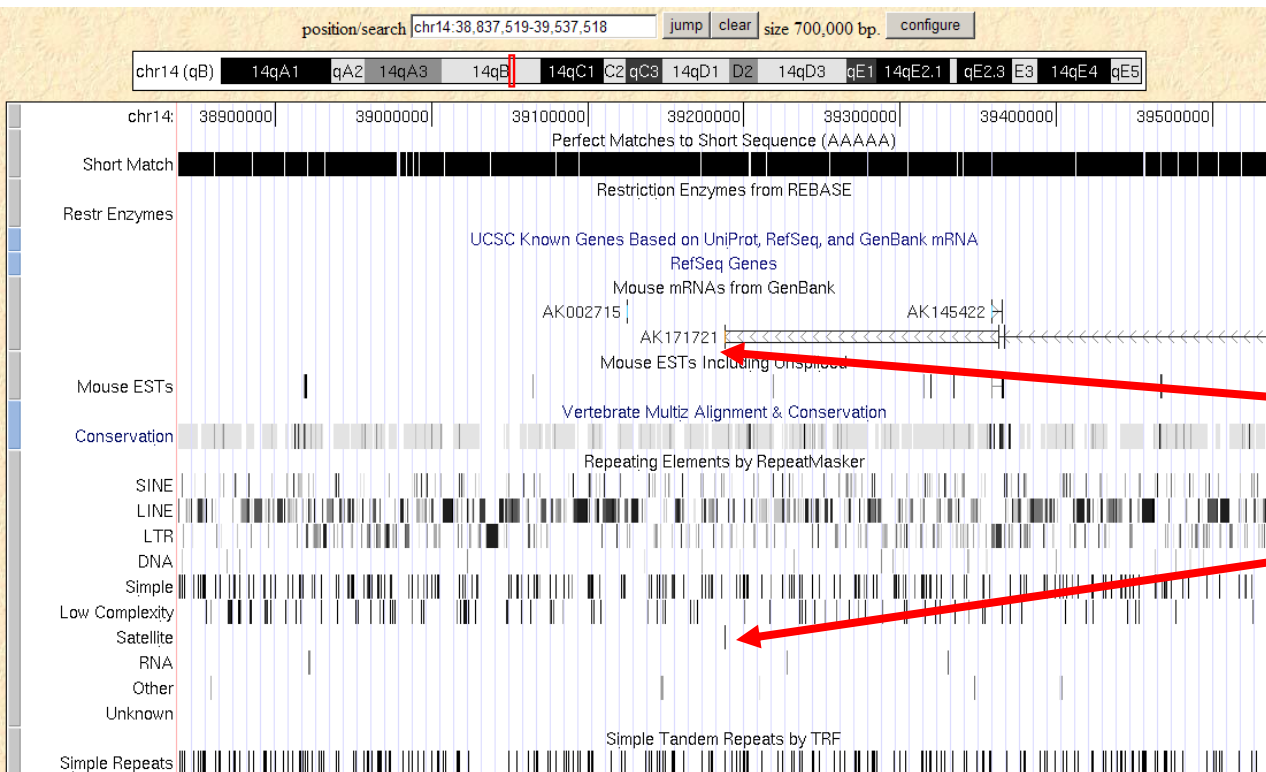
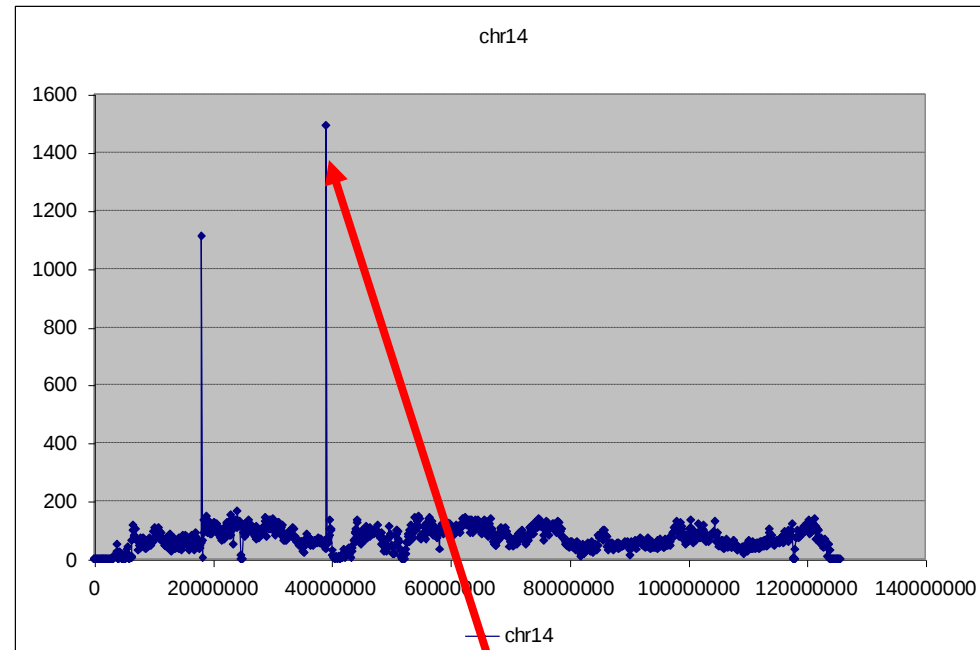
Huge cluster

Satellite repeat

Example. Analysis of artificial peaks

(too high and present in the control library)

chr14.39187519 783



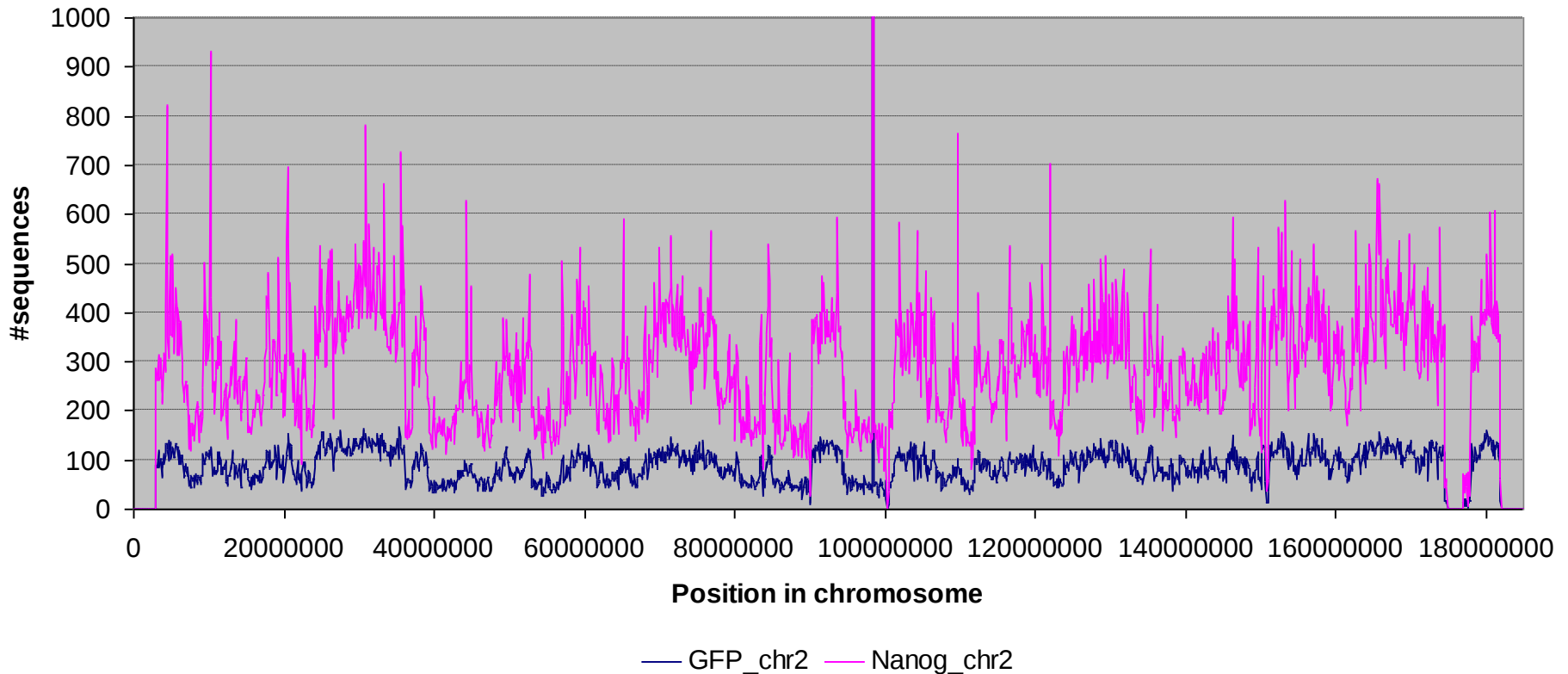
Huge cluster

No genes (no mRNA, no EST)

Satellite repeat

Специфичный и контрольный профиль ChIP-seq на хромосоме

Distributions of ChIP sequences in libraries correlate in chromosomes



Distribution of sequences follow chromatin density in chromosome.

Some extra large clusters (artifacts) still present in other libraries (maybe artifacts due to satellite repeats)

**Задача – отделить пики истинные пики
(обогащенные участки) от шума**



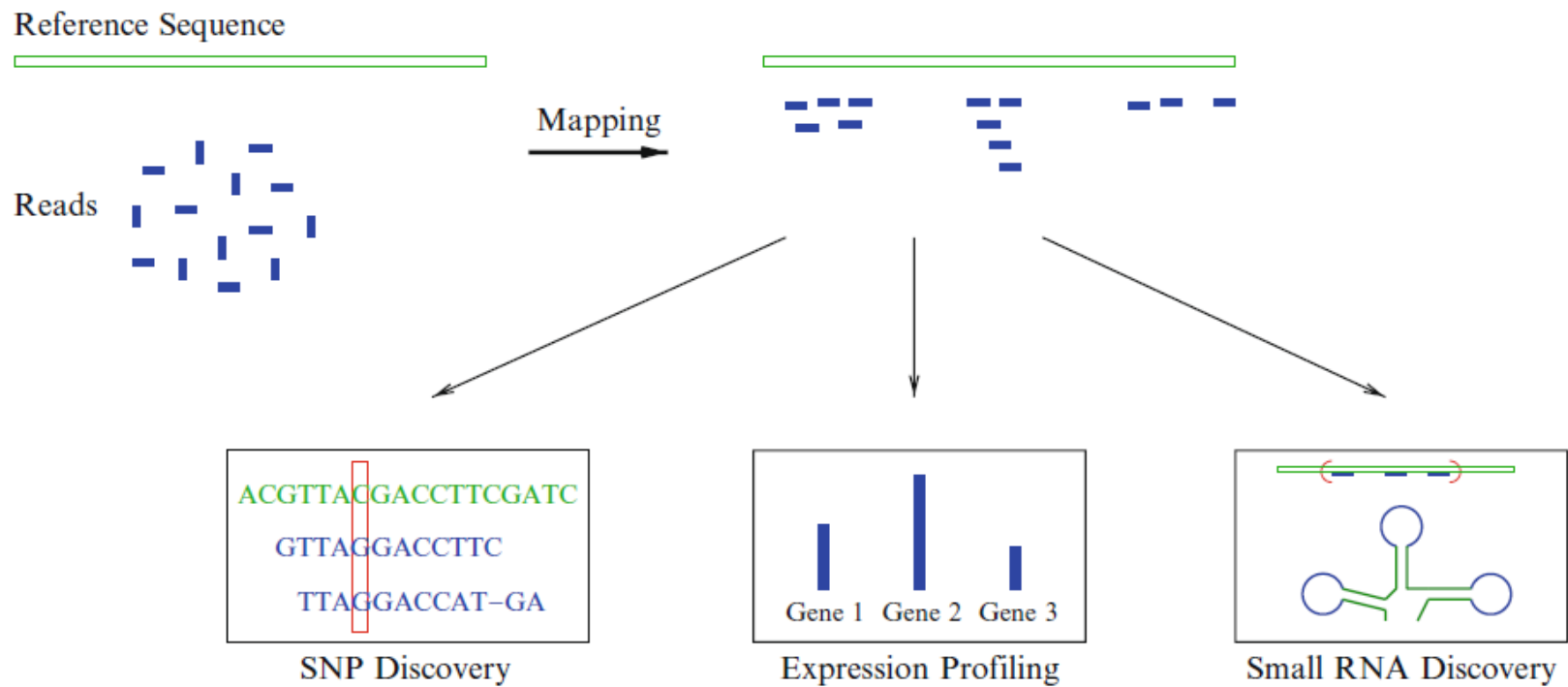
Ассемблирование сборка длинной последовательности (всего генома) из коротких (факультативно)

