Skoltech

Сколковский институт науки и технологий


Учебно-Научный Центр

Биоинформатика

Биология больших данных

Михаил Гельфанд

JPoint, 6 IV 2018

Краткий курс молекулярной биологии

- Центральная догма
  - Репликация, транскрипция, трансляция
- Регуляция
  - Сигнальные пути
  - Факторы транскрипции
  - Эпигенетика
  - Пространственная структура ДНК

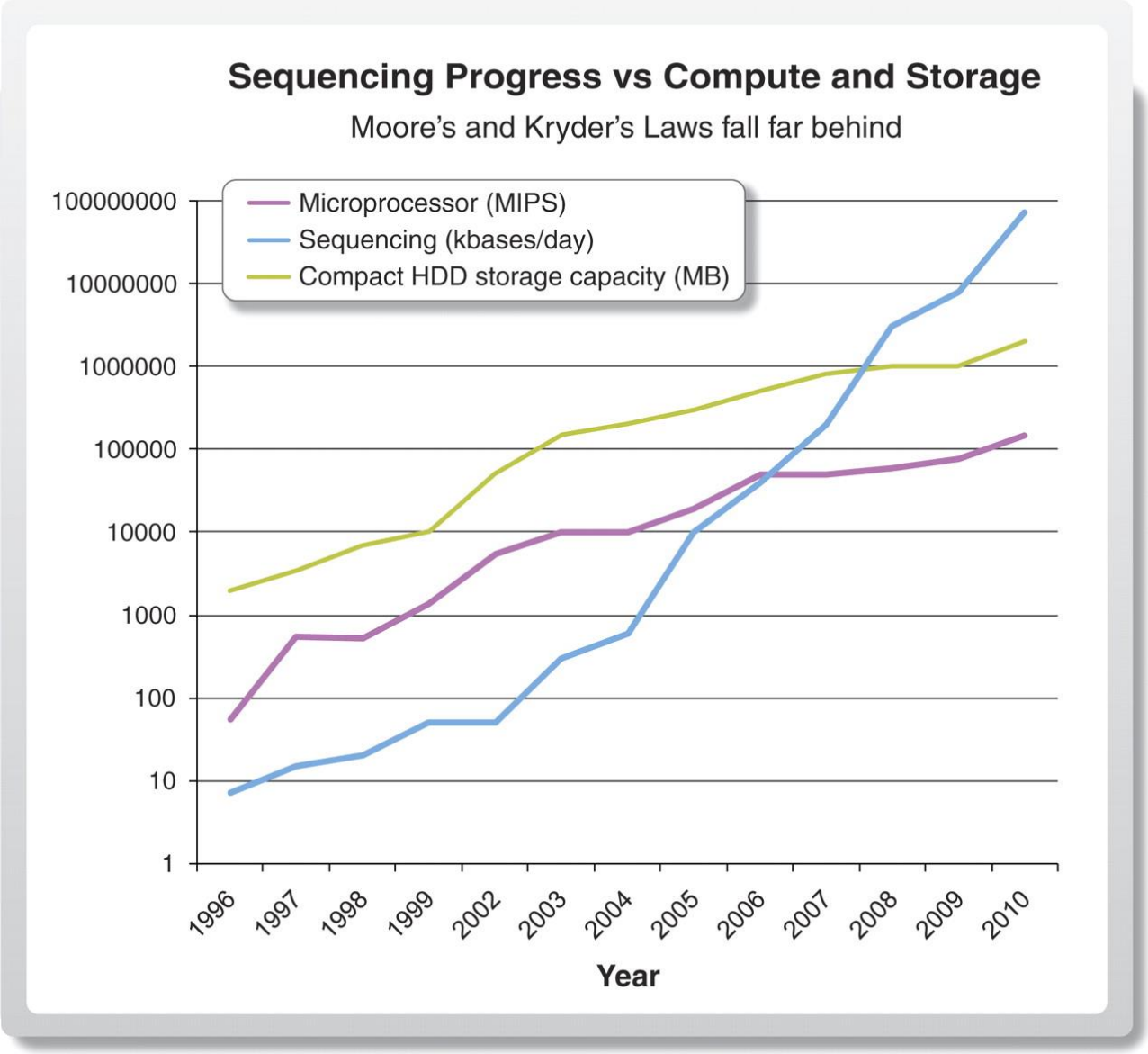
Данные

- Уровень экспрессии
  - Концентрации мРНК
  - Концентрации белков
  - Время жизни мРНК и белков
  - Концентрации метаболитов
- Взаимодействия
  - Белок-ДНКовые
  - Белок-белковые
- Структура генома
  - Метилирование и открытость ДНК
  - Положение нуклеосом и модификация гистонов
  - Пространственная структура
- Функционально-генетические
  - Летальность и фенотип мутаций
  - Синтетические летали

Многие методы основаны на секвенировании

- Уровень экспрессии
  - Концентрации мРНК:
    секвенирование транскриптома
- Взаимодействия
  - Белок-ДНКовые:
    ChIP-Seq
- Структура генома
  - Метилирование ДНК:
    бисульфитное секвенирование
  - Положение и модификация нуклеосом:
    ChIP-Seq
  - Пространственная структура:
    HiC

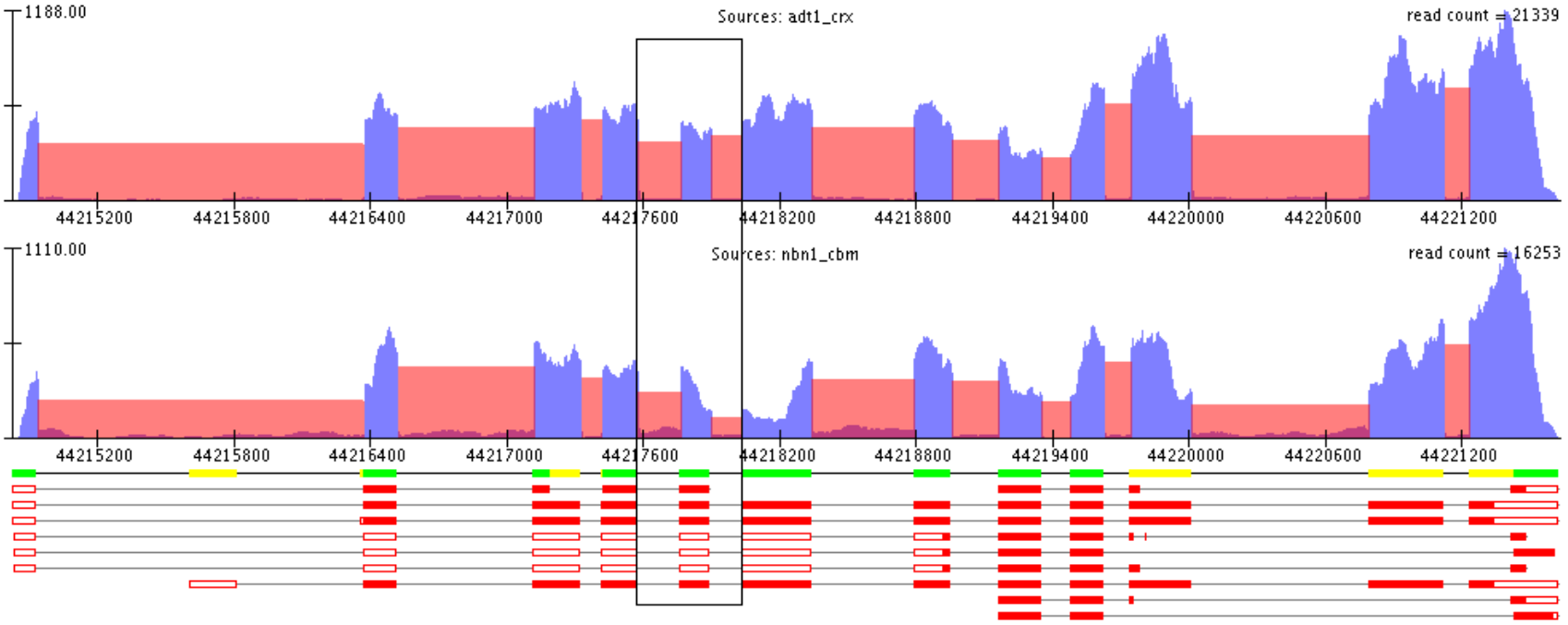
A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields



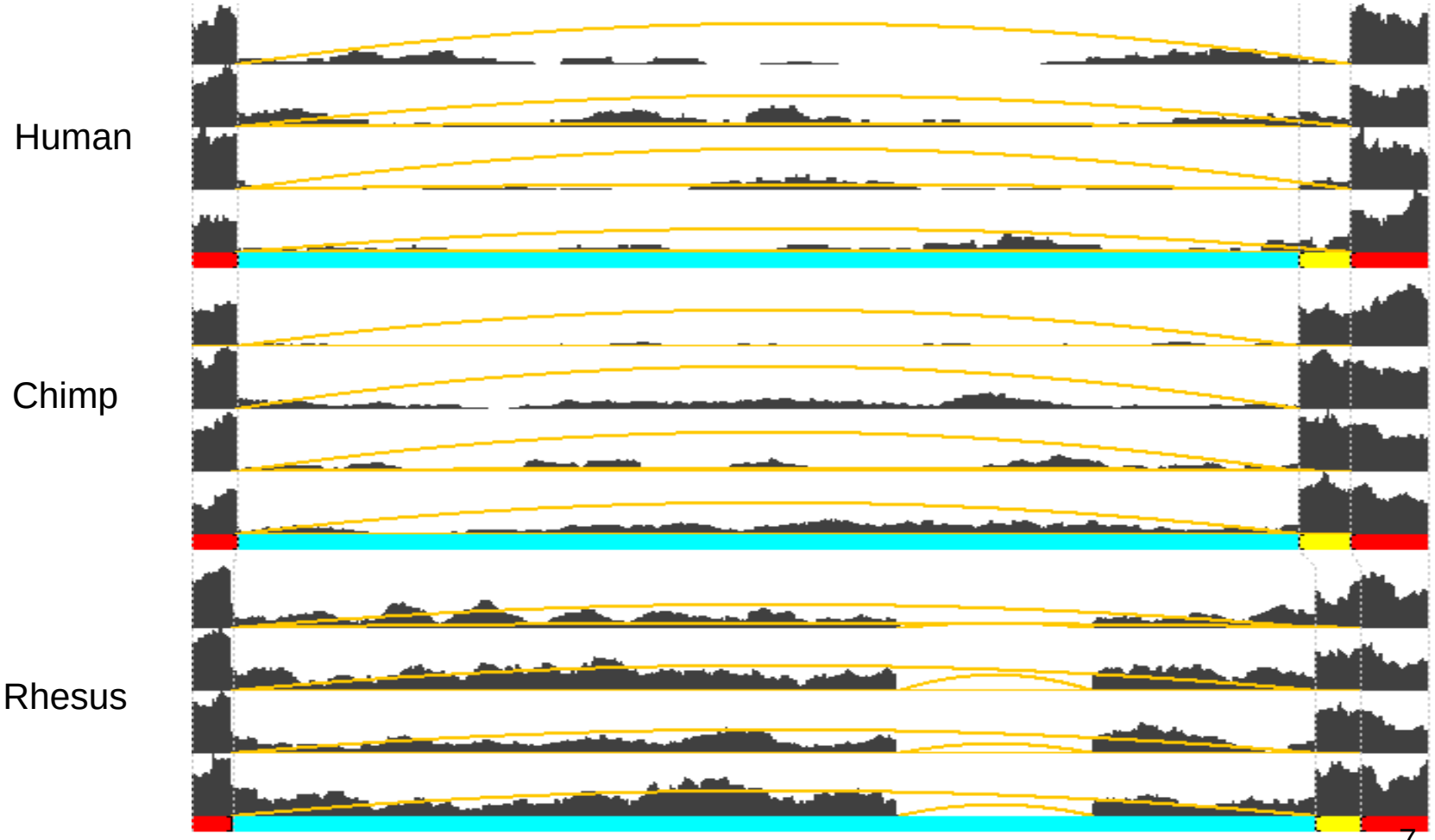
S D Kahn Science 2011;331:728-729

Science AAAS

Картирование транскриптов

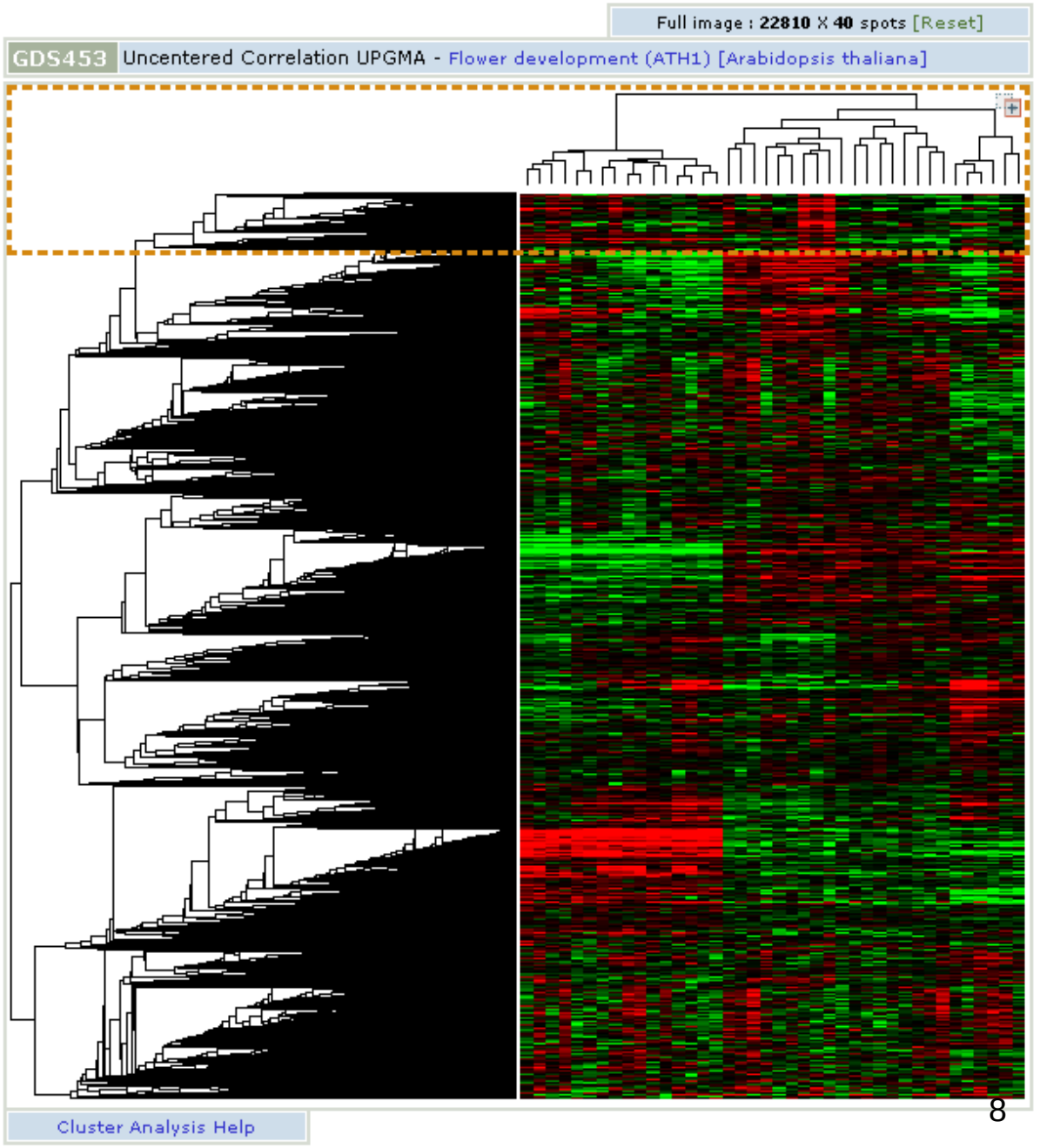


Картирование транскриптов



Экспрессия генов – 1. Развитие цветка резуховидки Таля

двойная кластеризация – на генах и на условиях



Экспрессия генов – 2. Цикл развития малярийного плазмодия

The Transcriptome
of the Intraerythrocytic Developmental Cycle
of *Plasmodium falciparum*

