JetPoint Meeting

JetBrains BioLabs
Шпынов Олег

6.03.2013

JetBrains

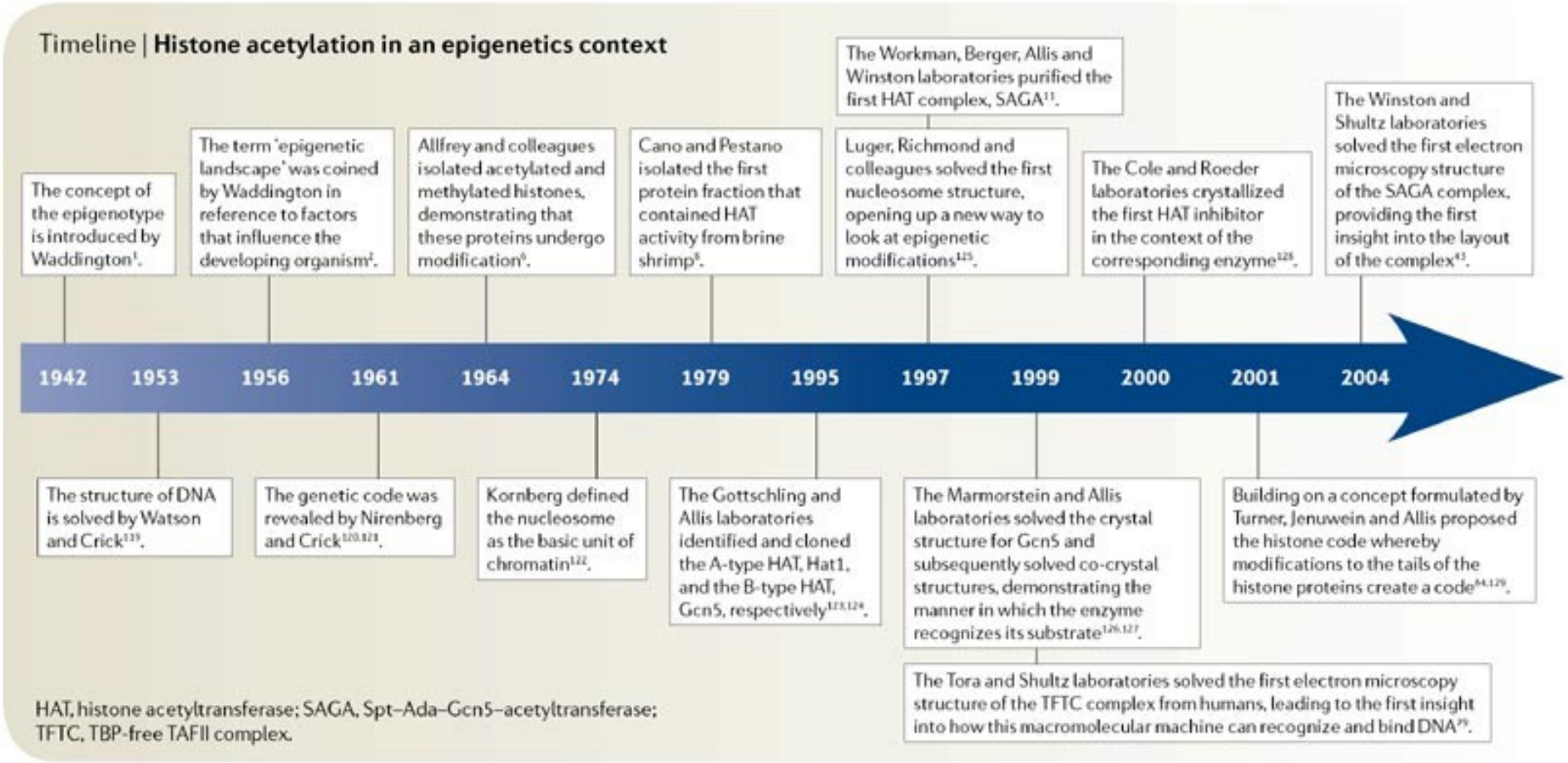
At JetBrains, we have a passion for making people more productive through smart software solutions that help them focus more on what they really want to accomplish, and less on mundane, repetitive "computer busy work".

Эпигенетика

Эпигенетика (греч. ἐπί — над, выше, внешний) — в биологии, в частности, в генетике представляет собой изучение закономерностей эпигенетического наследования — изменения экспрессии генов или фенотипа клетки, вызванных механизмами, не затрагивающими изменение последовательности ДНК.

История

- Термин «эпигенетика» был предложен Конрадом Уоддингтоном в 1942 году, как производное от слов генетика и эпигенез. Когда Уоддингтон ввел этот термин, физическая природа генов не была до конца известна, поэтому он использовал его в качестве концептуальной модели того, как гены могут взаимодействовать со своим окружением при формировании фенотипа.



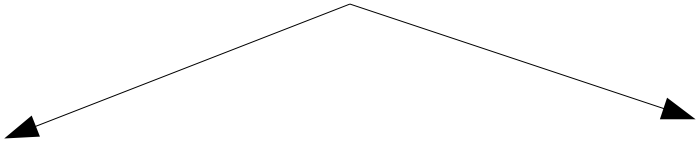
Информация

- **Генетическая** – ДНК, одинакова во всех клетках организма
- **Эпигенетическая** – специфична для конкретной клетки

Каждый вид информации обеспечен своими системами:

- Кодирования
- Хранения
- Передачи

Изменения



Генетические

- Необратимы (мутации)
- Изменения последовательности ДНК
- Стабильно наследуемые

