Сравнительный анализ данных метилирования ДНК для клеточных линий разной степени дифференциации

Сергей Лебедев

Руководитель: Олег Шпынов

JetBrains

8 июня, 2013

Метилирование ДНК

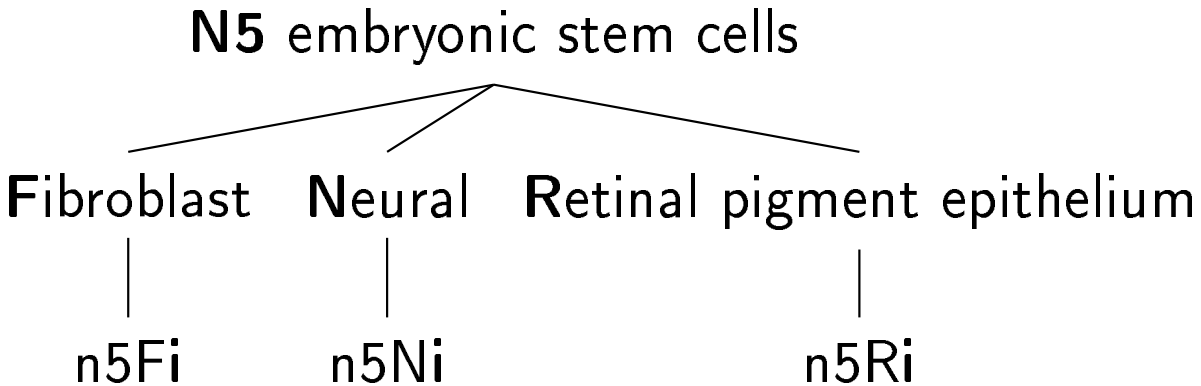
Химическая модификация, добавляющая метильную группу к цитозину или адениниу.

Индуцированные плюрипотентные стволовые клетки

- Могут быть получены из соматических клеток путем эпигенетического перепрограммирования с помощью т. н. факторов плюрипотентности (OCT4, KLF4, SOX2 и др.).
- Интересно исследовать эпигенетические различия, индуцируемые переходом клеток в плюрипотентное состояние.

- Данные предоставлены Институтом общей генетики им. Н. И. Вавилова РАН.
- Всего 24 образца, 2-3 биологических репликата для каждой клеточной линии.

Клеточные линии



Illumina Human Methylation 450K BeadChip[1]

- Микрочип содержит **два** вида проб – чтобы сравнивать результаты экспериментов, необходима нормализация.
- Для нормализации β-value одного вида проб на другой мы использовали TOST [TT12].

β-value – уровень метилирования ДНК

$$\beta = \frac{max(I_{methylated}, 0)}{max(I_{methylated}, 0) + max(I_{unmethylated}, 0) + \alpha}$$