Zbynek Bozdech[1☯], Manuel Llinás[1☯], Brian Lee Pulliam[1], Edith D. Wong[1], Jingchun Zhu[2], Joseph L. DeRisi[1*]

Транскриптомы: типичные задачи

- Когорты пациентов
- Опухоль – ткань

- Дифференциальный диагноз
- Прогноз
- Подбор лечения

- Деконволюция

Single cell геномика и транскриптомика

- Геномика: история клеток
  - Нейроны
  - Рак
- Транскриптомика: различия клеток
  - Intrinsic noise
  - Импринтинг
  - Эмбриология
  - Рак: клоны

Проблемы

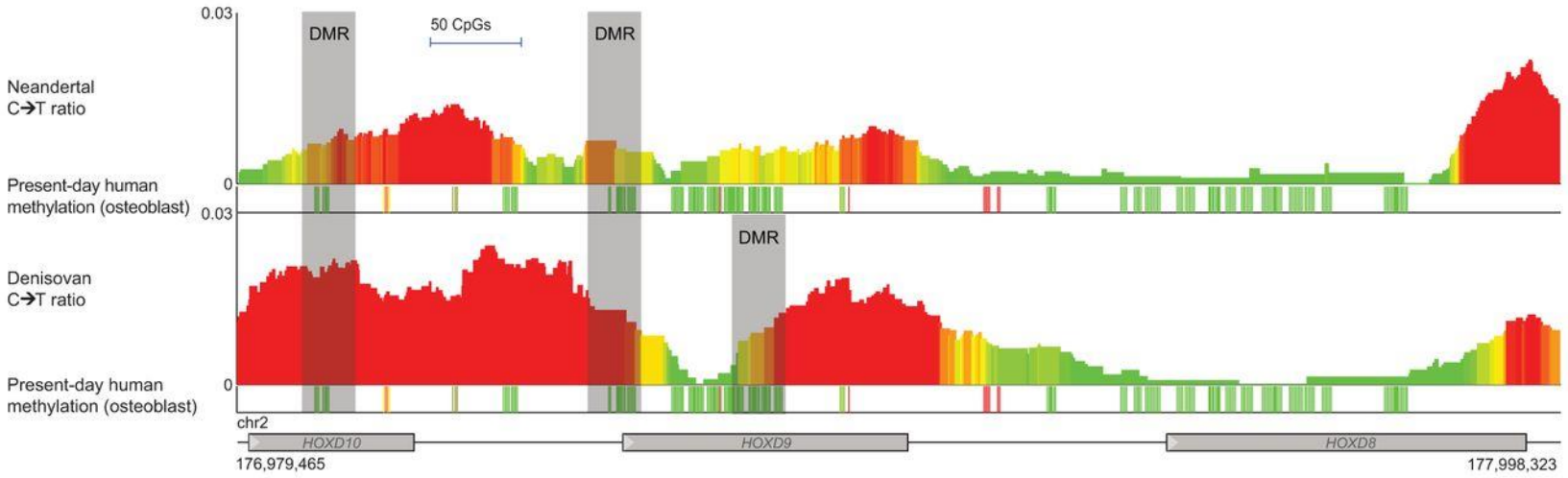
- Шум: каждое отдельное наблюдение ненадежно
- Множественное тестирование. FDR
- Большая значимость – слабый эффект
- Взаимозависимость параметров
- Абстракция: работаем с конструктами, а не с реальными объектами

- Предсказание (медицина)

vs

- Feature extraction (биология)

Метилирование



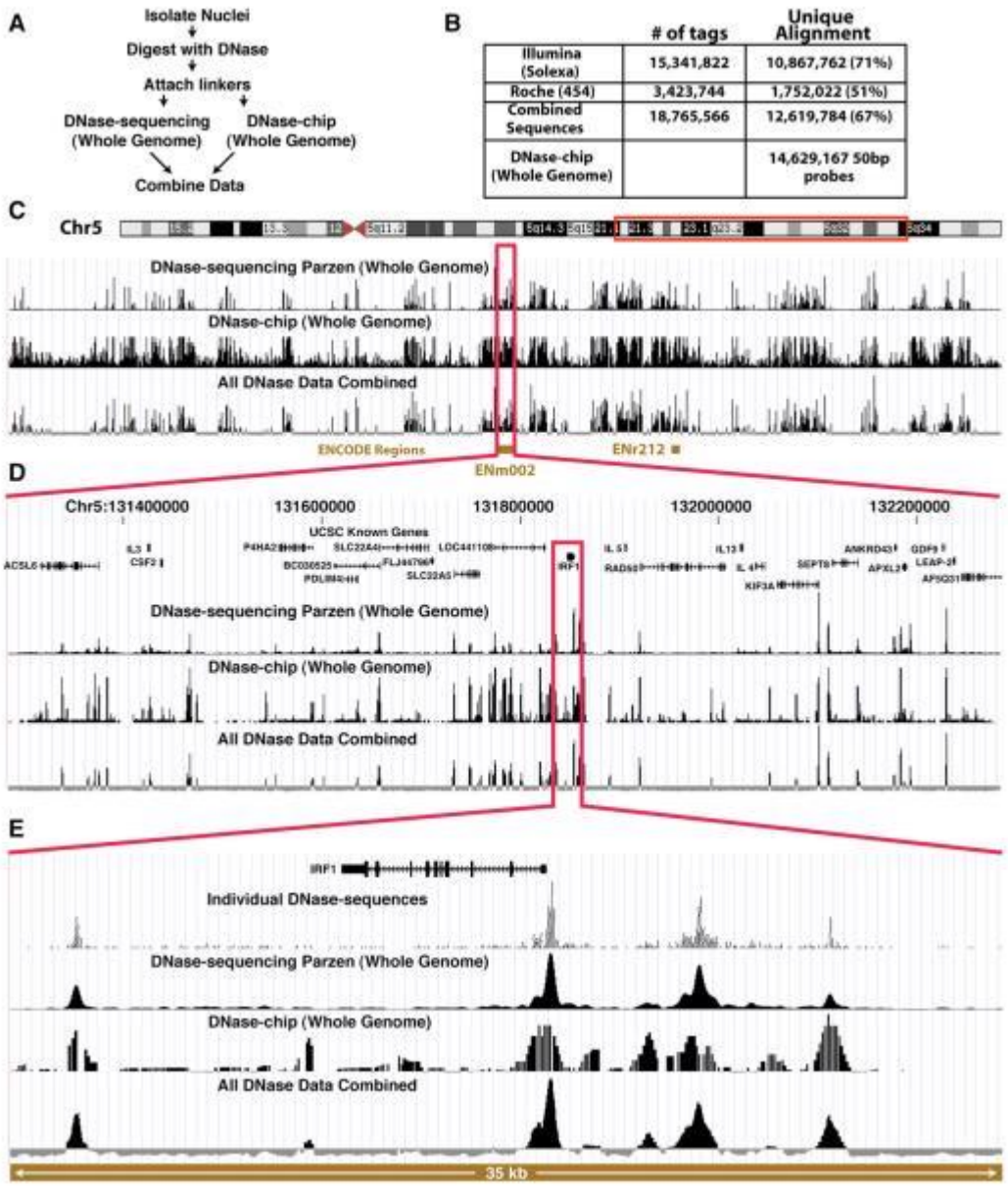
The HOXD cluster is hypermethylated in archaic humans.

C→T ratio in archaic humans, and present-day human methylation

Science
AAAS

D Gokhman et al. Science 2014;science.1250368

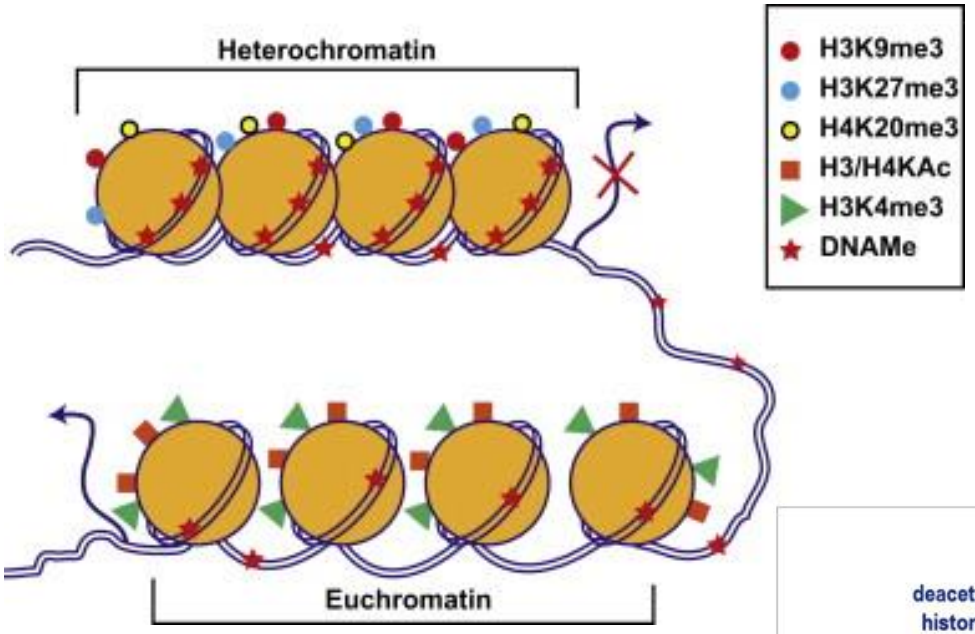
DNAse I – открытый хроматин

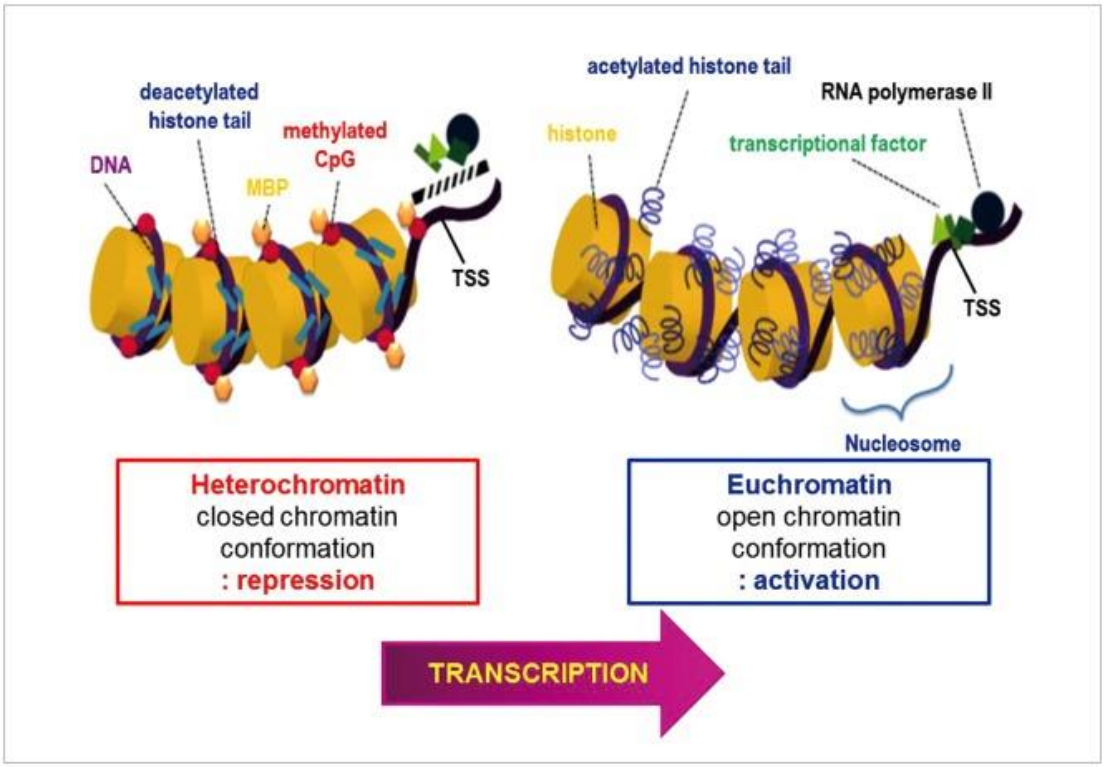


	# of tags	Unique Alignment
Illumina (Solexa)	15,341,822	10,867,762 (71%)
Roche (454)	3,423,744	1,752,022 (51%)
Combined Sequences	18,765,566	12,619,784 (67%)
DNase-chip (Whole Genome)		14,629,167 50bp probes

