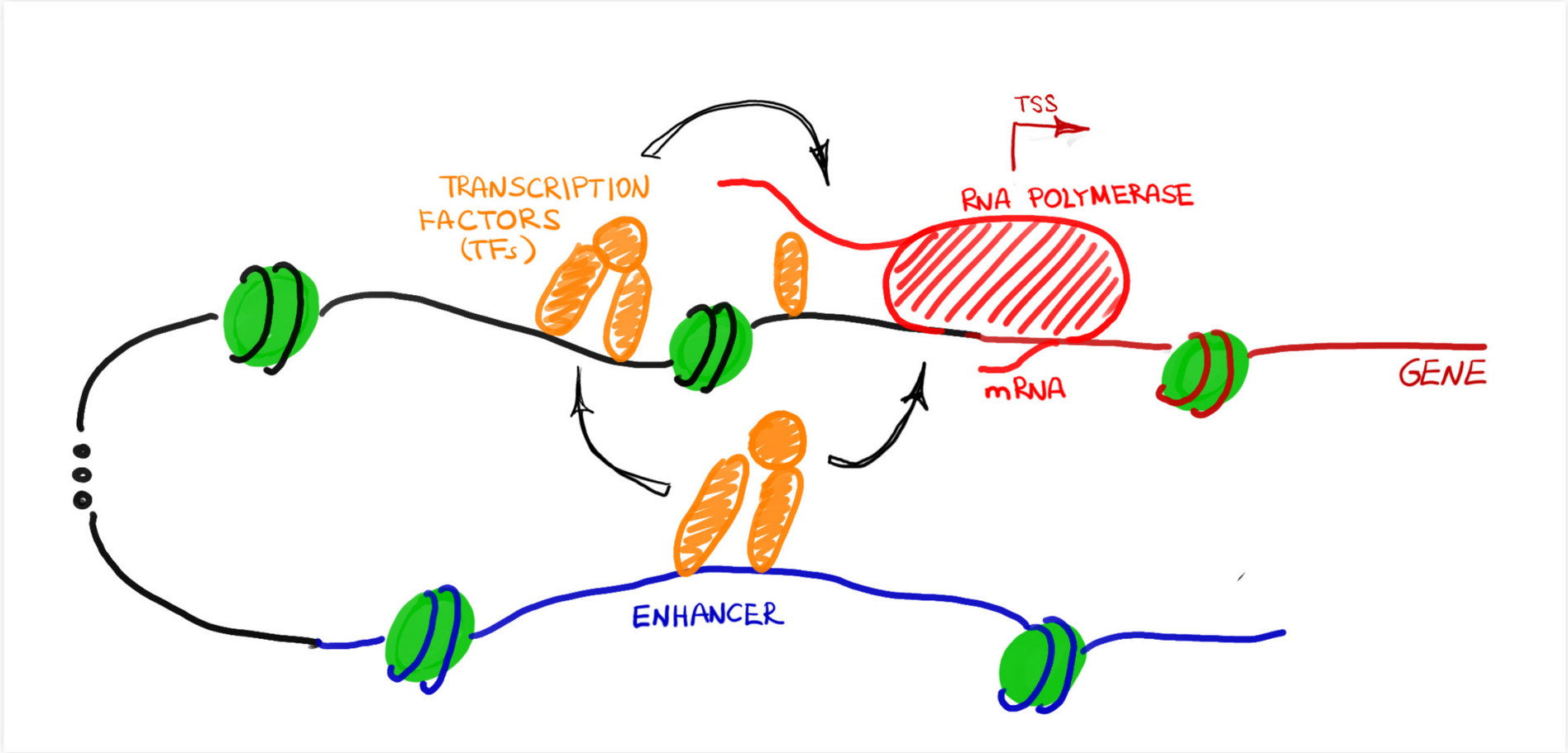
Ваня Кулаковский
ИМБ РАН, ИОГен РАН, Сколтех, autosome.ru

Анализ ДНК-белкового узнавания с помощью ChIP-Seq и полногеномное предсказание сайтов связывания: по следам соревнования DREAM-ENCODE


TSS
TRANSCRIPTION FACTORS (TFs)
RNA POLYMERASE
mRNA
GENE
ENHANCER

VIEWPOINT

The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium*†

The ENCyclopedia Of DNA Elements (ENCODE) Project aims to identify all functional elements in the human genome sequence. The pilot phase of the Project is focused on a specified 30 megabases (~1%) of the human genome sequence and is organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The results of this pilot phase will guide future efforts to analyze the entire human genome.

With the complete human genome sequence now in hand (*1*–*3*), we face the enormous challenge of interpreting it and learning how to use that information to understand the biology of human health and disease. The ENCyclopedia Of DNA Elements (ENCODE) Project is predicated on the belief that a comprehensive catalog of the structural and functional components encoded in the human genome sequence will be critical for understanding human biology well enough to address those fundamental aims of biomedical research. Such a complete catalog, or "parts list," would include protein-coding namics; undoubtedly, additional, yet-to-be-defined types of functional sequences will also need to be included.

To illustrate the magnitude of the challenge involved, it only needs to be pointed out that an inventory of the best-defined functional components in the human genome—the protein-coding sequences—is still incomplete for a number of reasons, including the fragmented nature of human genes. Even with essentially all of the human genome sequence in hand, the number of protein-coding genes can still only be estimated (currently 20,000 to 25,000) (*3*). Non–protein-coding genes are approaches, such as cDNA-cloning efforts (*4*, *5*) and chip-based transcriptome analyses (*6*, *7*), have revealed the existence of many transcribed sequences of unknown function. As a reflection of this complexity, about 5% of the human genome is evolutionarily conserved with respect to rodent genomic sequences, and therefore is inferred to be functionally important (*8*, *9*). Yet only about one-third of the sequence under such selection is predicted to encode proteins (*1*, *2*). Our collective knowledge about putative functional, noncoding elements, which represent the majority of the remaining functional sequences in the human genome, is remarkably underdeveloped at the present time.

An added level of complexity is that many functional genomic elements are only active or expressed in a restricted fashion—for example, in certain cell types or at particular developmental stages. Thus, one could envision that a truly comprehensive inventory of

Vol 447|14 June 2007|doi:10.1038/nature05874

nature

ARTICLES

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

We report the generation and analysis of functional data from multiple, diverse experiments performed on a targeted 1% of the human genome as part of the pilot phase of the ENCODE Project. These data have been further integrated and augmented by a number of evolutionary and computational analyses. Together, our results advance the collective knowledge about human genome function in several major areas. First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another. Second, systematic examination of transcriptional regulation has yielded new understanding about transcription start sites, including their relationship to specific regulatory sequences and features of chromatin accessibility and histone modification. Third, a more sophisticated view of chromatin structure has emerged, including its inter-relationship with DNA replication and transcriptional regulation. Finally, integration of these new sources of information, in particular with respect to mammalian evolution based on inter- and intra-species sequence comparisons, has yielded new mechanistic and evolutionary insights concerning the functional landscape of the human genome. Together, these studies are defining a path for pursuit of a more comprehensive characterization of human genome function.

ARTICLE

doi:10.1038/nature11247

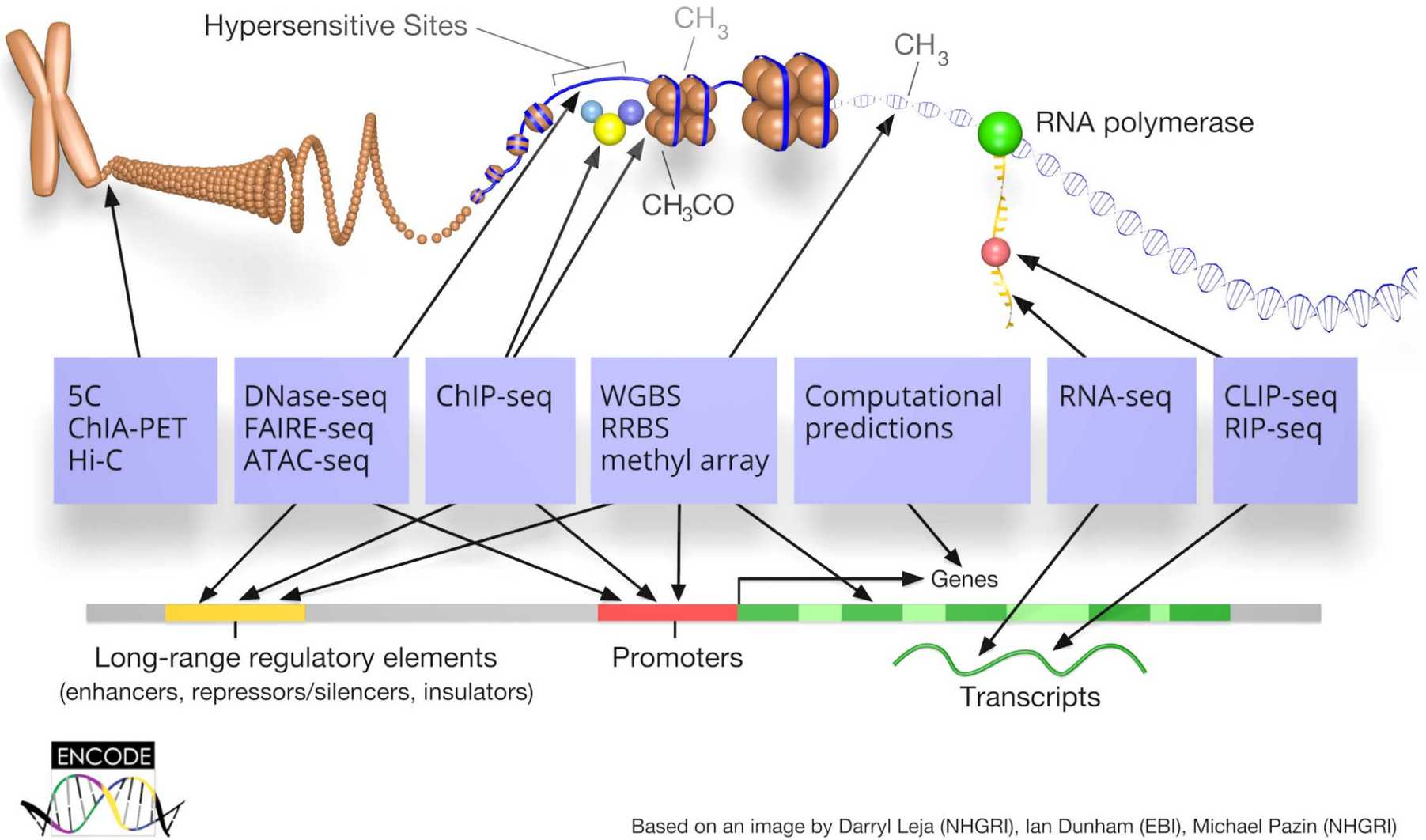
An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with

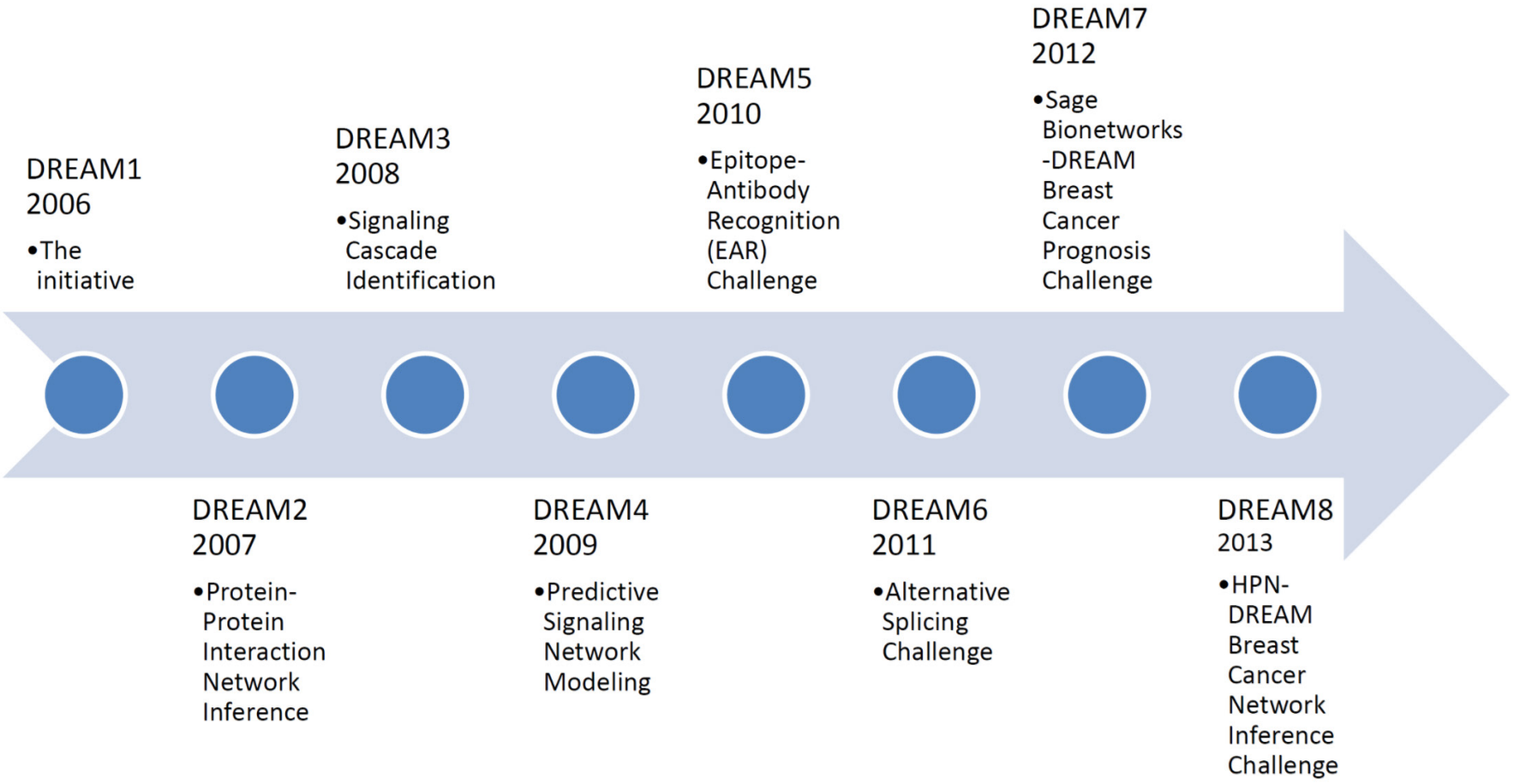
95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

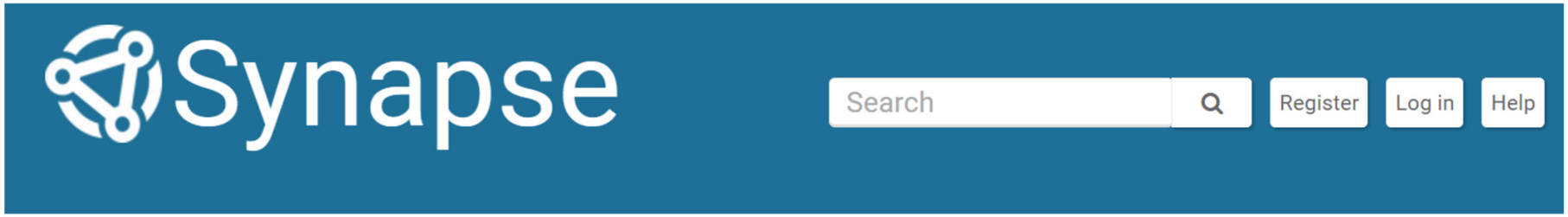


Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

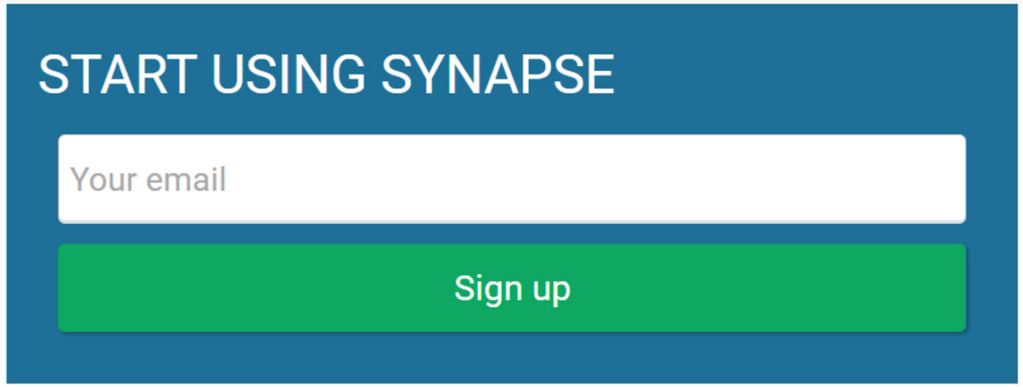
DREAM

The Dialogue on Reverse Engineering Assessment and Methods Project, founded in 2006 by Andrea Califano (Columbia University) and Gustavo Stolovitzky (IBM), was originally conceived as an initiative to advance the nascent field of network biology through the organization of Challenges on network reconstruction and pathway inference.


DREAM1
2006
•The initiative
DREAM2
2007
•Protein-Protein Interaction Network Inference
DREAM3
2008
•Signaling Cascade Identification
DREAM4
2009
•Predictive Signaling Network Modeling
DREAM5
2010
•Epitope-Antibody Recognition (EAR) Challenge
DREAM6
2011
•Alternative Splicing Challenge
DREAM7
2012
•Sage Bionetworks -DREAM Breast Cancer Prognosis Challenge
DREAM8
2013
•HPN-DREAM Breast Cancer Network Inference Challenge

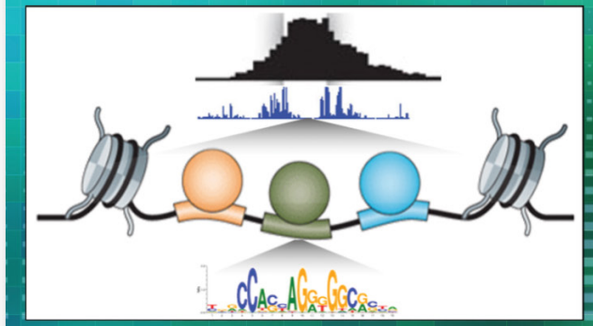


In 2013, Sage Bionetworks joined with the DREAM community to co-lead a new generation of Challenges that leverage collaborative data hosting and analysis tools available on Synapse (www.synapse.org), Sage Bionetworks' open bioinformatics compute space.