= Перерыв =

Картирование коротких последовательностей («рид» reads, фрагменты) и приложения (пики профилей ChIP-seq

Sequencing and Genome Assembly Using Next-Generation Technologies

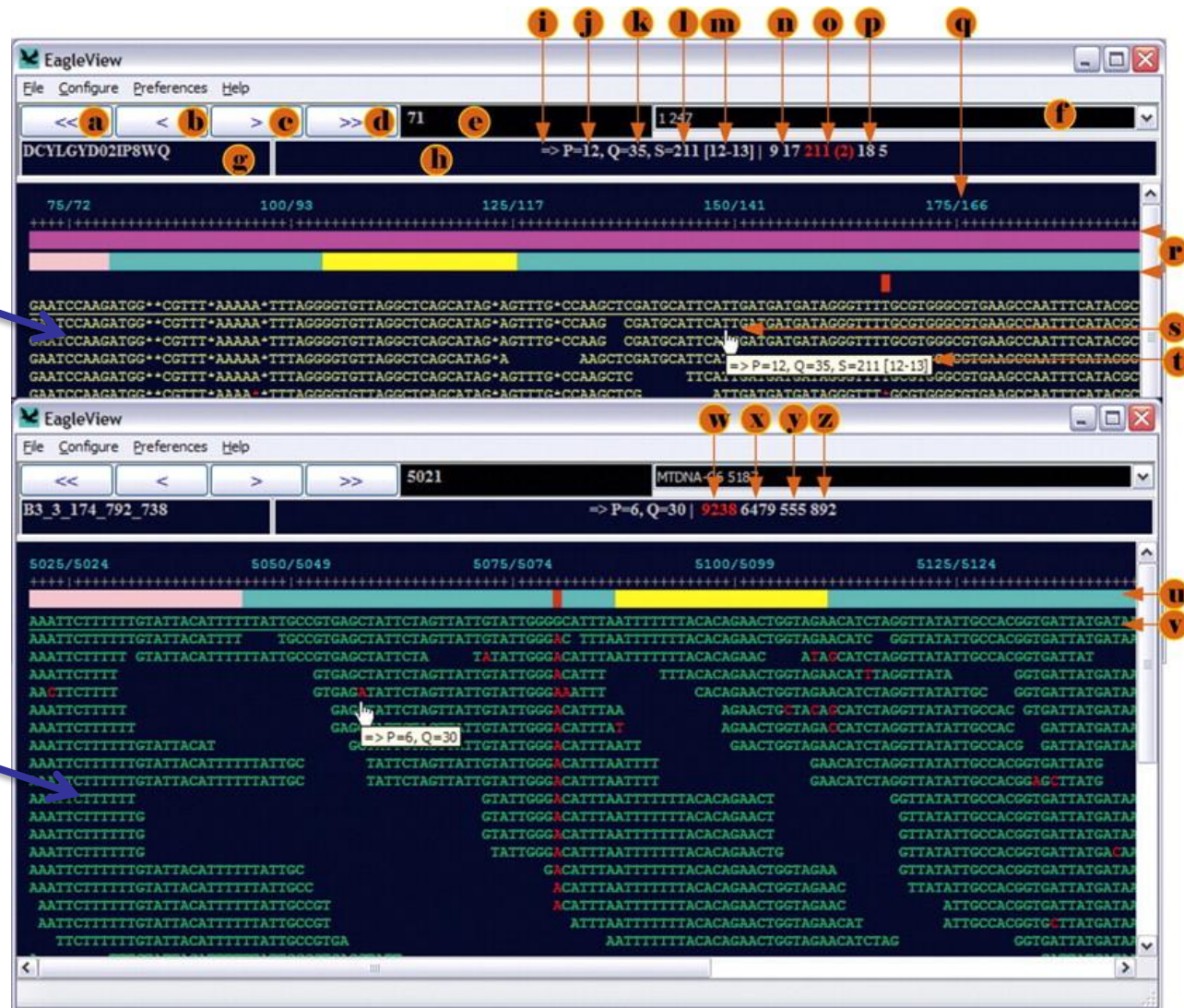
Niranjan Nagarajan and Mihai Pop



Read mapping and its applications. Mapping programs are widely used to align reads to a reference while allowing some flexibility in terms of mismatches and indels and a policy for handling ambiguous matches. The matches are then processed in different ways.

Illustration of EagleView features.

Программа визуализации картирования – как выглядит карта коротких фрагментов ДНК



Длинные
фрагменты
(454 Roche)

Короткие
«риды»
(Illumina)

Первая ступень анализа данных секвенирования – картирование на геном (на последовательности хромосом)

MUMmer and BLAT –
длинные контиги

MAQ (*Mapping and Assembly with Quality*)

SOAP (Short Oligonucleotide Alignment) Program indexes the reference instead of the reads

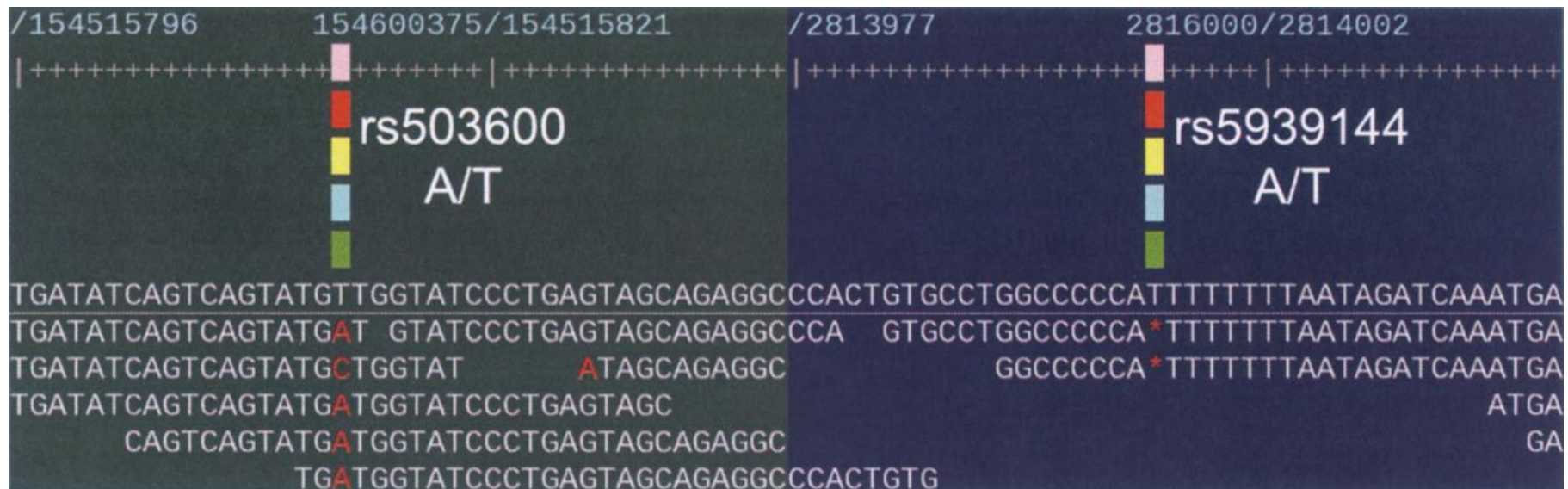
SHRiMP (SOLiD)

ELAND

Bowtie (compact Burrows–Wheeler index - BWT) 20M/час

SeqMap, RMAP, ZOOM, AMOScmp
Synamatix

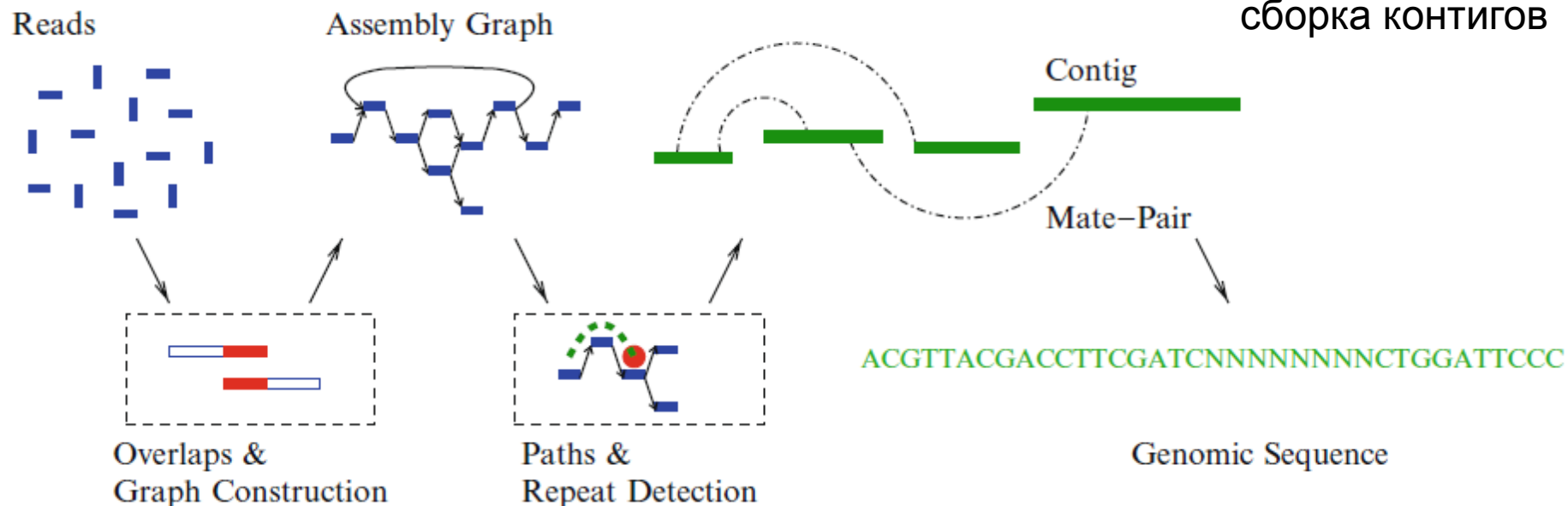
Возникающие задачи и применение
– поиск полиморфизмов, мутаций.
EagleView
Два примера (геном человека)



Two examples of SNP genotype discrepancies between the HapMap project and the new assembly data from the 1000 Genomes Project.

Сборка длинной последовательности из коротких фрагментов de novo – ассемблирование (assembly)

Идеал – один контиг (вся последовательность)
Проблема работы с короткими ~35нп – неполная сборка контигов



Понятие рид – короткая последовательность,

Контиг (составная)

Путь (path) – вариант сборки контигов

Скаффолд (составление контигов)

An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. It groups reads into contigs and contigs into scaffolds. Contigs provide a multiple sequence alignment of reads plus the consensus sequence. The scaffolds, sometimes called supercontigs or metacontigs, define the contig order and orientation.