Эпигенетические

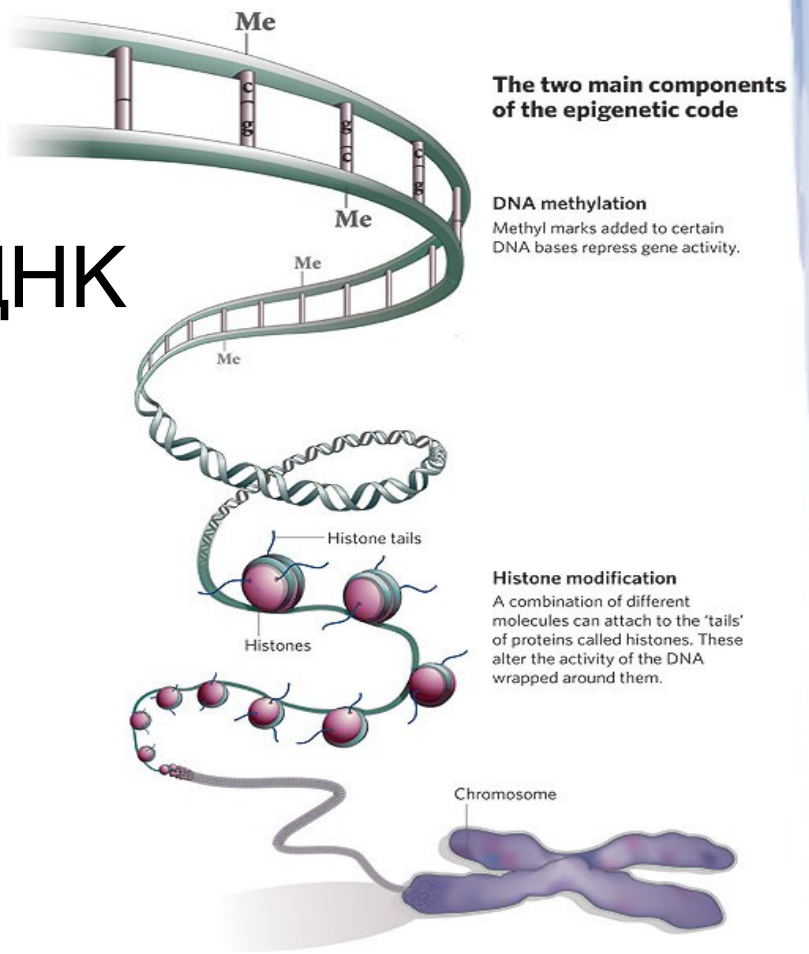
- Обратимы
- Не затрагивают изменений последовательности ДНК
- Долговременные или кратковременные

Эпигеном

Эпигеном - это совокупность всех эпигенетических маркеров, обусловливающих экспрессию генов в данной клетке.

Виды эпигенетических модификаций

- Метилирование ДНК
- Модификации гистонов
- Гидроксиметилирование ДНК
- ?



Связь

- Метилирование ДНК -> деацетилирование гистонов -> образование гетерохроматина
- Деметилирование ДНК -> ацетилирование гистонов -> образование эухроматина