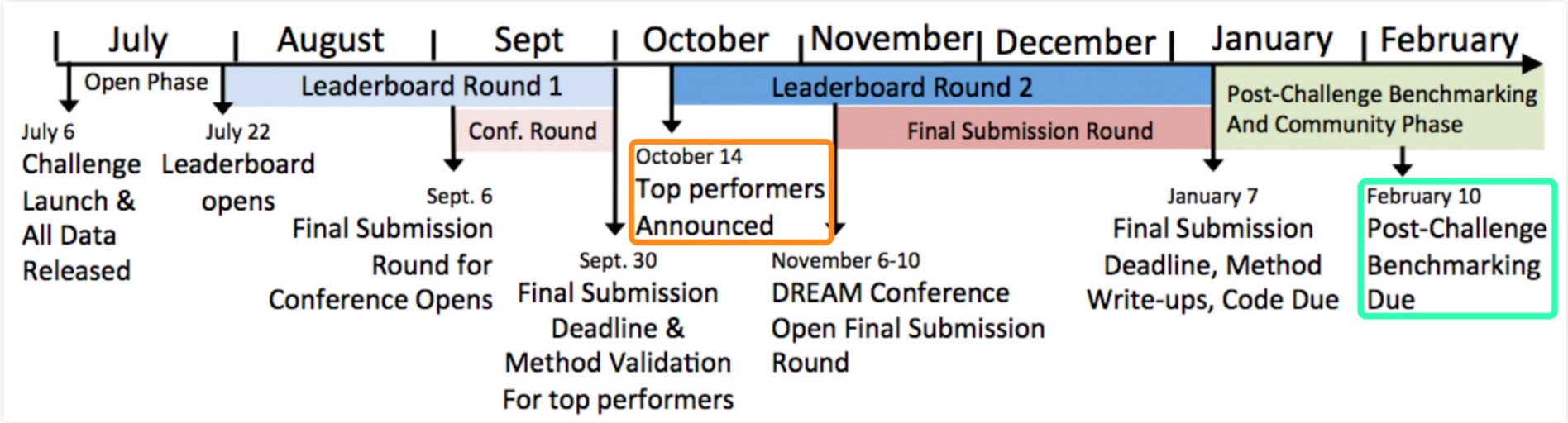


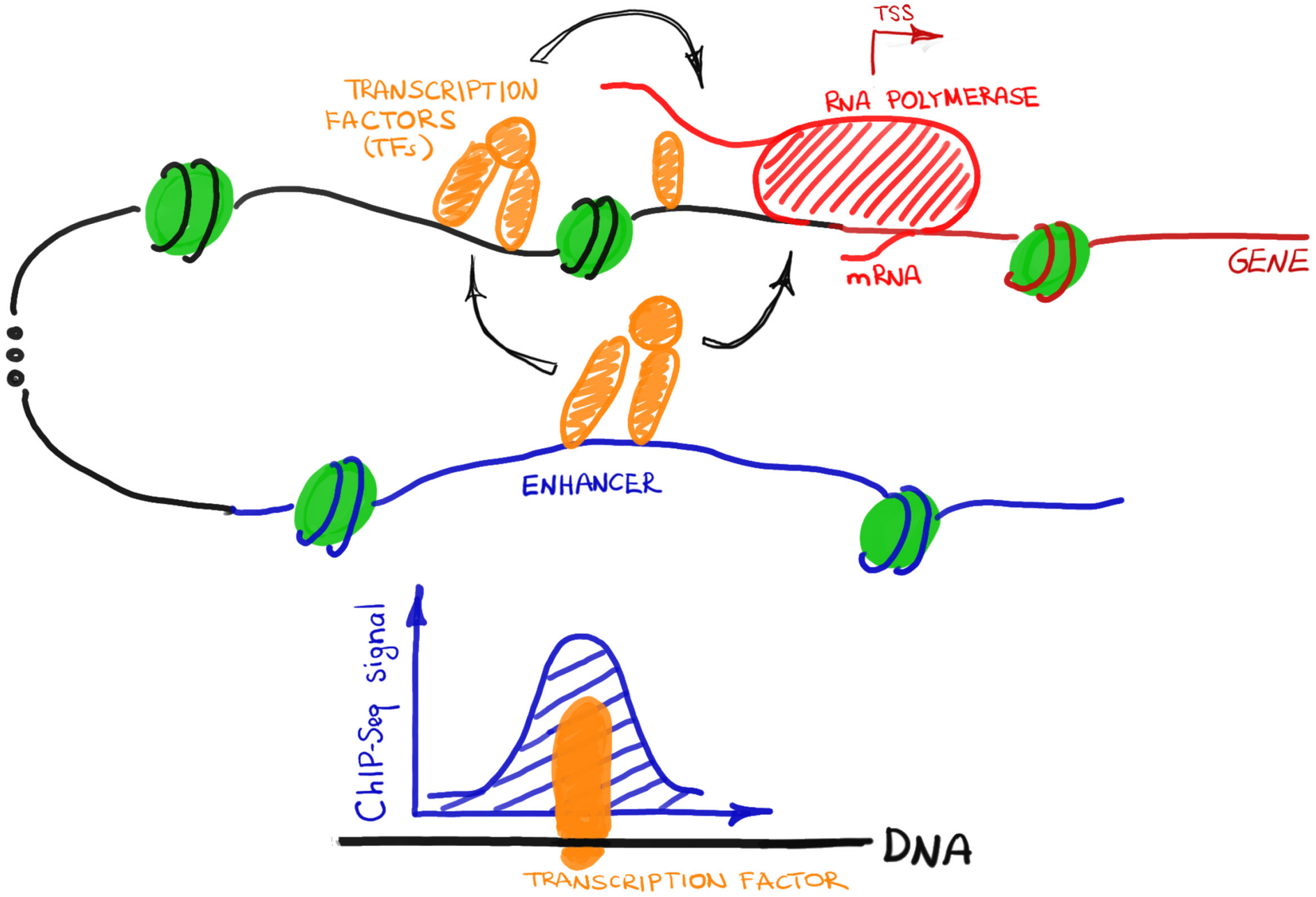
DREAM-ENCODE Challenge (2016-2017)

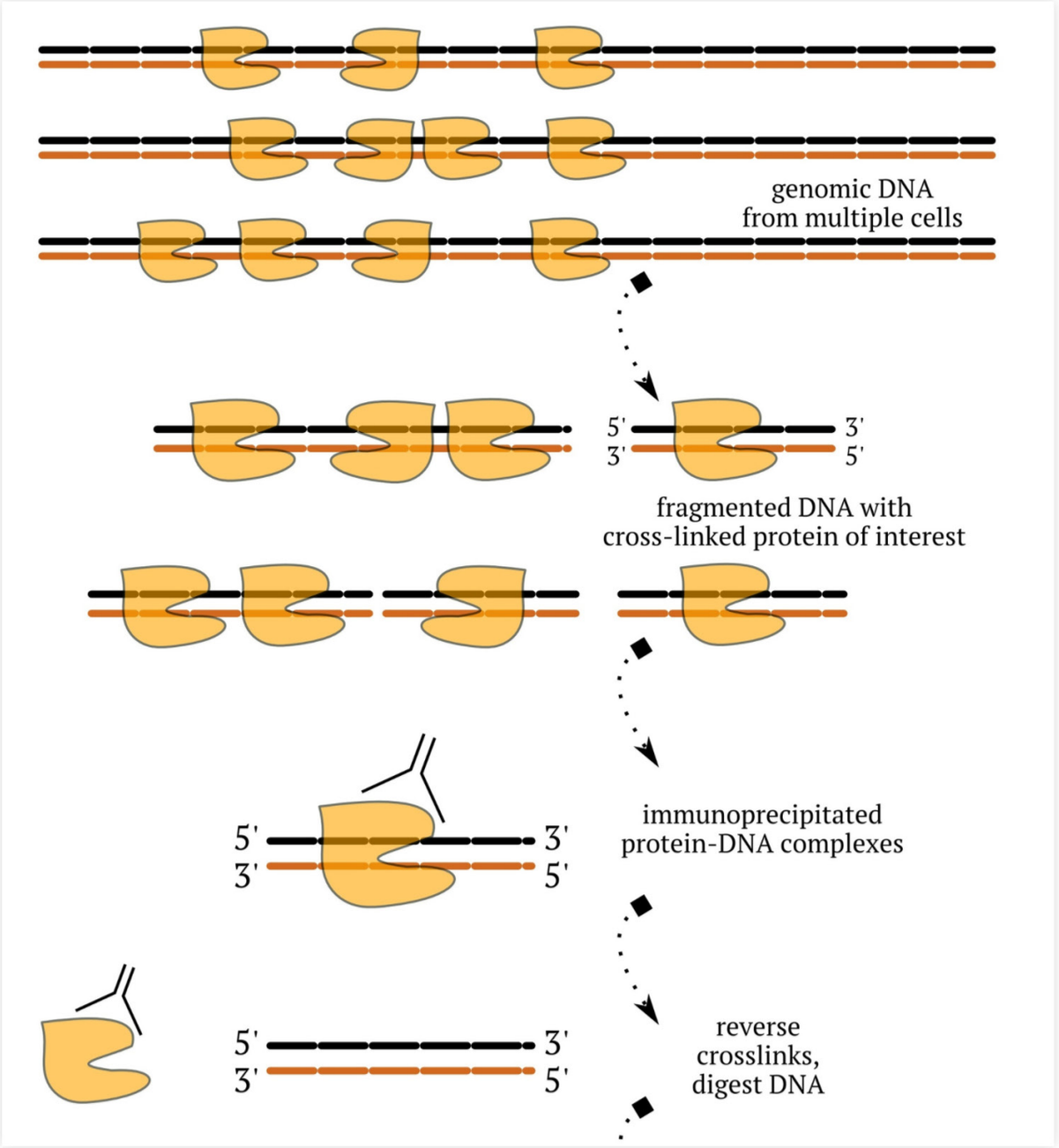
ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge

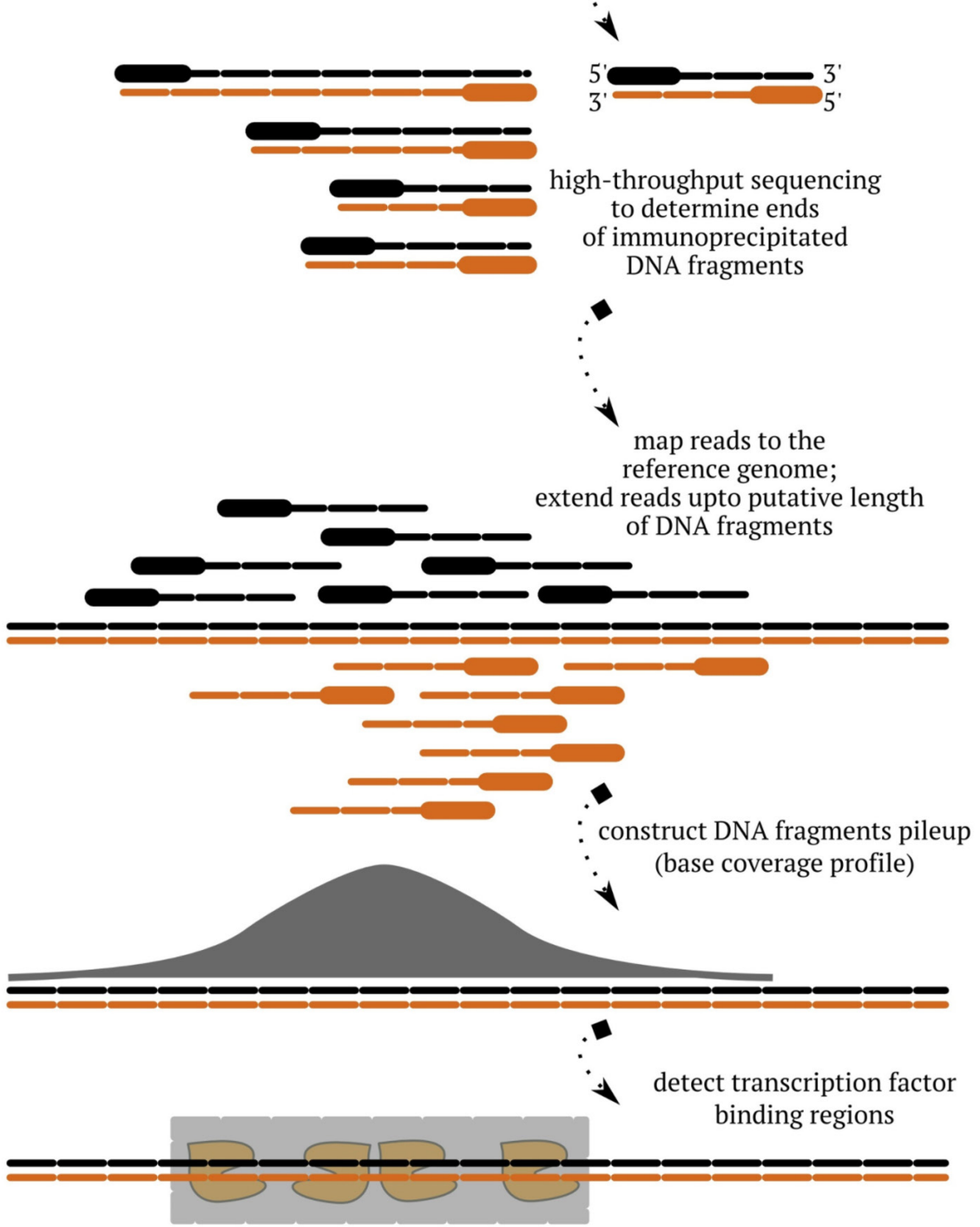


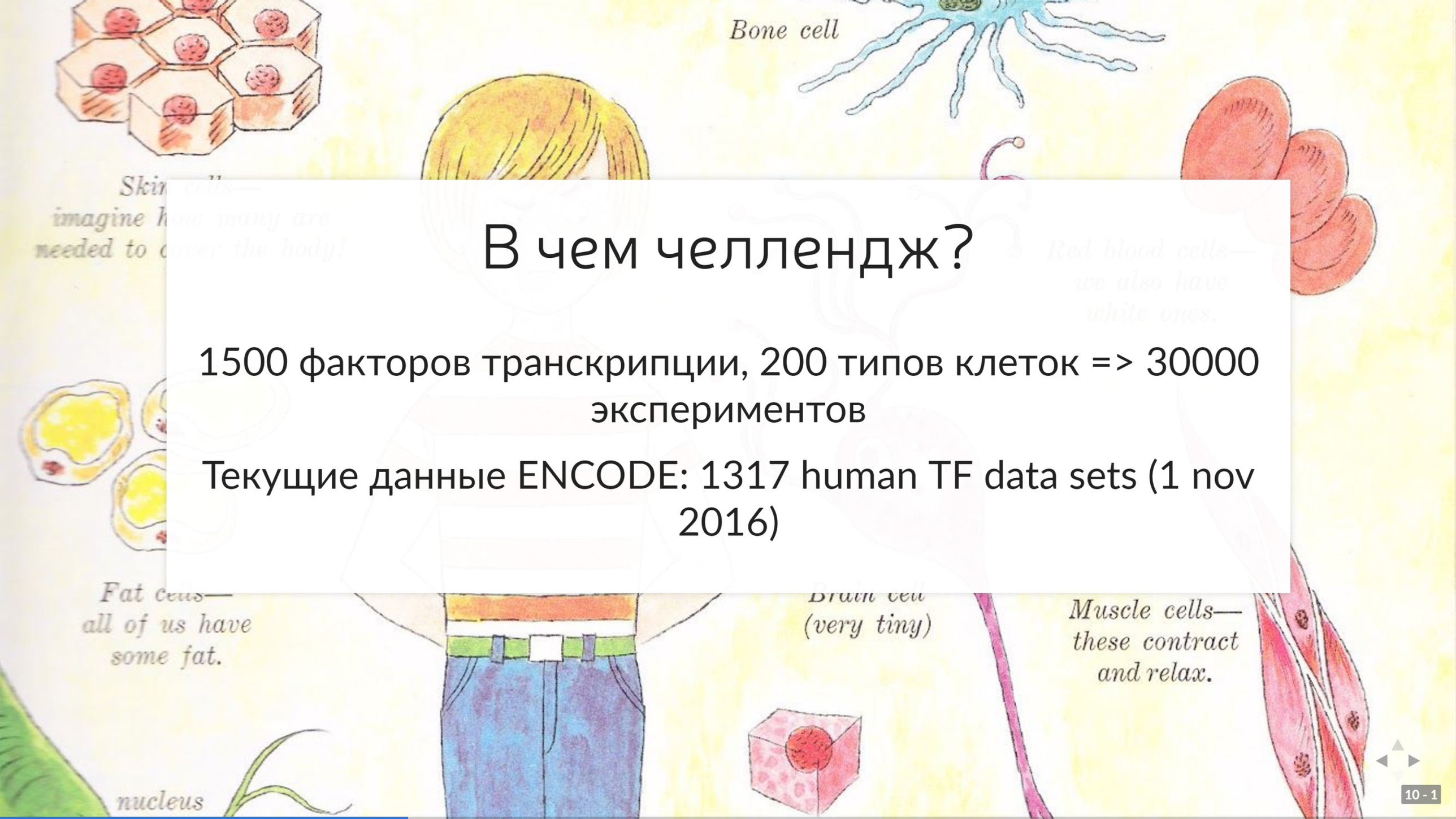
DREAM CHALLENGES powered by Sage Bionetworks · Sage BIONETWORKS · Stanford University · OHSU · IBM Research · HelmholtzZentrum münchen Deutsches Forschungszentrum für Gesundheit und Umwelt · School of Medicine Department of Pharmacology University of Colorado Anschutz Medical Campus




TSS
TRANSCRIPTION FACTORS (TFs)
RNA POLYMERASE
mRNA
GENE
ENHANCER
ChIP-Seq signal
DNA
TRANSCRIPTION FACTOR


genomic DNA
from multiple cells
5'
3'
3'
5'
fragmented DNA with
cross-linked protein of interest
5'
3'
3'
5'
immunoprecipitated
protein-DNA complexes
5'
3'
3'
5'
reverse
crosslinks,
digest DNA


5'
3'
3'
5'
high-throughput sequencing
to determine ends
of immunoprecipitated
DNA fragments
map reads to the
reference genome;
extend reads upto putative length
of DNA fragments
construct DNA fragments pileup
(base coverage profile)
detect transcription factor
binding regions



В чем челлендж?

1500 факторов транскрипции, 200 типов клеток => 30000 экспериментов

Текущие данные ENCODE: 1317 human TF data sets (1 nov 2016)

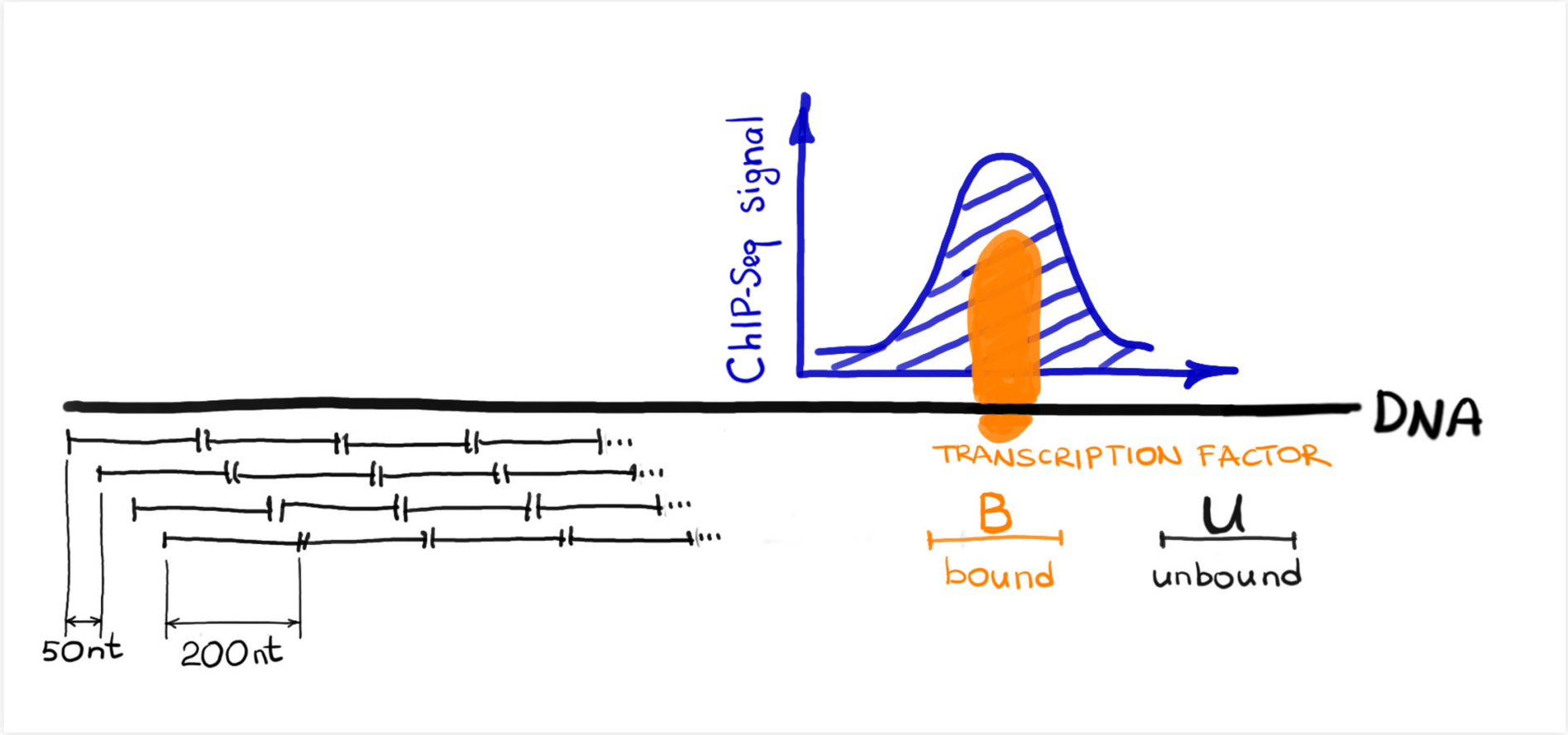
Challenge: предсказание полногеномного связывания факторов транскрипции in vivo по даным:

- ДНКазной доступности (DNase-Seq)
- Паттернов в геномной последовательности (sequence motifs)
- Экспрессии генов (RNA-Seq)
- ~~Локальных особенностей формы спирали ДНК (DNA shape)~~