Ассемблирование коротких последовательностей = составление одной (длинной, супердлинной) последовательности текста из коротких (20-100 нп)

Проблема NP-полна

J.R. Miller et al. / Genomics 95 (2010) 315–327

The Challenge of Assembly
Metagenomics

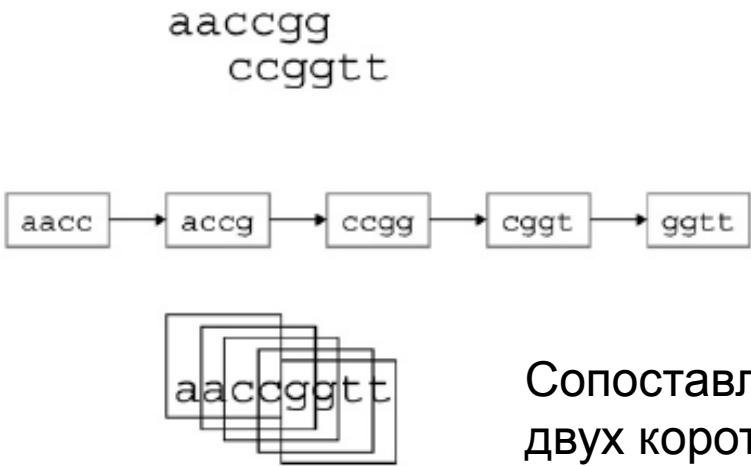
Graph Algorithms for Assembly

Greedy Graph-based Assemblers

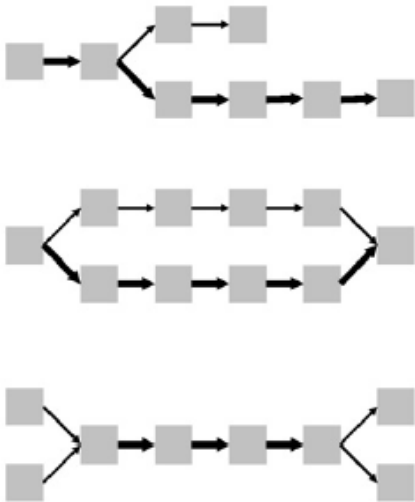
Overlap/Layout/Consensus Assemblers

The de Bruijn Graph Approach

- Euler
- Velvet
- ABySS
- AllPaths
- SOAPdenovo
- Other Software



Сопоставление
двух коротких
ридов в виде графа



Возникающие
проблемы сборки
(графа): тупики
(dead end),

Пузыри (bubbles),

Концы (веревки)

**Профиль ChIP-seq – секвенирование фрагментов ДНК,
связанных с белком методом иммунопреципитации,
Построение профиля (piled stock)**

Собственные разработки

Published:

SISSRs (Keji Zhao's)

MACs (Shirley Liu's)

PeakSeq (new, Snyder&Gerstein's)

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (**MACS**).

Genome Biol. 2008;9(9):R137. Epub 2008 Sep 17.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. **PeakSeq** enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol.* 2009 Jan;27(1):66-75. Epub 2009 Jan 4.

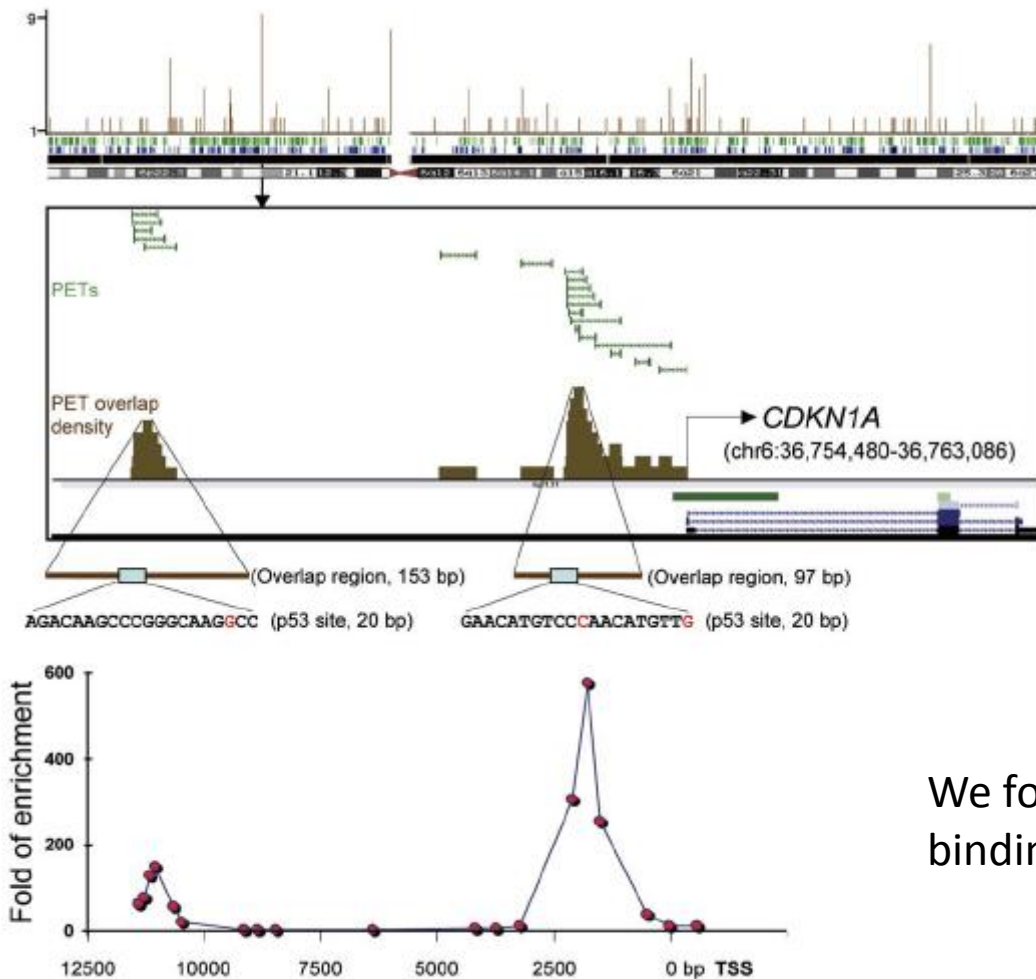
F-Seq: a feature density estimator for high-throughput sequence tags.

Boyle AP, Guinney J, Crawford GE, Furey TS. *Bioinformatics.* 2008 Nov 1;24(21):2537-8. Epub 2008 Sep 10.

QuEST -

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglu S, Myers RM, Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods.* 2008 Sep;5(9):829-34.

Список публикаций – программ анализа
профиля ChIP-seq и выделения пиков



Validation of PET-Cluster-Identified p53 Binding Loci

The whole-chromosome view of p53 ChIP-PETs mapping to chr6.

A genomic span of 23 kb that contains the CDKN1A gene and its 50 region is enlarged. CDKN1A was localized by two PET clusters; one contained 5 PETs, and the other contained 13 PETs. Both PET overlaps contained recognizable p53 binding motifs.

We found strong correlations between binding motifs and cluster size.

Большинство пиков содержат выраженные мотивы (консенсусные последовательности) сайтов связывания ТФ

Но это не соответствие один к одному (много подходящих нукл.последовательностей в геноме, не все реализованы)

Список возникающих теоретических задач связанных с ChIP-seq

Определение пиков связывания (положение сайтов) на хромосоме – peak calling (MACS, SISSoRs, PeakFinder..)

Поиск мотива (лого) в пиках – использование баз данных (TRANSFAC, JASPAR)

Сатурация (определение полноты эксперимента, насыщенности данных – все ли сайты в геноме определены в эксперименте? (через симуляционную модель или статистически)

Сравнение с данными по экспрессии генов, определение генов-мишеней, кластеризация профилей

Сравнения с нуклеосомной упаковкой (открытый хроматин, сайты разрезания ДНКазой, FAIRE и эпигенетическими маркерами – метилирование/ацетилирование гистонов (кластеризация, предсказание

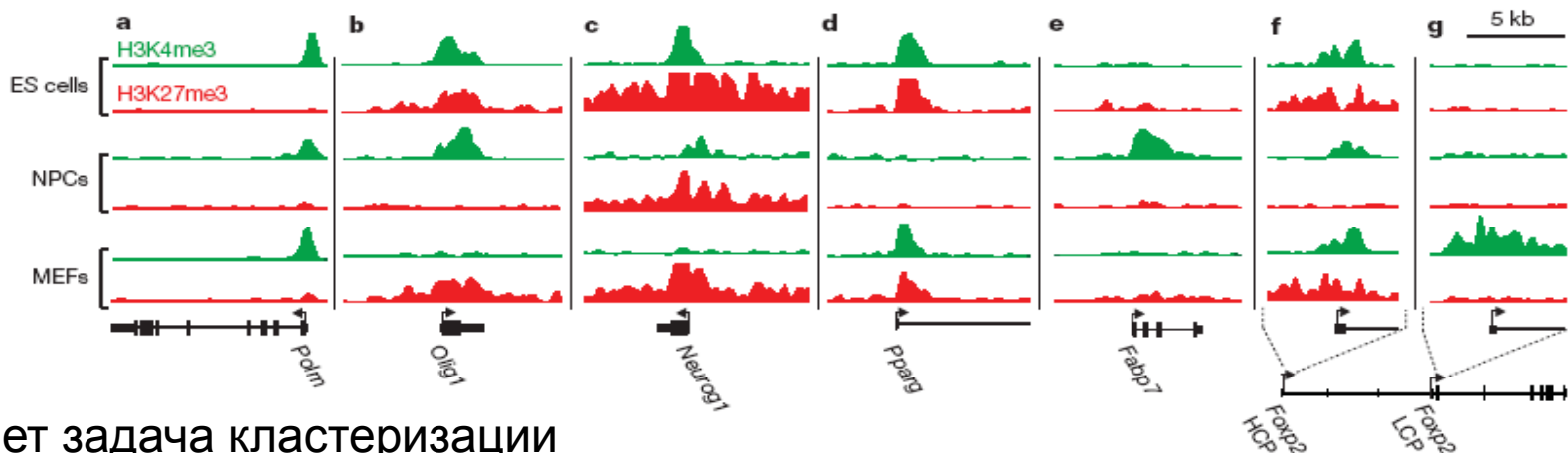
ChIP-Seq. Кластеры профилей модификаций гистонов

- Mikkelsen, et. al. *Nature*.

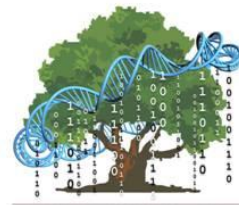
Histone modification in ES, NPC, MEF

Two methods: binned approach for sharply peaked pattern like H3K4me3; HMM based approach for broad pattern like H3K27me3.

Comparison of histone modification types among cell-lines.