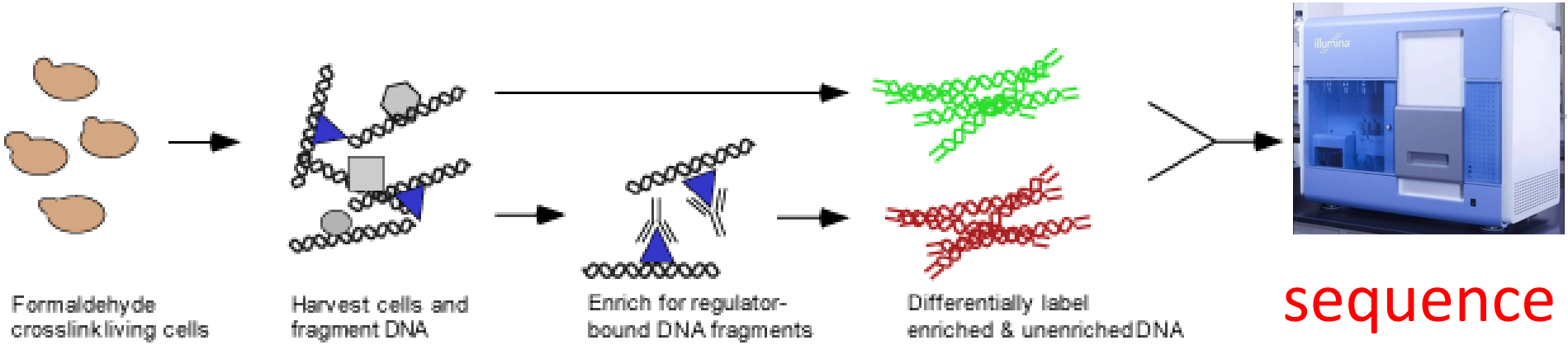
Boyle et al. 2008

Модификации гистонов

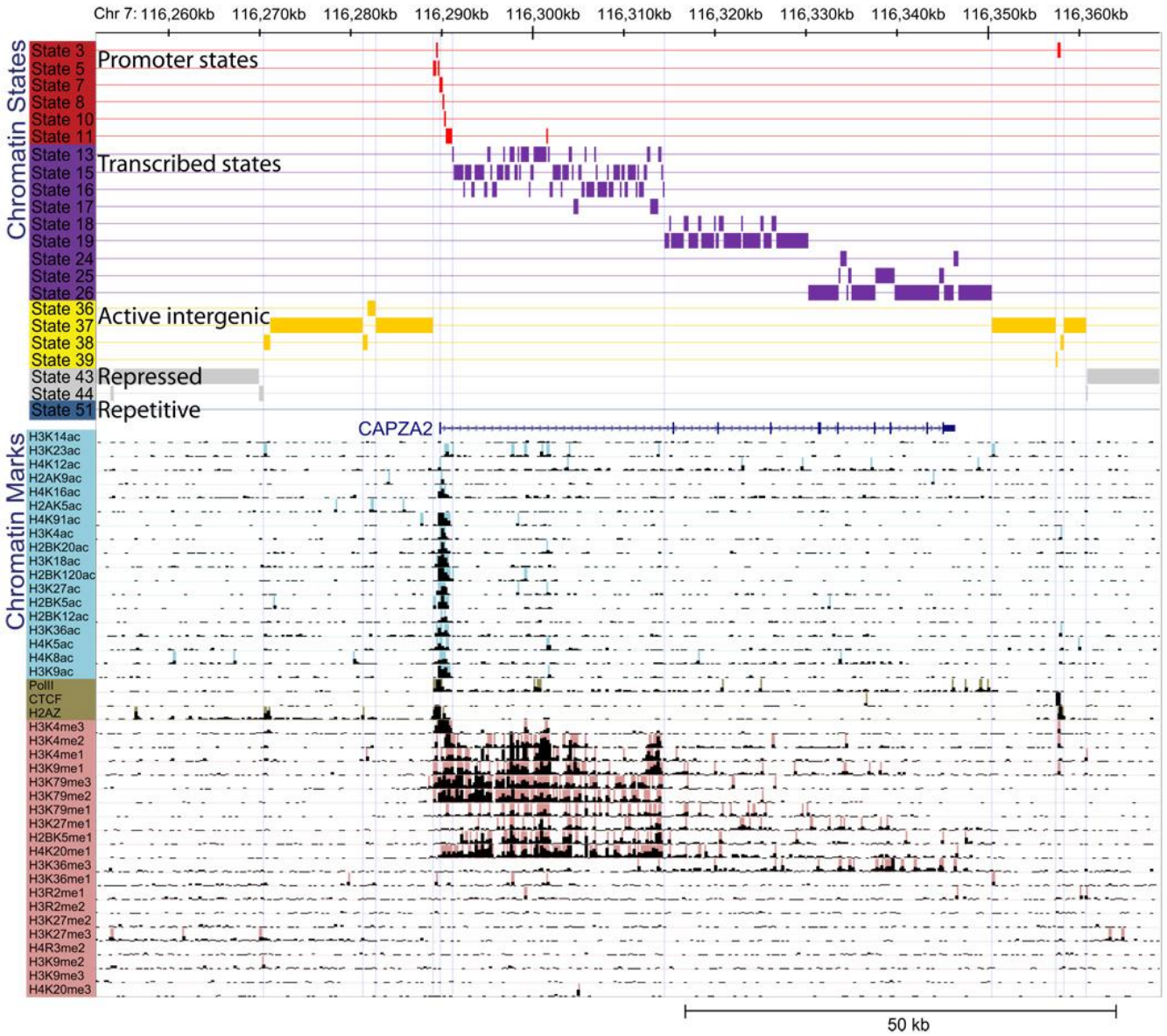




ChIP-Seq

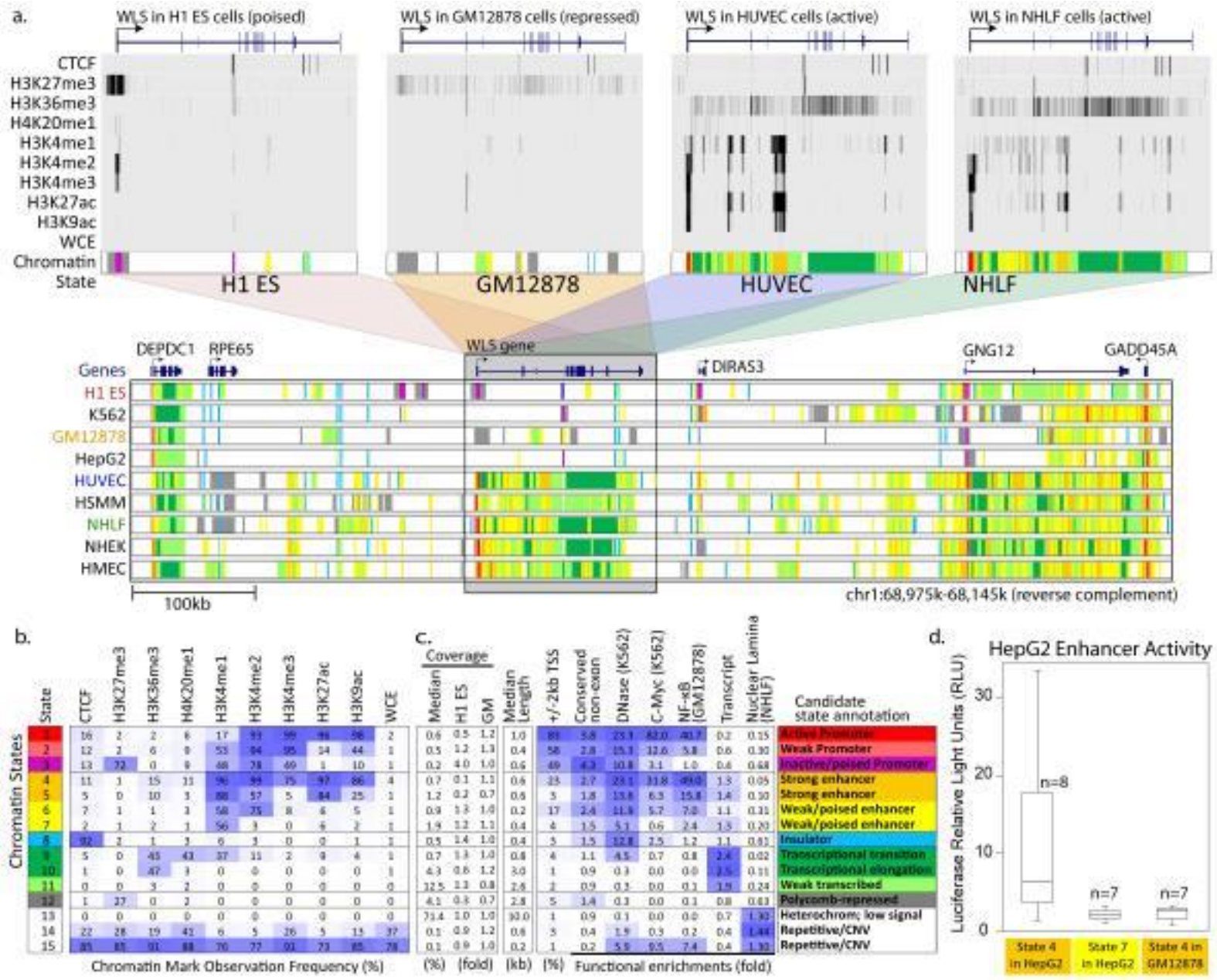

Formaldehyde crosslink living cells
Harvest cells and fragment DNA
Enrich for regulator-bound DNA fragments
Differentially label enriched & unenriched DNA
sequence