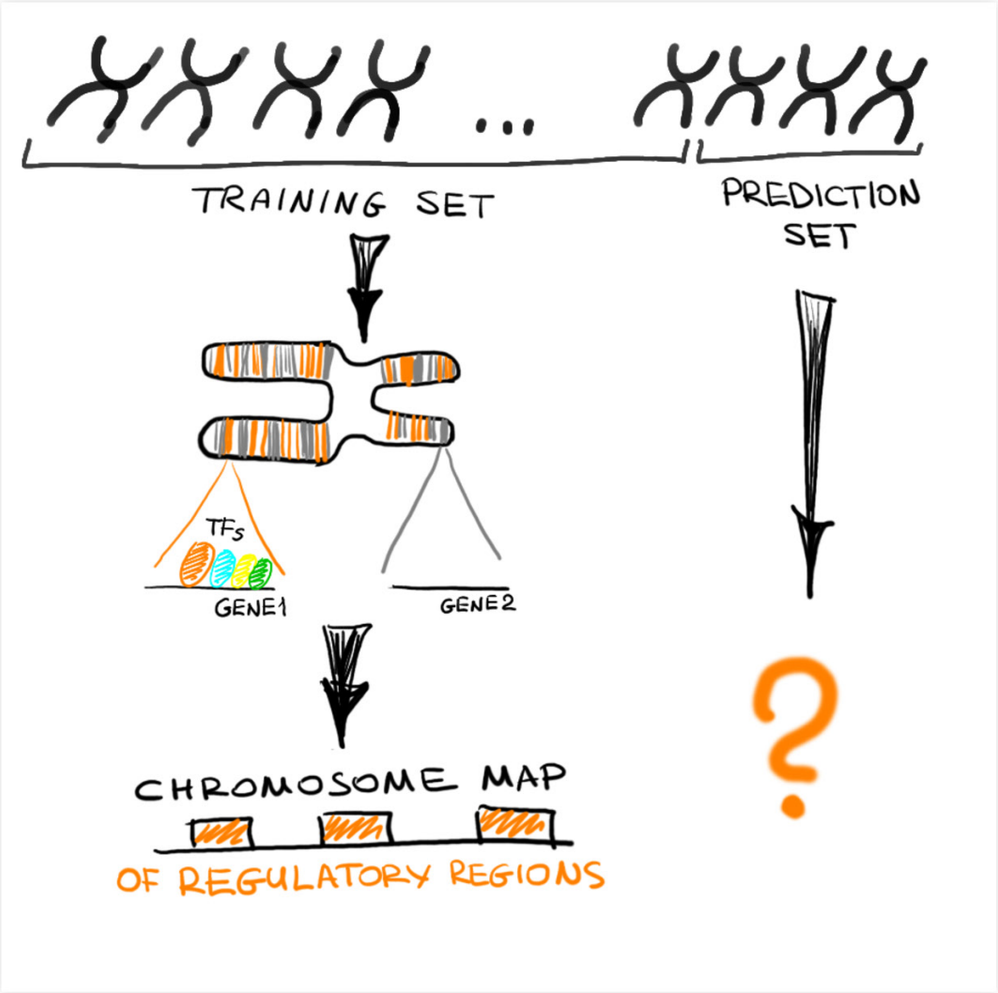
32 фактора транскрипции, 9 типов клеток

	ARID3A	ATF2	ATF3	ATF7	CEBPB	CREB1	CTCF	E2F1	E2F6	EGR1	EP300	FOXA1	FOXA2	GABPA	GATA3	HNF4A	JUND	MAFK	MAX	MYC	NANOG	REST	RFX5	SPI1	SRF	STAT3	TAF1	TCF12	TCF7L2	TEAD4	YY1	ZNF143
H1-hESC		T	T		T	T	T		T	T	T			T				T	T		T	T			T		T	T		T	T	T
GM12878		T		T		T	L	T		T	T			T				T	T				T	T	T	L	T	T			T	T
K562	L	L	T	T	T	T	T	F	L	L	T			L			T	L	T	T		L		L	T		T	L		T	L	L
IMR-90					T		T											T															
A549					T		T		T						T				T	T										T			
HCT116			T		T					T							T		T						T			T	T	T	T	
HeLa-S3					T		T	T	T		T			T			T	T	T	T		T	T			T	T		T			T
HepG2	T	F	T	T	T	T	T				T	T	T	T		T	T	T	T	L		T	L		T		L			T	T	T
MCF-7		T		L	L	L	T			T	L	L		T	L		T	L	L	T		T	T		L			T	L	L		
Panc1																						T							T				
SK-N-SH											T			T	T		T		T			T	T				T	T		T	T		
liver			L							F		F	F	F		F	F		F			F					F						
PC-3							F																										
iPS							F													F													

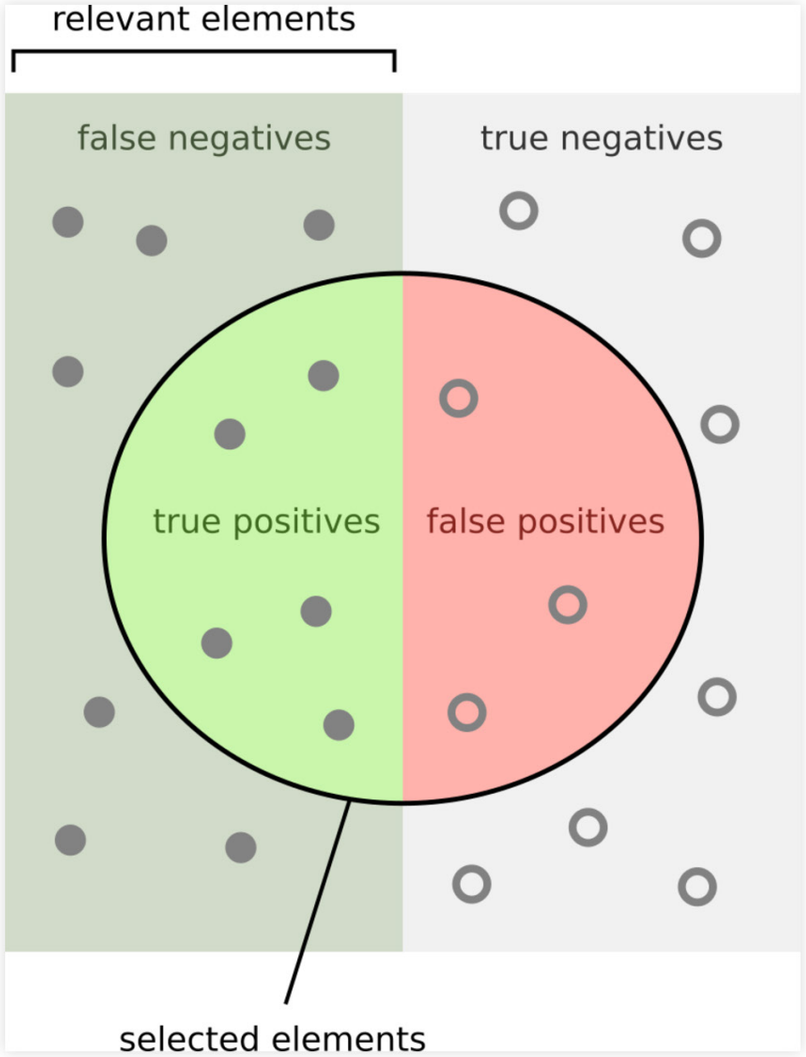
Задача бинарной классификации



Held-out chromosomes: "не учите координаты"



О метриках качества



О метриках качества

How many selected items are relevant?

Precision = 

How many relevant items are selected?

Recall =   = True Positive Rate

How many non-relevant items are selected?

False Positive Rate = 

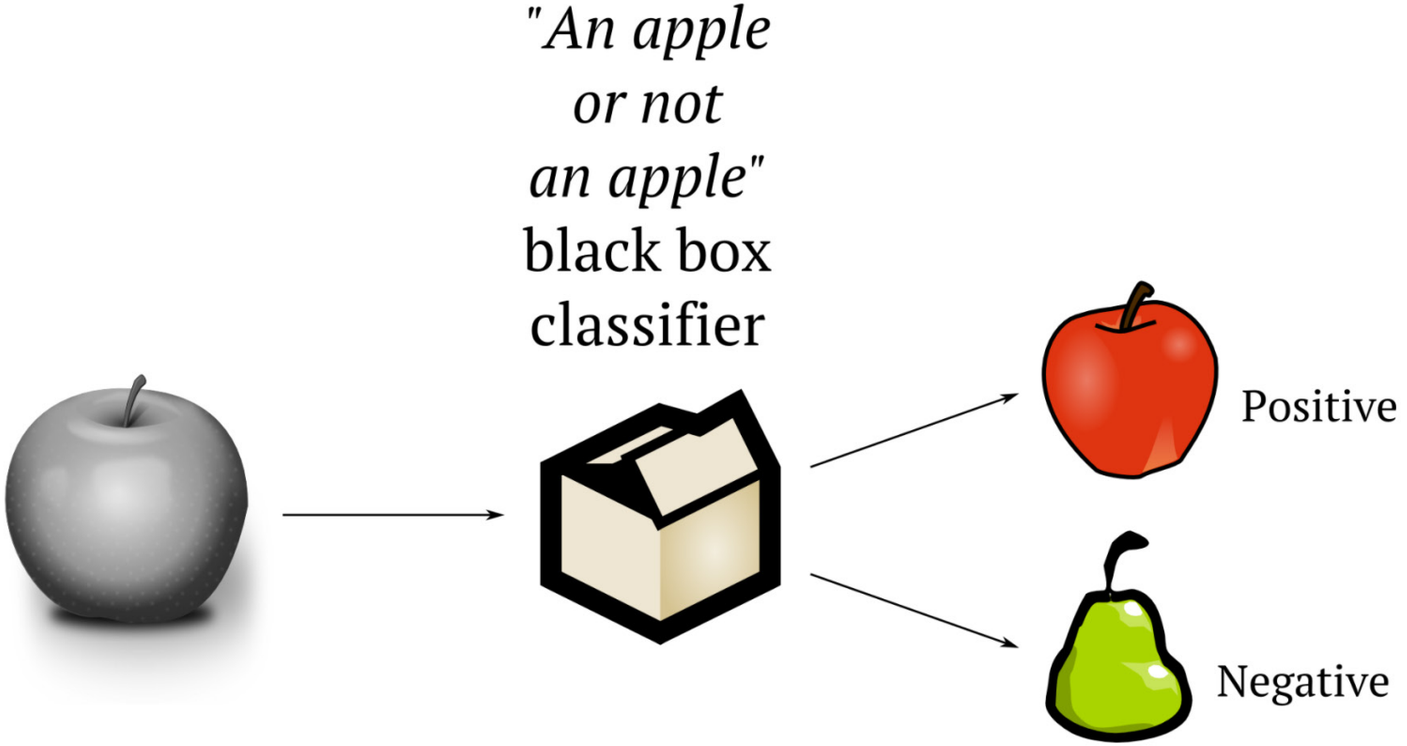
Classifier

Precision and Recall : PR curve
TRP and FPR : Receiver Operating Characteristic (ROC)

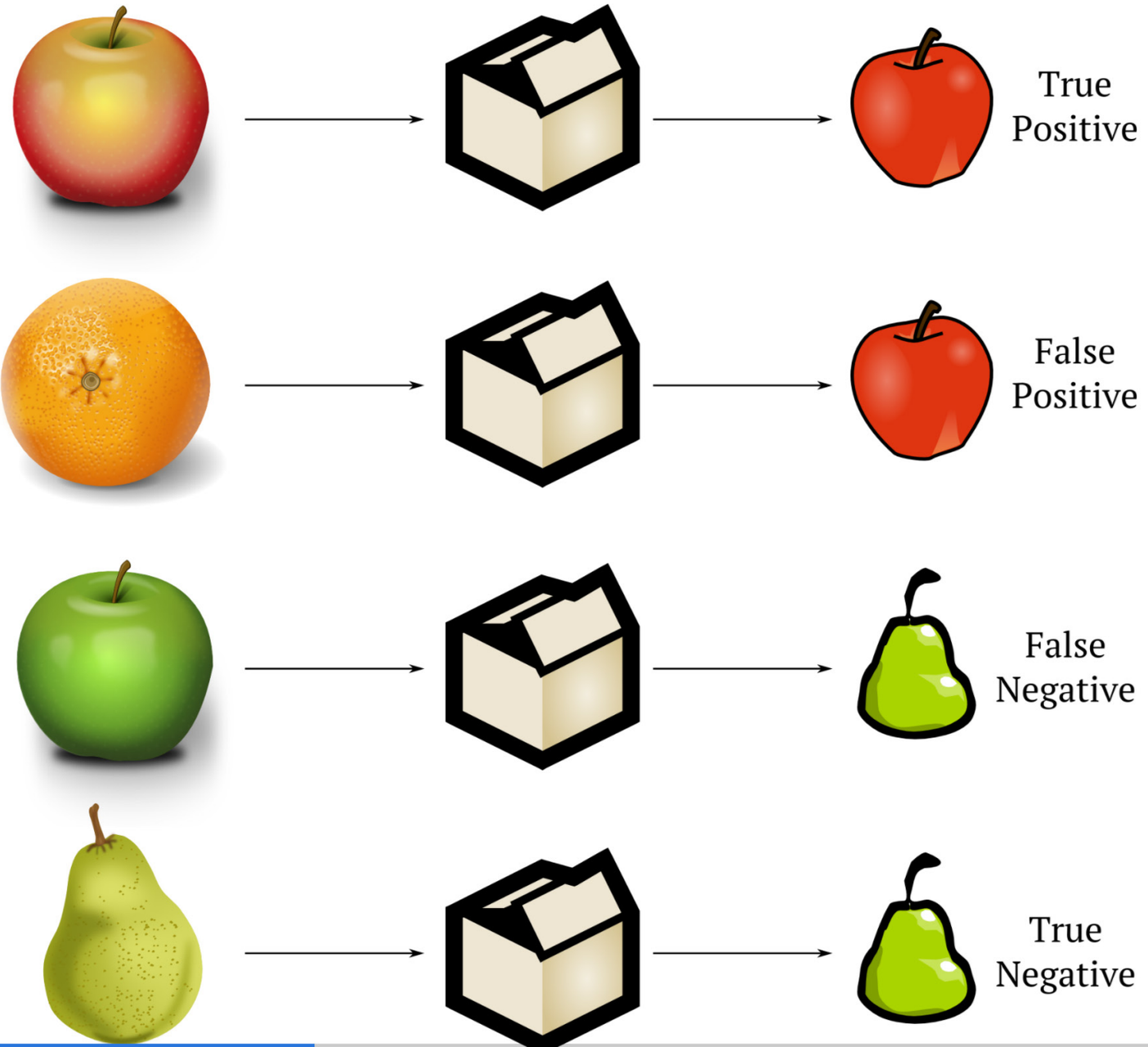
TPR / FPR = LR+ Positive likelihood ratio

adapted from Wikipedia (original image by Walber)

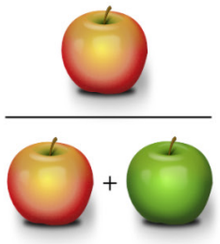
Черный ящик


"An apple
or not
an apple"
black box
classifier
Positive
Negative

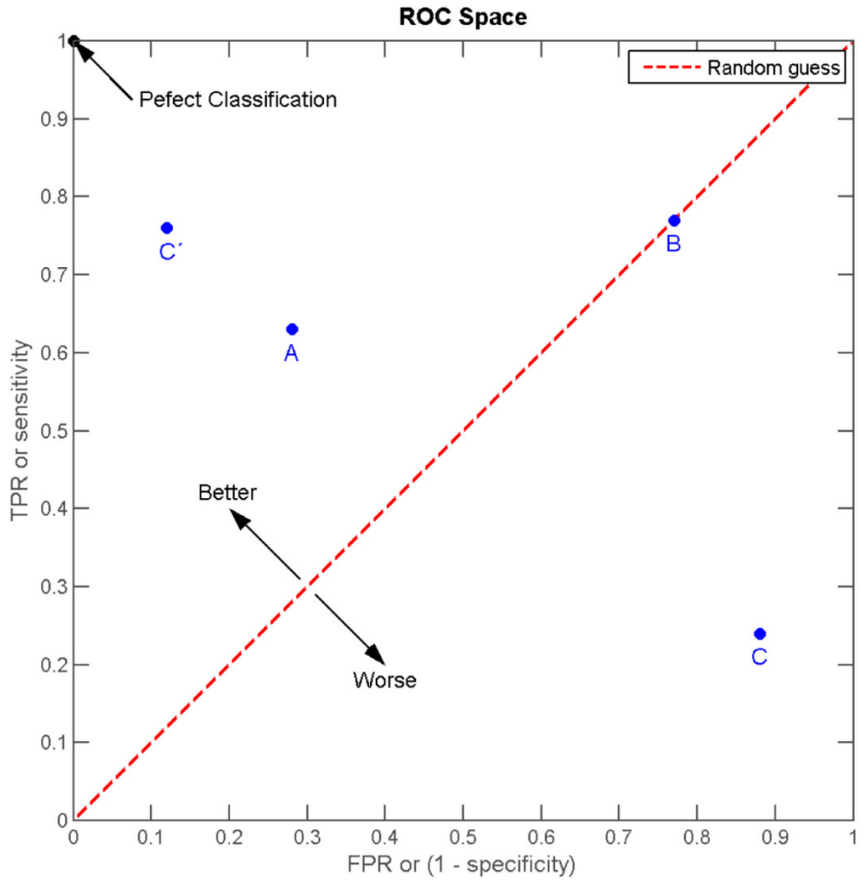
Черный ящик


True
Positive
False
Positive
False
Negative
True
Negative

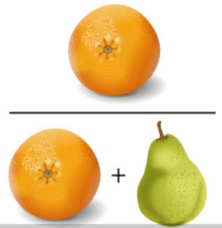
Черный ящик

TPR = 

True Positive Rate =
True Positive / (True Positive + False Negative)



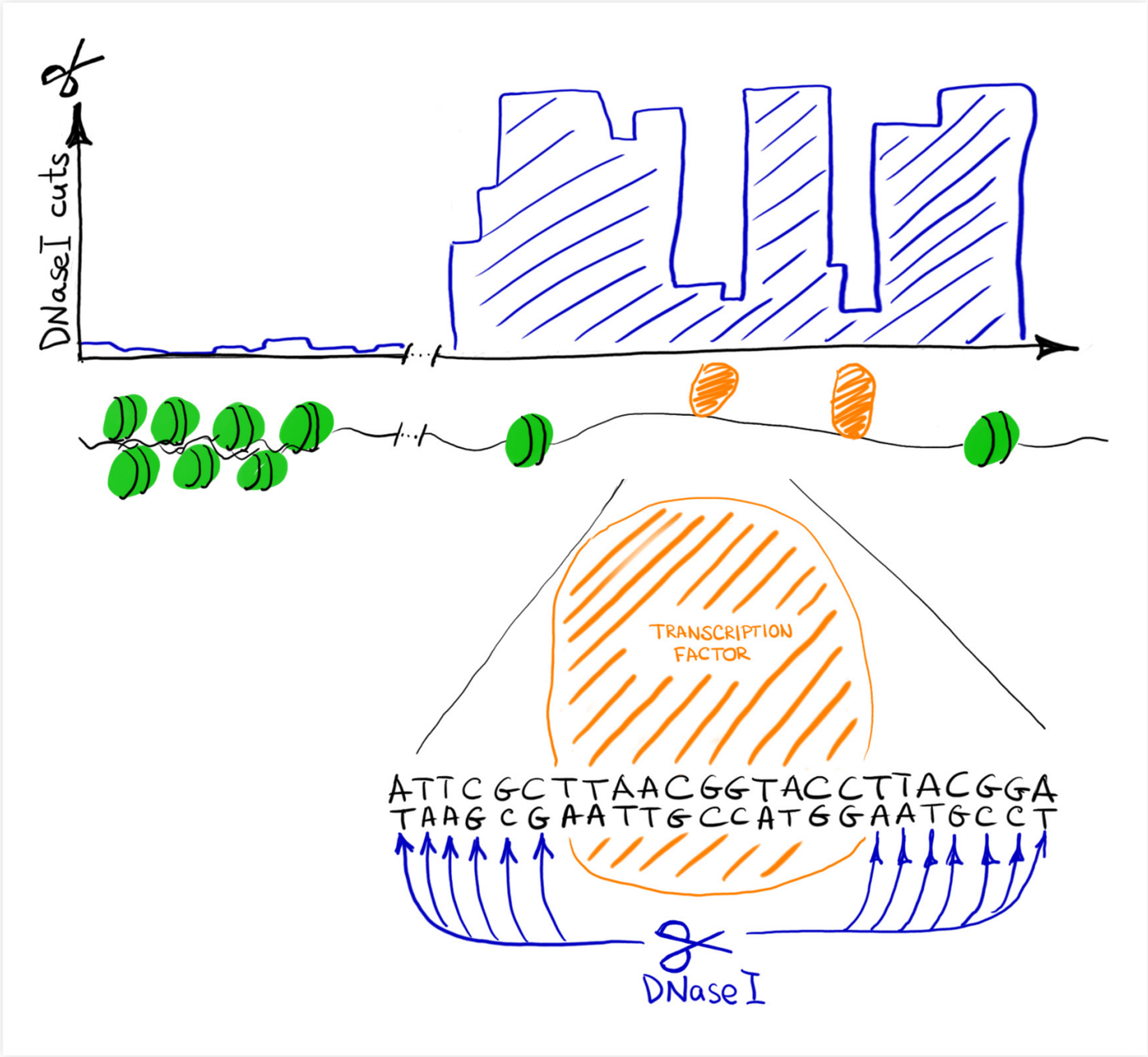
False Positive Rate =
False Positive / (False Positive + True Negative)