Встает задача кластеризации



Профили связывания ТФ в эмбриональных стволовых клетках мыши (по данным ChIP-seq) (факультативно)

Работа по определению 15 профилей ТФ в геноме мыши

Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells

Xi Chen,^{1,2,6} Han Xu,^{3,6} Ping Yuan,¹ Fang Fang,^{1,2} Mikael Huss,⁴ Vinsensius B. Vega,³ Eleanor Wong,⁵ Yuriy L. Orlov,⁴ Weiwei Zhang,^{1,2} Jianming Jiang,^{1,2} Yui-Han Loh,^{1,2} Hock Chuan Yeo,⁴ Zhen Xuan Yeo,⁴ Vipin Narang,³ Kunde Ramamoorthy Govindarajan,³ Bernard Leong,³ Atif Shahab,³ Yijun Ruan,⁵ Guillaume Bourque,³ Wing-Kin Sung,³ Neil D. Clarke,⁴ Chia-Lin Wei,^{5,*} and Huck-Hui Ng^{1,2,*}

¹Gene Regulation Laboratory, Genome Institute of Singapore, Singapore 138672

²Department of Biological Sciences, National University of Singapore, Singapore 117543

³Computational and Mathematical Biology

⁴Computational and Systems Biology Group

⁵Genome Technology and Biology Group

Genome Institute of Singapore, Singapore 138672

**Статья Chen et al., Cell, 2008 –
пример массового
полногеномного анализа
профилей ChIP-seq**

Пошаговая реализация задач

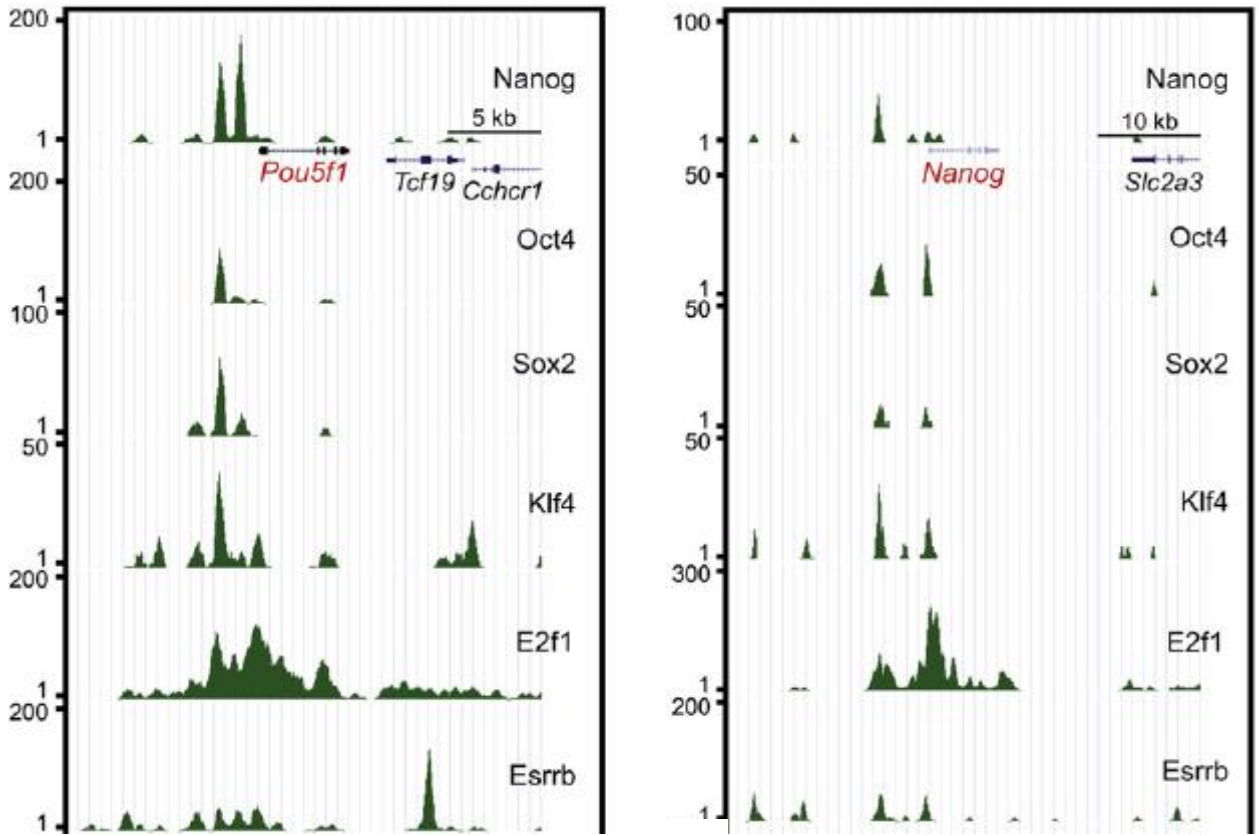
- Extending each tag by 200bp towards its direction to generate a pseudo-PET and overlapping pseudo-PETs into an genome-wide intensity profile;
- Applying peak finding algorithm on the intensity profile;
- A cut-off threshold determined by randomly placing tags in the genome and calculating false discovery rate (FDR);
- Removing trivial peaks by the fold-change against control library (<5 fold).

Как выглядят профили на хромосоме (сравнение пиков связывания для различных белков – транскрипционных факторов)

Вопрос – сколько сайтов в геноме (человека)? Локализуются ли они вместе?

Большое число сайтов в геноме определено экспериментально с помощью ChIP-seq (от 5,000 до 40,000 в зависимости от белка).

Невозможно проверить их все экспериментально традиционными методами (PCR)

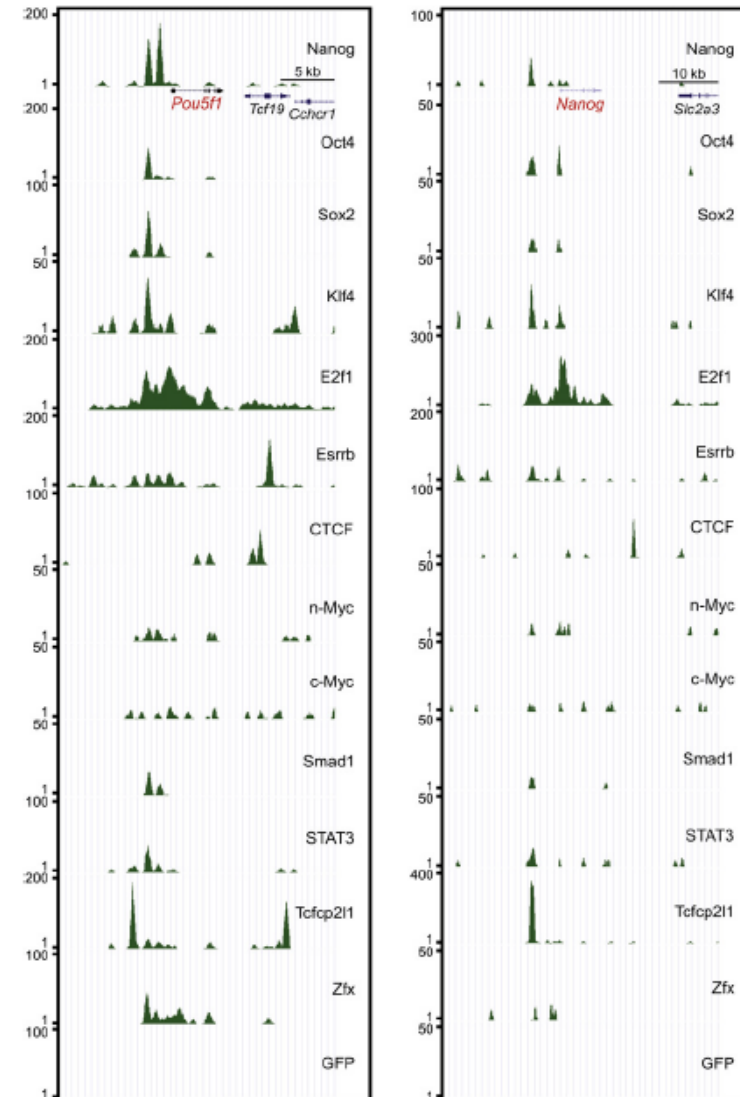


TFBS profiles for the sequence-specific transcription factors at the Pou5f1 and Nanog gene loci

Примеры
расположения пиков
профилей ChIPseq в
локусах генов
Pou5f1(Oct4) и Nanog
(в геноме мыши)

Суммарная таблица (объем данных – сотни мегабайт только профили ChIP-seq)

	total uniquely mapped loci, excluding chrY	peak number	false discovery rate	positive rate by PCR validation (> 3 fold)
Nanog	8424102	10343	0.01	100%
Oct4	4911144	3761	0.01	>92%
Sox2	4821446	4526	0.01	100%
Smad1	3338896	1126	0.05	96%
P300	5424357	524	0.05	>90%
E2f1	8449181	20699	0.01	100%
Tcfcp2i1	8787961	26910	0.01	>92%
Suz12	5986520	4215	0.01	100%
CTCF	3686056	39609	0.01	>95%
Zfx	3844429	10338	0.01	100%
Stat3	5351116	2546	0.01	>97%
Klf4	3807970	10875	0.01	>95%
Nmyc	4823212	7182	0.01	100%
Cmyc	6637404	3422	0.01	>95%
Esrrb	7148528	47888	0.01	<50%
Tbp	3528988	7405	0.05	>95%
GFP (control)	3578466	-	-	

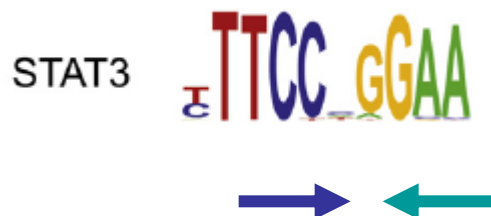


Два примера профилей – на 57
хромосомах мыши для всех ТФ

Весовые матрицы, определенные по пикам ChIP-seq (тысячи последовательностей)



Examples of similar (overlapping) motifs



Example of strict motif (symmetry)

Весовые матрицы – лого – и возникающие задачи:

Определение именно по данным ChIP-seq – подгонка под пики, выбор вершины пика

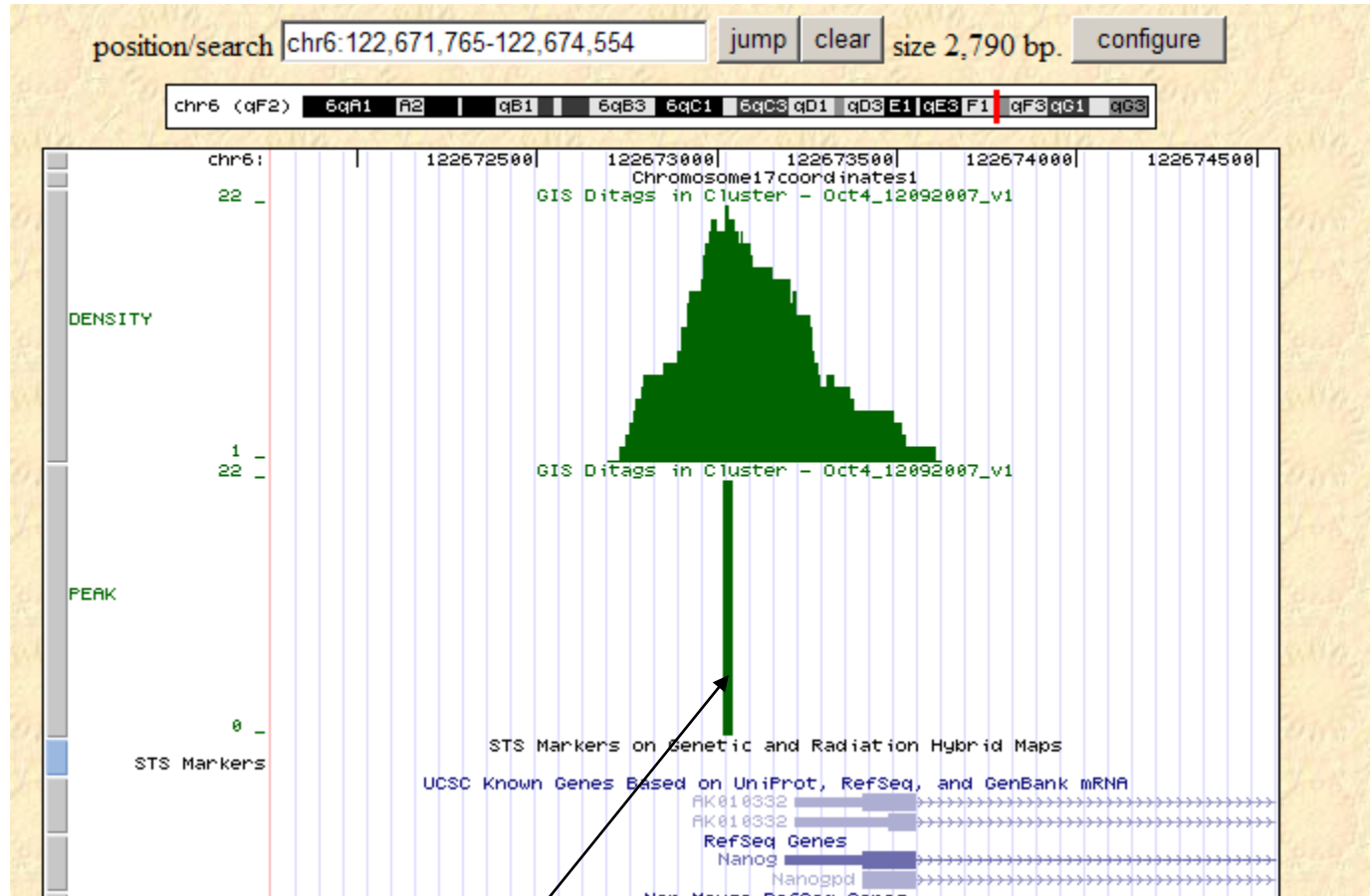
Перекрывание сайтов (кофакторы, piggy-back)