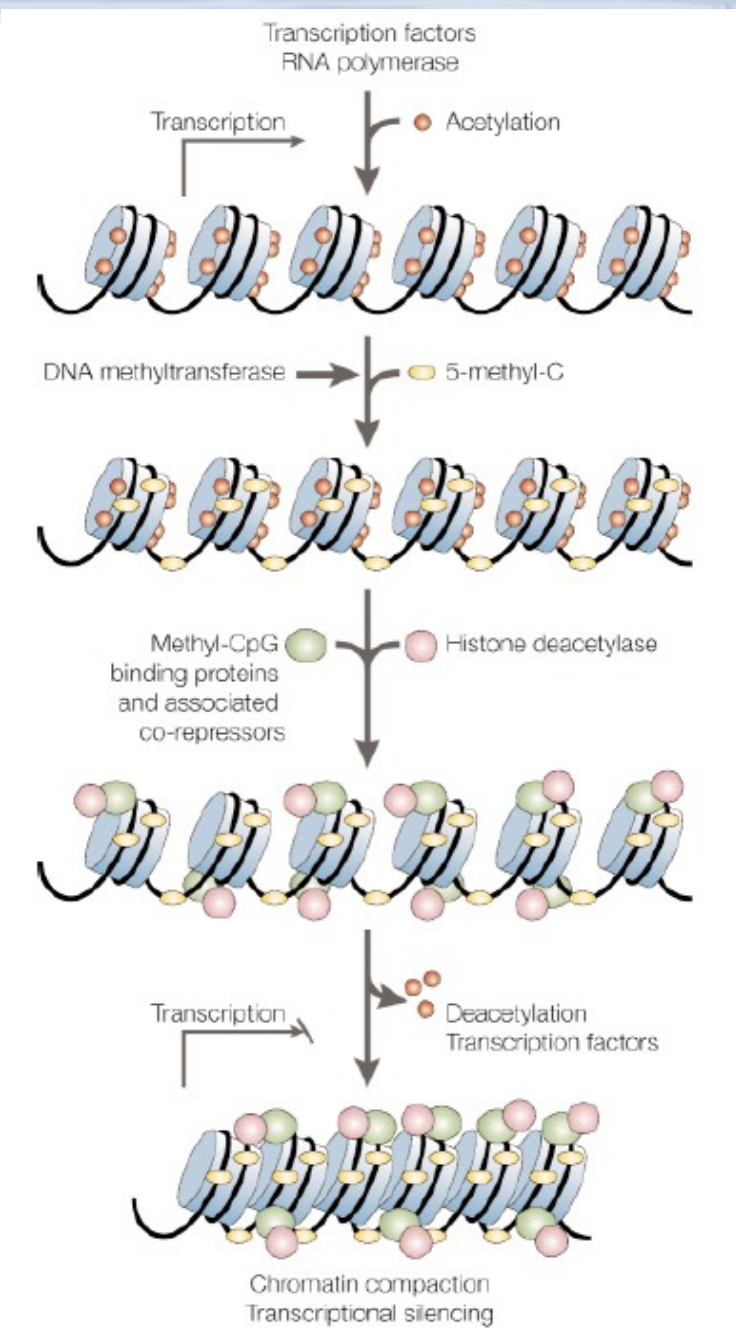
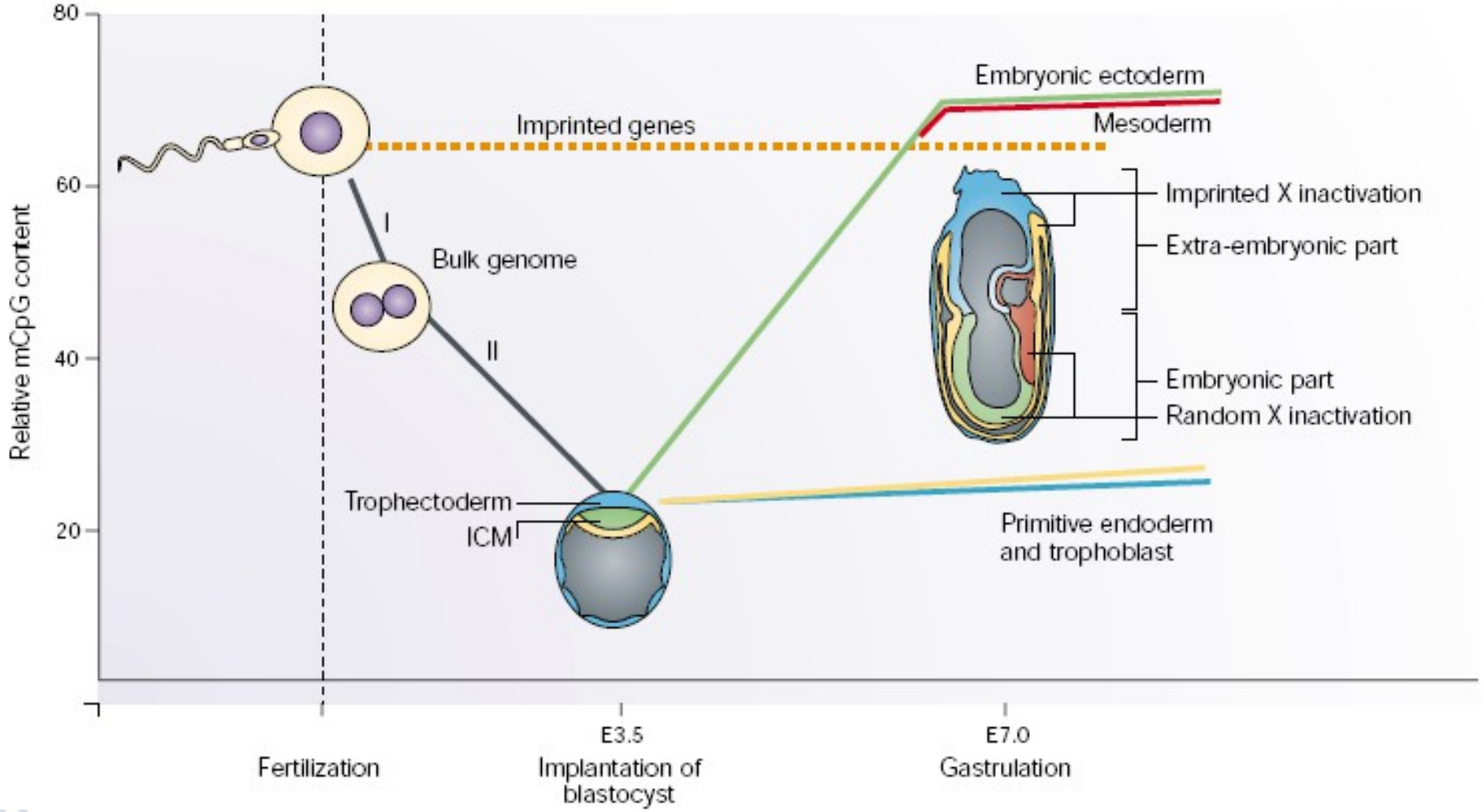


Figure 4 | **The mechanism whereby DNA methylation and histone deacetylation cooperate to repress transcription.**

Эмбриогенез



Эпигенетика – СИЛА!

- Эмбриогенез
- Дифференциация
- Регуляция
- Защита

- Старение?
- Рак?
- ???

Методы исследования

- **Метилирование ДНК**
  BS-seq
  ChIP-seq
  Illumina27/450K

- **Модификации гистонов**
  ChIP-seq

- **ДНК + гистоны**
  ChIP-BS-Seq

Open Data

- Локальность исследований
- Часто очень шумные
- Часто не верифицируемы

+ Много данных в открытом доступе

Wet Labs problems

- Загрязнения проб
- Несоблюдение протоколов
- Использование просроченных реагентов или их заменителей

Academic software

- Много низкокачественного софта, нужного только для публикации.
- Есть реальные примеры софта, в котором отсутствует заявленная функциональность, но на который есть ссылки в статьях.

- A Farewell to Bioinformatics
  http://madhadron.com/a-farewell-to-bioinformatics
  “Fuck you, bioinformatics. Eat shit and die.”