[1] http://www.smd.qmul.ac.uk/gc/Services/IlluminaMeth/index.html

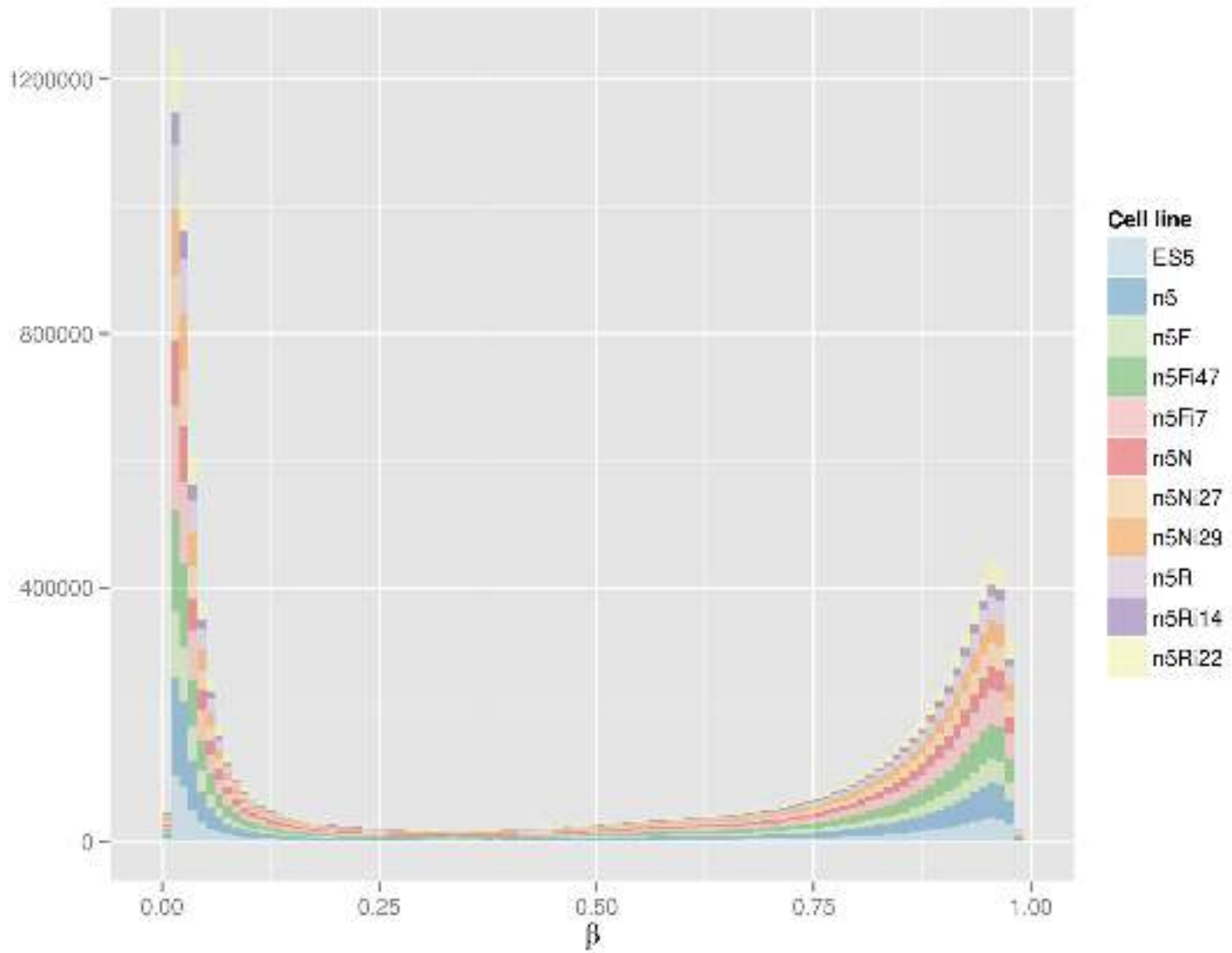
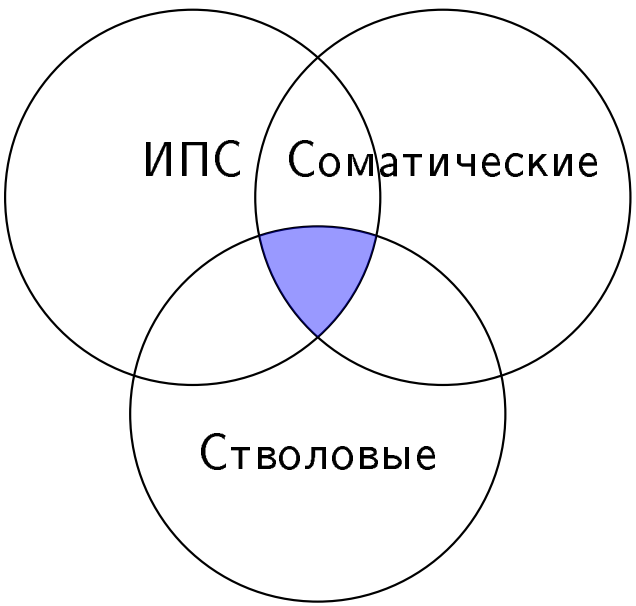