Индивидуальные маркеры и интегрированные состояния



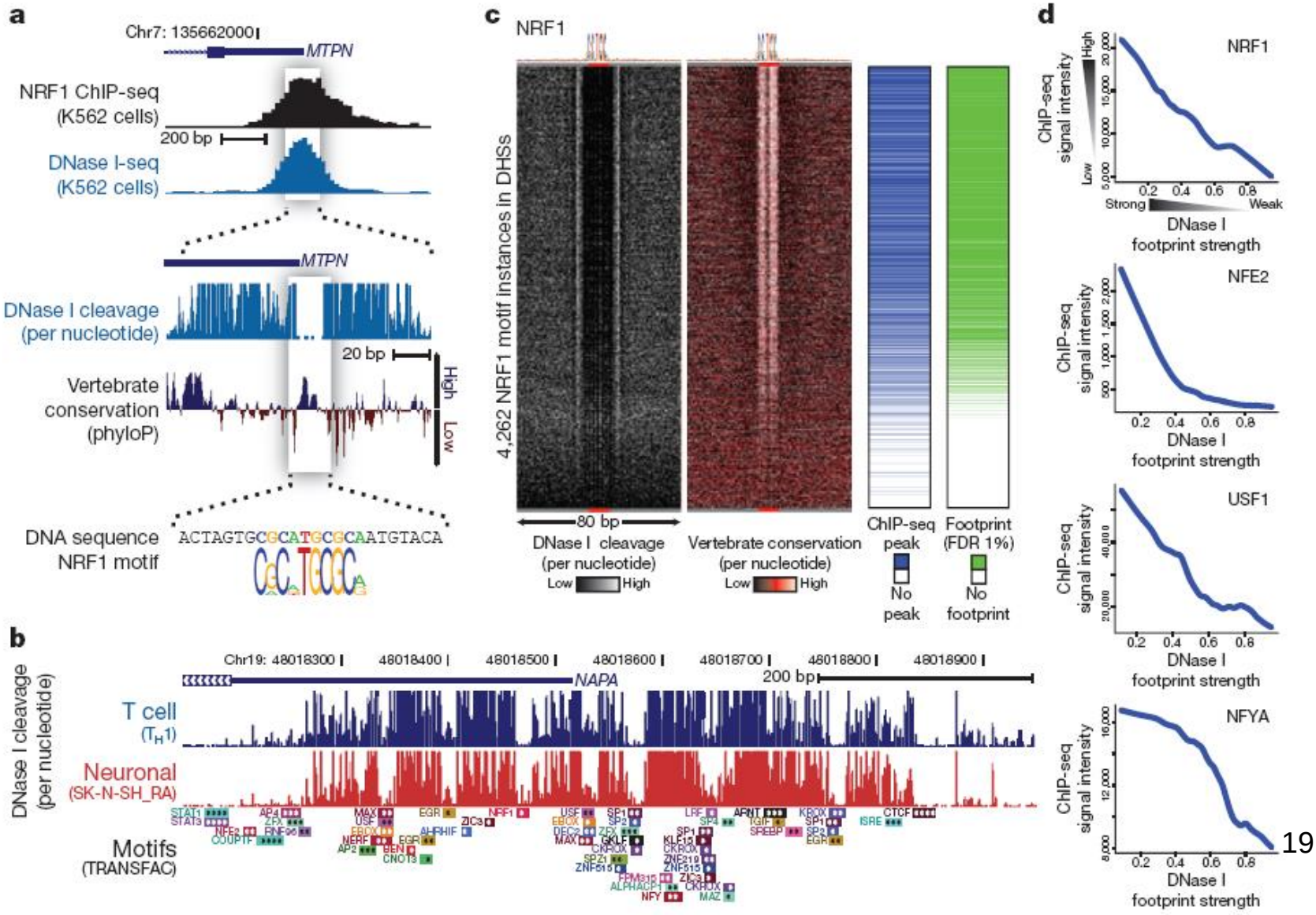
Ernst & Kellis, 2010

Другой пример: сравнение клеточных линий

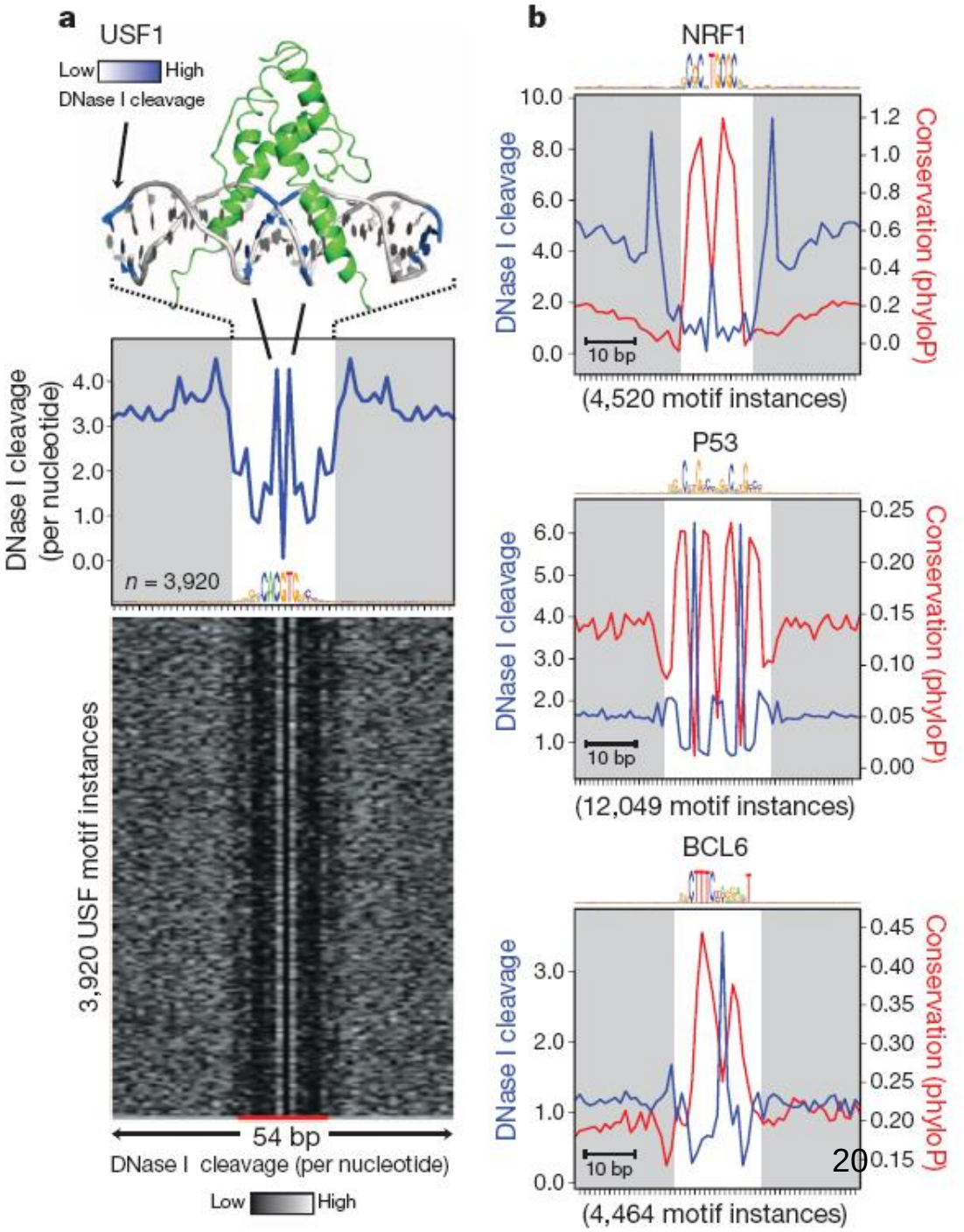


Ernst et al., 2011

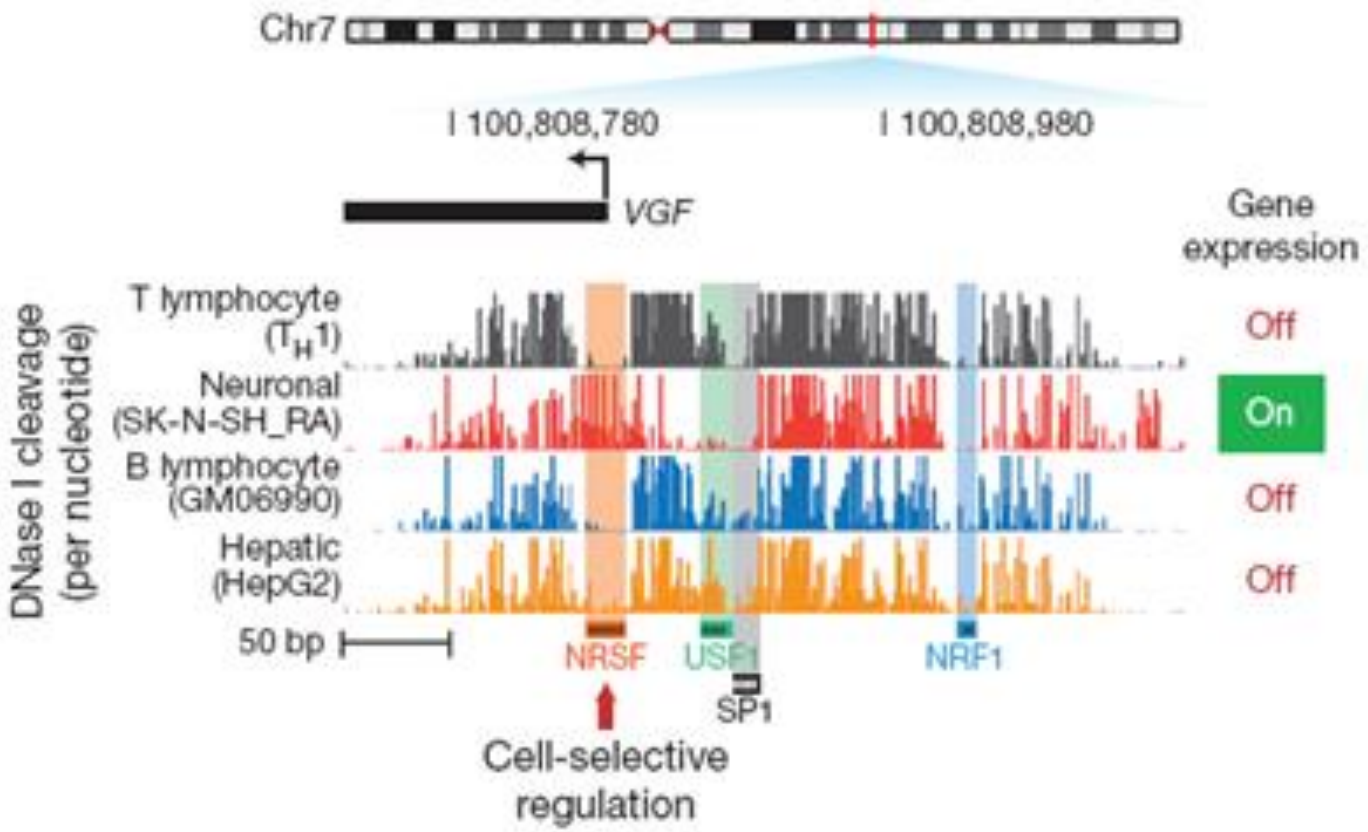
Доступность хроматина для ДНКазы I



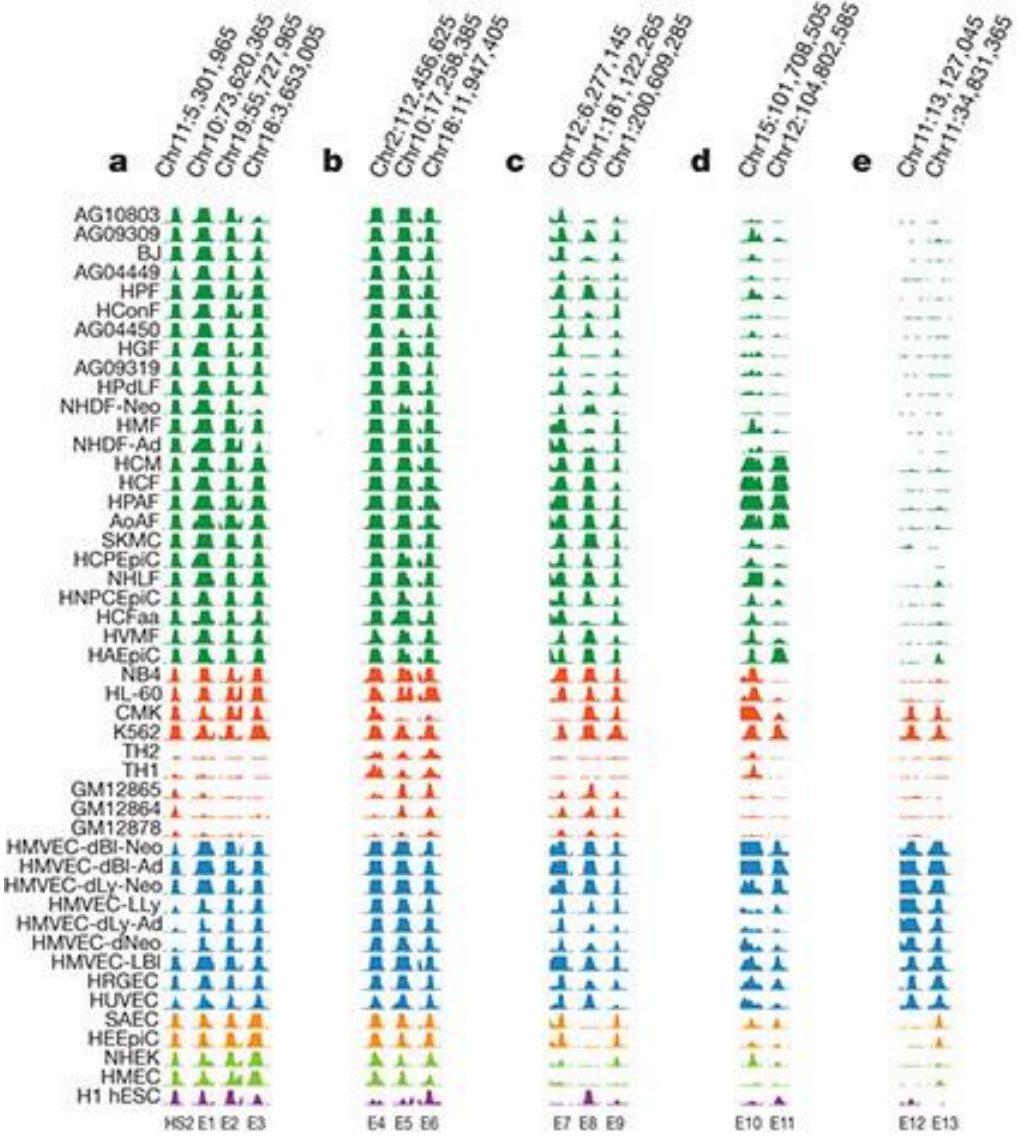
Доступность хроматина для ДНКазы I и структура комплекса фактор транскрипции - ДНК



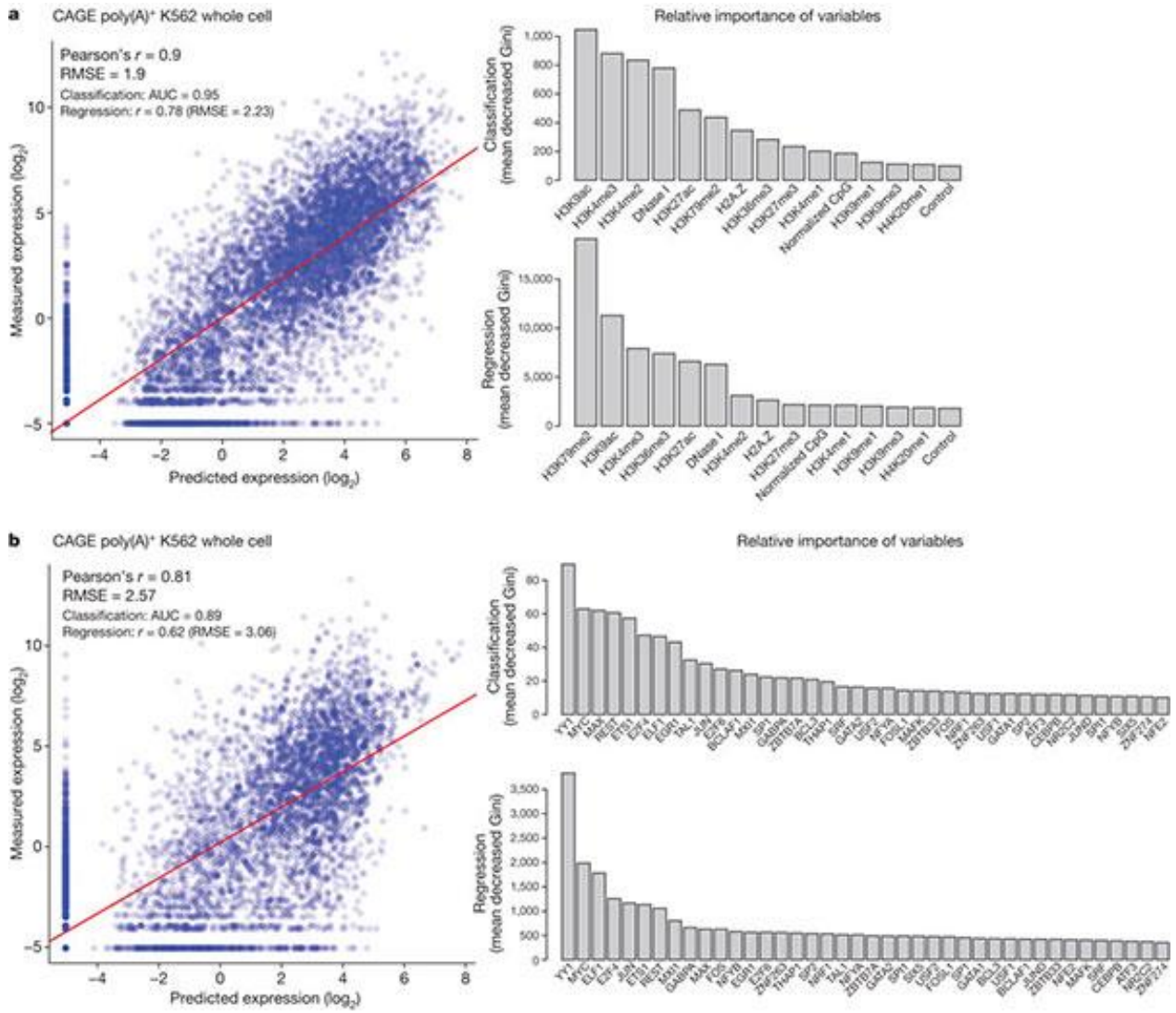
Тканевая специфичность



Кластеризация энхансеров по открытости в разных клеточных линиях

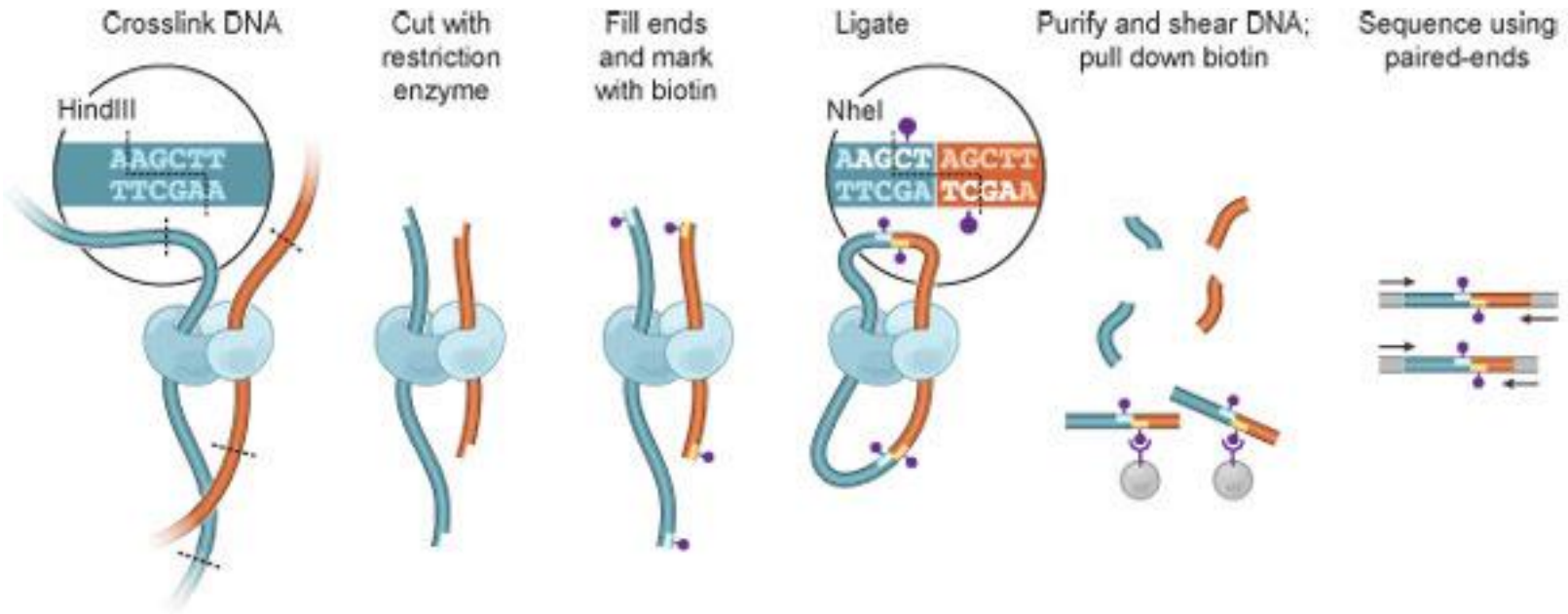


Предсказание уровня экспрессии по модификациям гистонов и по связыванию факторов транскрипции

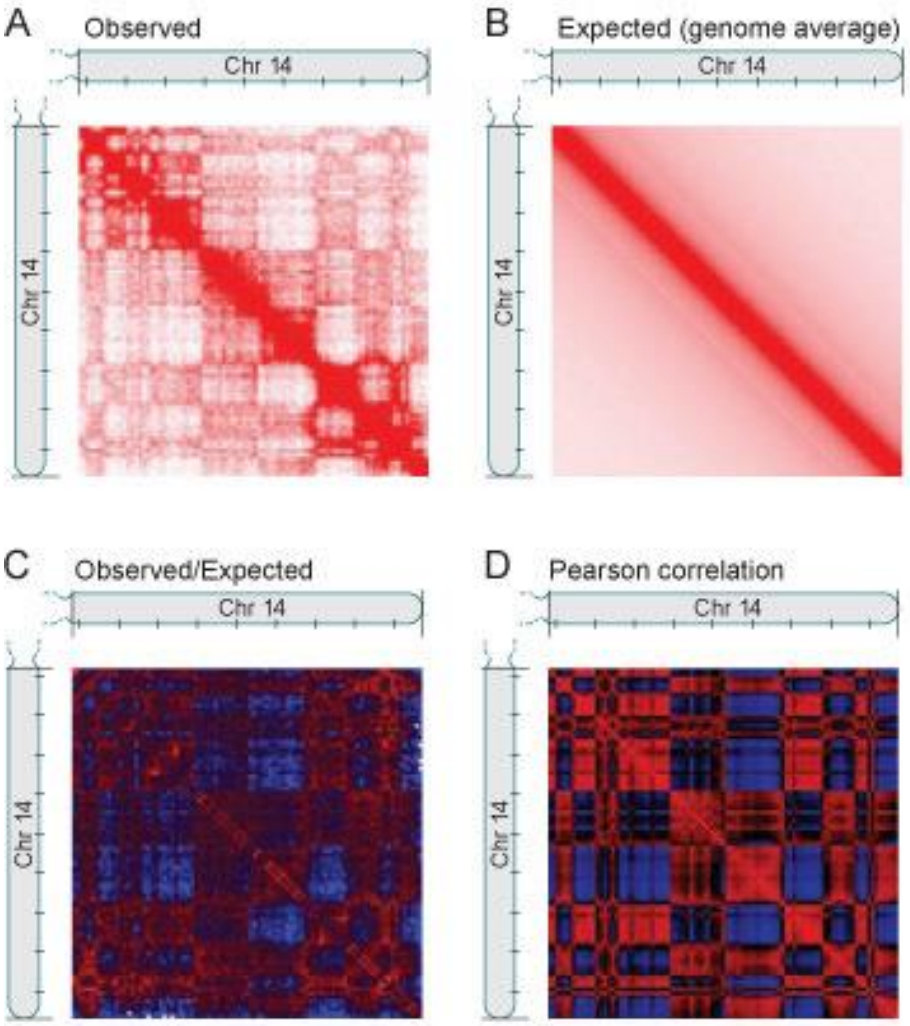


HiC

E.Lieberman-Aiden … J.Dekker, Science, 2009



Карты HiC

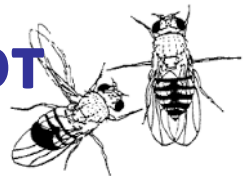


Dekker's group

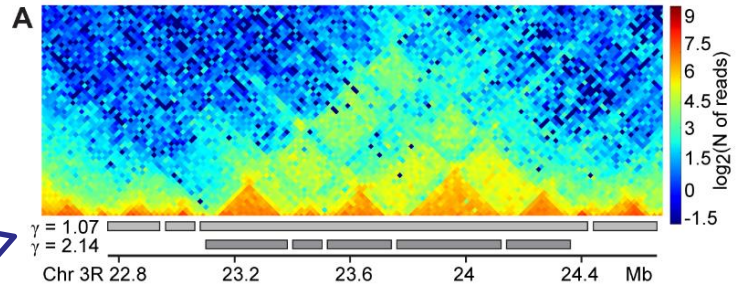
NB: это не геометрическая близость в 3D, а частота контактов.