JetBrains BioLabs

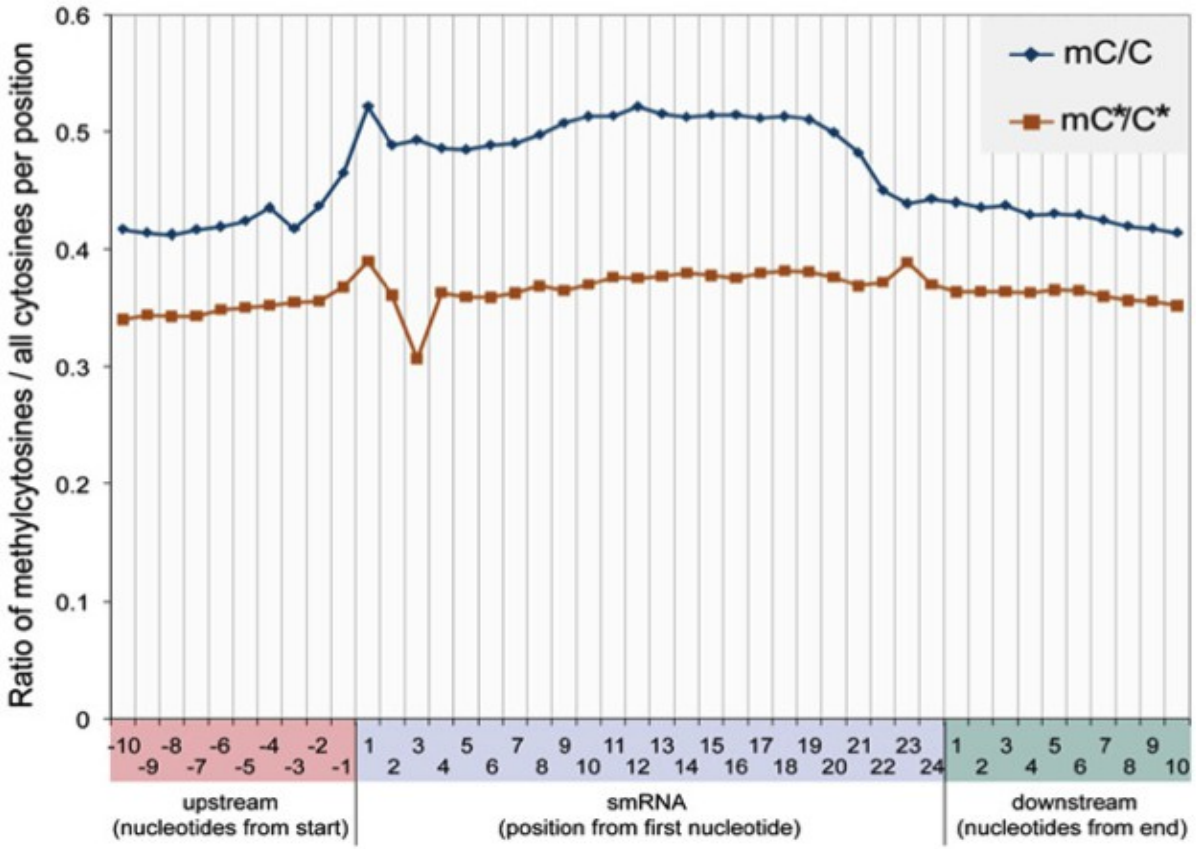
Мы пытаемся применять методы статистики и машинного обучения для выявления фундаментальных эпигенетических механизмов

Гипотеза

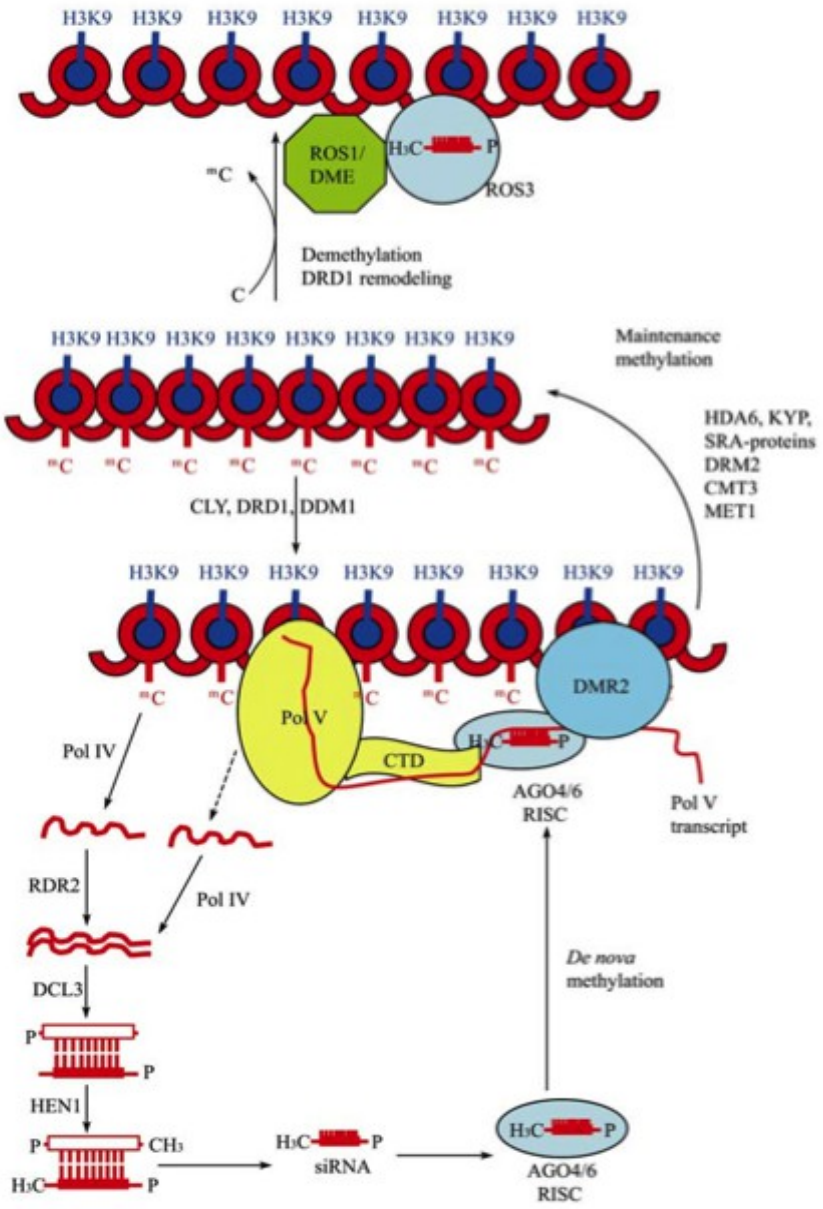
Механизм эпигенетических модификаций управляется последовательностью ДНК

Мотивация?