Рис. : Распределение β-value для исследуемых клеточных линий

1. Внутри каждой группы {n5F, n5N, n5R} для каждой пары клеточных линий найти гены со статистически различным уровнем метилирования ДНК.
2. Определить гены, специфичные для плюрипотентного состояния:



$$\textit{ИПС-специфичные} =$$
$$(\textit{ИПС-Стволовые}_{<} \cap \textit{ИПС-Соматические}_{<}) \cup$$
$$(\textit{ИПС-Стволовые}_{>} \cap \textit{ИПС-Соматические}_{>})$$

Сравнение: непараметрические тесты

- Будем сравнивать уровни метилирования для двух клеточных линий “поточечно” – по отдельности для каждого гена.
- Распределение β-value сильно отклоняется от нормального, поэтому мы ограничены непараметрическими тестами.
- Можно ли считать, что для каждого гена уровни метилирования в двух различных клеточных линиях **независимы**?
  - Да
  - Нет

Да: U-критерий Манна-Витни

- **все** наблюдения из обеих выборок независимы;
- нулевая гипотеза: медиана уровня метилирования у двух сравниваемых клеточных линий одинакова.

Нет: T-критерий Уилкоксона

- наблюдения парные, причем **все пары** наблюдений независимы;
- нулевая гипотеза: медиана разности уровней метилирования для двух сравниваемых клеточных линий равна нулю.

Проблемы: эффект множественных сравнений

- Зафиксируем некоторый уровень значимости $\alpha = Pr\{FP\}$, тогда вероятность ошибки первого рода в *хотя бы одном* из m сравнений:

$$FWER = 1 - (1 - \alpha)^m$$

- Что делать?
  - Поправка Бонферрони: $FWER \leq \frac{\alpha}{m}$;
  - Метод Бенджамини-Хохберга: $FDR = \frac{FP}{TP+FP} \leq q$;
  - Q-value [ST03], минимальный FDR, при котором нулевая гипотеза для теста с P-value p_i отвергается:

$$\hat{q}_i = \min_{t \geq p_i} \widehat{FDR}(t)$$

Если для гена X Q-value равняется 0.013, то 1.3% генов с меньшим или равным P-value – это ошибки первого рода.