FPR = 

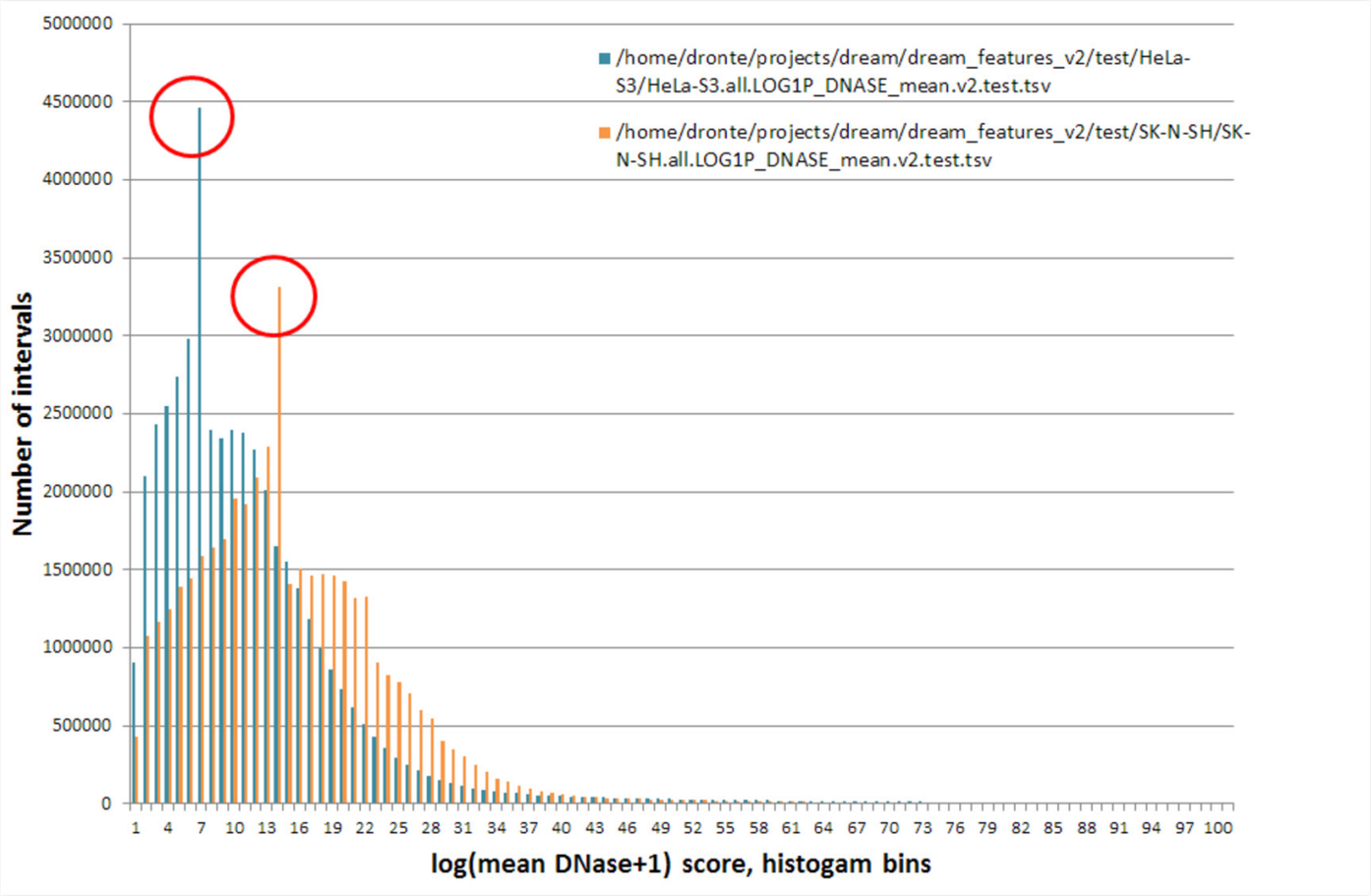
Превращаем данные в признаки ("фичи")

	Feature 1	Feature 2	. . .	Value
chr10:600-800				U
chr10:650-850				U
chr10:700-900				U
⋮	⋮	⋮		⋮
chr10:138700-138900				B
⋮	⋮	⋮		⋮

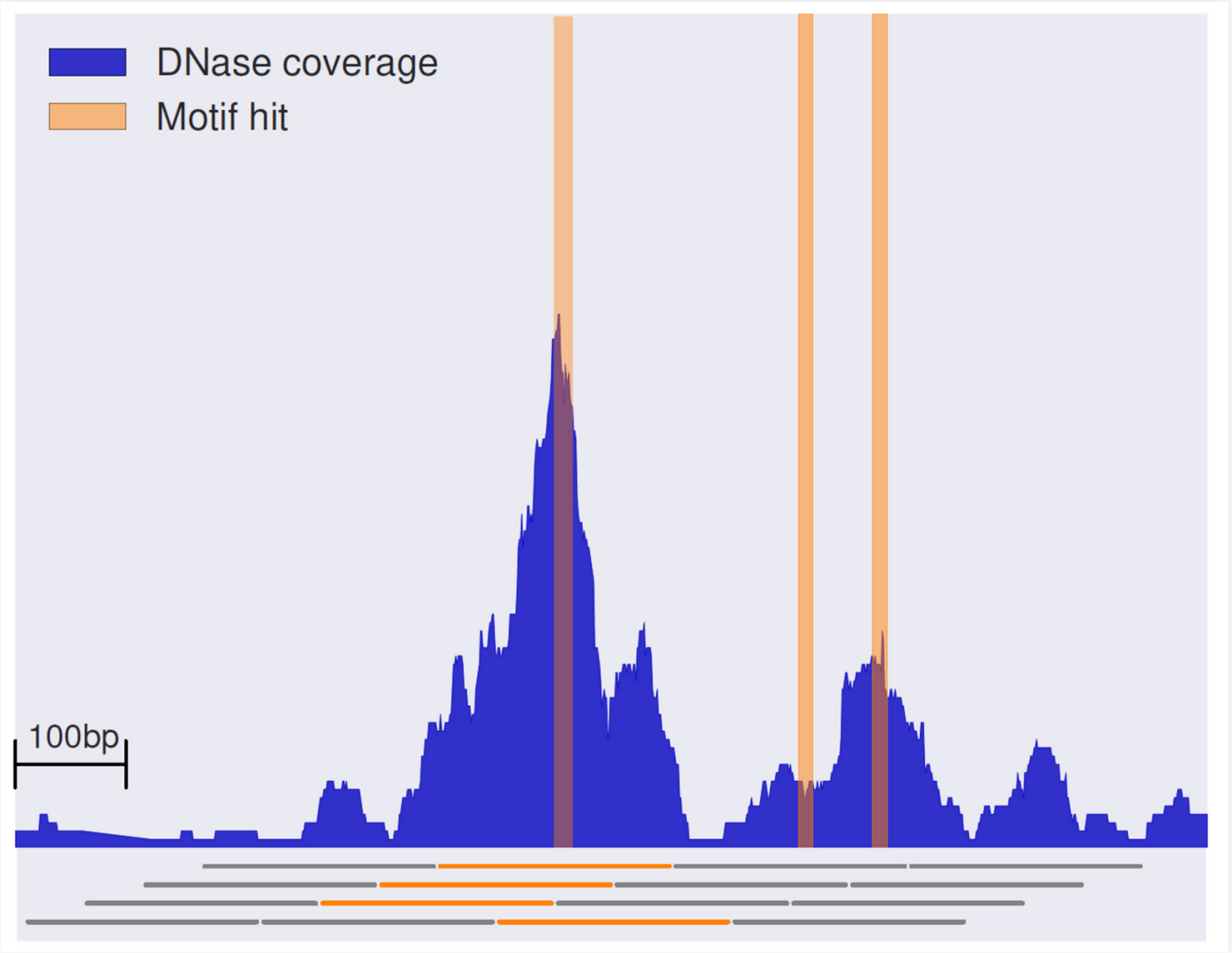
ДНКазная доступность как она есть



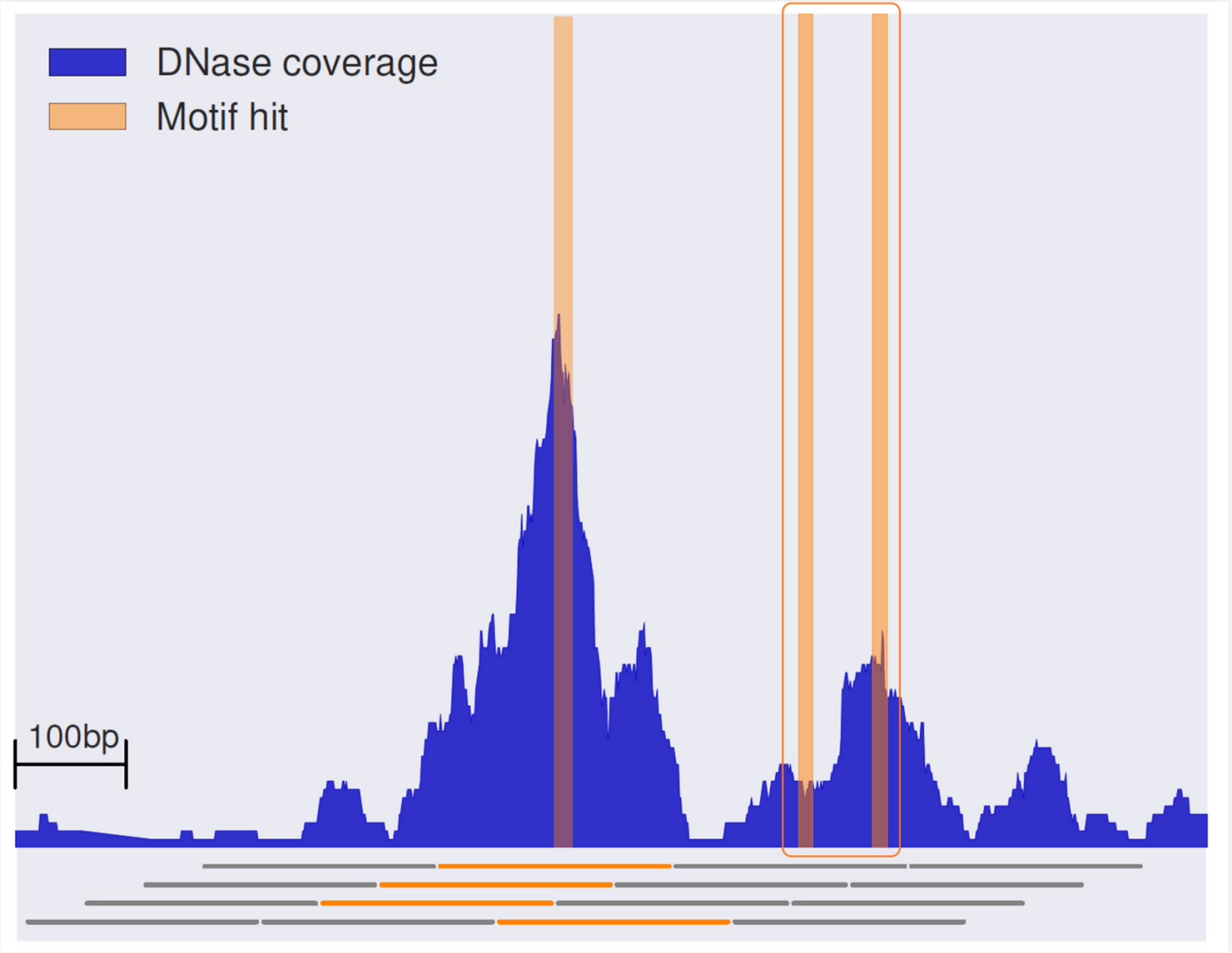
Проблема нормализации между клеточными типами



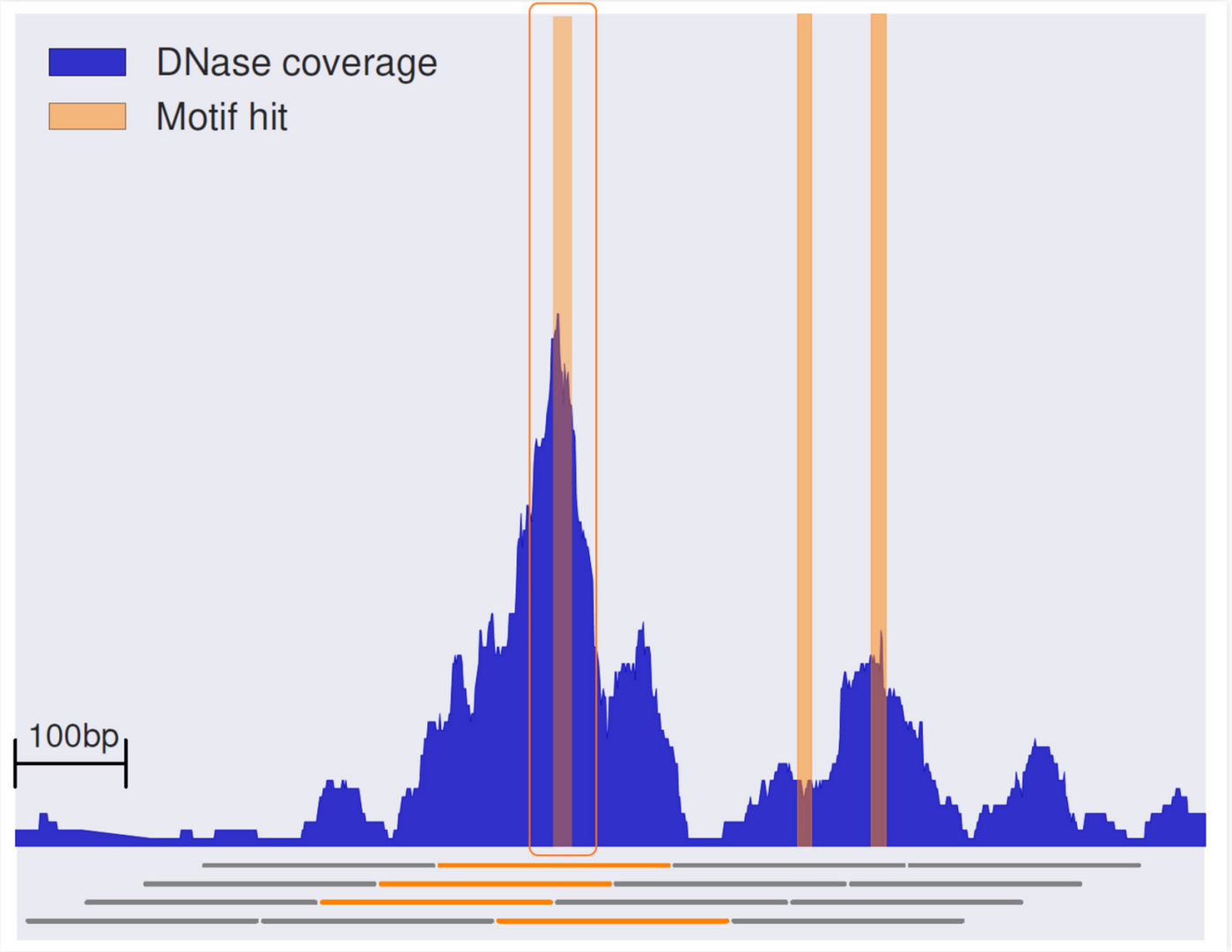
От свойств в фичам



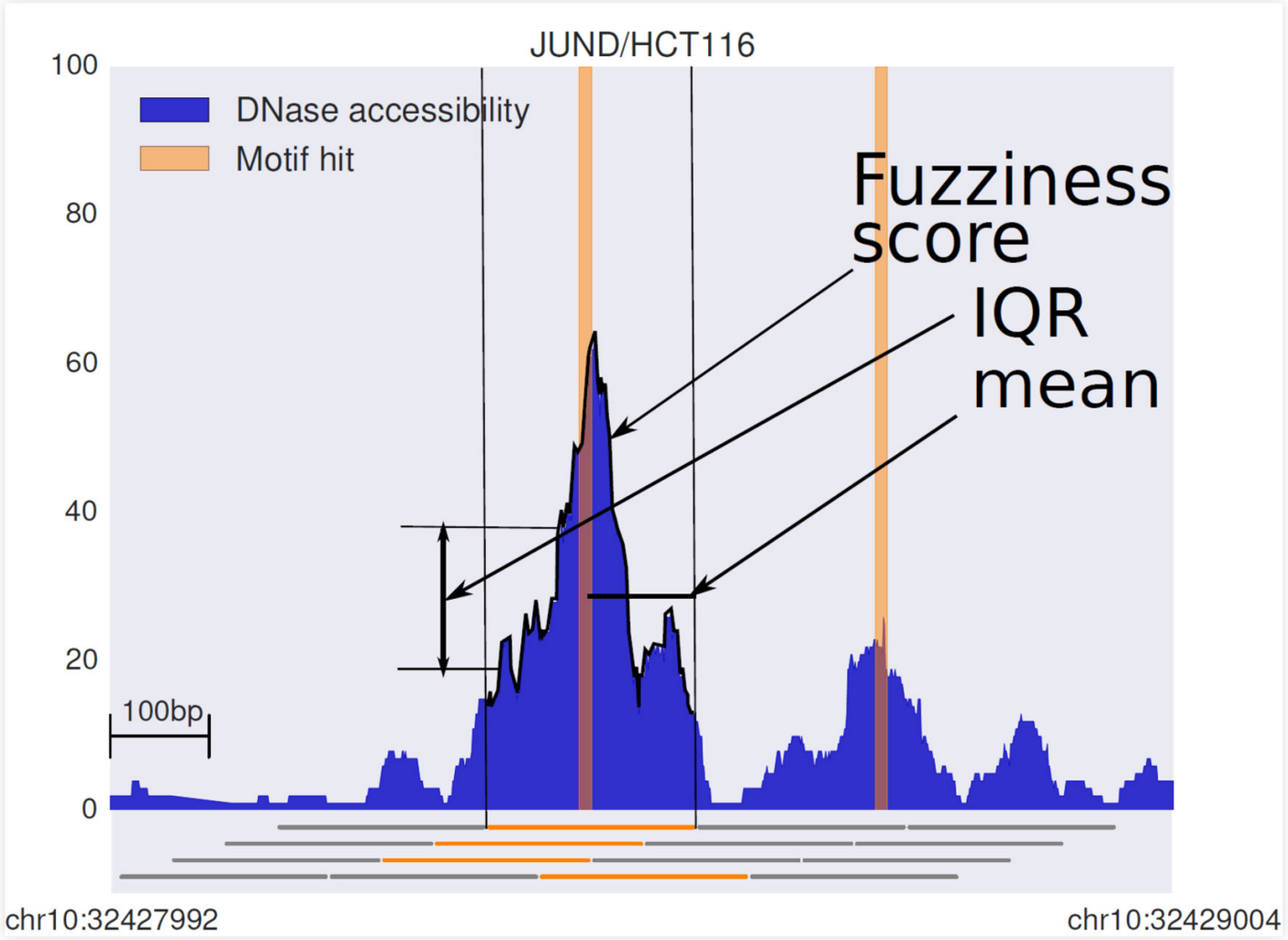
Число вхождений мотива



Наилучшее вхождение мотива

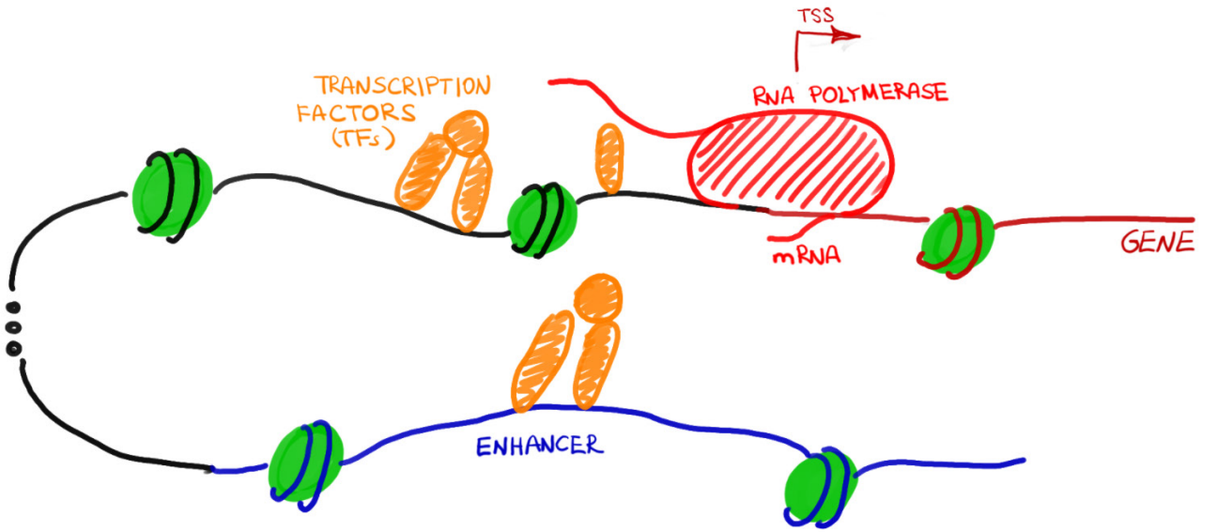


Форма профиля ДНКазной доступности



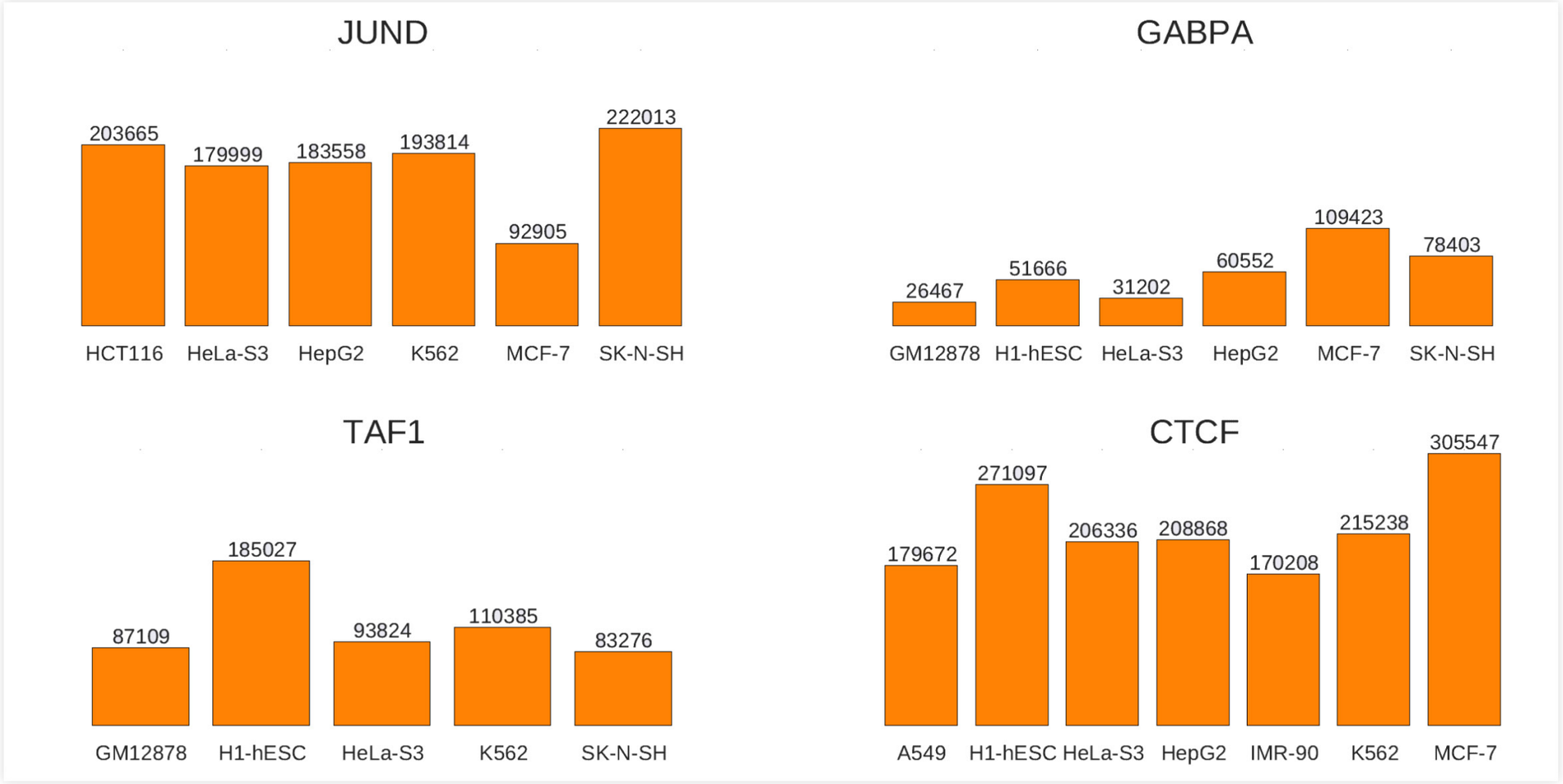
Дополнительные свойства

- Расстояние до ближайшего гена
- Экспрессия ближайшего гена
- Вхождения мотивов потенциальных кофакторов
- Вторичные мотивы связывания