- RNA-directed DNA methylation in Arabidopsis



Механизм


H3K9
ROS1/
DME
ROS3
Demethylation
DRD1 remodeling
Maintenance
methylation
HDA6, KYP,
SRA-proteins
DRM2
CMT3
MET1
CLY, DRD1, DDM1
DMR2
Pol V
CTD
Pol IV
AGO4/6
RISC
Pol V
transcript
RDR2
DCL3
HEN1
siRNA
De nova
methylation

Задачи

- Исследование закономерностей в геноме
- Анализ данных метилирования
- Анализ данных гистонных модификаций
- Анализ причинно-следственных связей
- Разработка системы экспериментов

Подходы к изучению

- Построение адекватных математических моделей по имеющимся данным
- Применение техник машинного обучения для описания регионов генома, где происходят важные с биологической точки зрения события.
- Верификация данных с помощью коллег -биологов

Исследование промоутеров

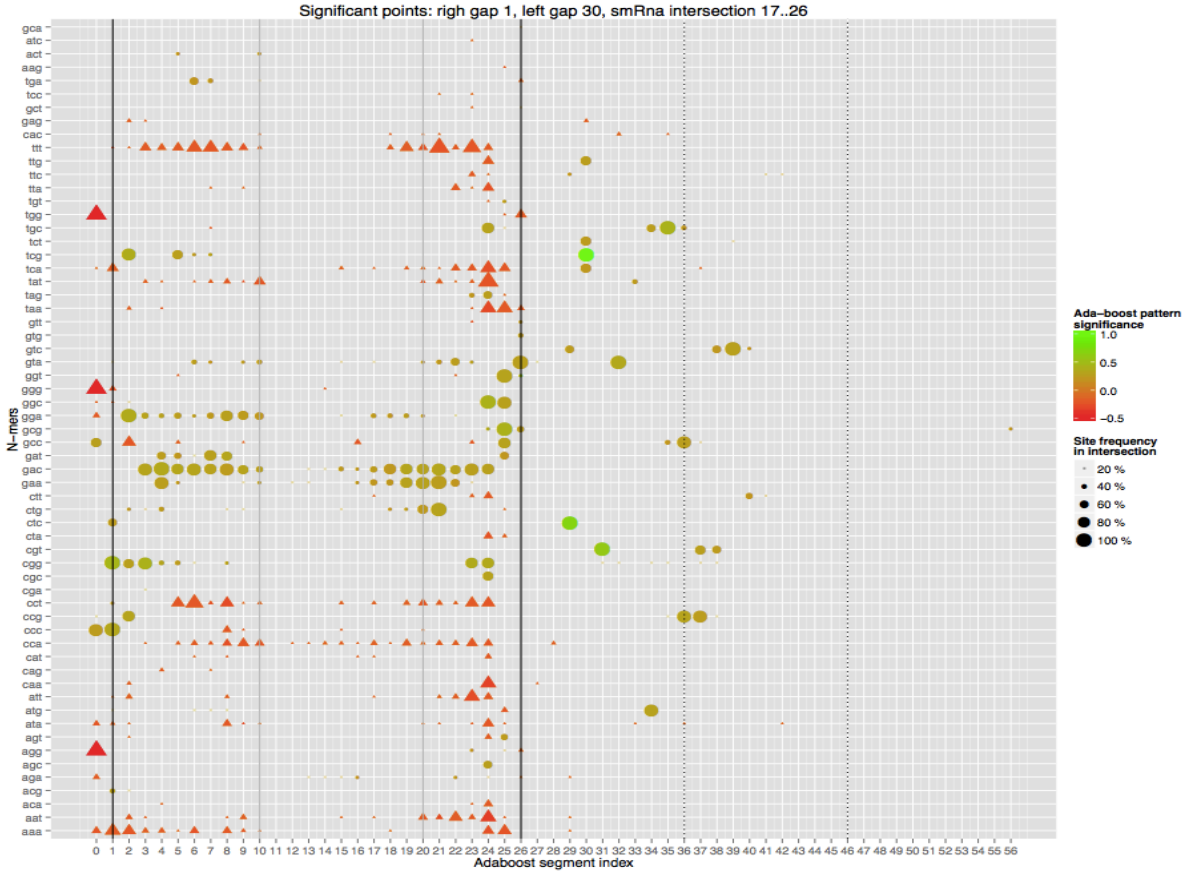
- Вычислительная задача, не имеющая точного решения
- SVM + Ada Boost ML. Простейшие классификаторы – n-мер и его позиция на участке. Обучение и верификация на реальных данных.
- Tradeoff: полнота и точность
- Точность ~ 80%

Экзон-интрон

- Proof of concept для AdaBoost
- Точность ~ 99%
- ML подход – работает!