Проблемы: неравномерное покрытие генов

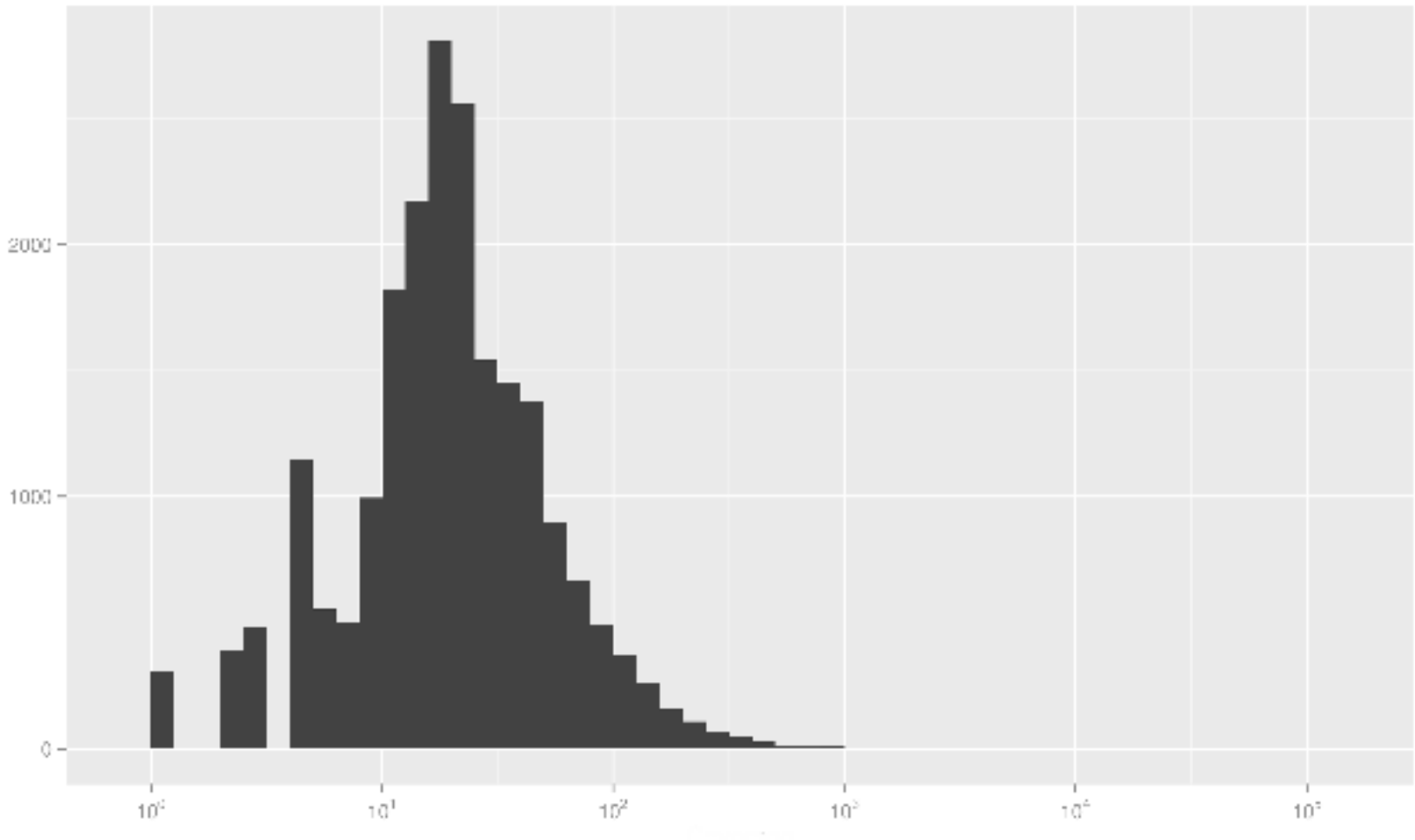


Рис. : Гистограмма количества проб для гена на Illumina Human Methylation 450K BeadChip