Перепредставленные (простые) последовательности

Пример взаимной регуляции (Oct4 -> NANOG)

Сайт связывания Oct4 находится в промоторе гена Nanog

Oct4 binding site near Nanog gene

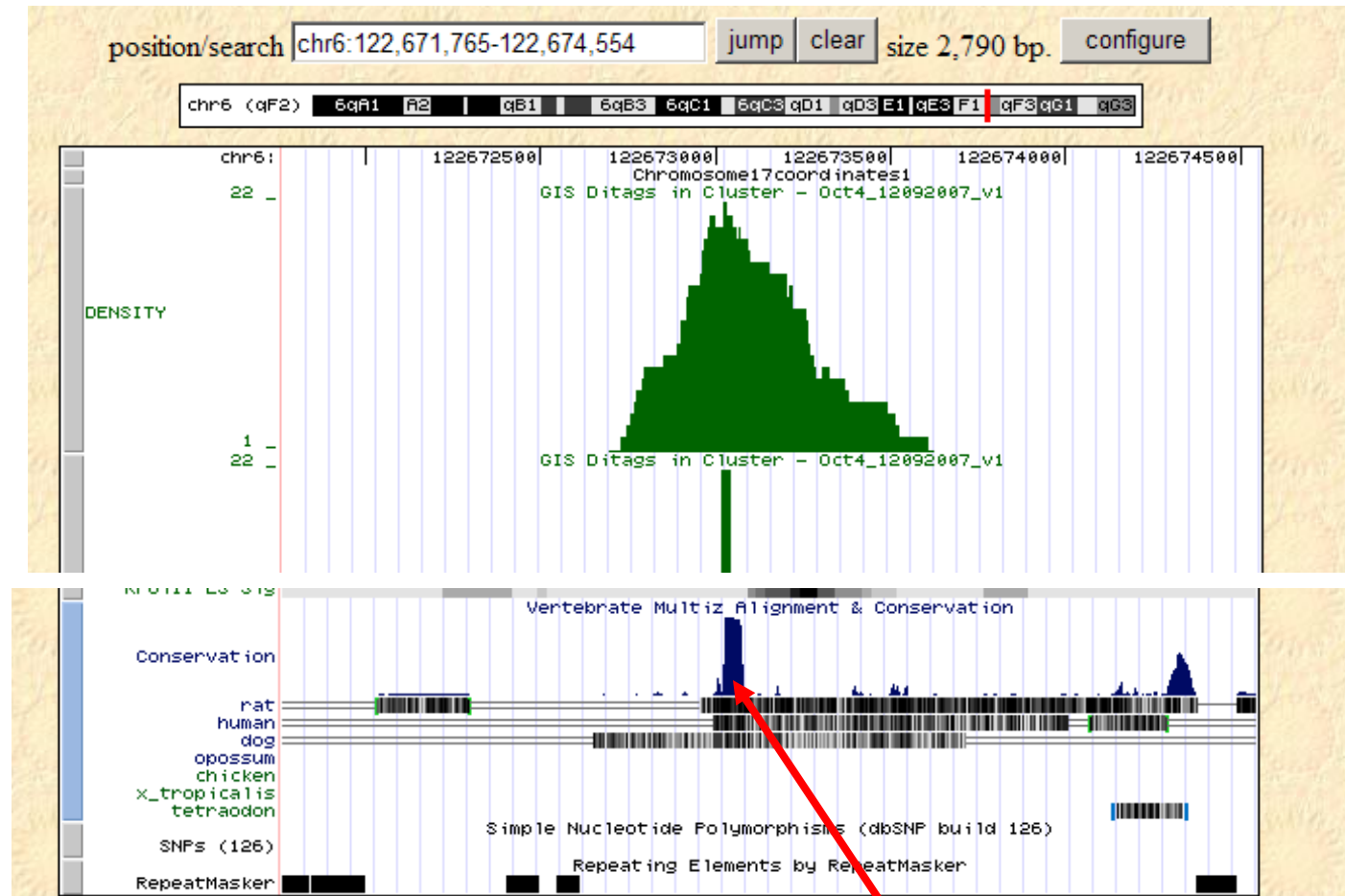


Oct4

Transcription start of Nanog gene

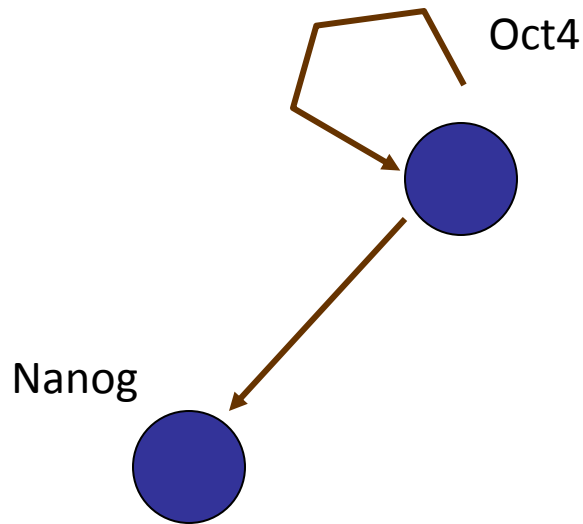
Кроме того, данный сайт консервативен
(у мыши и других организмов)

Oct4 binding site near Nanog gene



Evolutionary conservation of
Oct4 binding site

Примеры регуляторных взаимодействий по профилям ChIP-seq



Oct4 регулирует Nanog (через связывание в промоторе) и регулирует сам себя

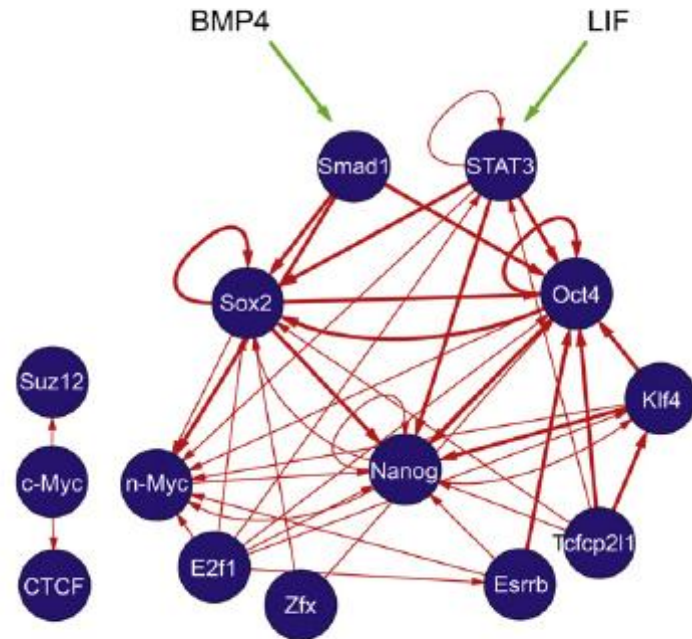


Figure 7. Transcriptional Regulatory Network in ES Cells

Восстановление регуляторных контуров («кто кого» регулирует по расположению в промоторном районе соответствующего гена,

Дополнительно (толщина стрелок) – по дифференциальной экспрессии на микрочипах

Задача надежности определения сайтов

- 1) Сколько из найденных по профилю связывания пиков
Содержат сайты связывания транскрипционных факторов**
- 2) Сколько сайтов в геноме?**

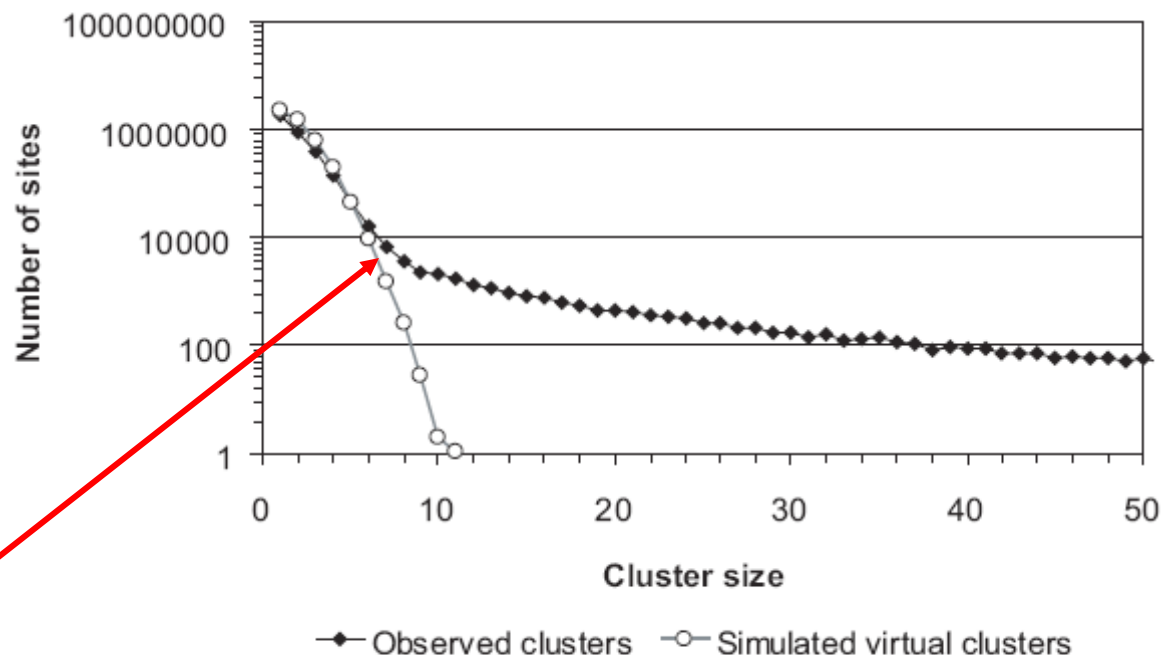
Задача надежности определения сайтов

1) Сколько из найденных по профилю связывания пиков

Содержат сайты связывания транскрипционных факторов

2) Сколько сайтов в геноме?

Observed and simulated
cluster size distributions in
Nanog
ChIP-seq library (numbers
are after GFP filtering).