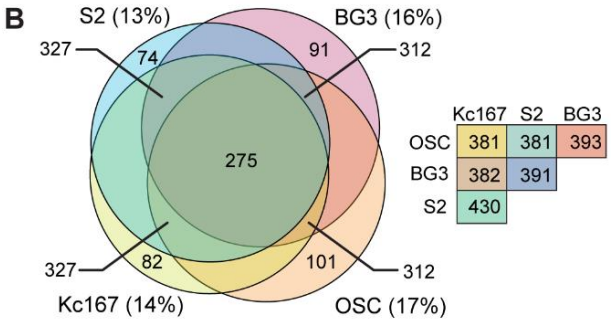
Корреляция (общие соседи) как прокси для геометрии

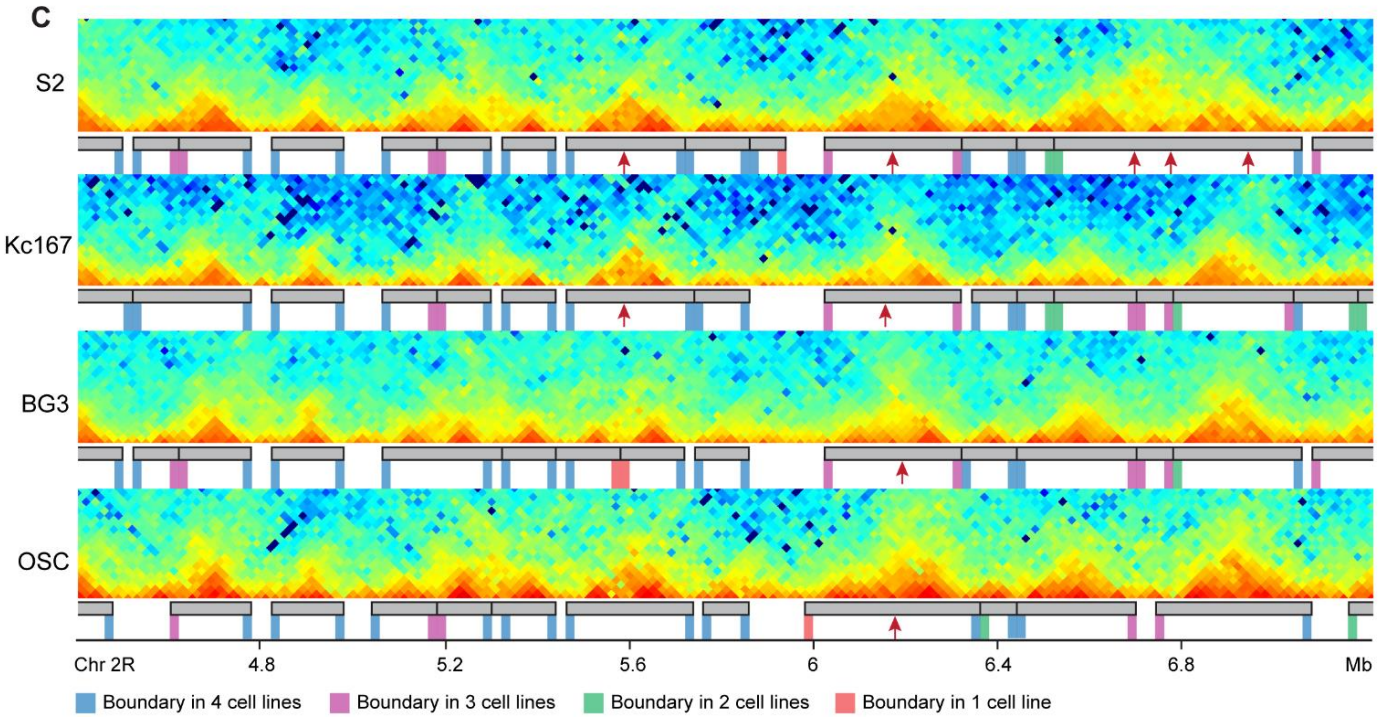
Границы ТАДов в основном совпадают между клеточными линиями



Definition of TADs depends on a parameter (γ). Low γ => larger TADs

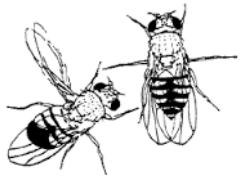


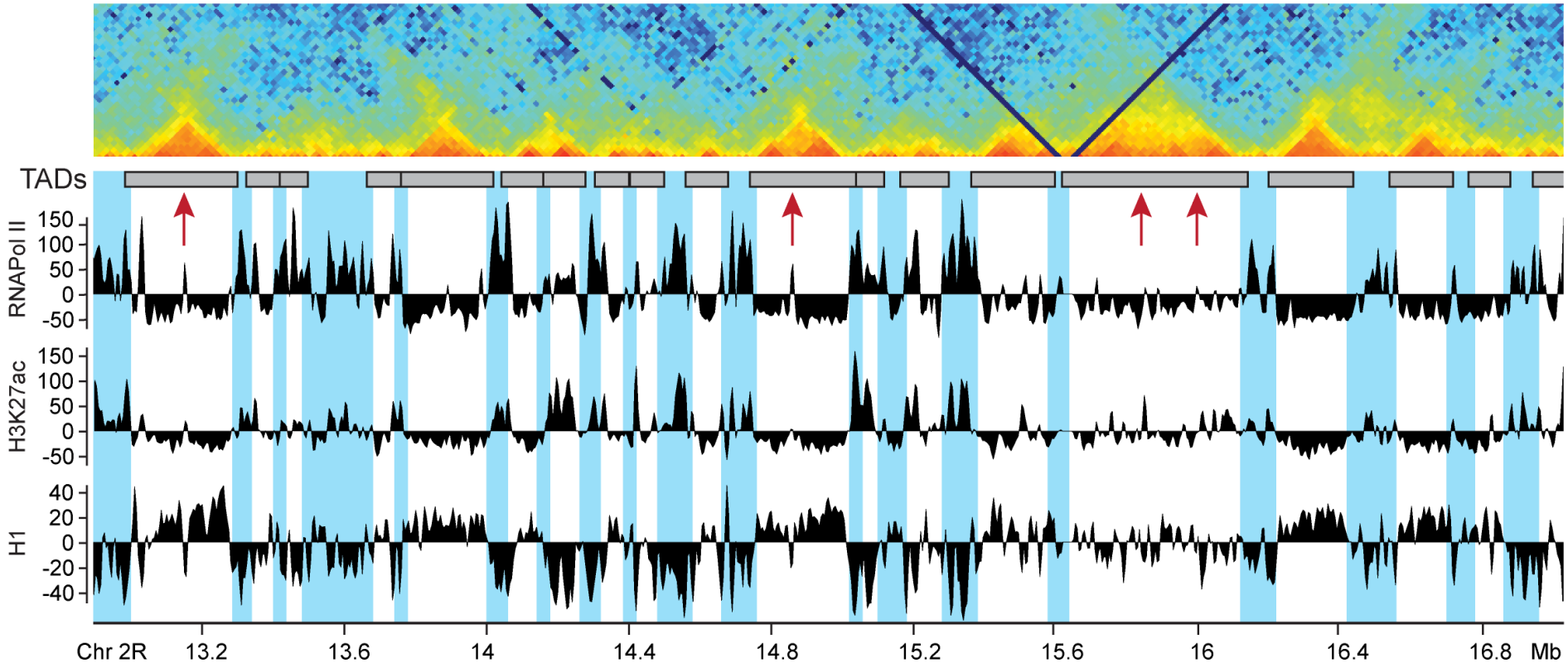




GENOME RESEARCH 2016

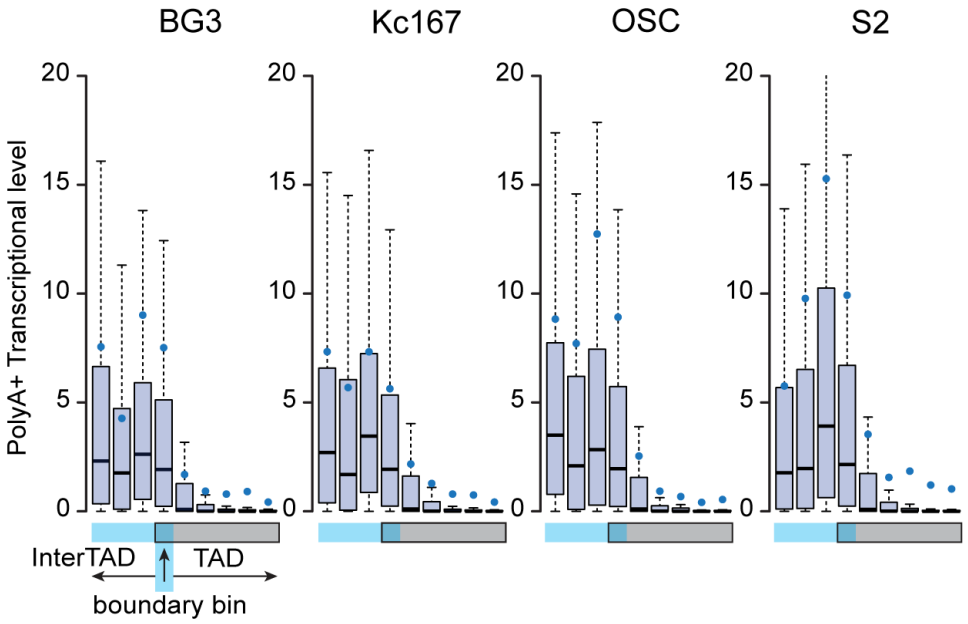
Меж-ТАДы обогащены активными хроматиновыми метками и pol II и обеднены репрессивными метками