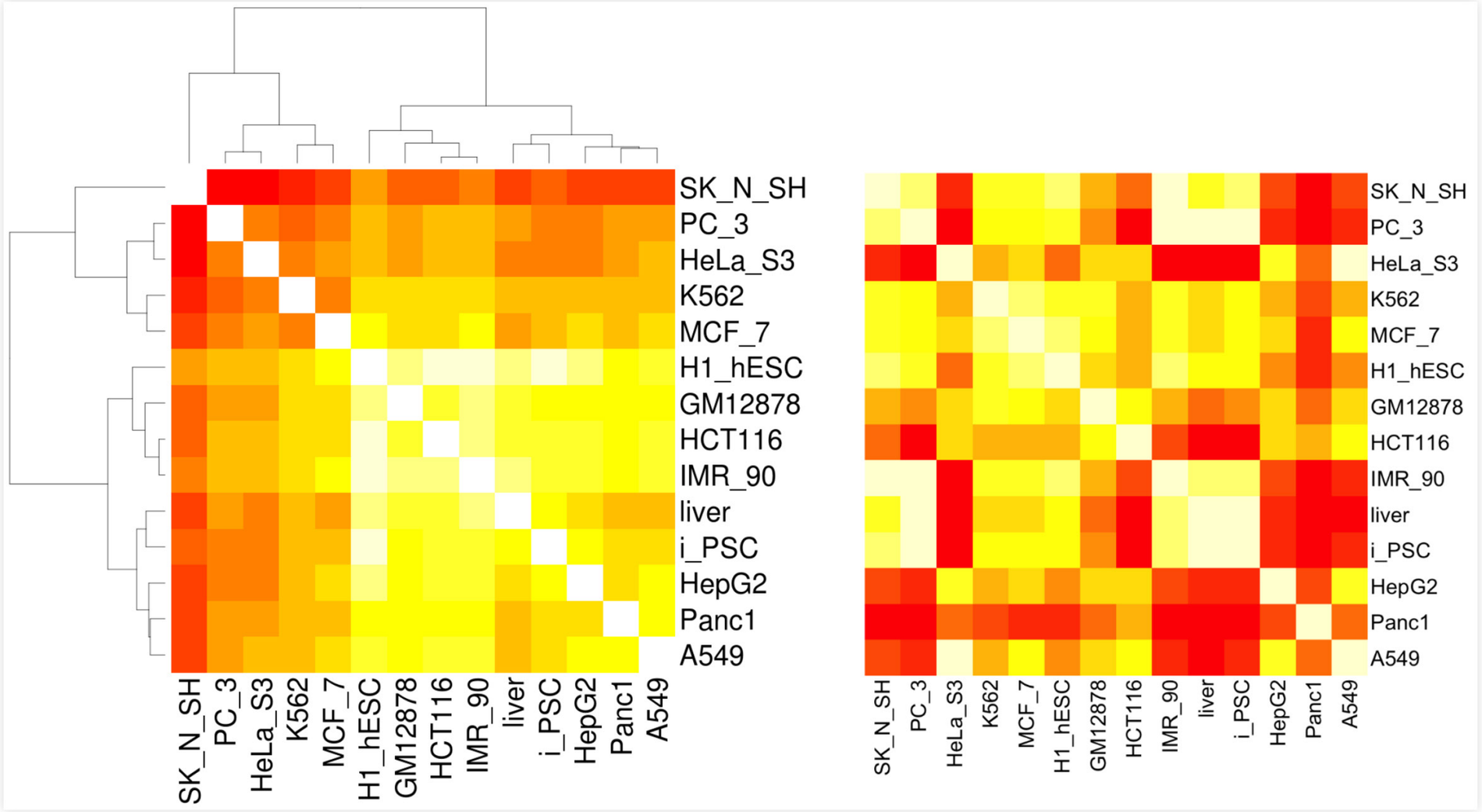


Насколько похожи разные типы клеток?

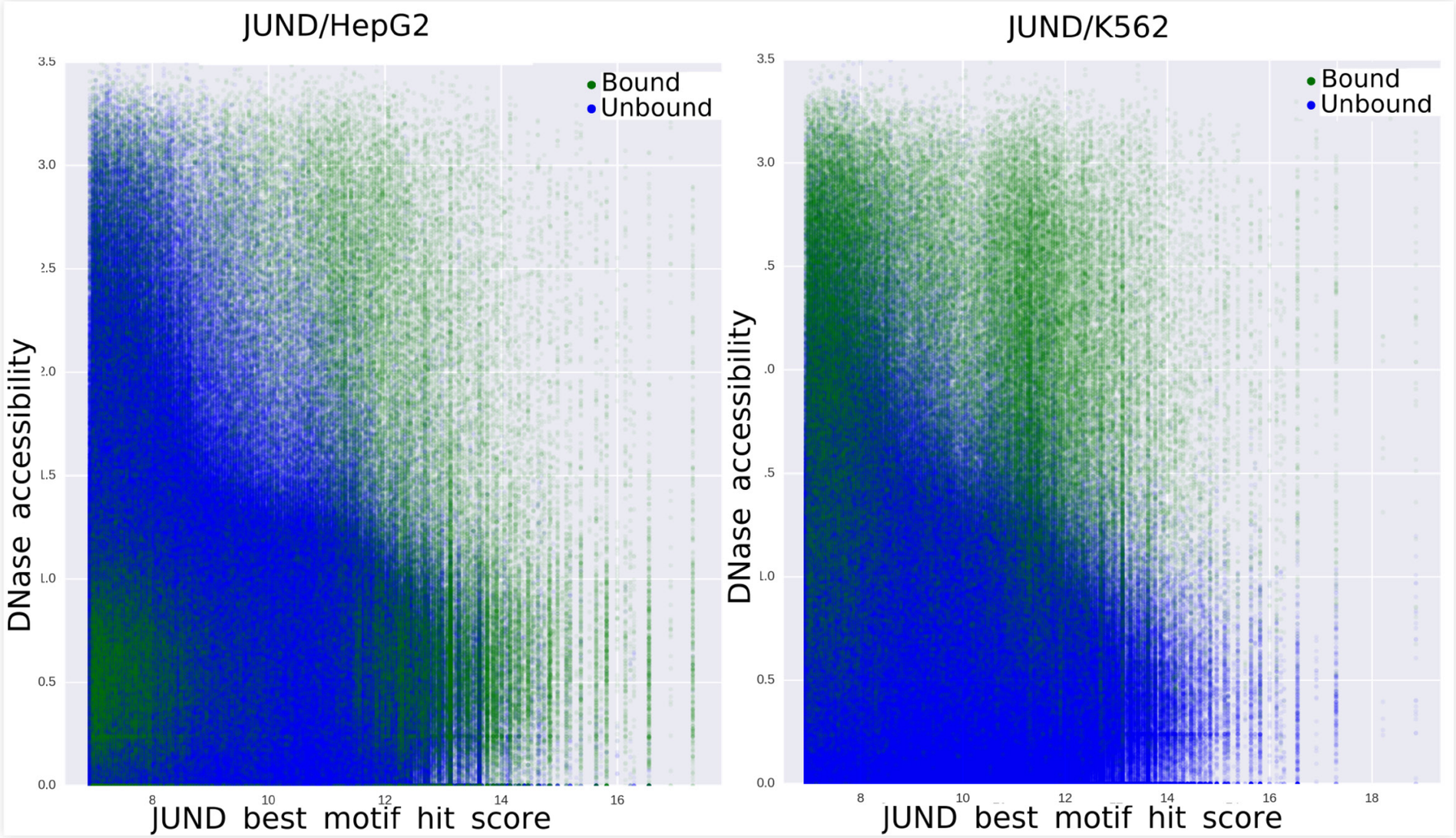
Абсолютное число B относительно



Сходство профилей ДНКазной доступности не согласуется со сходством экспрессии генов



Ткань-специфичные сайты не всегда находятся в открытом хроматине



Можно ли придумать стабильную метрику похожести типов клеток?

Отношение правдоподобия для положительных предсказаний (positive-likelyhood ratio, LR+)

How many selected items are relevant?

Precision = 

How many relevant items are selected?

Recall =  = True Positive Rate

How many non-relevant items are selected?

False Positive Rate = 

Classifier

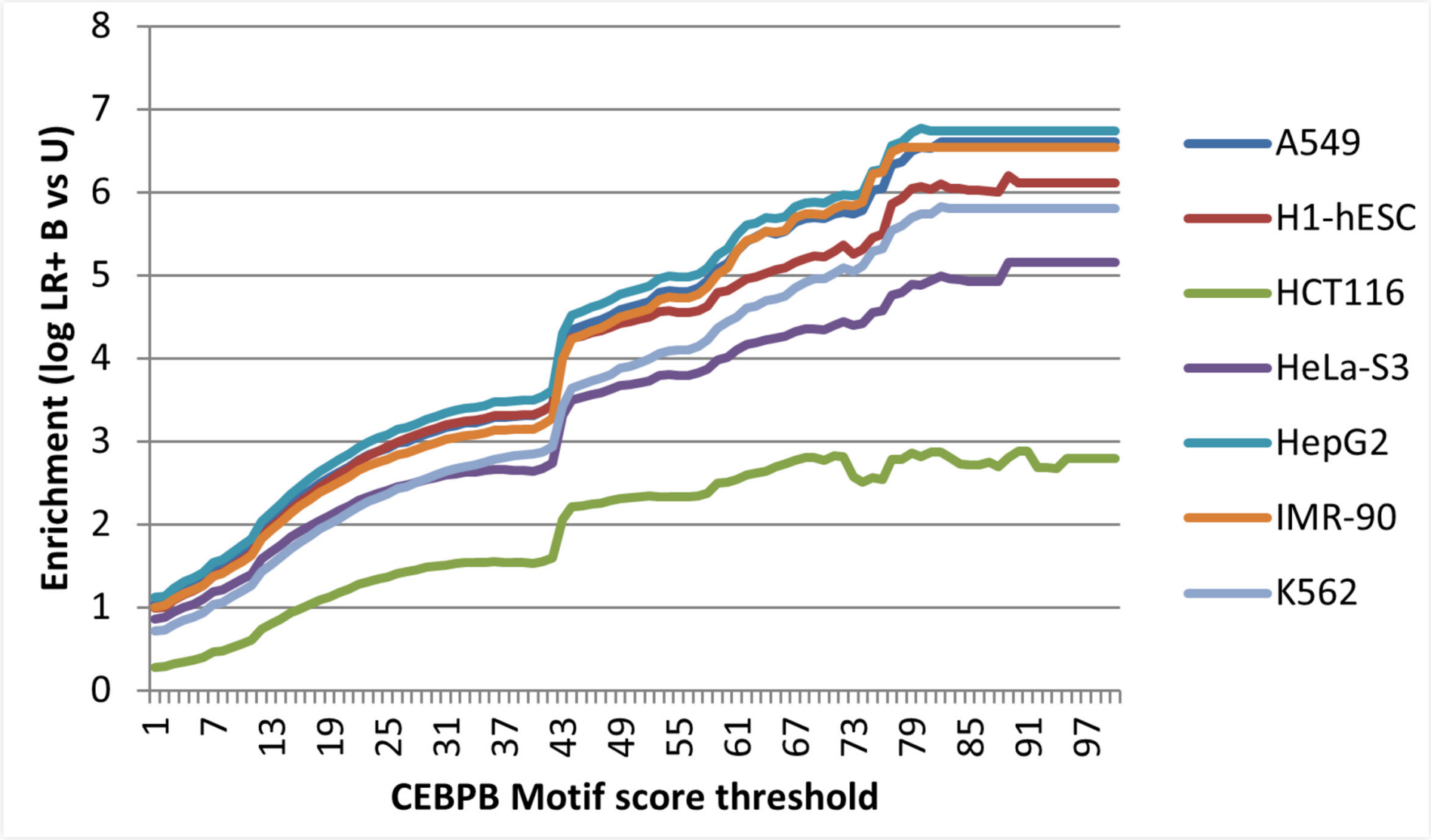
Precision and Recall : PR curve
TRP and FPR : Receiver Operating Characteristic (ROC)

TPR / FPR = LR+ Positive likelihood ratio

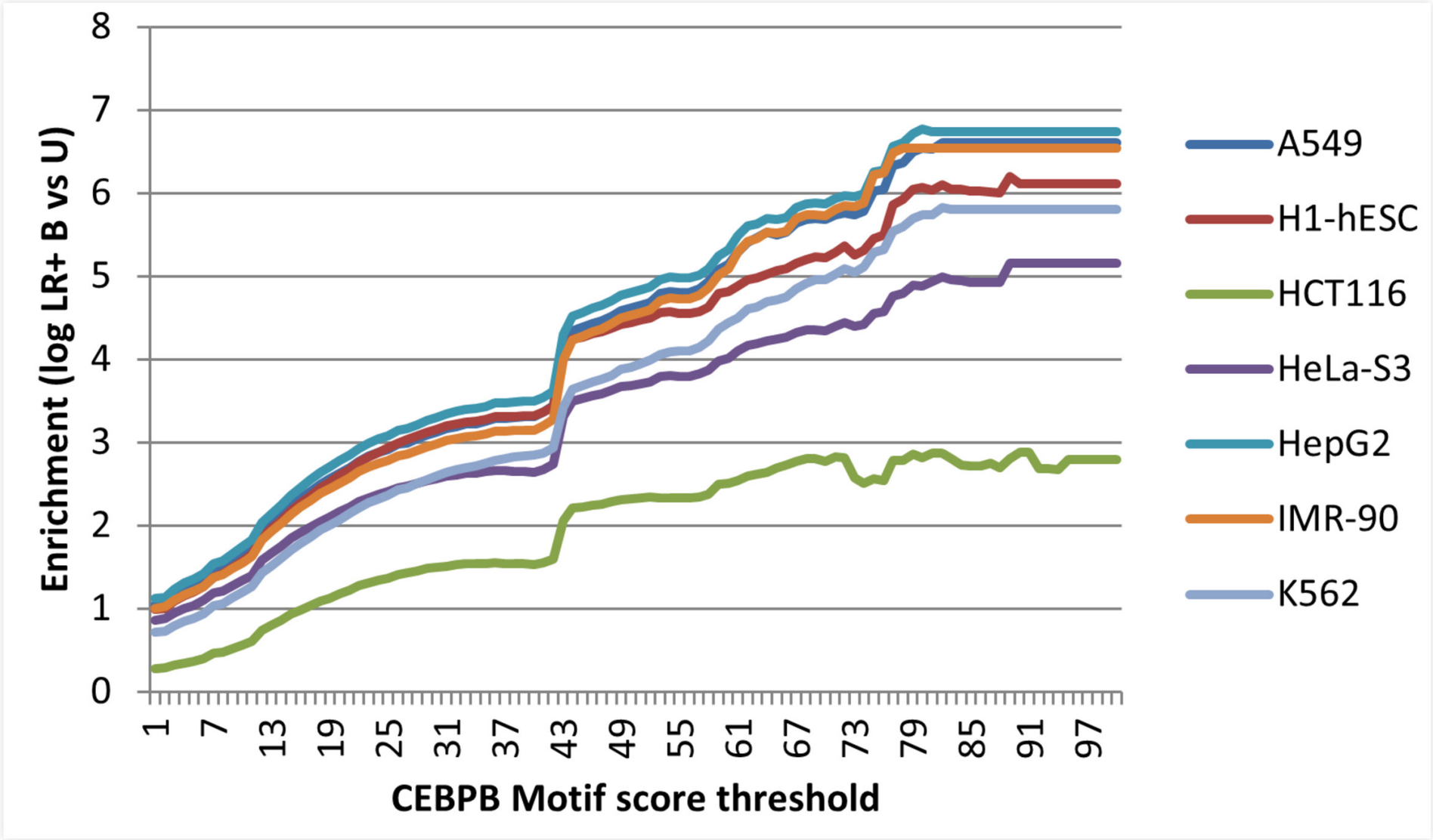
Отношение правдоподобия для положительных предсказаний (positive-likelyhood ratio, LR+)

$$E(feature \geq t) = log\left(\frac{\%B}{\%U}\right) \text{ where } \%B = \frac{\#B(feature \geq t)}{\#B(feature \geq 0)} \text{ and } \%U = \frac{\#U(feature \geq t)}{\#U(feature \geq 0)}$$

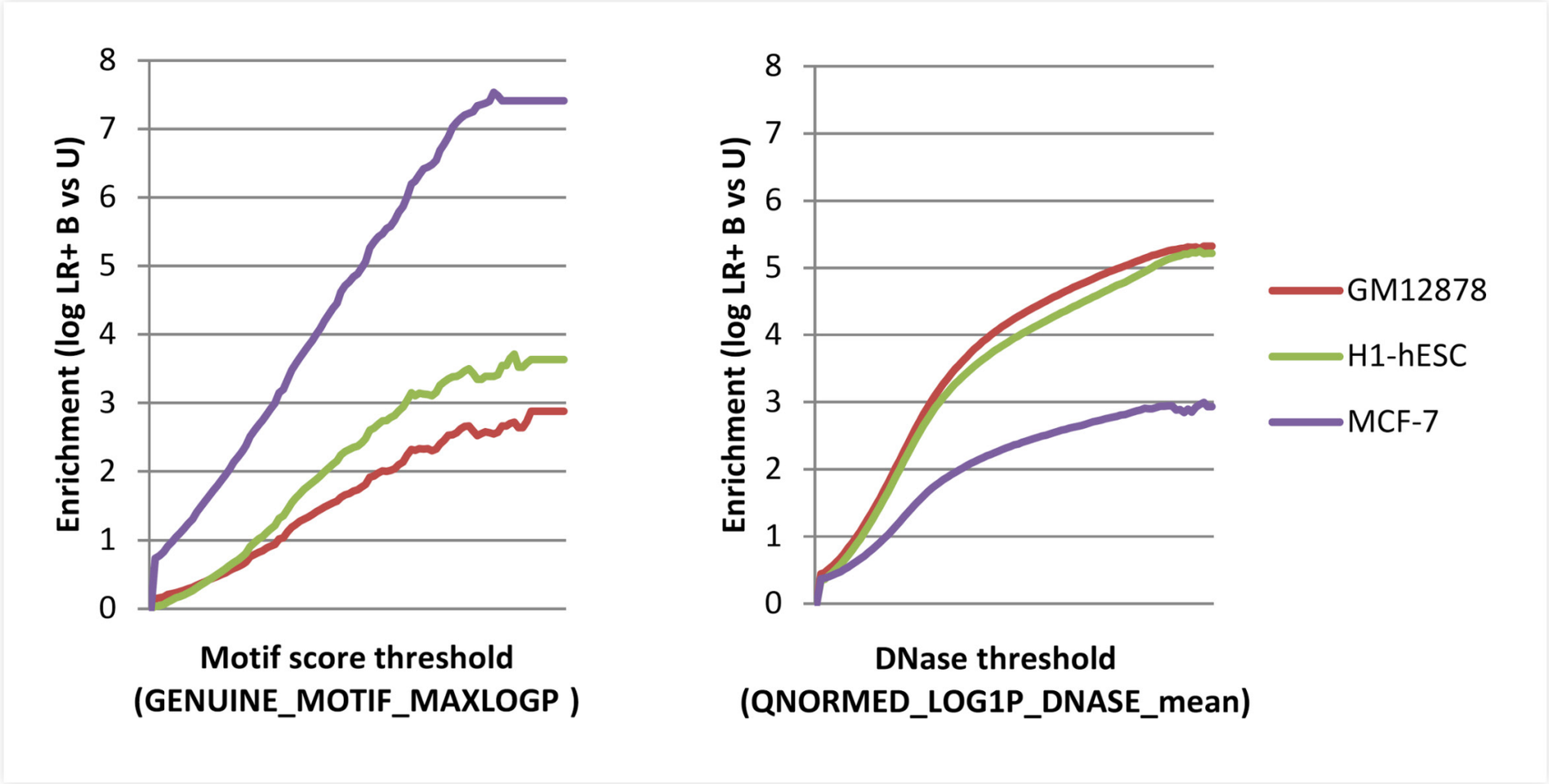
$$E(feature \geq t) = log\left(\frac{\%B}{\%U}\right) \text{ where } \%B = \frac{\#B(feature \geq t)}{\#B(feature \geq 0)} \text{ and } \%U = \frac{\#U(feature \geq t)}{\#U(feature \geq 0)}$$