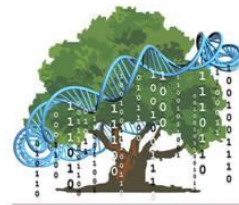


Noise filtering cut-off value

Cut-off value for ChIP-seq peaks – определение
порогового значения высоты пика

Статистическая задача – определить пороги и отношение
между высотой пика при специфическом секвенировании
и при контрольном чтобы минимизировать ошибку

(в качестве контроля – компьютерное симулирование и
секвенирование IgG, GFP, ДНК без связи с белками)



Профили связывания ТФ и эпигенетика: Эстрадиоловый рецептор (ER)

Integrative model of genomic factors for determining binding site selection by estrogen receptor- α

Roy Joseph¹, Yuriy L Orlov¹, Mikael Huss¹, Wenjie Sun, Say Li Kong, Leena Ukil, You Fu Pan, Guoliang Li, Michael Lim, Jane S Thomsen, Yijun Ruan, Neil D Clarke, Shyam Prabhakar, Edwin Cheung and Edison T Liu^{1,*}

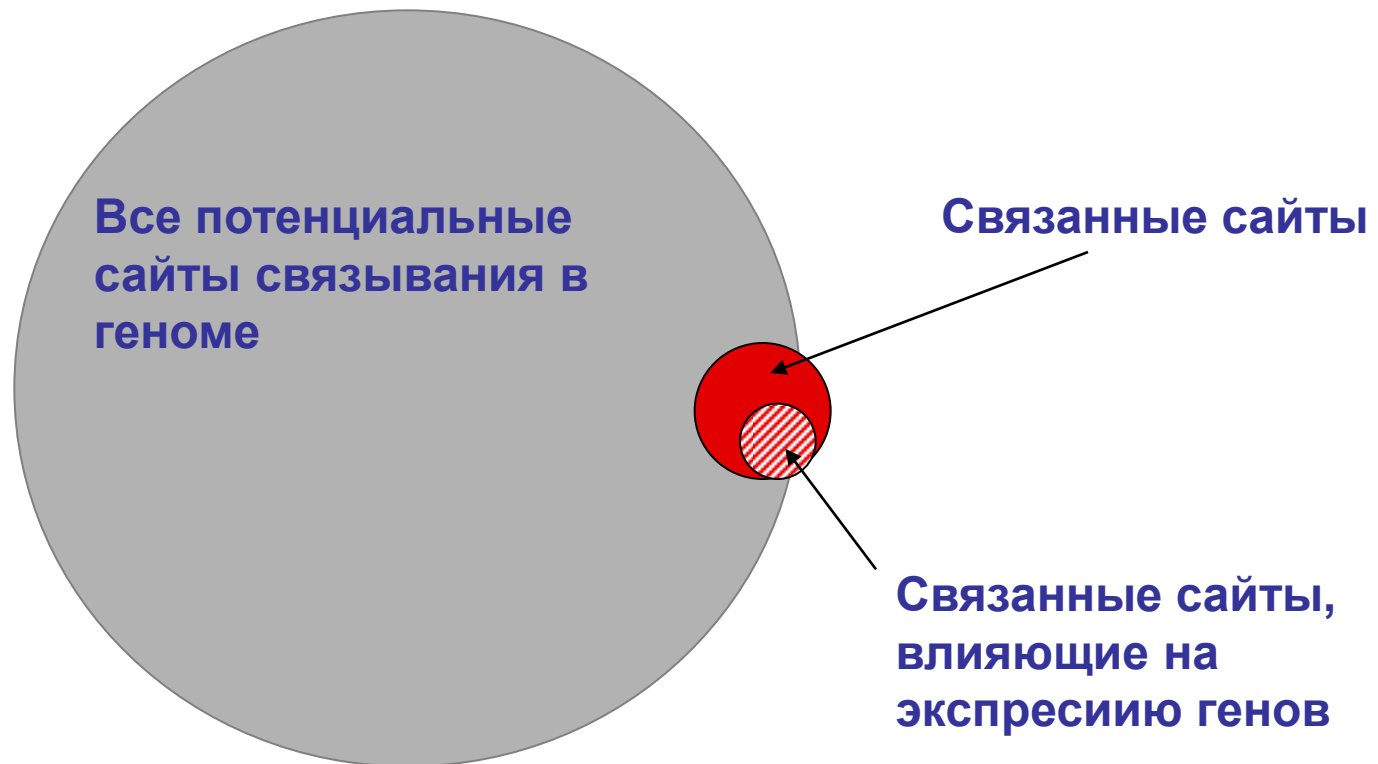
Обзор работы по статье (декабрь 2010)

По анализу сайтов связывания ТФ эстроген-рецептора в геноме человека – экспериментальные данные получены с помощью высокопроизводительного секвенирования – ChIP-seq

Задача – исследование почему именно эти сайты в геноме функциональны – интегрирующая модель



Общая идея – из огромного набора потенциальных сайтов связывания ТФ в геноме (определенном только по последовательности букв 6-10 нт например матрицей связывания (PWM)) лишь малая часть действительно связывается с ТФ (1-2%) как подтверждено прямым экспериментом по секвенированию



Рассмотрим 3 работы - Joseph, Orlov et al (2010)

Ernst et al (2010)

John, Sabo et al (2011)

Предсказание сайтов связывания транскрипционных факторов в геноме и их экспериментальная проверка (валидация)

Сайты связывания ER (геном человека)

Только по последовательности ДНК –

~1 млн сайтов в геноме человека по консенсусу (с несовпадениями)

32 тыс.сайтов по позиционной весовой матрице

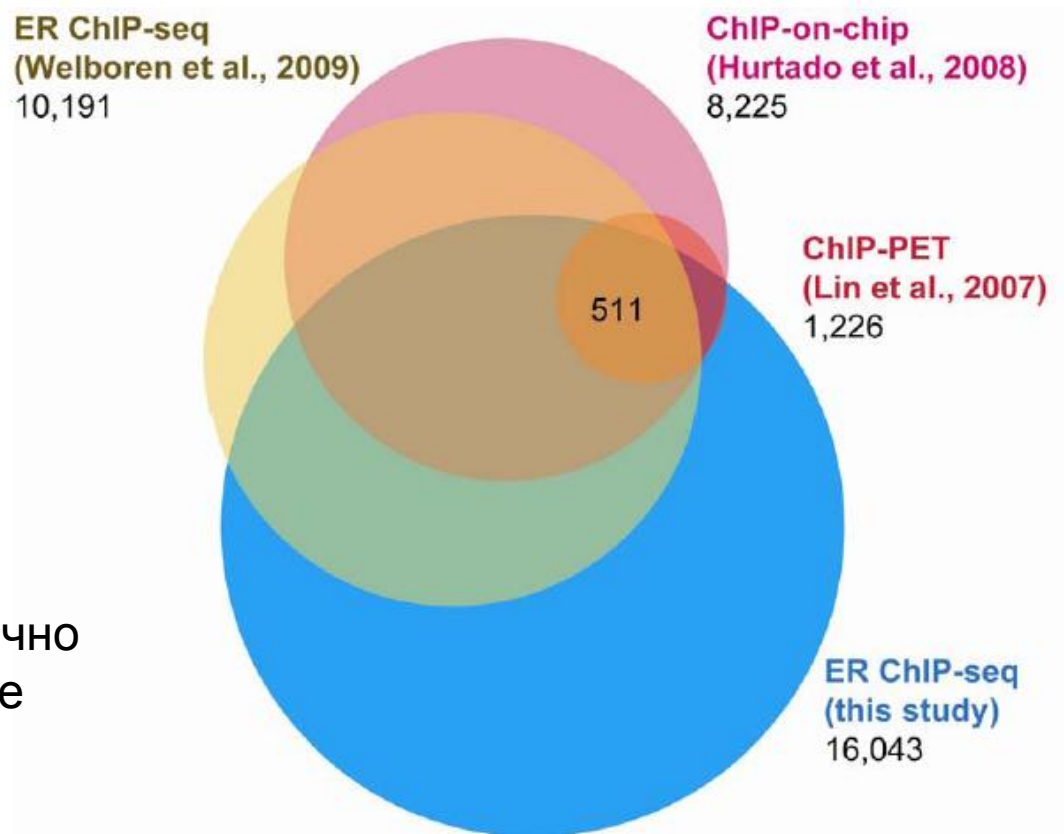
17 тыс. сайтов – найдено через секвенирование сопряженное с иммунопреципитацией (ChIP-seq) в культуре раковых клеток MCF-7

Около 100 сайтов подтверждено экспериментально с помощью qPCR (ручная технология, обычно проверяют десятки сайтов)

Эксперимент микроэrray (Microarray, платформа Affymetrix)

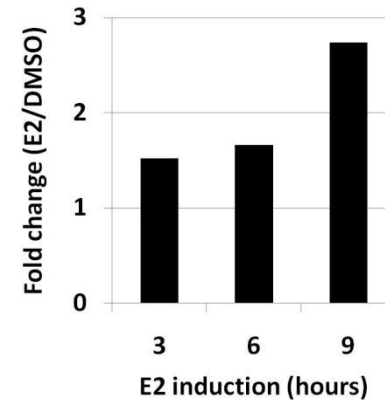
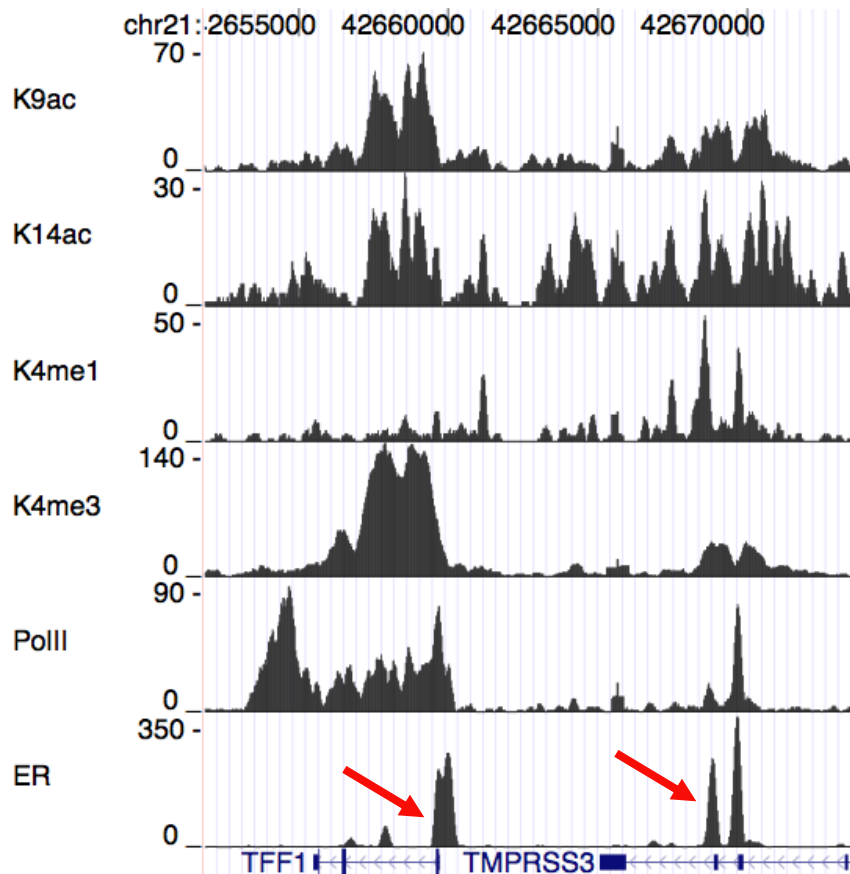
Сайты могут быть специфичны к данному типу клеток (тканеспецифичность)

Воспроизводимость ChIP-seq экспериментов (сколько всего сайтов в геноме может быть определено через технологии секвенирования - оценки для разных ТФ с 2005 года)



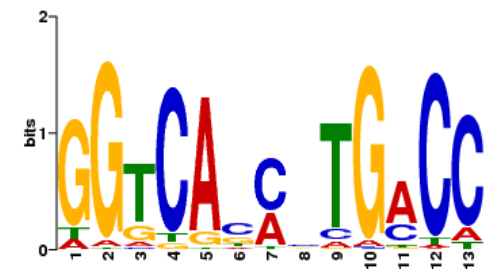
Различные
технологии
секвенирования
дают лишь частично
воспроизводимые
результаты

Пример расположения маркеров модификаций гистона H3 на хромосоме и сайтов связывания ER по ChIP-seq

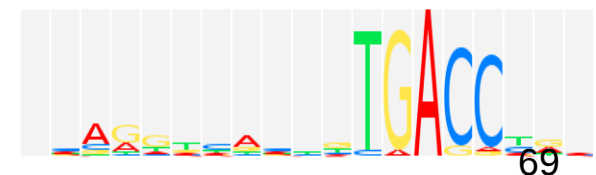


Экспрессия гена TFF1 после активации эстрогеном E2 (увеличение в разы, логарифм. шкала)