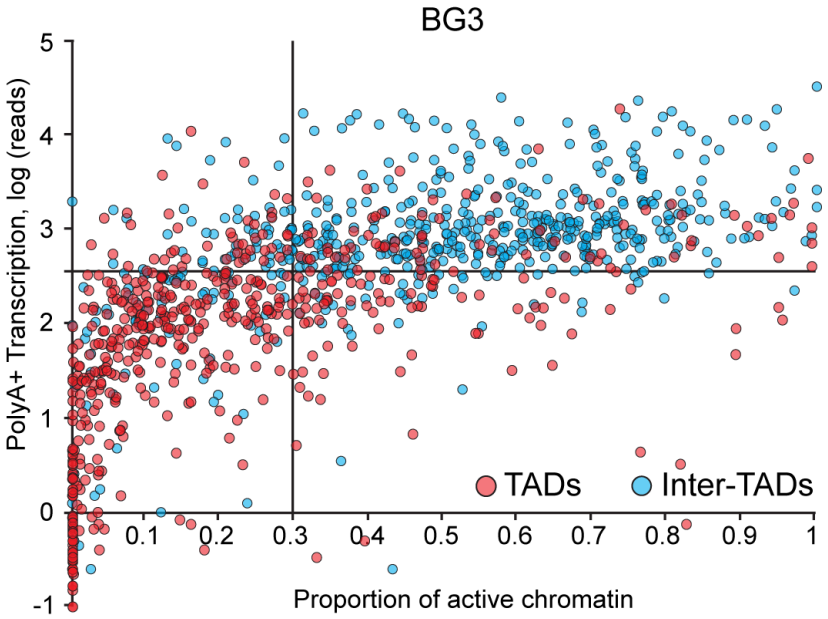




GENOME RESEARCH 2016

Уровень транскрипции и доля активного хроматина в ТАДах и меж-ТАДах

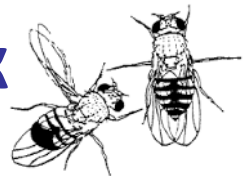


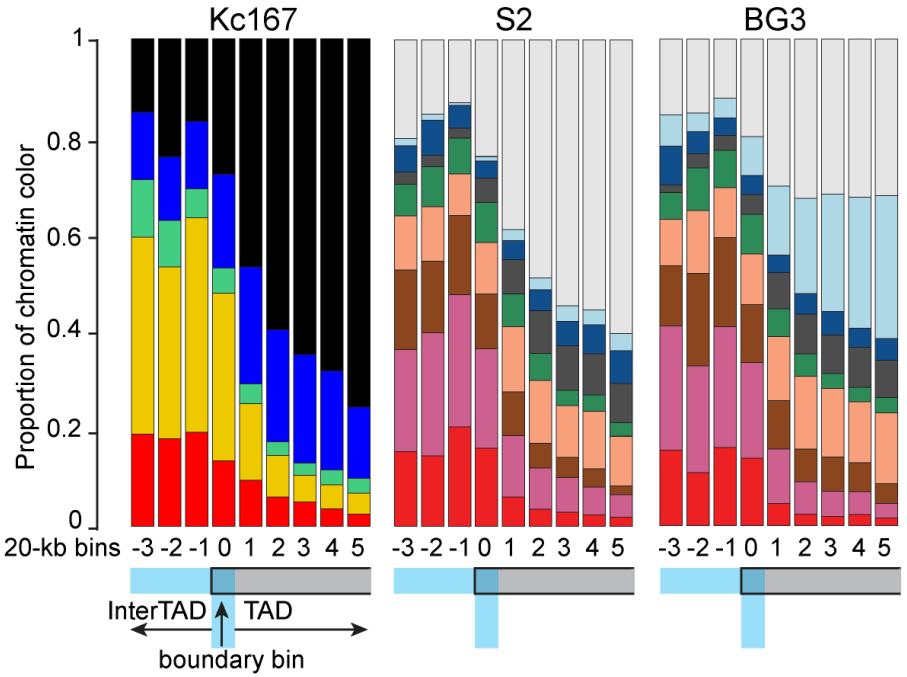


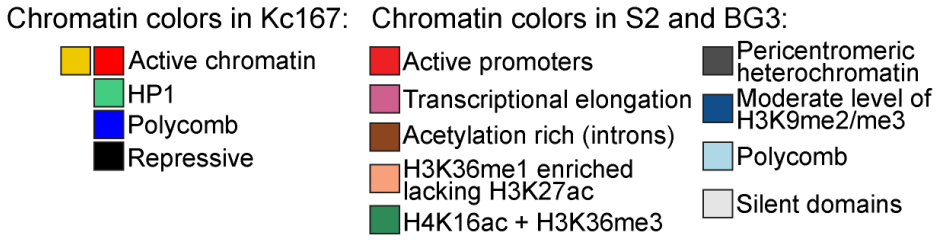
Log10 normalized read counts

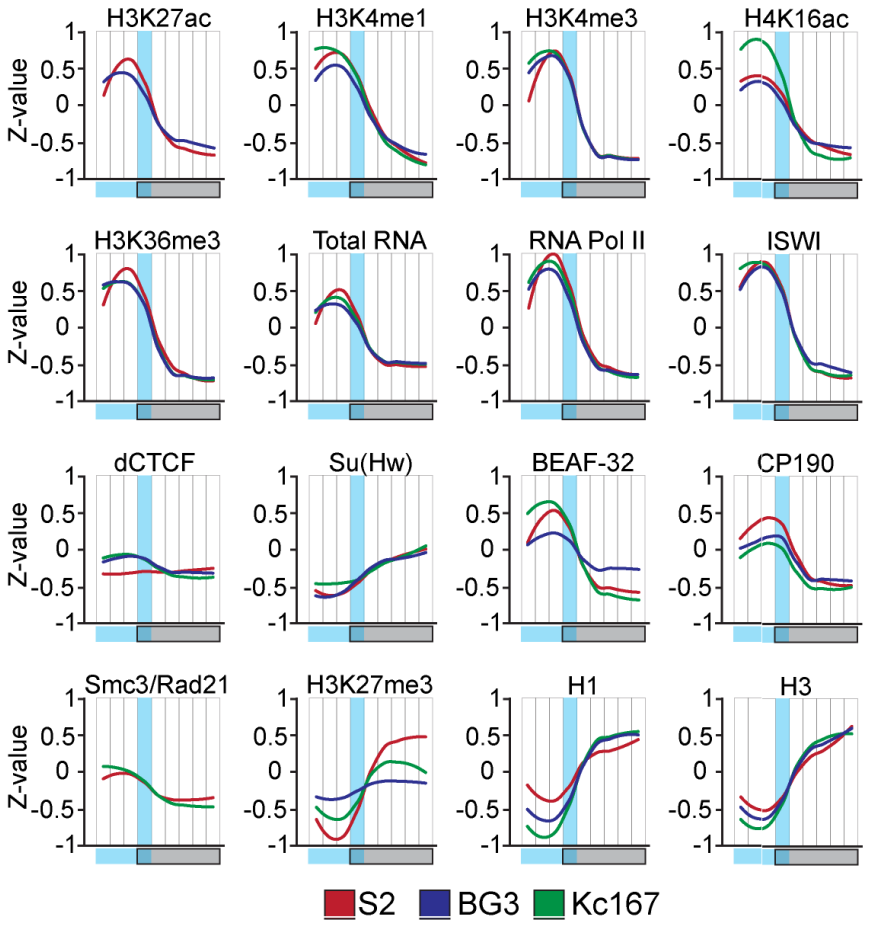
GENOME RESEARCH **2016**

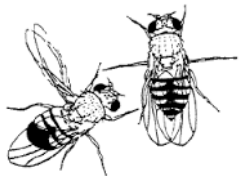
Распределение хроматиновых меток на границах ТАДов





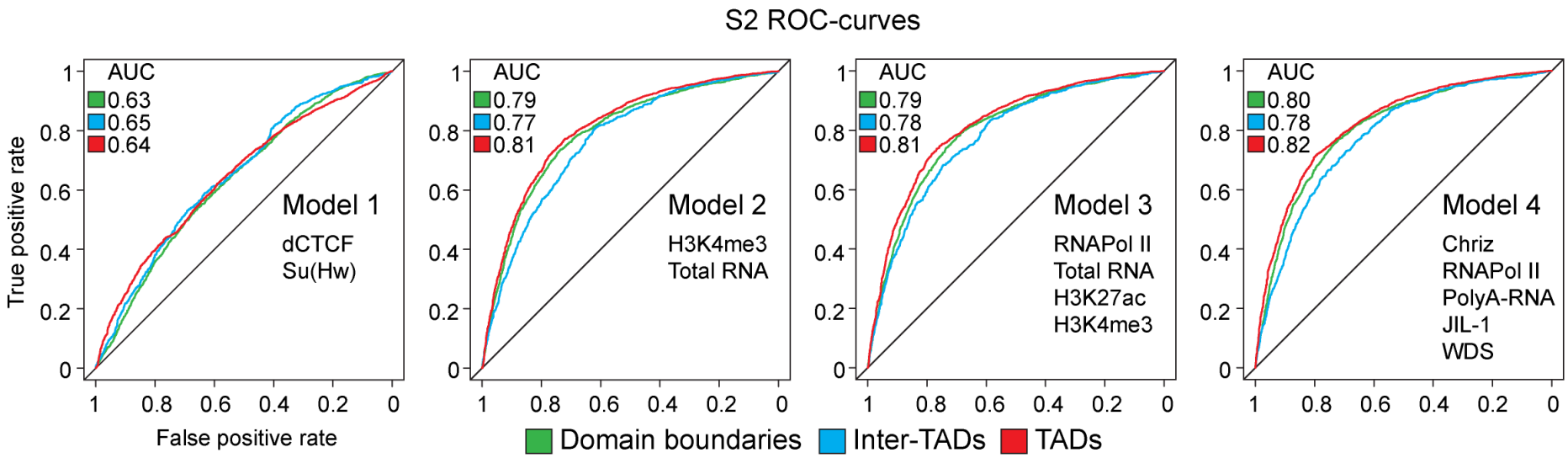






Предсказание границ по меткам и сайтам связывания инсуляторов

(попытка feature extraction)



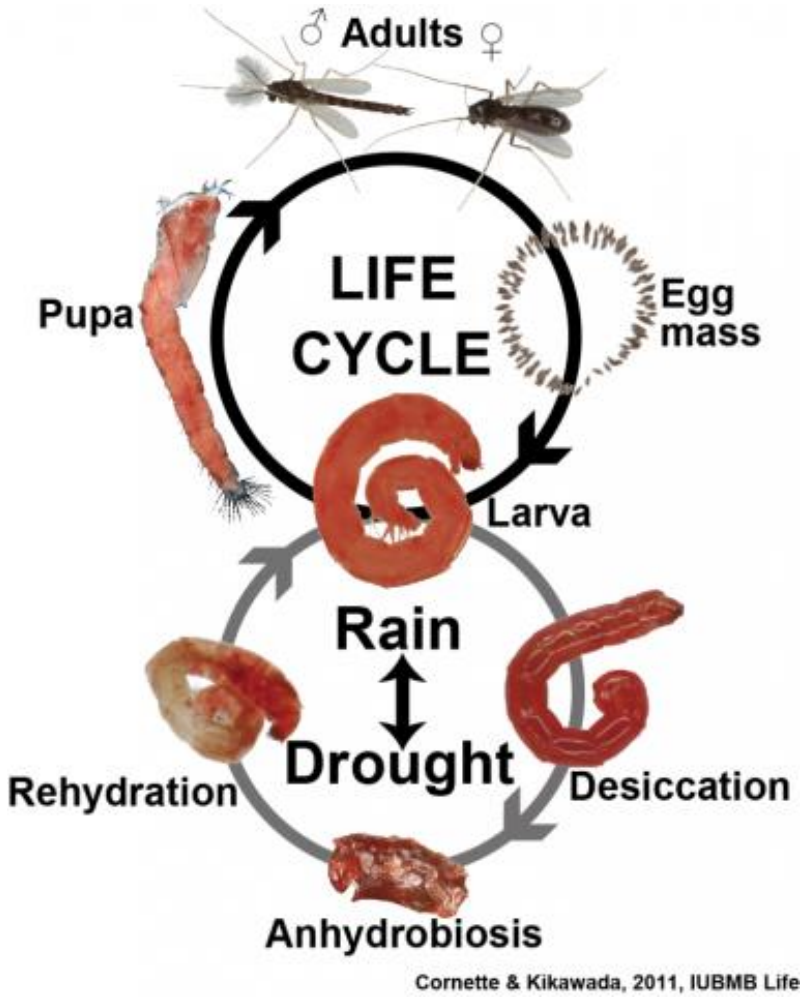
letters to nature

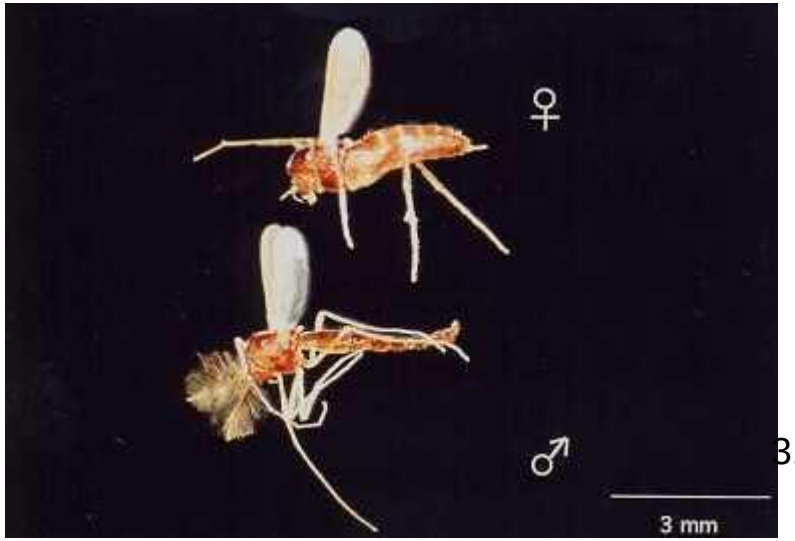
Nature **188**, 336 - 337 (22 October 1960); doi:10.1038/188336a0

A Fly Larva that tolerates Dehydration and Temperatures of −270° to +102° C.

H. E. HINTON

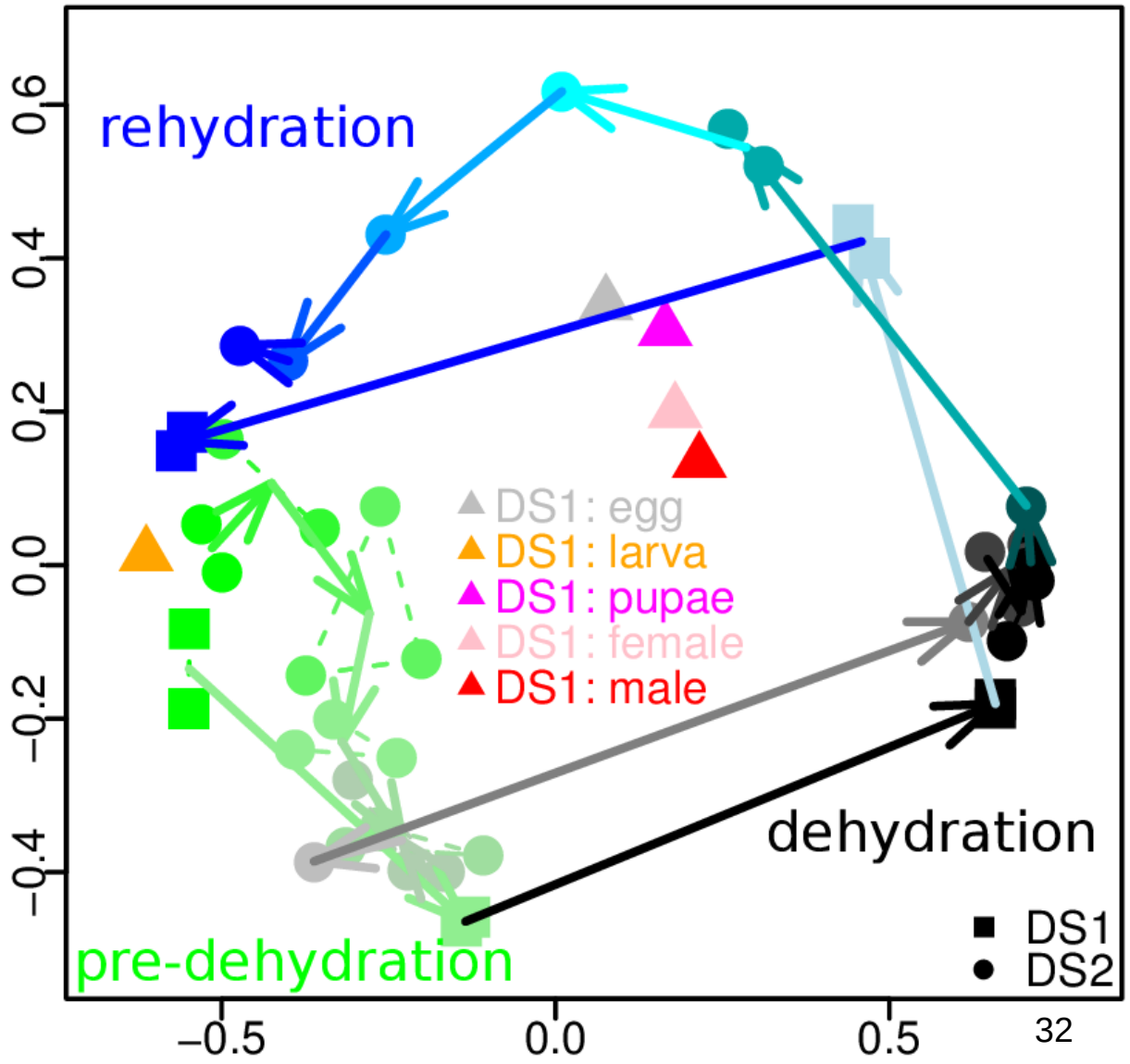
Department of Zoology, University of Bristol.