ML для регионов smRNA

Значимость различных простейших классификаторов



Исследование метилирования

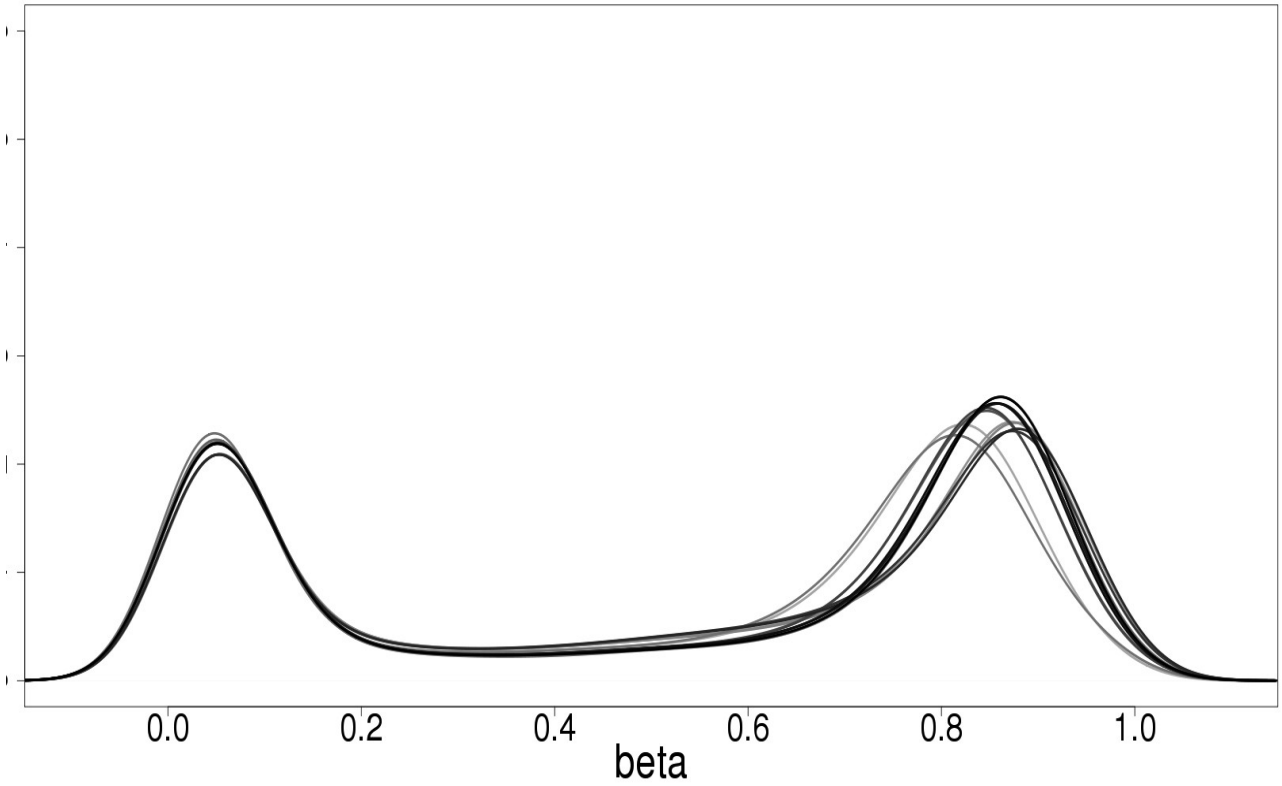
- Исследование BS-Seq данных – выявление паттернов метилирования
- Исследование паттернов в метилировании в различных регионах генома, smRNA, PiRNA, lncRNA, etc
- Корреляция метилирования и других эпигенетических модификаций
- Исследование различий метилирования в гомологичных участках разных животных
- Построение математических моделей, которые описывают метилирование в клетке
- Сравнение разных клеточных линий

Illumina450K

- Infinium Methylation 450K is a hybrid of two different assays, Infinium I and II.
- Due to its design, Infinium Methylation 450K technology generates a dataset that should be viewed as two distinct datasets. Infinium II data are less accurate and reproducible than Infinium I data.
- Peak-based correction makes it possible to treat Infinium I and Infinium II data as a single dataset.
- Infinium Methylation 450K is one of the most attractive powerful and cost-effective tool currently available for generating quantitative DNA methylomes for health and disease, notably in the framework of large biomarker discovery studies.

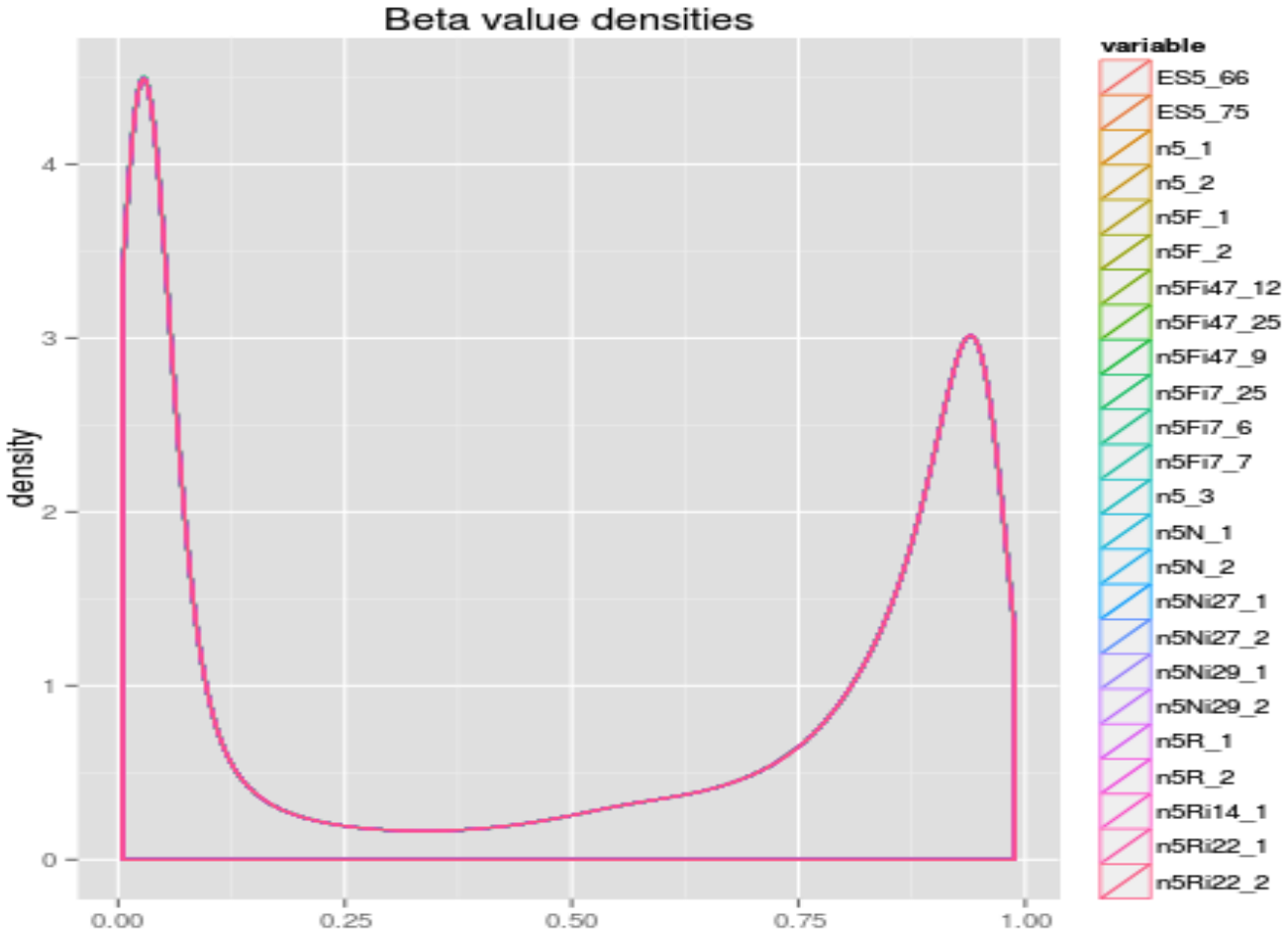
Illumina450K

- Beta = methylated / (methylated + unmethylated)