Уточнение мотива связывания по данным ChIP-seq



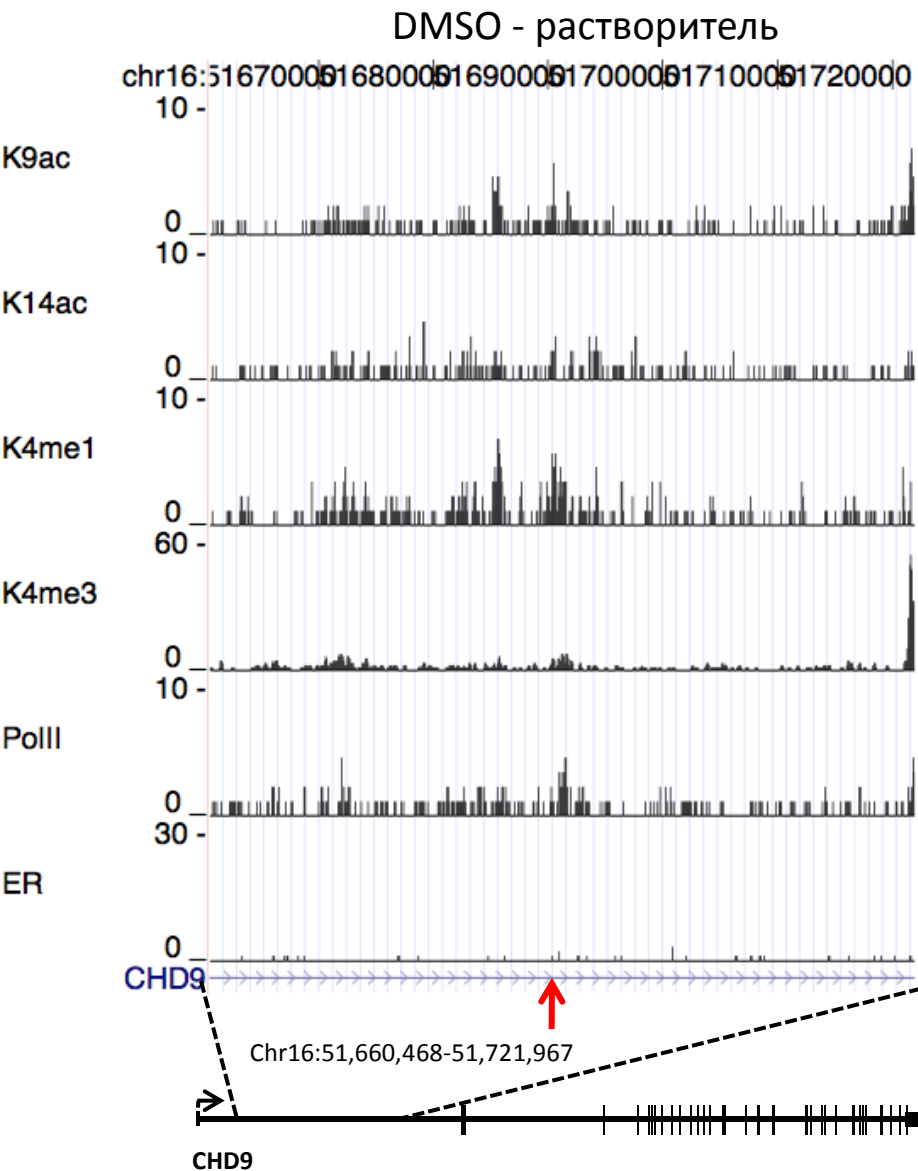
ERE motif (ChIP-seq)



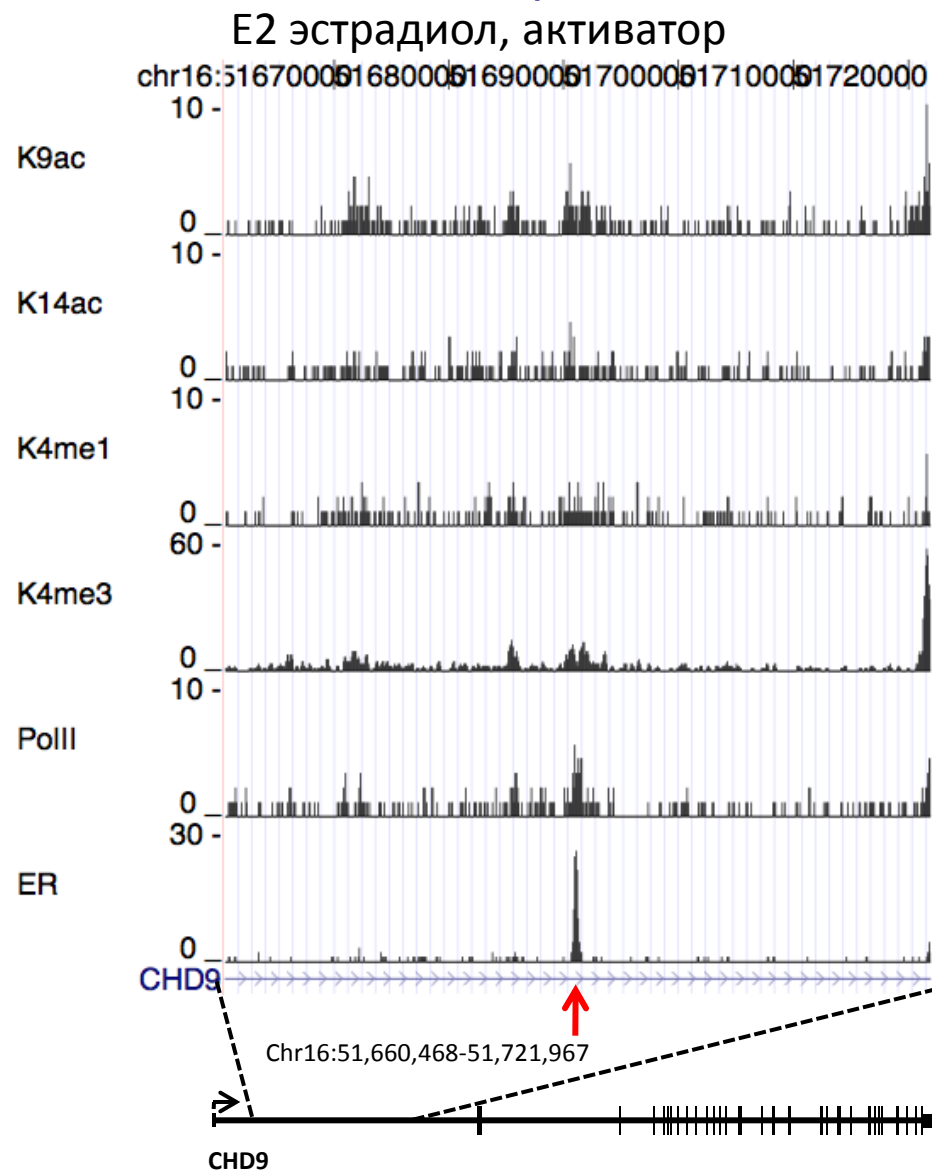
TRANSFAC logo

Сайты связывания ER в регуляторных районах гена TFF1 (соответствуют пику модификаций хроматина, гистона H3 - H3K9ac, H3K4me3 и др.)

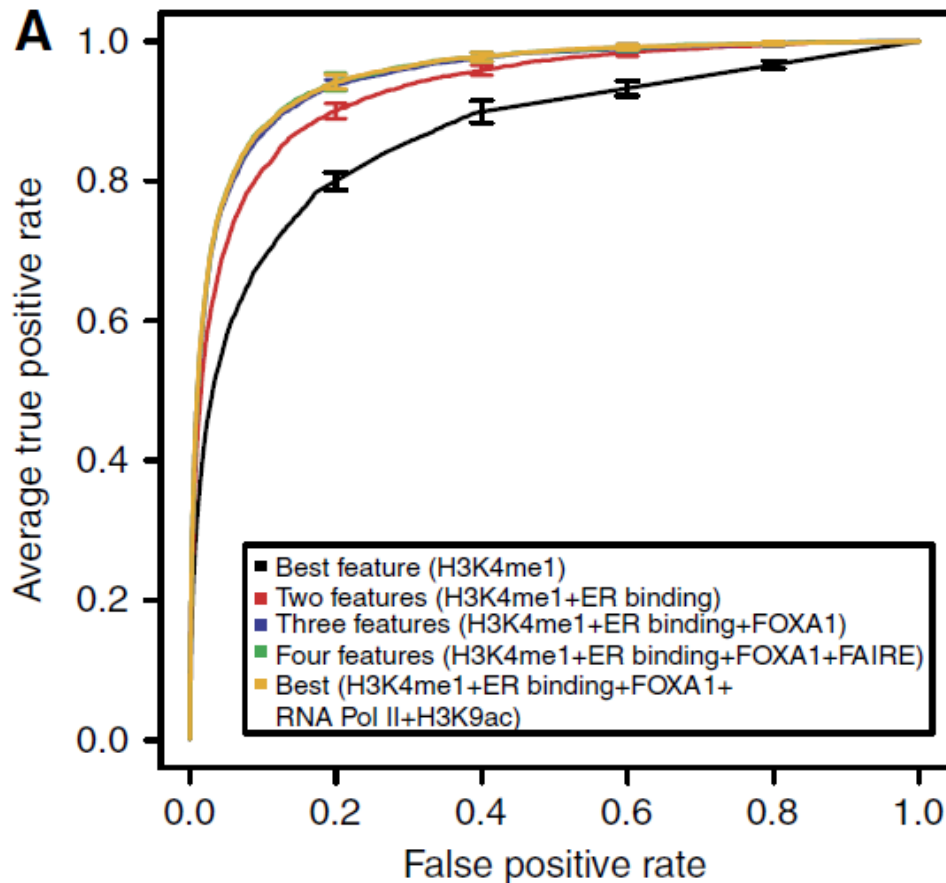
Клетки «в покое»



Клетки после активации эстрадиолом

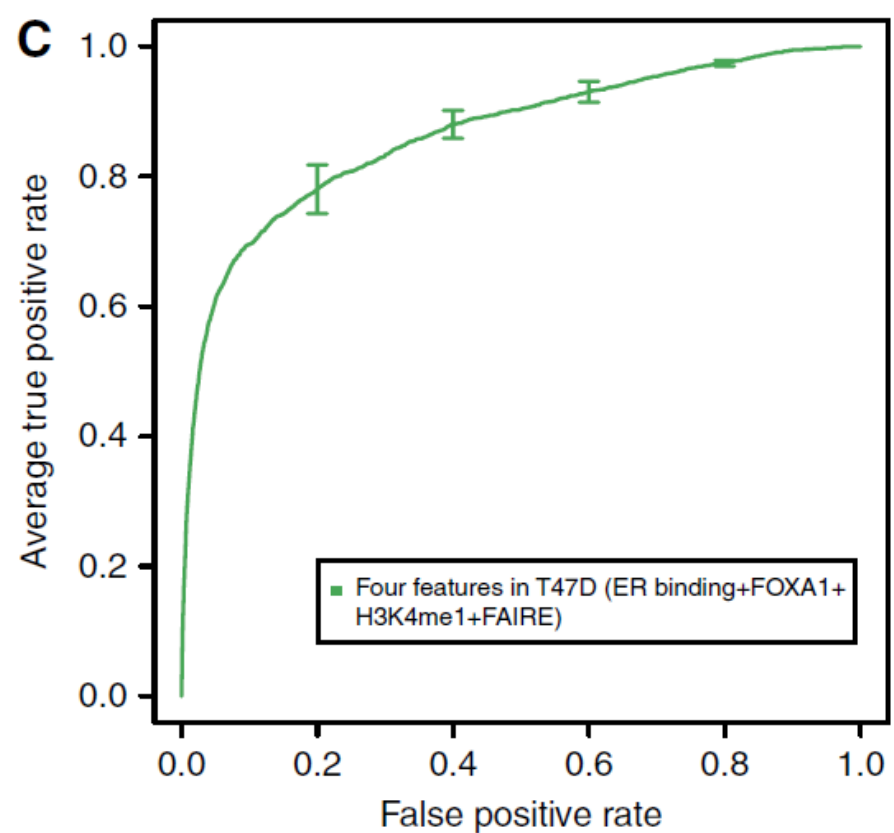
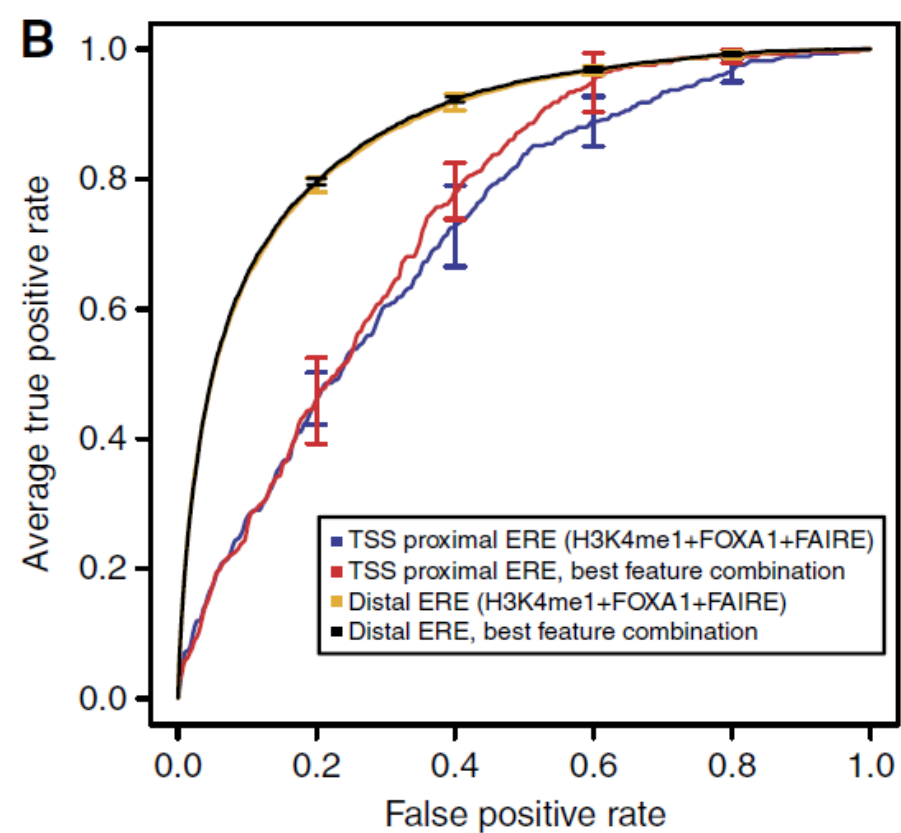