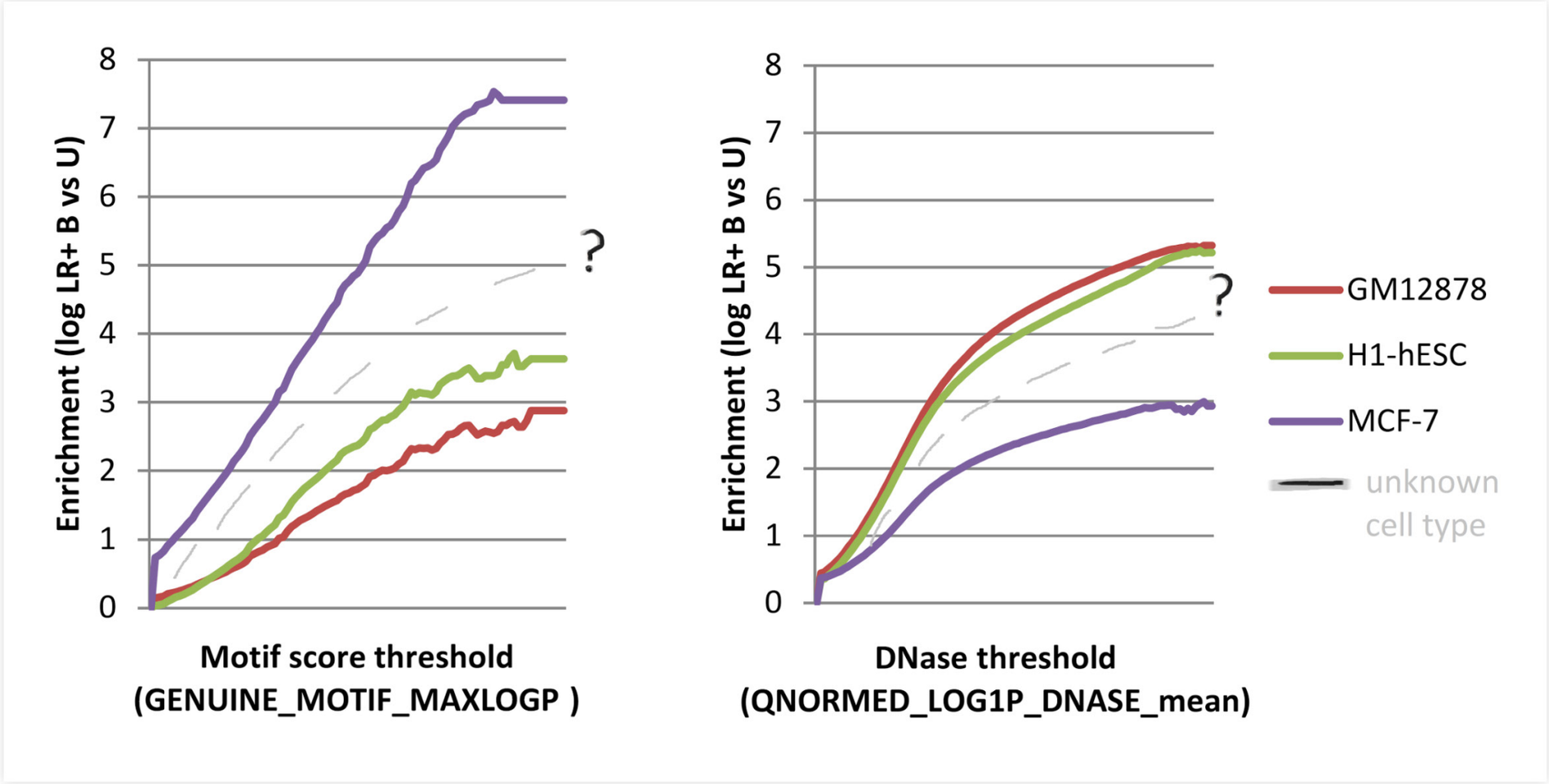
Форма E-кривой - инвариант между типами клеток



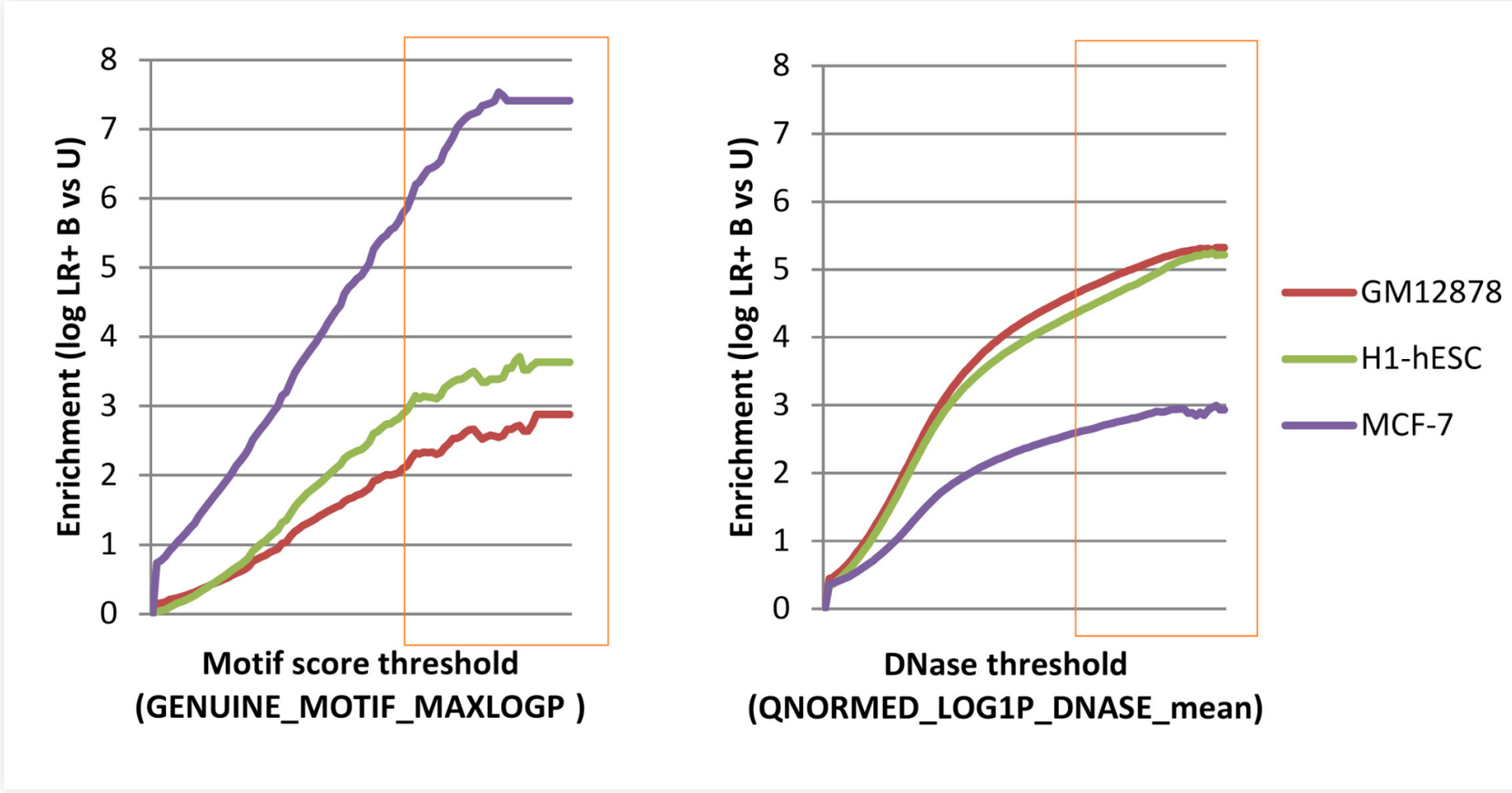
А масштабный коэффициент зависит от типа клеток



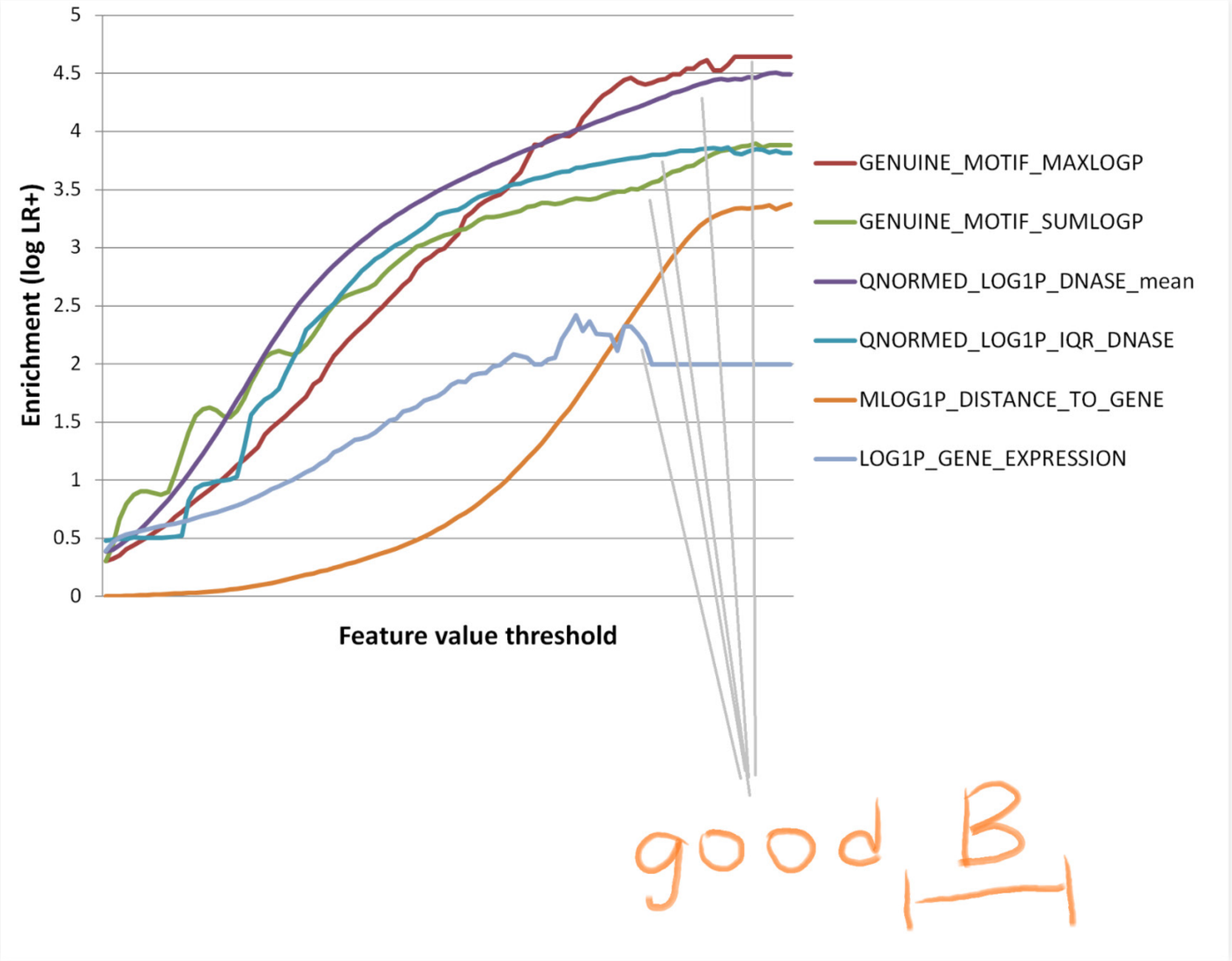
Определение E-кривой для неизвестной ткани



Достаточно оценить масштабный коэффициент в области больших значений

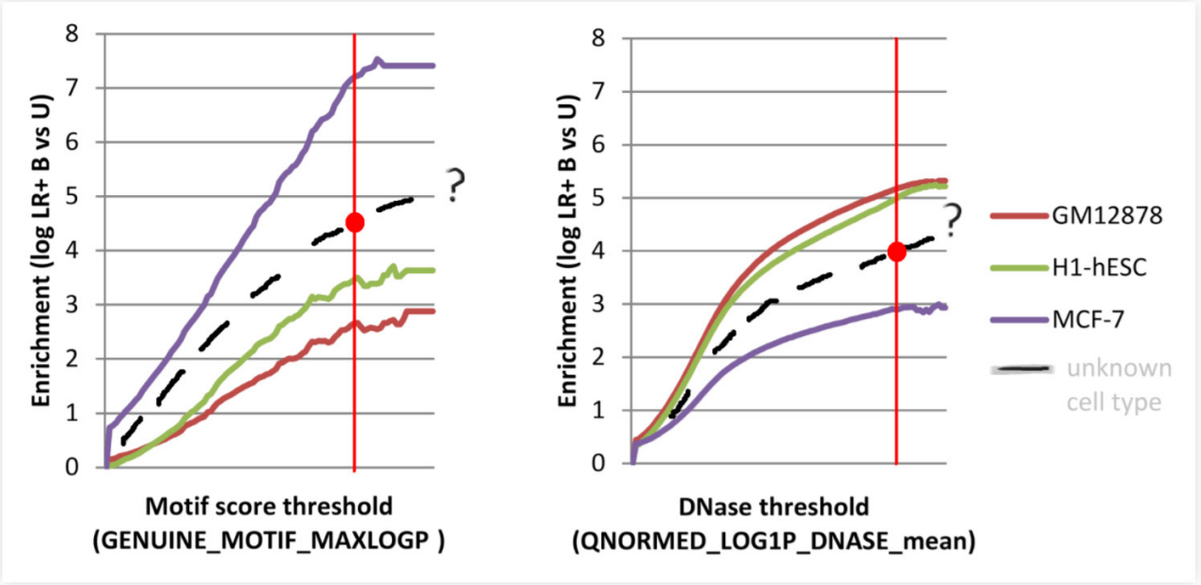


Большие значения свойств = хорошие B



Следовательно:

- Можно предсказать хорошие B в новом типе клеток
- Оценить для них кривую обогащения свойством
- Оценить масштаб обогащения для больших значений каждого свойства
- Узнать на что похож новый тип клеток



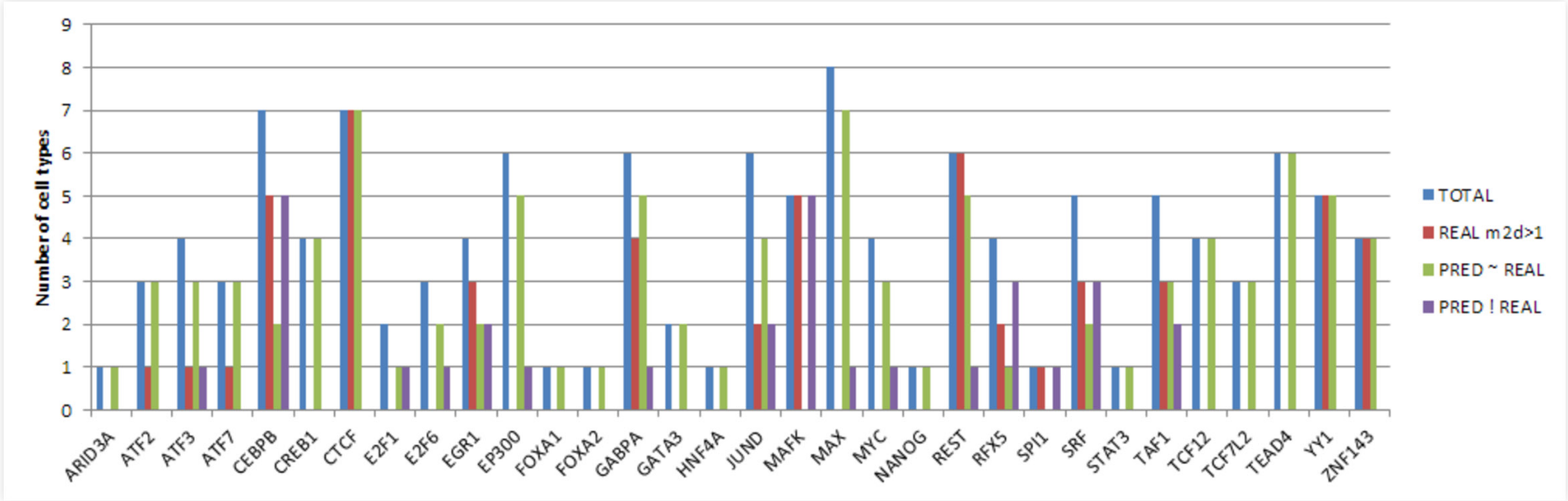
Получающиеся по предсказанным B абсолютные цифры

не соответствуют реальным (оцененным по реальным B).