Illumina 450K

- Фильтрация + subset quantile normalization



Illumina450K

- Загрузка
- Фильтрация
- SNP-процессинг
- Subset Quantile Normalization
- Batch effects
- Сравнение локусов (genes, gene regions, etc) с использованием Mann-Whitney U-test
- Результат: NDA

Исследование гистонов

- Построение математических моделей модификаций гистонов
- Сравнение разных клеточных линий
- Связь модификаций гистонов с другими организмами
- Поиск схожих паттернов модификаций гистонов

Математические модели модификаций гистонов

- Данные – покрытие генома после ChIP-seq
- Большинство генома не покрыто
- Рассматриваем покрытие по корзинам
- Можно предполагать, что покрытие разных корзин порождено независимыми случайными величинами
- Плотность распределения

Poisson Mixture

- Бимодальное распределение
- Рассматриваем как смесь двух Пуассоновских распределений
- Методом оценки максимального правдоподобия получаем скрытые состояния корзин
- Скрытые состояния – есть гистонная модификация или нет?

Poisson Mixture + HMM

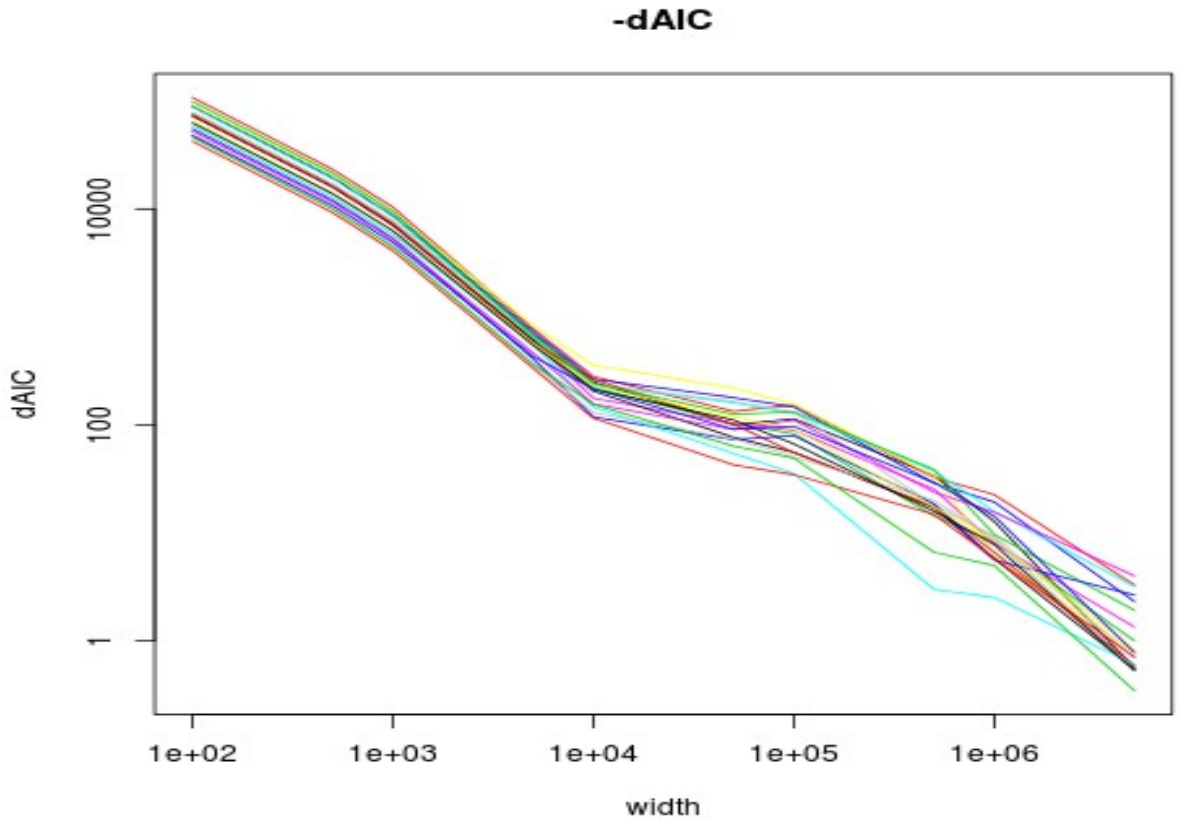
- А вдруг соседние корзины не независимы? Введем скрытую Марковскую цепь с вероятностями переходов.
- Оценка методом максимального правдоподобия + алгоритм Виттерби для оценки всех параметров системы
- Есть и более сложные модели, например для сравнения двух измерений