Пример расположения маркеров модификаций гистонов на хромосоме и связывание ER (из геномного браузера)
Видно появление пика ER (красная стрелка)



Пример предсказания
сайтов ER в геноме
человека только по данным
модификаций хроматина
ROC-AUC – общее значение
достаточно большое, 0.7-0.8

Предсказание сайтов связывания с помощью различных факторов модели (факторы не связанные с мотивом (нуклеотидной последовательностью), а только с геномным окружением и эпигенетическими факторами



Предсказание сайтов связывания для
промоторных районов

И для другого типа клеток (T47D) с помощью той
же модели



Предсказание сайтов в геноме: к общей теории и общим свойствам района хромосом



Общий профиль предсказания связывания транскрипционных факторов в геноме (по профилям модификаций хроматина ChIP-seq и другим геномным особенностям). Статья Ernst и соавторов.

Method

Integrating multiple evidence sources to predict transcription factor binding in the human genome

Jason Ernst,^{1,4} Heather L. Plasterer,^{2,5} Itamar Simon,³ and Ziv Bar-Joseph^{1,6}

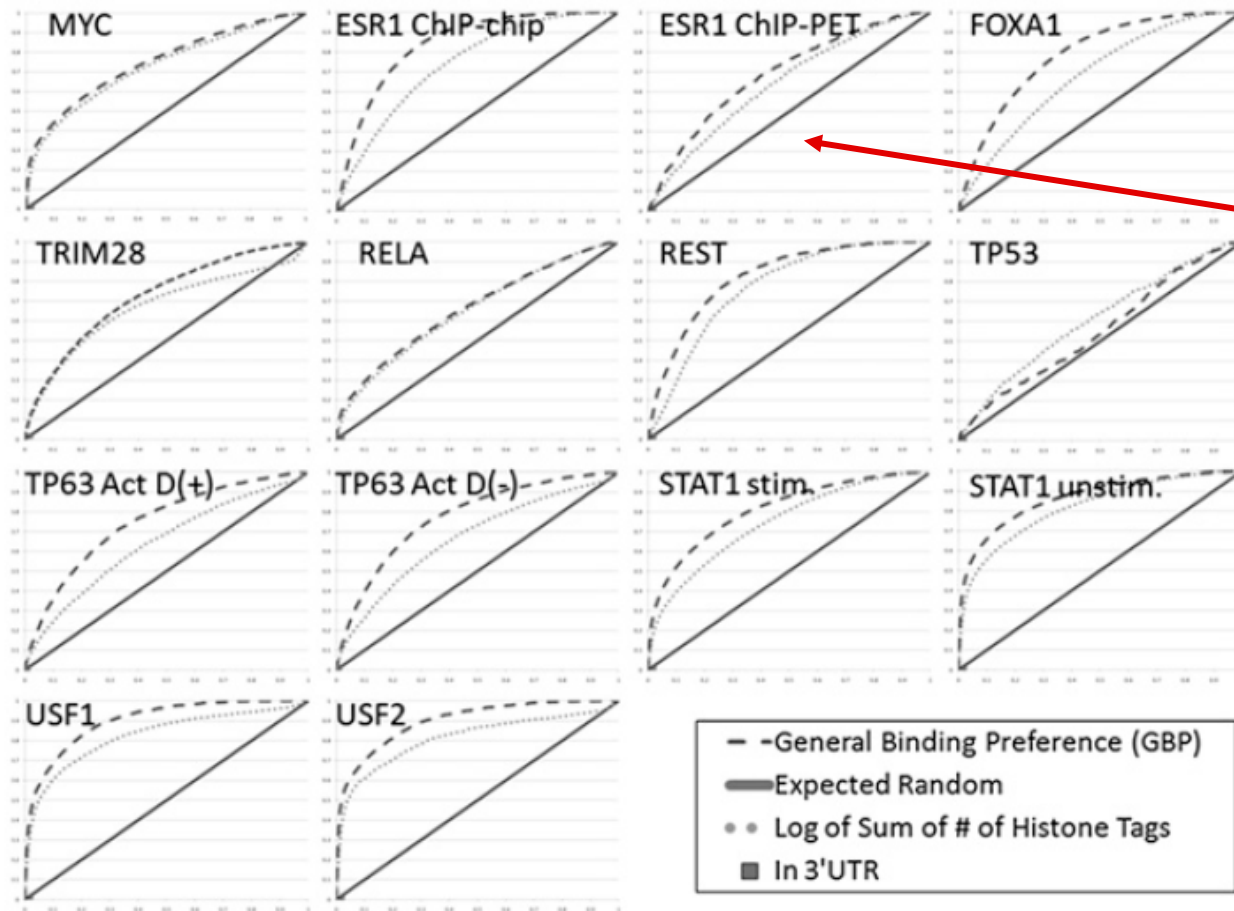
¹Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA;

²Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; ³Department of Molecular Biology, Hebrew University Medical School, Jerusalem 91120, Israel

$$P(b|f_b) = \frac{1}{C} \times \sum_{c=1}^C P_c(b|f_b)$$

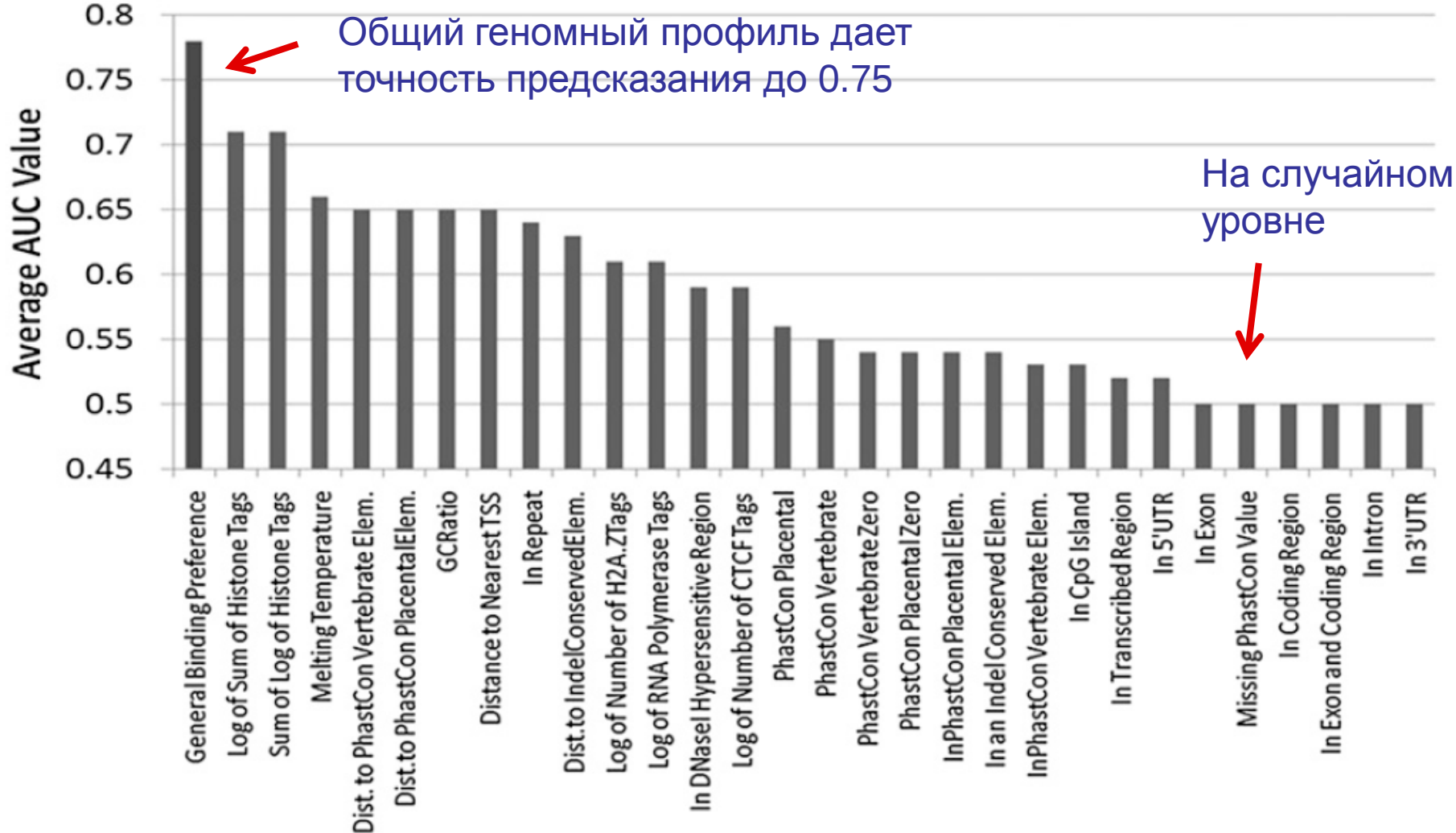
f_b is a vector of feature values specific to genome location b (see above for a discussion of the set of features used), and $P(b|f_b)$ is the probability a logistic regression classifier gives that a location with a set of features f_b is bound by a transcription factor.

Предсказания сайтов с помощью геномного «профиля предпочтения» для различных транскрипционных факторов



Тот же
фактор ER

Общая точность – значение ROC-AUC относительно невелико, около 0.7, в то же время результат неплох, учитывая что ни мотивы ТФ, ни секвенирование ChIP-seq не использовалось, только общий профиль



Точность распознавания сайтов без информации о хроматине (гистоны) в целом невелика (0.6). Район-специфичность (интрон/экзон) практически незначимо ROC-AUC = ~0.5

Спасибо за внимание!

Вопросы?



ЦЕНТР КОЛЛЕКТИВНОГО ПОЛЬЗОВАНИЯ СО РАН
“БИОИНФОРМАТИКА”