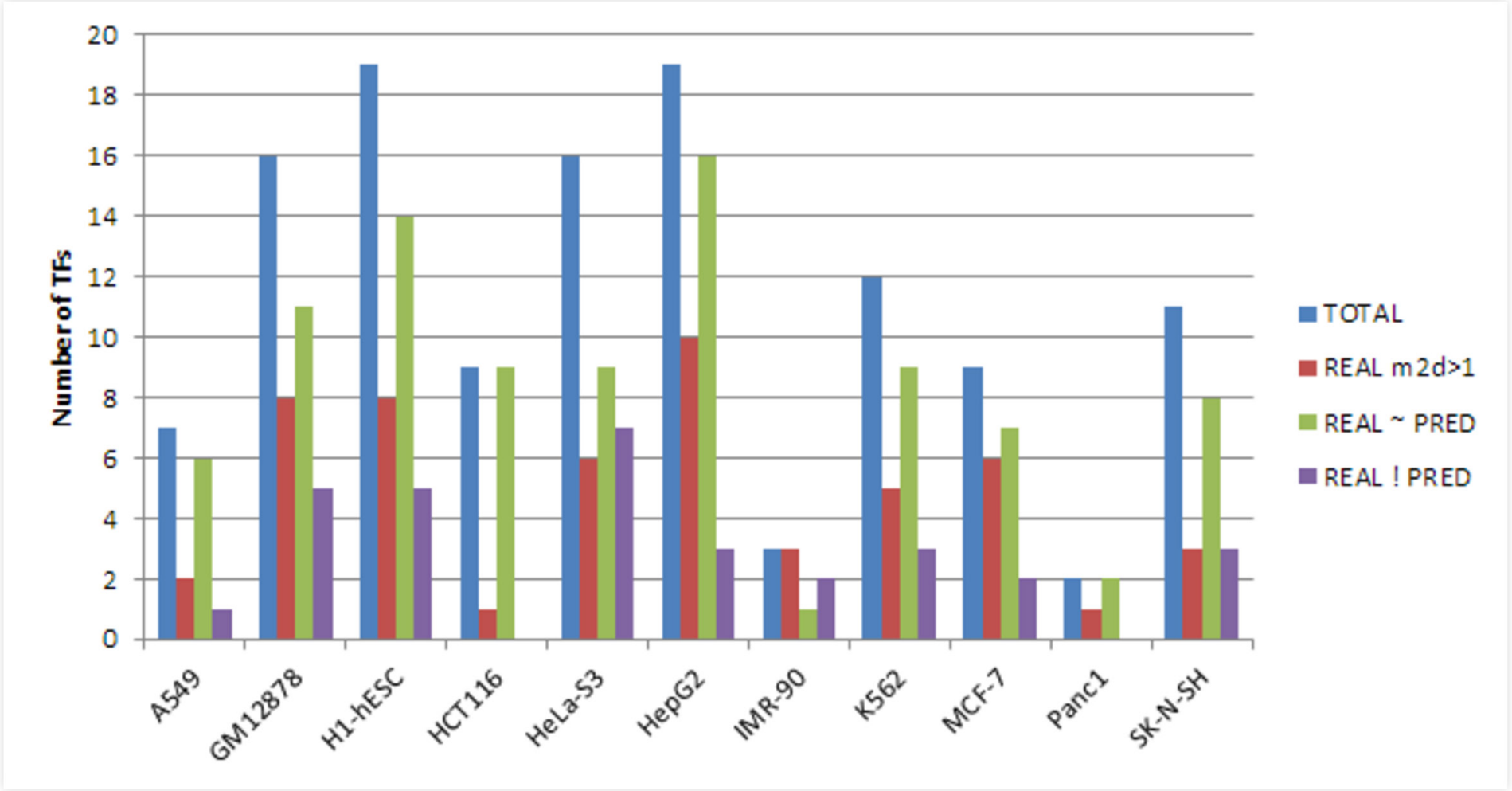
Решение: использовать относительные E для надежных свойств

$$m2d = E\text{-}motif / E\text{-}DNase$$

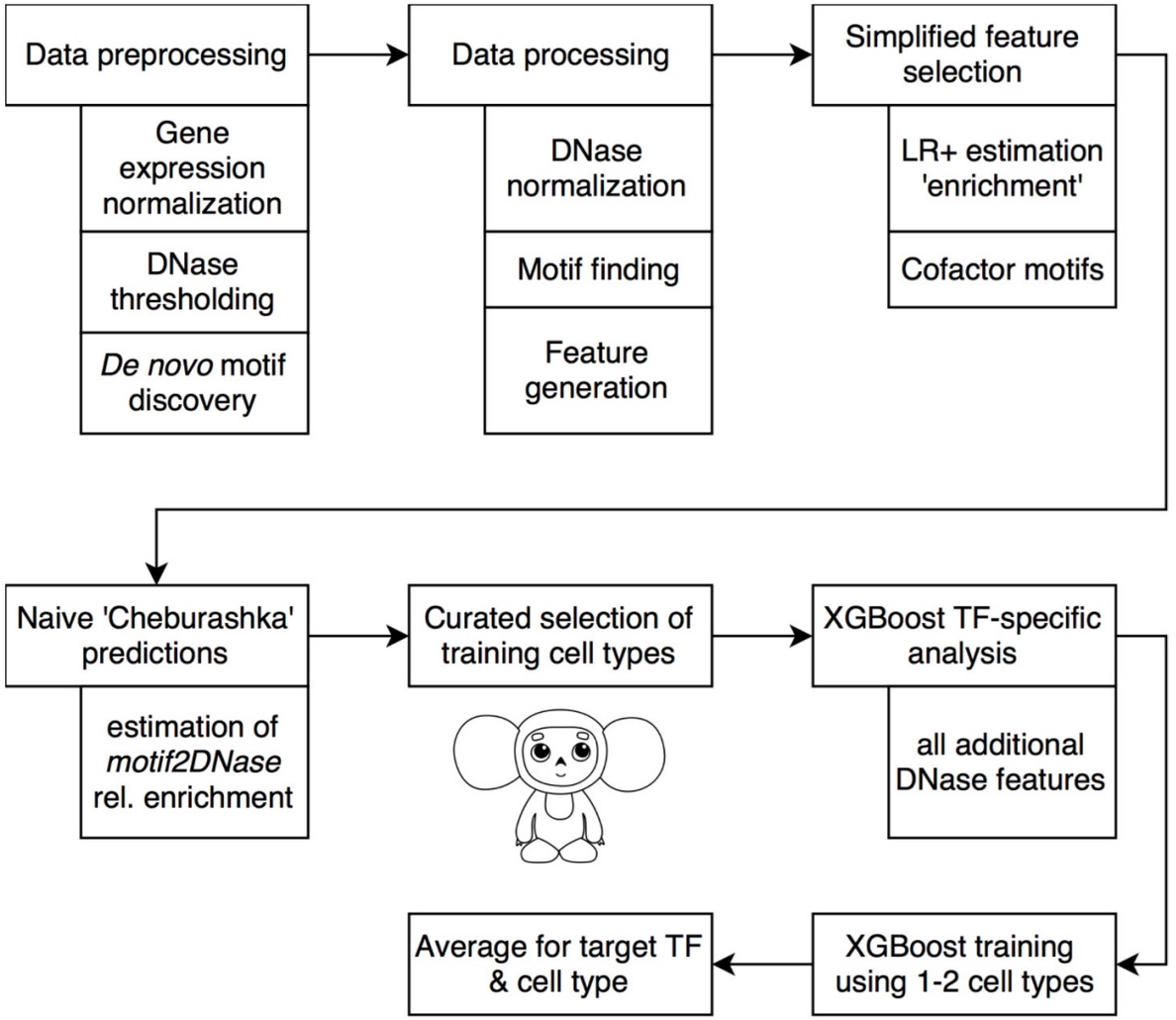
Угадывание "знака" m2d зависит от фактора транскрипции



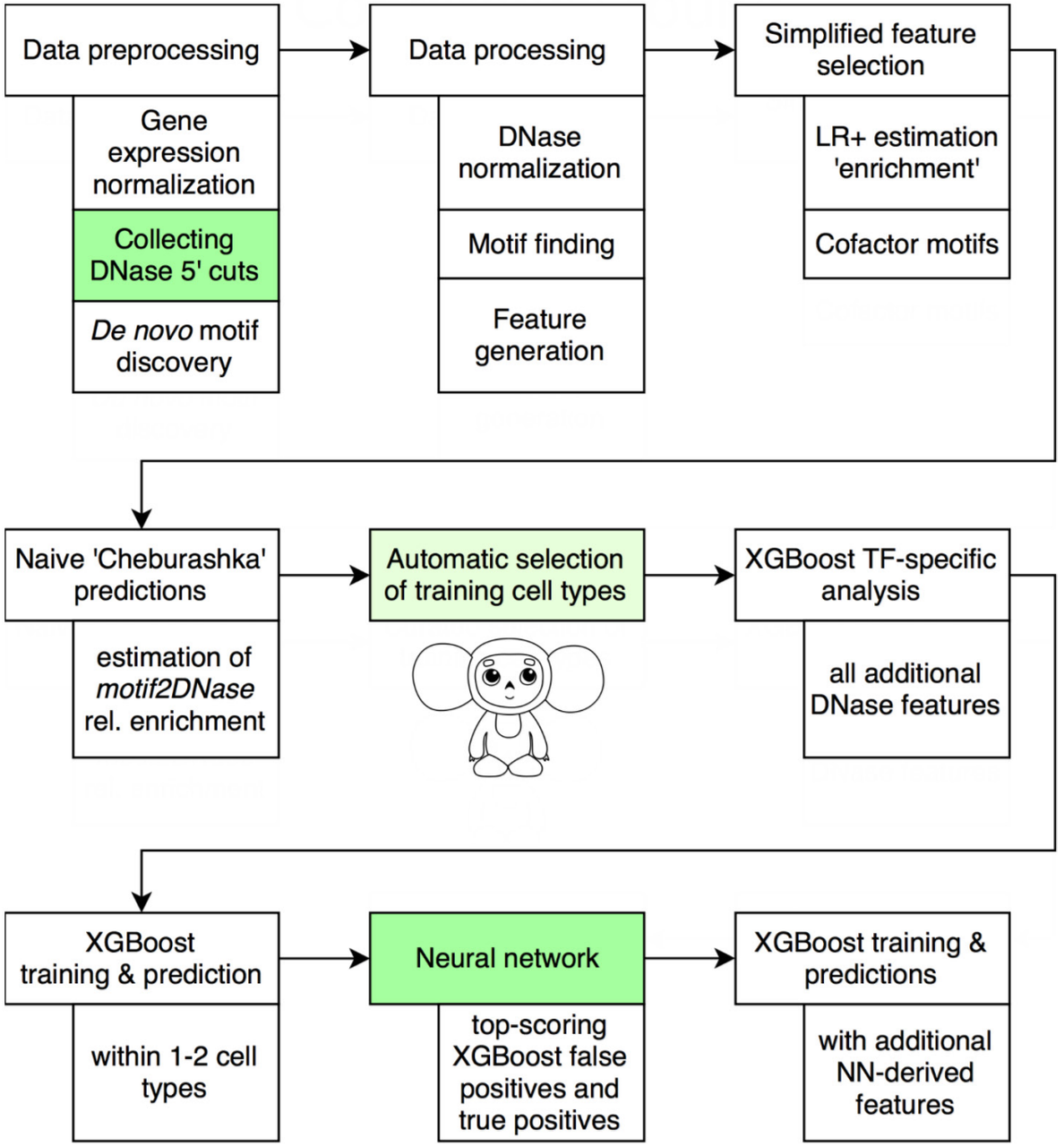
Но не зависит от типа клеток!



Conference round



Final round


Data preprocessing
Gene expression normalization
Collecting DNase 5' cuts
De novo motif discovery
Data processing
DNase normalization
Motif finding
Feature generation
Simplified feature selection
LR+ estimation 'enrichment'
Cofactor motifs
Naive 'Cheburashka' predictions
estimation of *motif2DNase* rel. enrichment
Automatic selection of training cell types
XGBoost TF-specific analysis
all additional DNase features
XGBoost training & prediction
within 1-2 cell types
Neural network
top-scoring XGBoost false positives and true positives
XGBoost training & predictions
with additional NN-derived features

Результаты соревнования

Challenge Scoring

auPRC is used as ranking measure in leaderboard round

Combined score in final round:

$$\sum_i \sum_j -log_{10}[min(0.5, rank(score_{i,j})/(N_j + 1))]$$

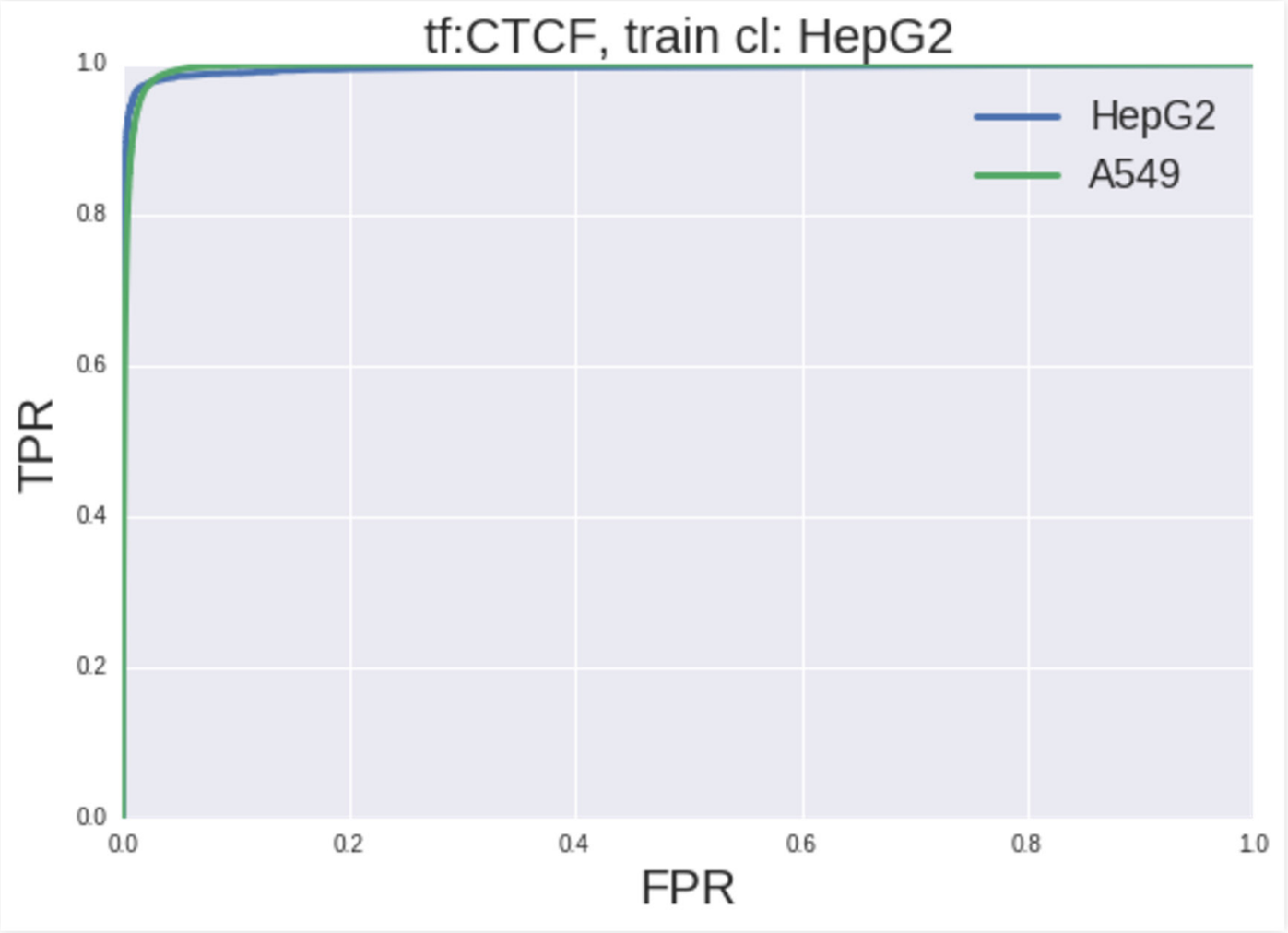
$score_{i,j}$: score for measure *i* in TF-cell type combination *j*
$rank(score_{i,j})$: rank among all submissions for $score_{i,j}$
N_j: the total number of submissions for TF-cell type *j*

where:

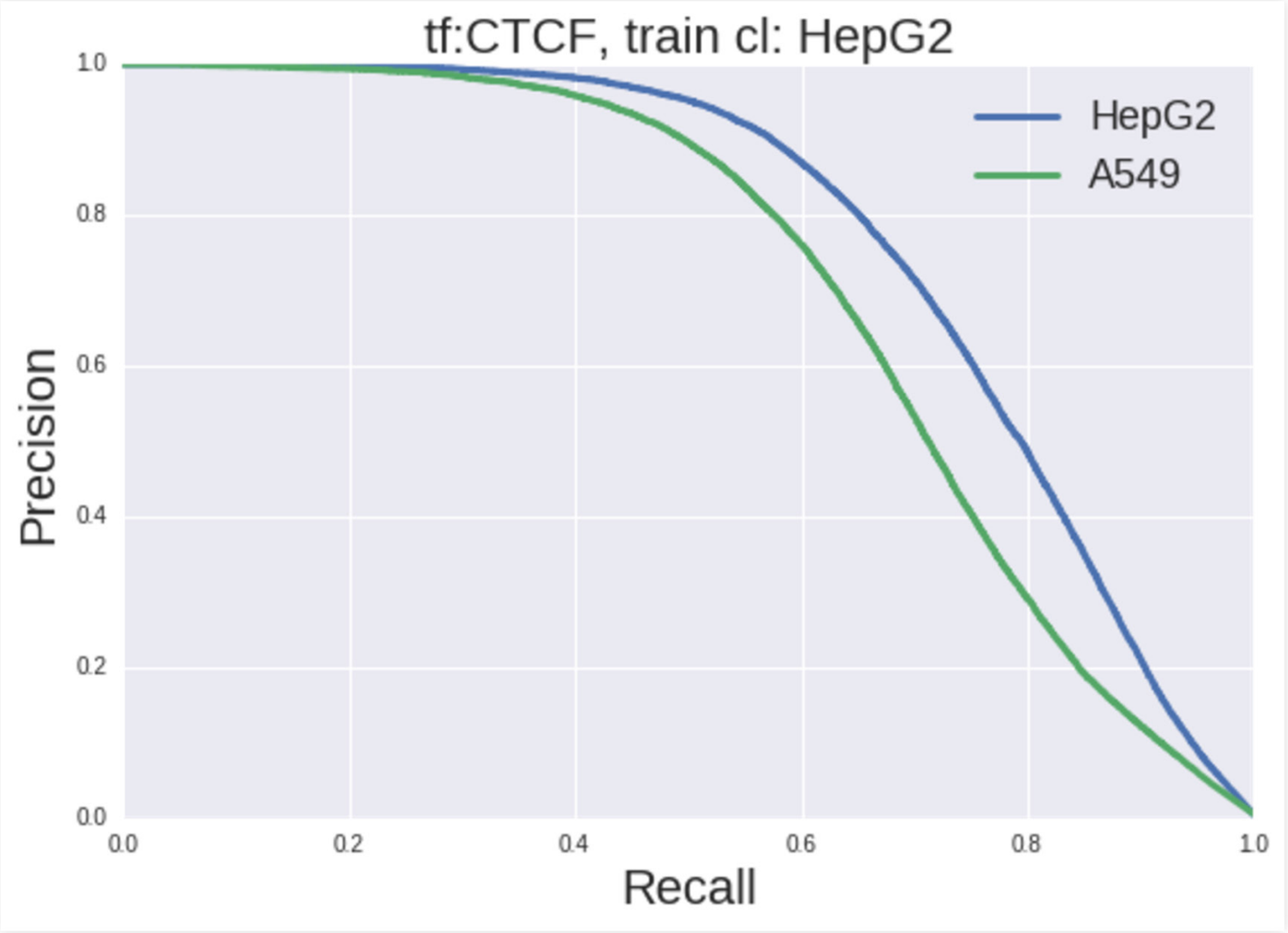
i is taken over all four measures
auROC, auPRC, recall 10% FDR, recall 50% FDR
j is taken over all TF-cell type combination in the final set

False discovery rate (FDR)
$= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
=1-Precision