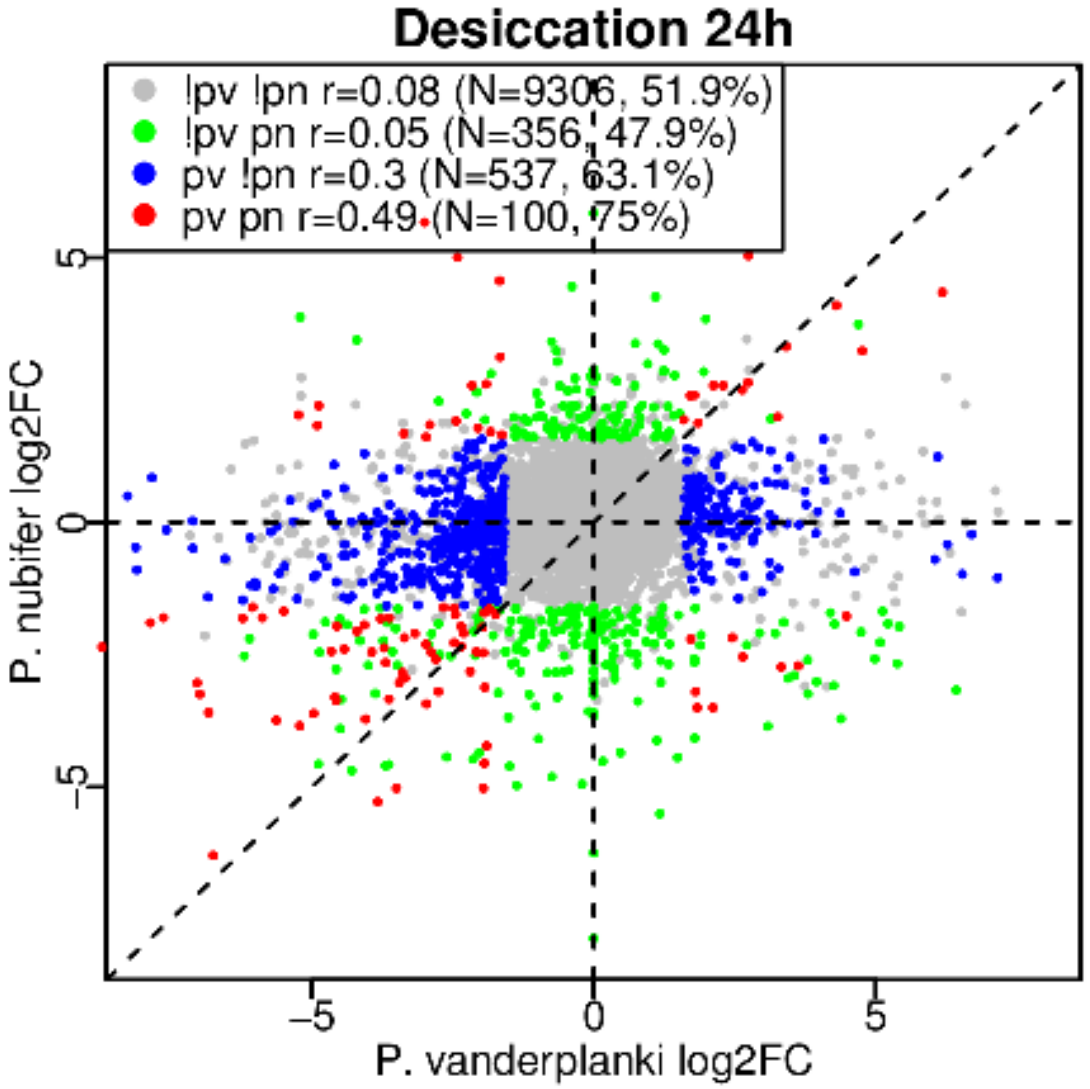


RNA-seq

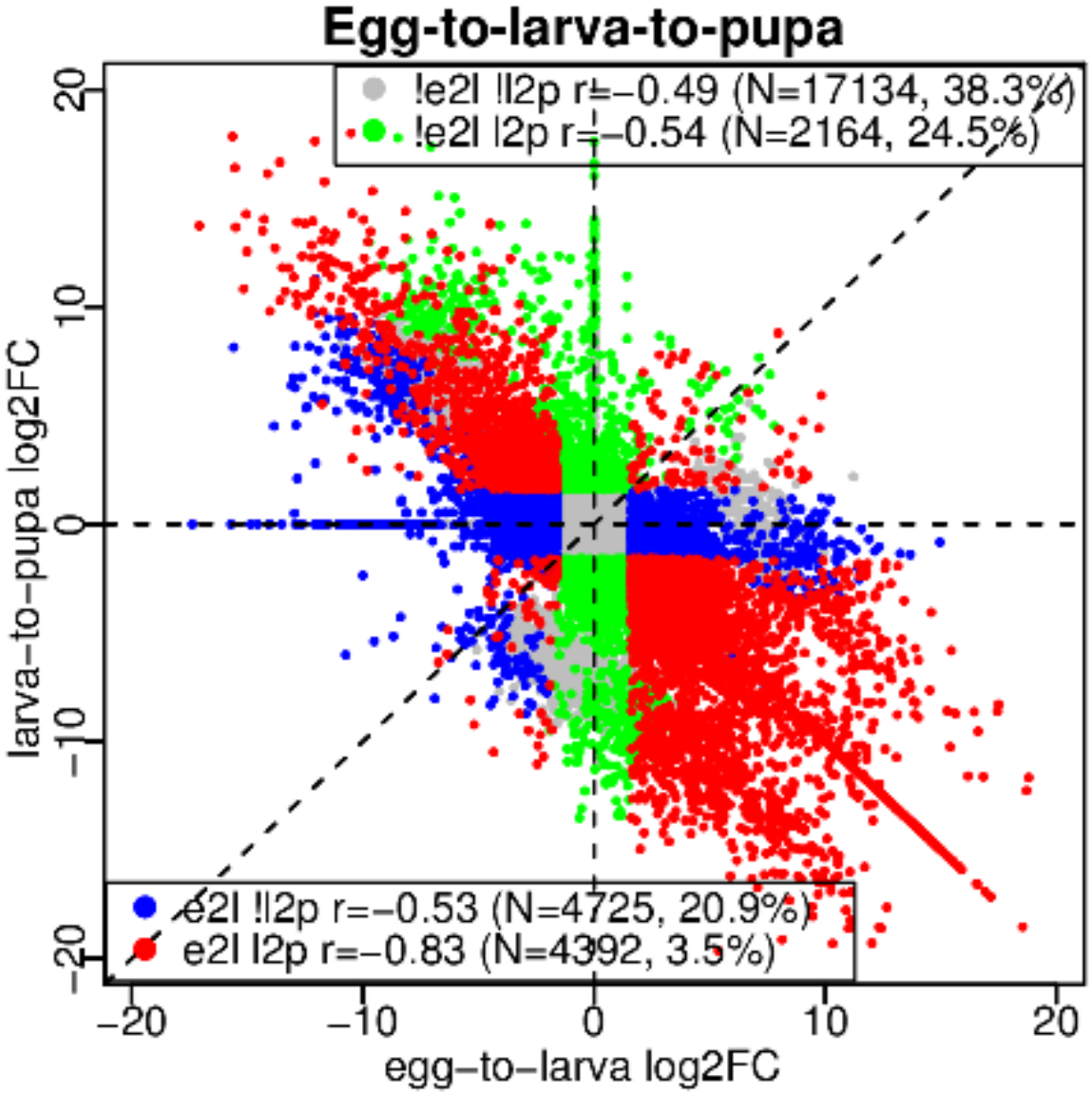
- Жизненный цикл
  - яйцо
  - личинка
  - куколка
  - самец и самка
- Цикл высыхание – размачивание
  - высыхание: 0, 24, 48 ч.
  - размачивание: 3, 24 ч.
  - *P. nubifer*: 0 и 24 ч.



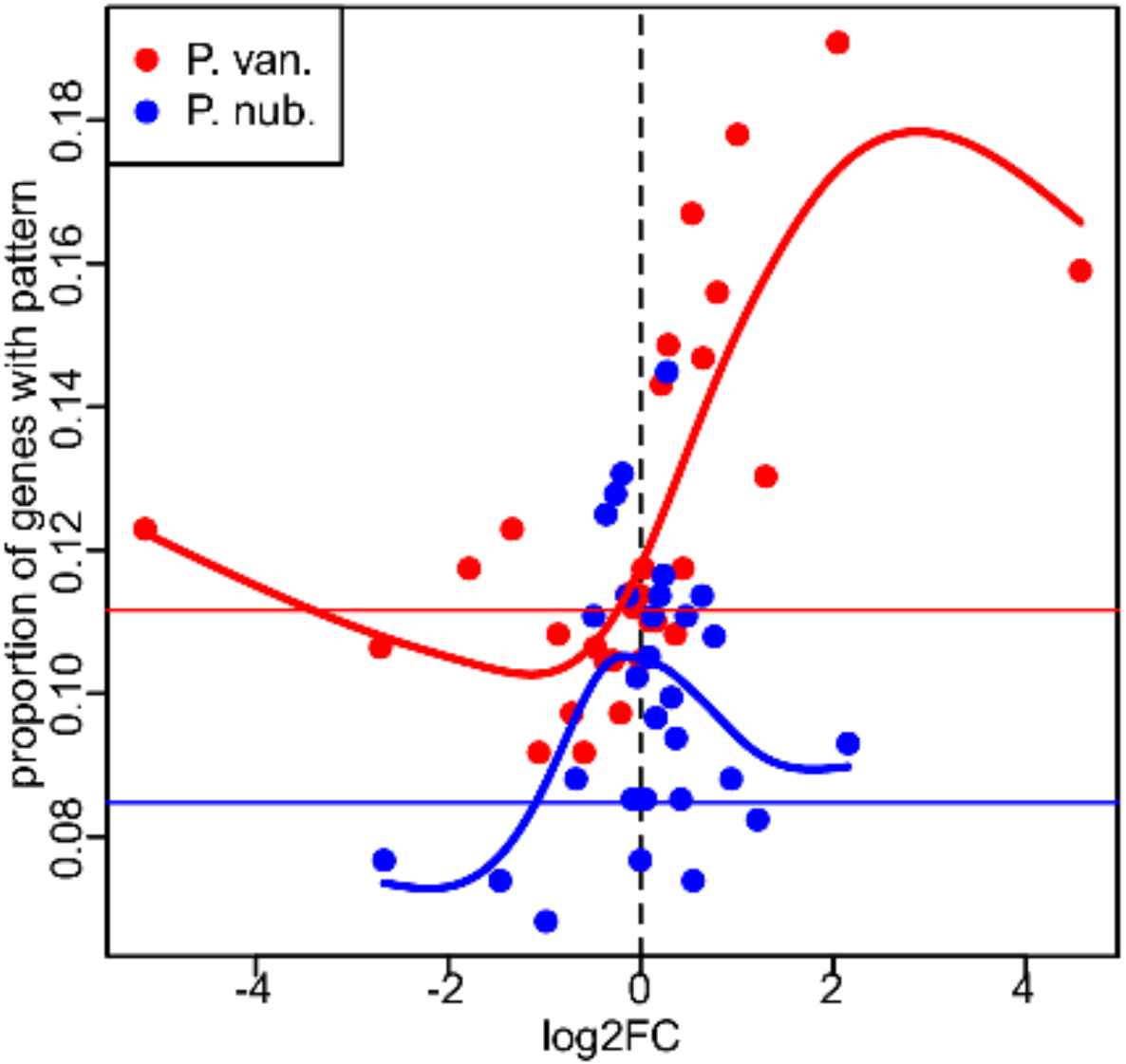
Высыхание – два комара



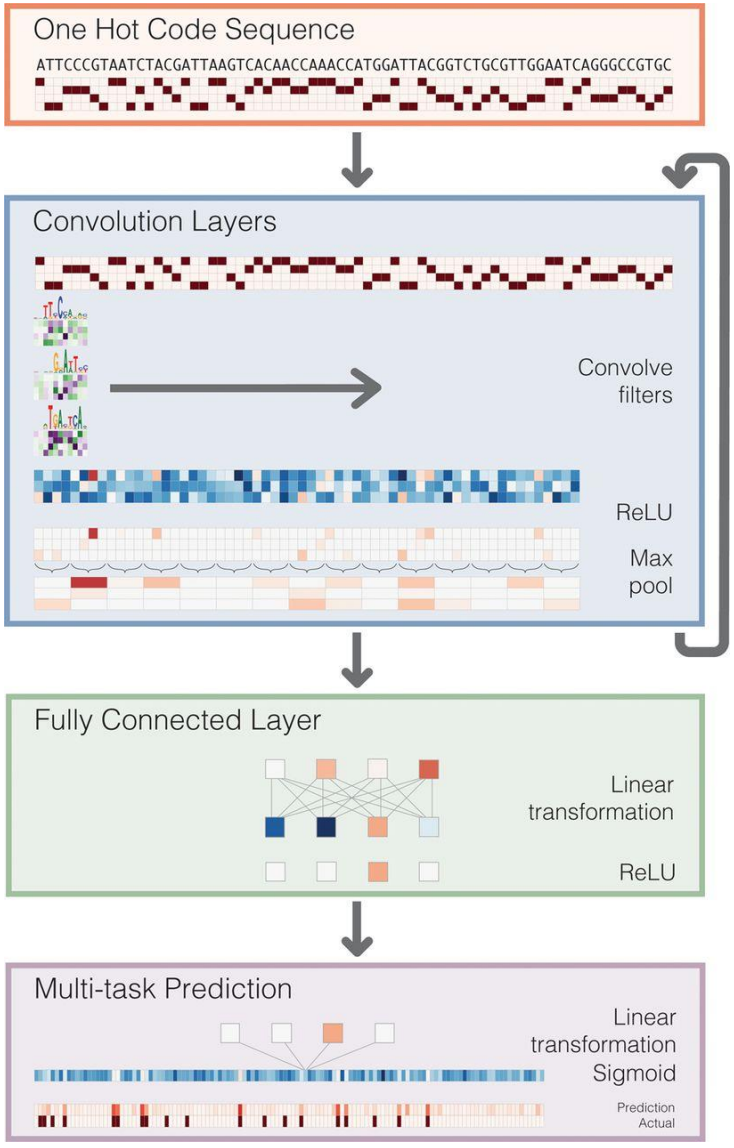
Развитие: яйцо – личинка – куколка



TCTAGAA => рост экспрессии при высыхании



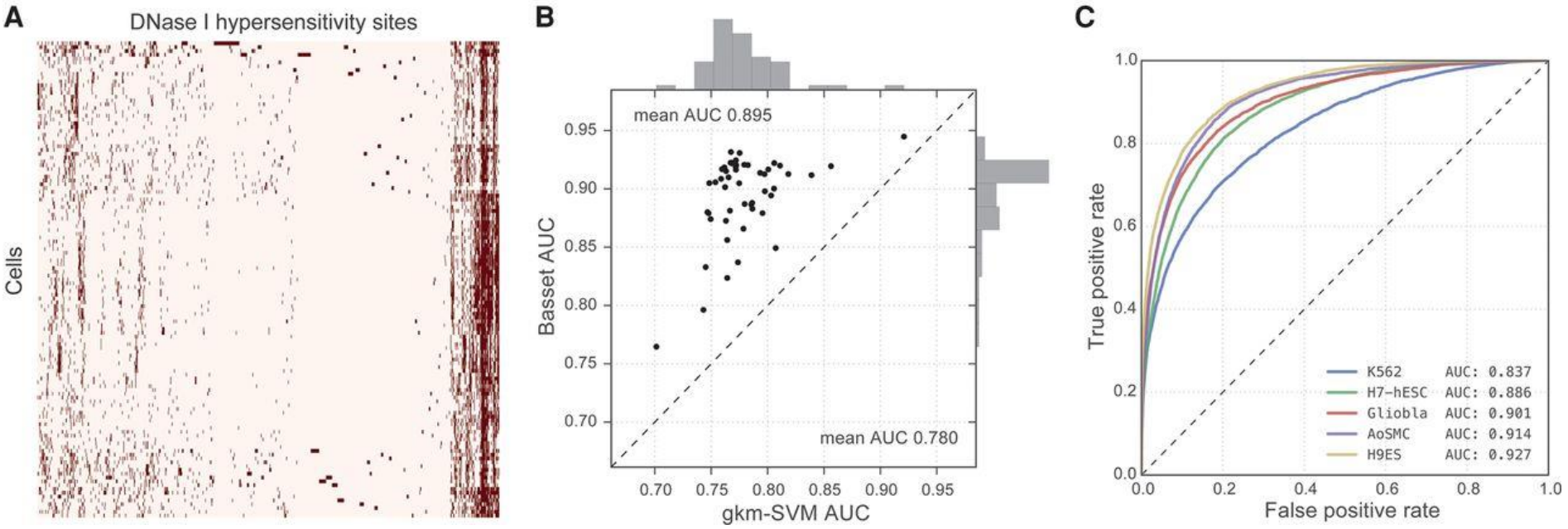
Глубокая сверточная нейронная сеть (CNN)



Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks

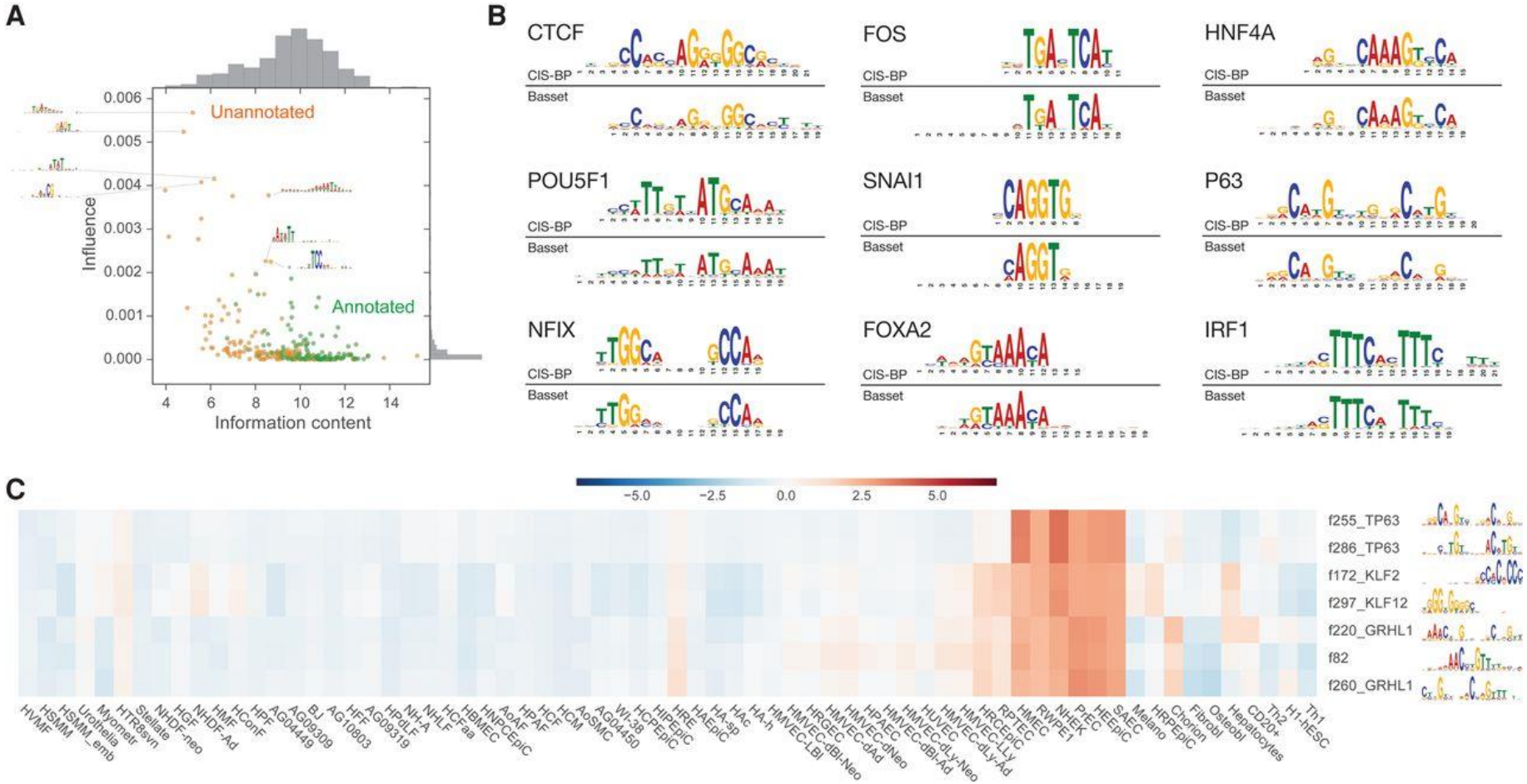
David R. Kelley et al. Genome Res. 2016;26:990-999

… предсказывает доступность хроматина в клеточных линиях



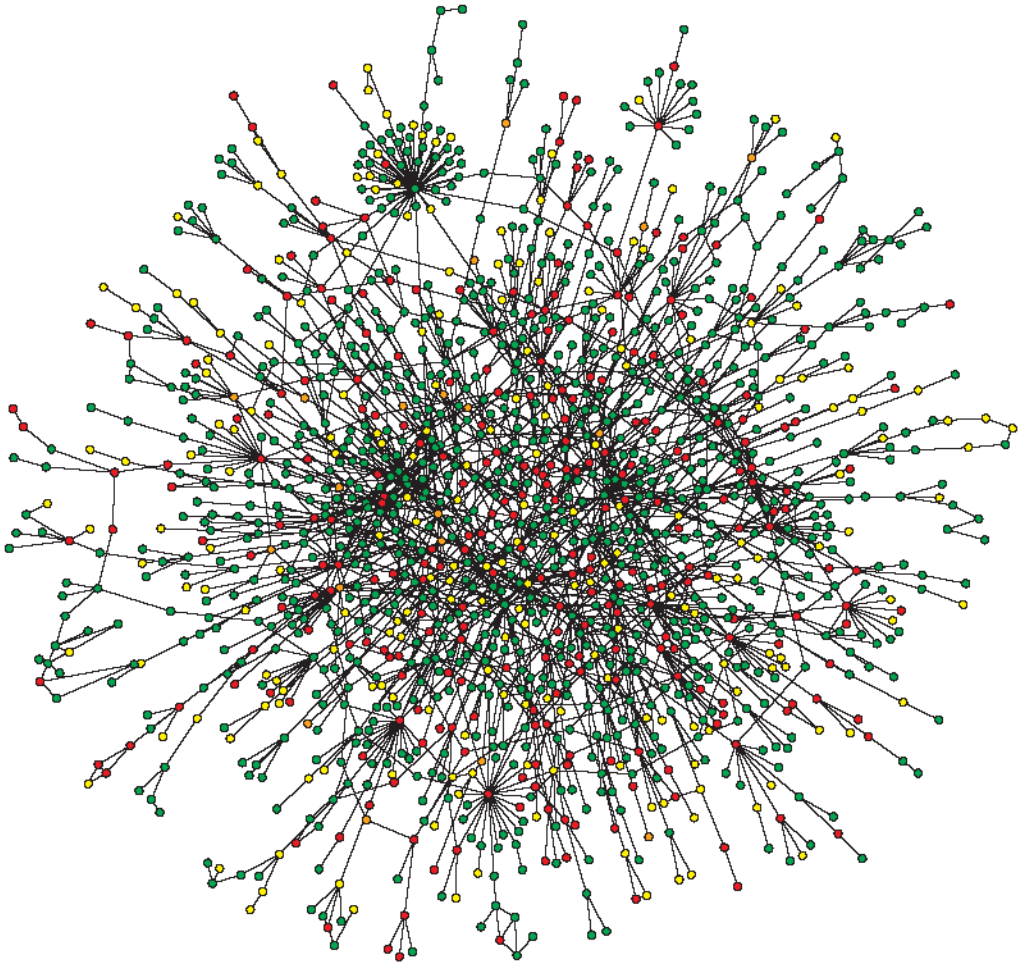
David R. Kelley et al. Genome Res. 2016;26:990-999

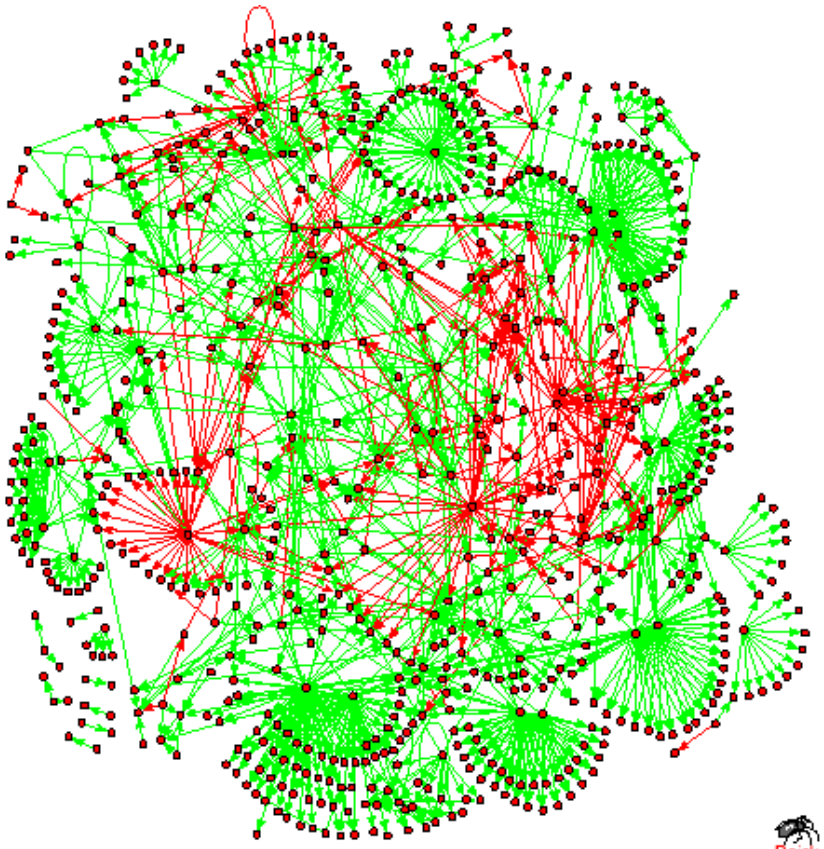
Первый сверточный слой – известные (и новые?) мотивы связывания факторов транскрипции



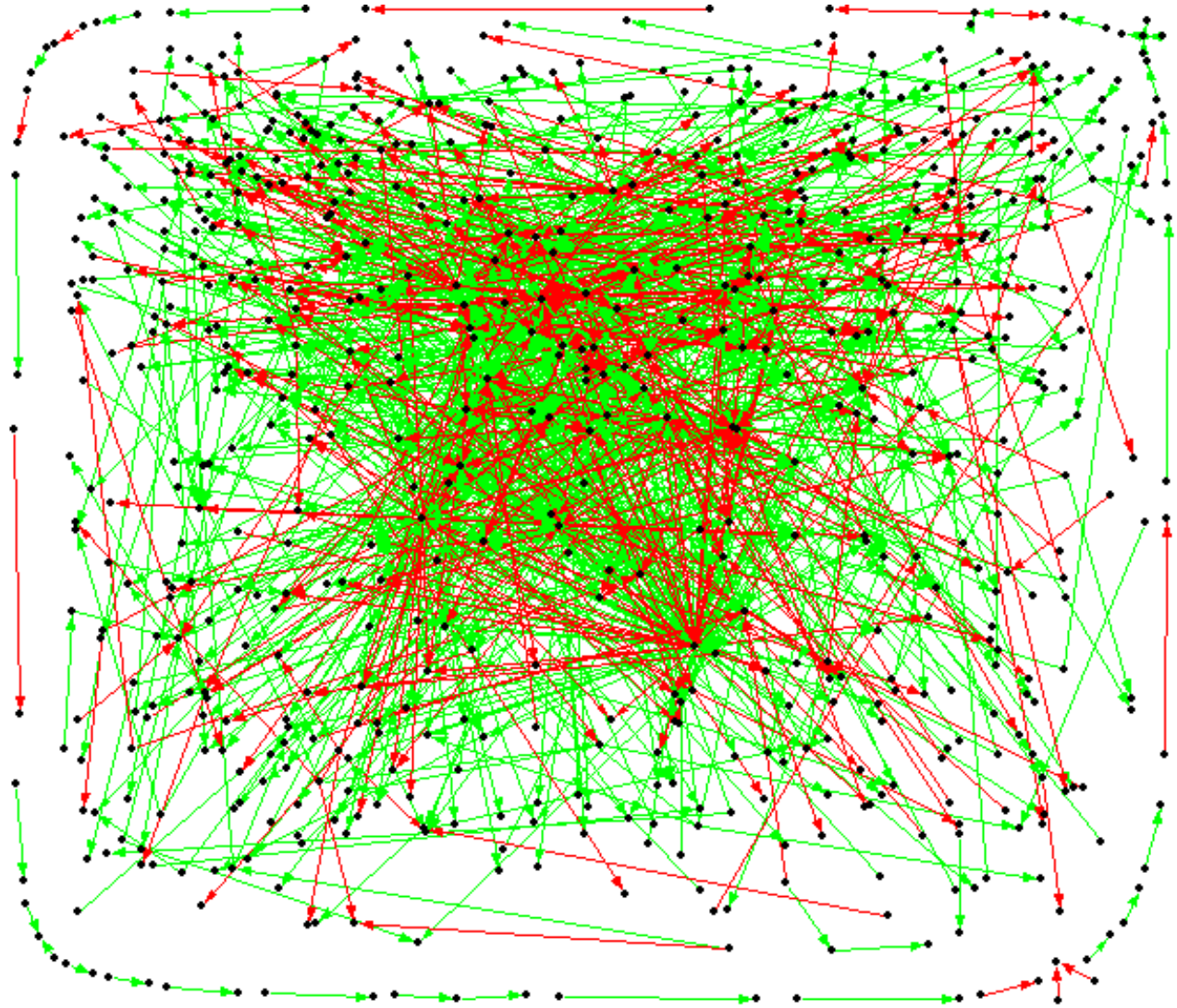
David R. Kelley et al. Genome Res. 2016;26:990-999

Белок-белковые (структурные, сигнальные и др.) **и белок-ДНКовые** (регуляторные) **взаимодействия в дрожжах**





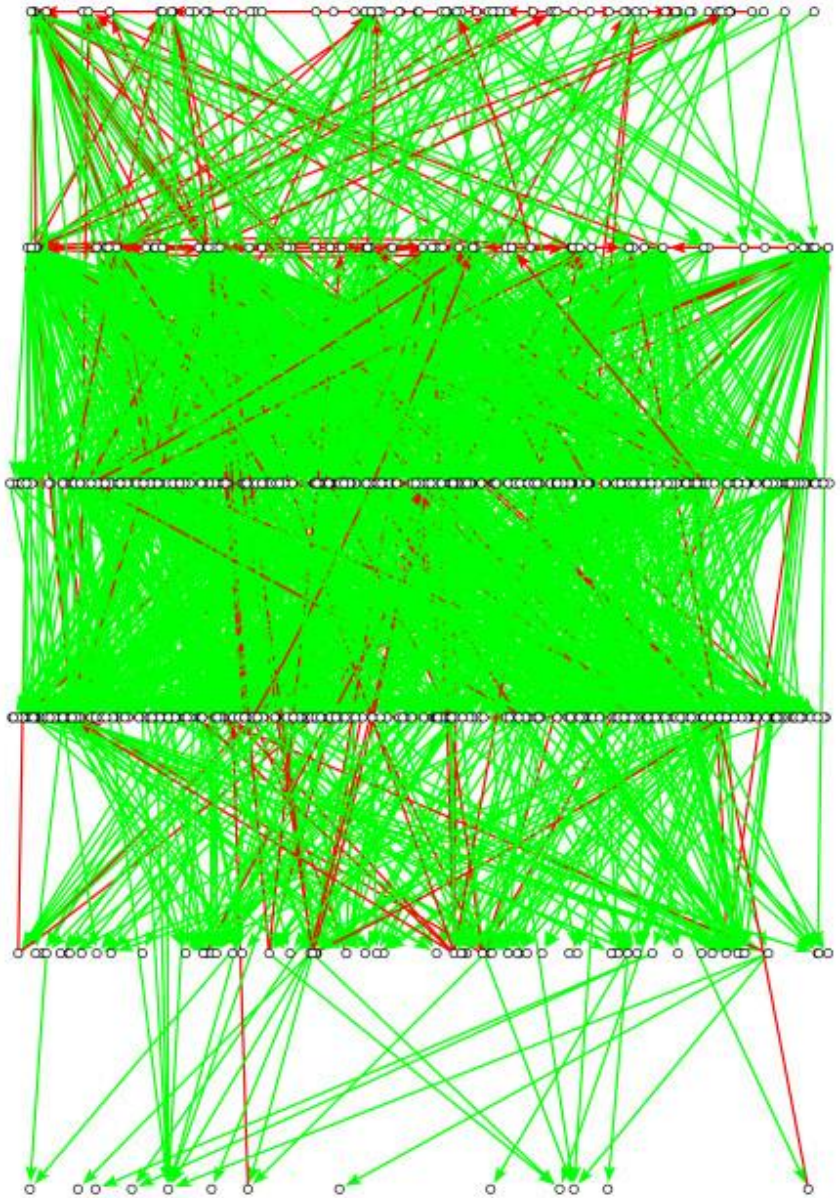
Регуляция транскрипции у человека



- 1449 взаимодействий между 689 генами
- Отношение «активаторы : репрессоры» = 3:1
- До 95 регулируемых генов, до 45 регуляторов.

Иерархия:
732 белков
(71 рецепторов),
1671 взаимодействий
(фосфорилирование,
дефосфорилировние,
гидролиз etc)

208 анти-
иерархических ребер



Динамика: активность транскрипционных взаимодействий в клеточных линиях