Сравнение моделей

- Критерий Акайке

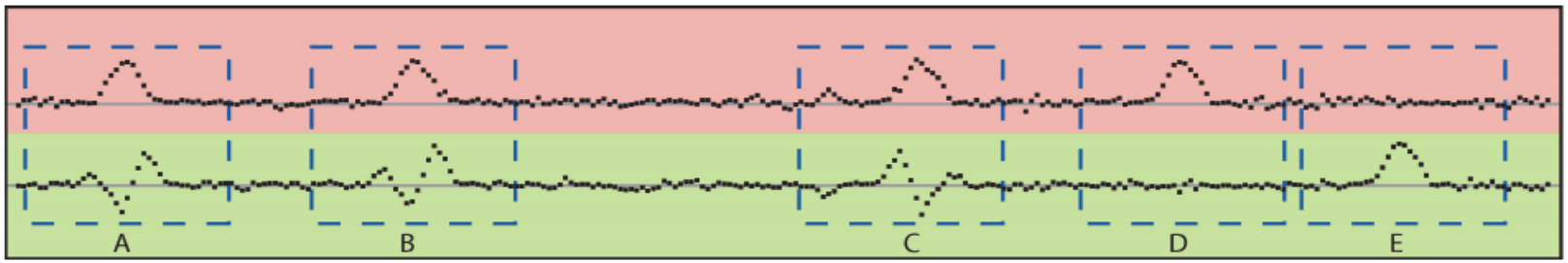
  AIC = 2*freedom_degrees – log(likelihood)

-



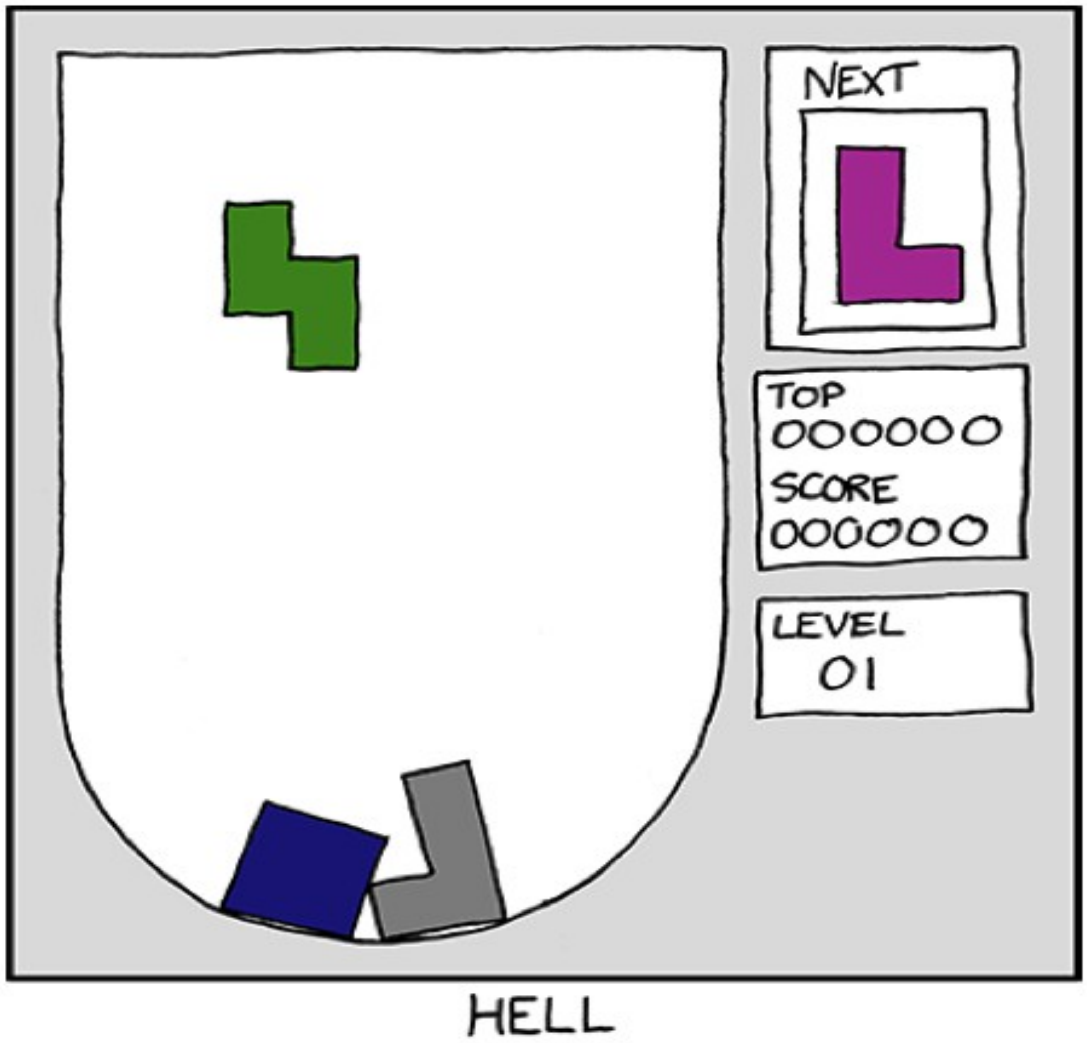
Chromasig

- Нахождение схожих паттернов метилирования и модификации гистонов



- Инструмент для поиска мотивов для ChIP-Seq данных - Chromasig

Реализация алгоритма из статьи



Анализ результатов Chromasig

- Онтологии генов участков генома
  - Функции
  - Компартменты клетки
  - Наличие у разных организмов

Разработка системы экспериментов

- **Данные**: описывать входные данные, с удобной системой хранения и доступа, разделять данные полученные нами и из сторонних источников, переиспользование данных
- **Эксперименты**: описание входные данных, описание экспериментов, формат для переиспользования
- Имеющиеся системы громоздки
- Не удовлетворяют запросам

Tools

- Java
- R
- Big server computations (Linux)
- Confluence, Bamboo, Crucible
- Continuous integration, tests

Проекты JetBrains в биоинформатике

- JetBrains BioLabs
- LabBook - электронный лабораторный журнал. Проблема разрозненности данных. Большинство отчетов в Excel. Несоответствие модели данных и инструментов.
- Genome query – студенческий проект.
- Genestack Platform - universal collaborative ecosystem for bioinformatics research and development. http://genestack.com

JetBrains BioLabs

- Алексей Диевский
- Сергей Дмитриев
- Евгений Курбацкий
- Сергей Лебедев
- Роман Чернятчик
- Олег Шпынов

Вопросы?

Спасибо за внимание!

Oleg.Shpynov@jetbrains.com

Twitter: oleg_s