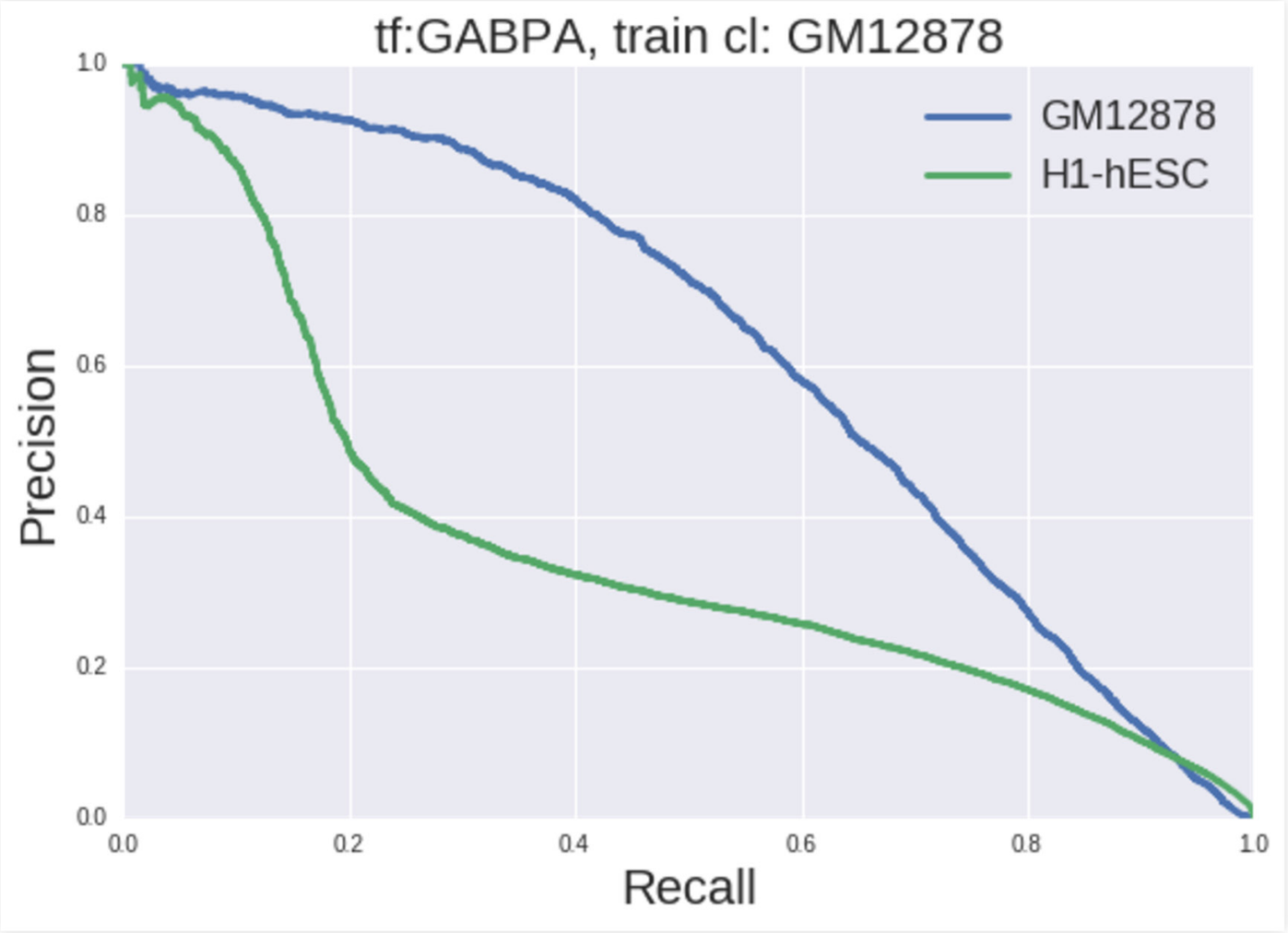
Как выглядит ROC-кривая



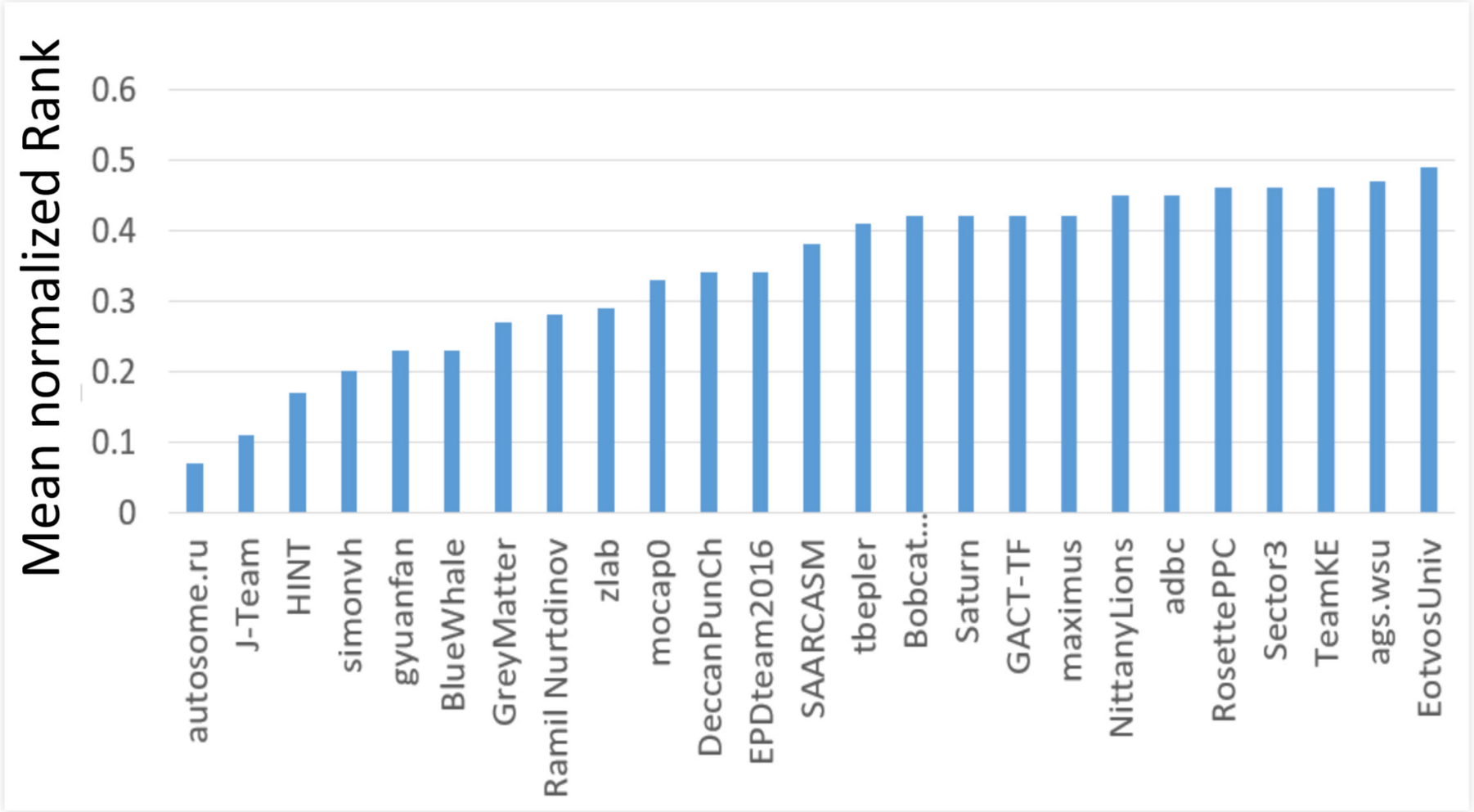
Как выглядит хорошая PR-кривая (CTCF)



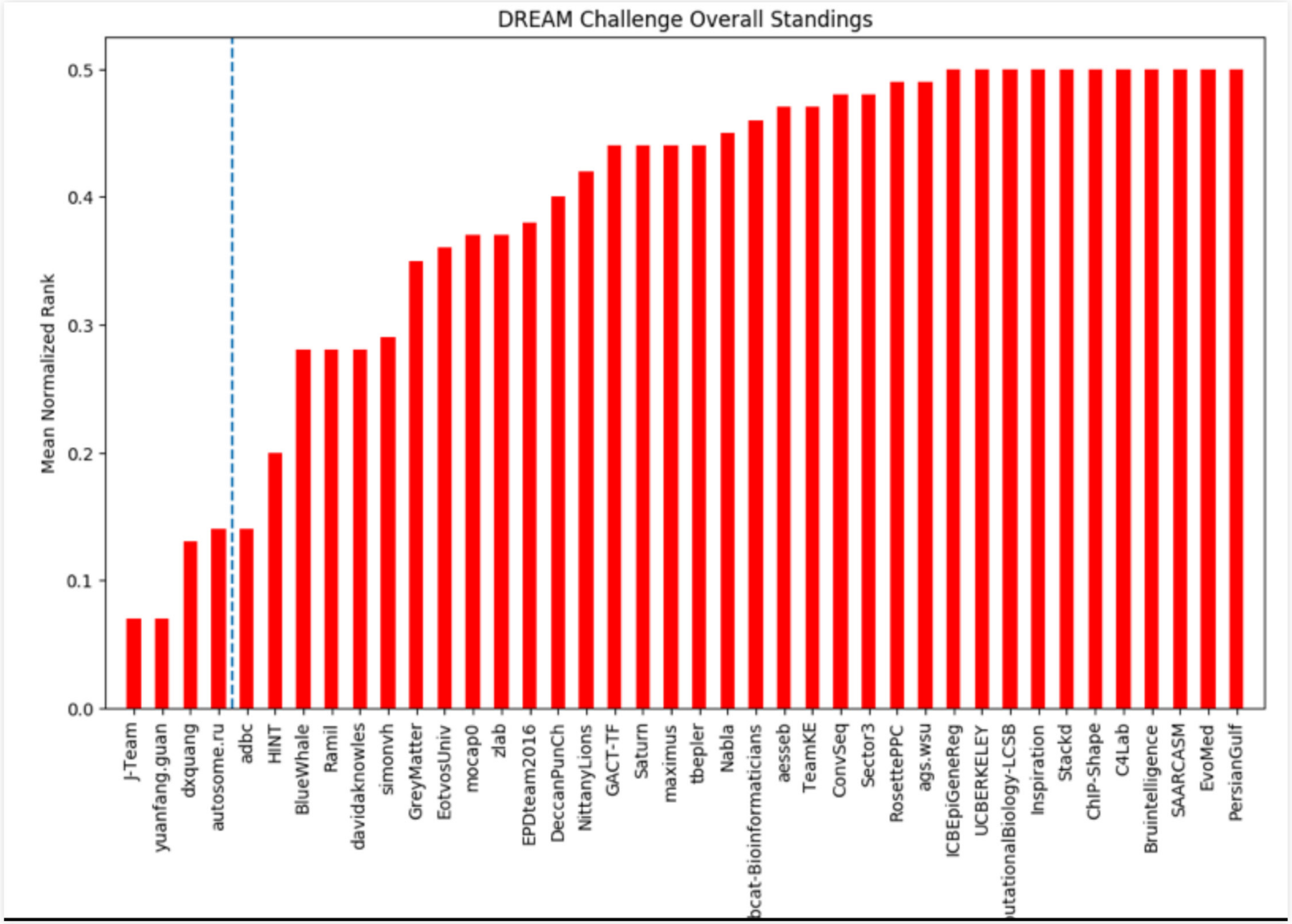
Как выглядит обычная PR-кривая (GABPA)



Результаты Conference-раунда



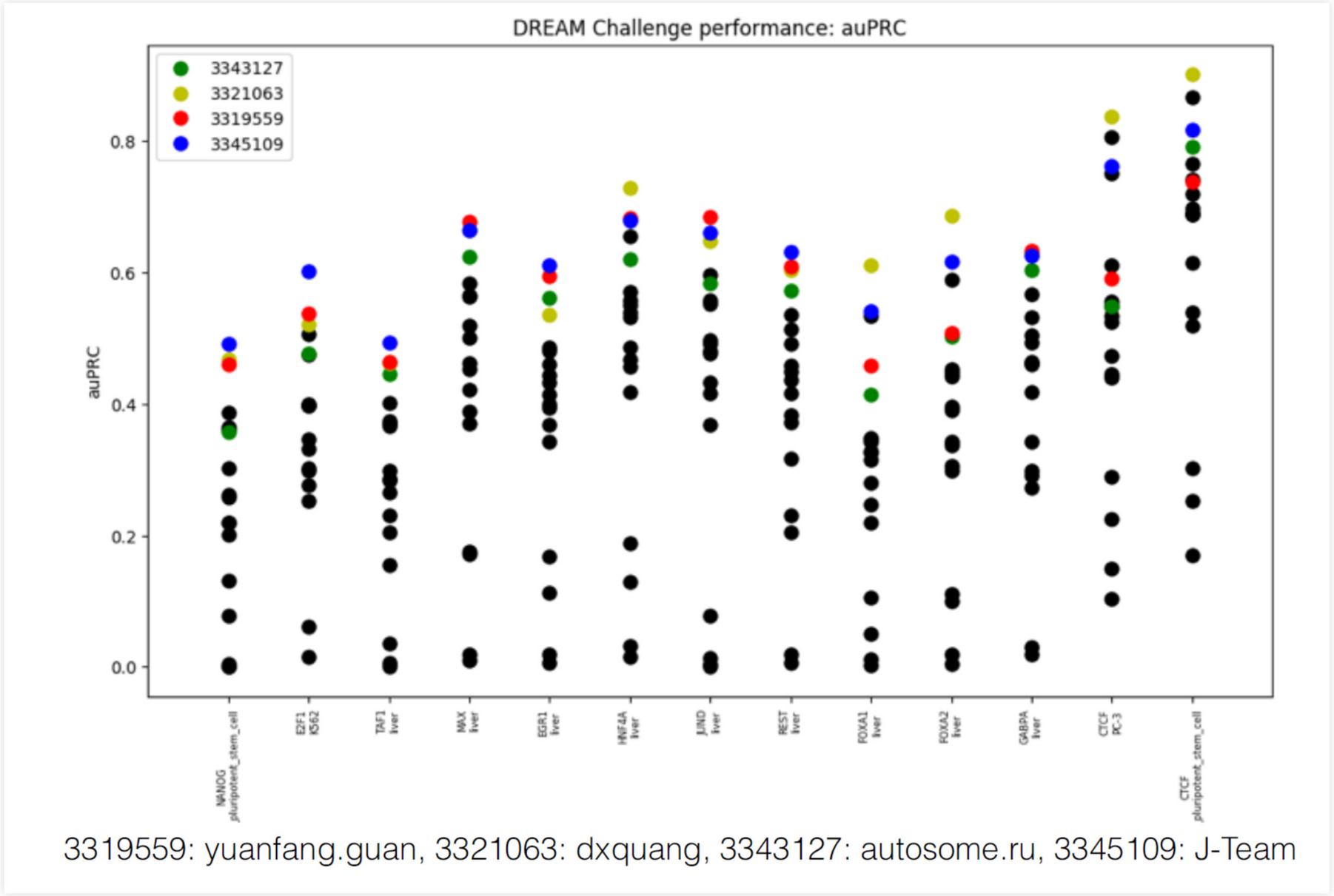
Результаты Final-раунда



Некоторые особенности решений

	J-Team	Yuanfang Guan	Daniel Quang	Autosome .ru	adbc	HINT
Summary	Winner in Rd 1 also; Iterative training on “difficult” examples	Co-winner, very few features but avoids overfitting celltype	Best deep learning submission	Winner in Rd 1; Chooses training celltypes using a simple learned celltype similarity classifier	Simple method, few features, sequential feature selection	Runs footprinting tool HINT-BC (authors)
Prediction algorithm	Bayesian model (max-likelihood; shallow model)	Gradient boosting (for logistic regression; xgboost)	Deep network (lower-layer seq features, other modalities higher).	Gradient boosting (for logistic regression)	Gradient boosting (for logistic regression)	Gradient boosting (for logistic regression)
Data sampling	Iterative training for background set (i.e. active learning)	Undersamples negatives for balanced labeled set	Undersamples negatives for balanced labeled set; no multitask training	Two-stage training, first stage picking training celltypes for second	Undersamples negatives for balanced labeled set	Chr2,9,22 train ensemble
Featurization	Many chosen features: SLIM motifs (authors), DNase-/RNA-Seq, dist. to closest TSS… all input modalities	Several top motif hits; cross-celltype DNase norm.; no RNA-Seq	Multitask (deep sequence features + shallow Dnase features, many GENCODE annotations)	“Basic” aggregated features in 1st stage, “advanced” in 2nd. Both carefully engineered	Sequential feature selection; a few engineered features (dist. to nearest 5 genes)	6 motifs/TF; DNase mostly with HINT-BC; no other modalities
Notes	Ensembled models: all from iterative training + all training task models	Final ensemble prediction averages many models (ad hoc cross-validation-like approach); auPRC optimization	Ensemble with same model on reverse complement. Conv+BiRNN architecture.	Long exploration of the data for model selection; many manually chosen features with all input modalities.	Ensemble with a sequentially chosen model - by exponentiating ensemble probabilities	One ensemble model for each training task

Достижимое качество предсказаний



3319559: yuanfang.guan, 3321063: dxquang, 3343127: autosome.ru, 3345109: J-Team

ARID3A Leaderboard

К вопросу о простых свойствах: GC% для ТФ ARID3A

			Quang				
7477123			Nathan Boley	SCORED	0.9157468895577268	0.20109382662746741	0.0	0.0

Round 1 (frozen)

ID	Date	name	team	status	auROC	auPRC	recall at 5%fdr
7247540	09/15/2016 05:14:24PM	L.ARID3A.K562.tab.gz	autosome.ru	SCORED	0.9935249455718399	0.46830990555055513	0.004552490212146044
7343749	10/04/2016 04:15:38PM	L.ARID3A.K562.tab.gz	zlab	SCORED	0.993067397681897	0.4090496652659795	0.0
7152512	08/21/2016 08:12:09AM	L.ARID3A.K562.tab.gz	maximus	SCORED	0.9152488844852229	0.3924180929559782	0.0
7187021	08/24/2016 12:16:34PM	L.ARID3A.K562.tab.gz	autosome.ru	SCORED	0.9913492026915374	0.36644384734602586	2.731494127287626E-4
7179922	08/23/2016 10:57:50AM	L.ARID3A.K562.tab.gz	autosome.ru	SCORED	0.9916850983548245	0.36568628288899213	2.731494127287626E-4
7187844	08/25/2016 02:26:14PM	L.ARID3A.K562.tab.gz	autosome.ru	SCORED	0.9919016273302999	0.3655484809837394	0.003915141582445597
7343210	10/04/2016 07:20:10AM	L.ARID3A.K562.tab.gz	zlab	SCORED	0.9924851096191286	0.3652387704129901	0.0

ARID3A ChIP-Seq не показывает характерного мотива

Factorbook human mouse

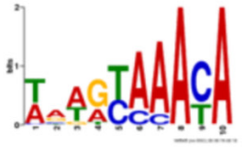
ARID3A Function Histone Profiles Motif Enrichment Histone Heatmap

Motifs Enriched in the Top 500 ChIP-seq Peaks

HepG2 - Snyder K562 - Snyder

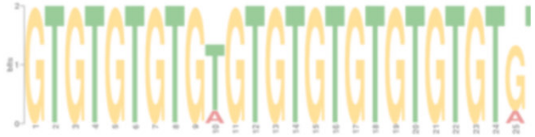
HepG2 - Snyder - ENCSR000EDP

1.
332 / 500
2.0e-94
TAAGTAAACA



2.
42 / 500
4.1e-34
AGTGCCATCTAGTGG

3.
6 / 500
3.1e-30
GTGTGTGTGTGTGTGTGTGTGTGTGTG



4.
53 / 500
1.5e-11
TGGACTTTGG

Factorbook human mouse

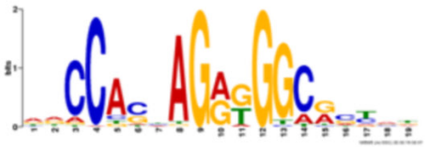
ARID3A Function Histone Profiles Motif Enrichment Histone Heatmaps TF

Motifs Enriched in the Top 500 ChIP-seq Peaks

HepG2 - Snyder K562 - Snyder

K562 - Snyder - ENCSR000EFY

1.
123 / 500
2.0e-160
AACCACTAGAGGGCGCTAT

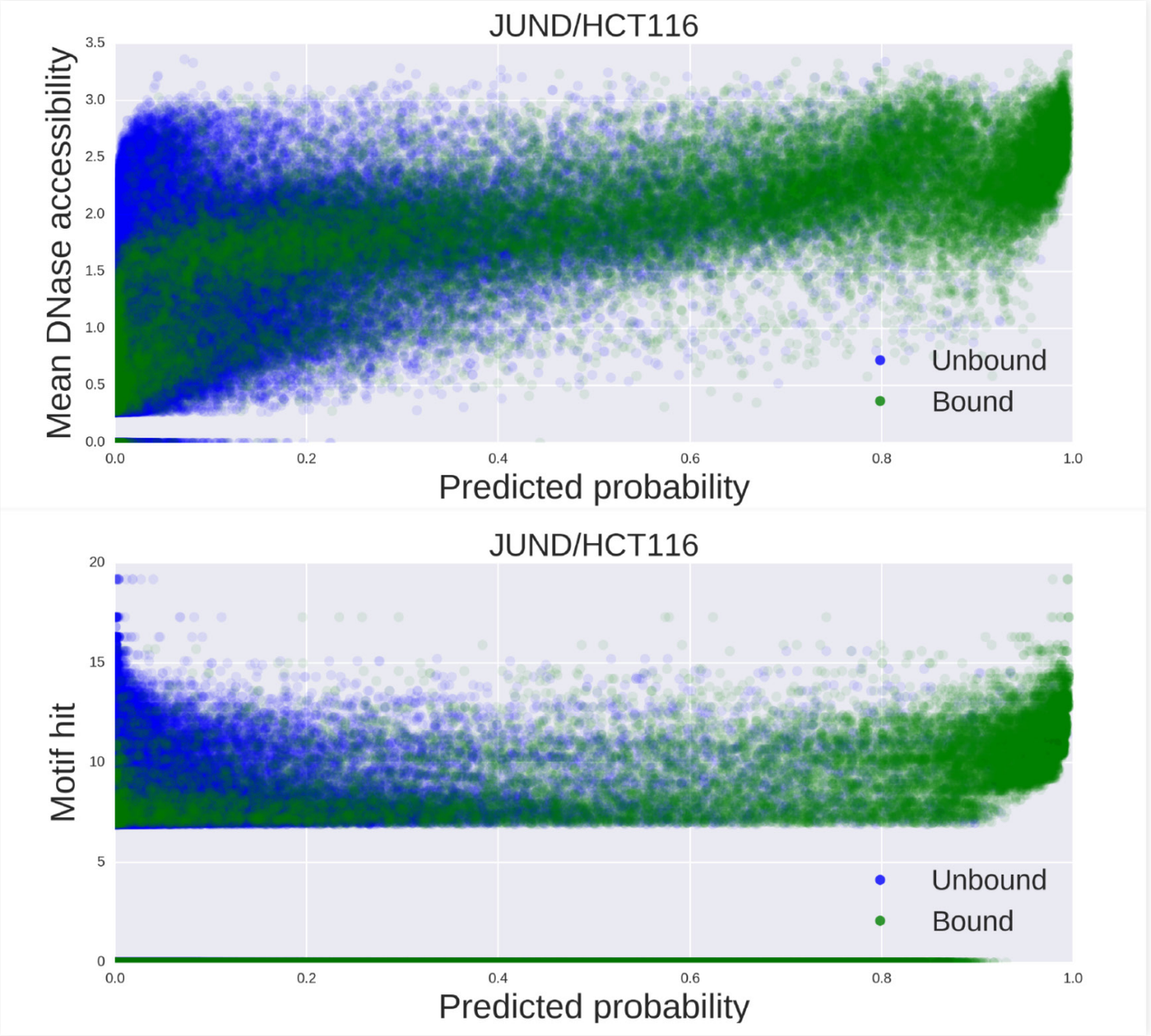


2.
166 / 500
1.7e-52
AGATAAGG

3.
12 / 500
1.2e-25
AGTGTGTGTGTGTGTGTGTGTGTGTGTG

4.
56 / 500
3.5e-20
AAAGGGATGAGTCAT

Решение еще далеко



На пути к идеальному предсказанию

- Метилирование ДНК
- Конкурентное связывание членами одного семейства
- Кооперативное связывание
- Непрямое связывание
- Транскрипционные факторы-пионеры
- Ткань-специфичные мотивы связывания

Предсказание полногеномного связывания in vivo

не то же самое, что предсказание результатов ChIP-Seq

...enrichment in a given ChIP experiment depends on many intractable parameters, likely including a phase of a moon.

A. Barski in *Genomic location analysis by ChIP-Seq, 2009*

autosome.ru team

- Ваня Кулаковский
- Валентина Боева
- Чебурашка
- Ирина Елисеева
- Всеволод Макеев
- Андрей Ландо
- Григорий Сапунов
- Илья Воронцов