Проблемы: неравномерное покрытие генов

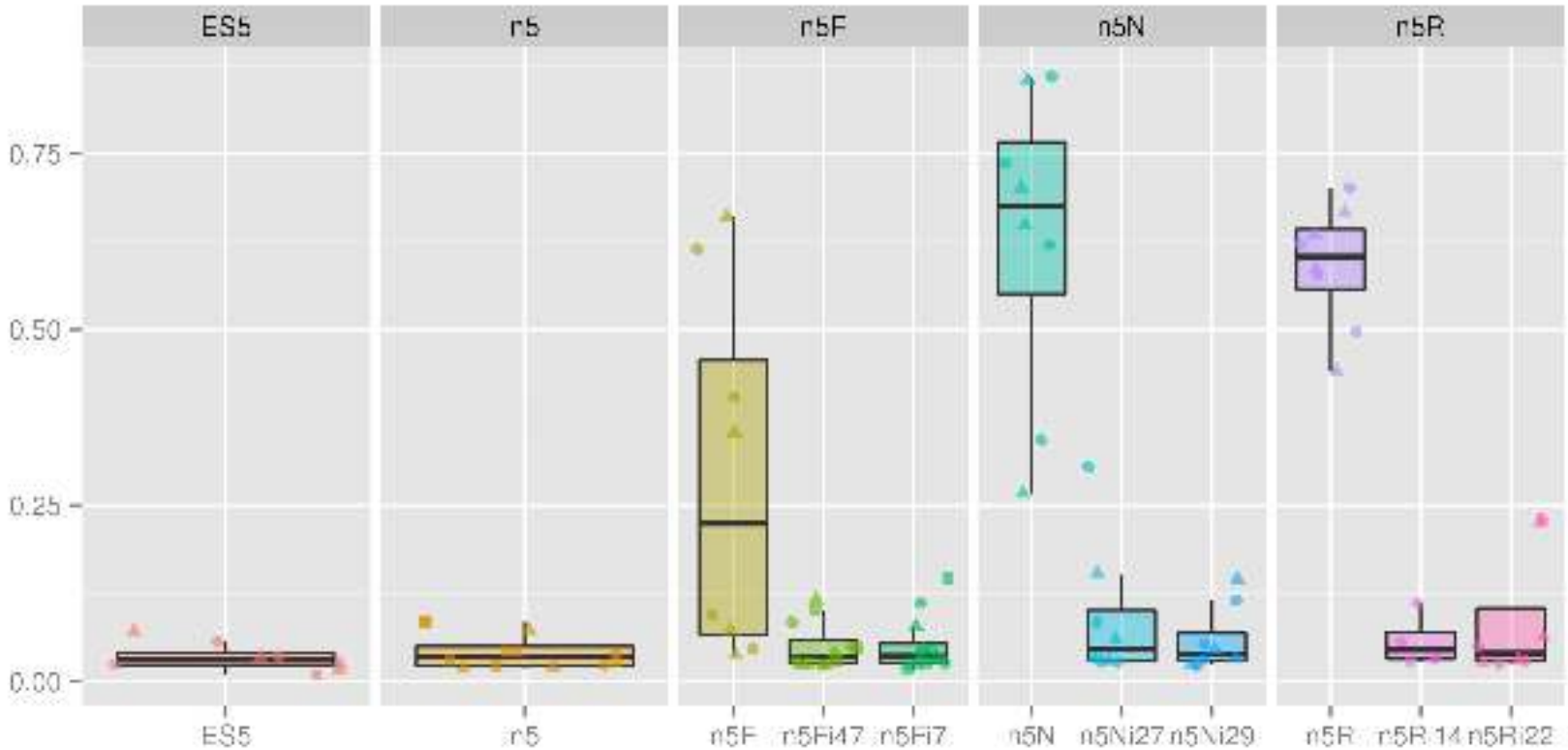


Рис. : Ящик с усами для уровня метилирования гена NANOG в исследуемых клеточных линиях

Клеточная линия	Q-value ≤ 0.05	Метод Б-Х, P-value ≤ 0.05
n5F	–	–
n5N	PTPRN2	PTPRN2
n5R	TBX5, SLC9A3	TBX5, SLC9A3

Таблица : ИПС-специфичные гены для исследуемых клеточных линий

Ожидаемое количество ошибок первого рода

$$FDR = \frac{FP}{TP + FP} \leq 0.05$$

$$TP + FP = 134453 \quad \Rightarrow FP \approx 6722$$

Планы

- Применить модельный подход к сравнению микрочипов, см. например [HCY+08].
- Сузить рассматриваемые в сравнениях множества до набора генов, участвующих в известных метаболических путях.
- Провести корреляционный анализ данных по метилированию ДНК с данными об экспрессии для исследуемых клеточных линий.

Вопросы?

Ссылки

Andres E Houseman, Brock Christensen, Ru-Fang Yeh, Carmen Marsit, Margaret Karagas, Margaret Wrensch, Heather Nelson, Joseph Wiemels, Shichun Zheng, John Wiencke, et al.

Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions.

Bmc Bioinformatics, 9(1):365, 2008.

John D Storey and Robert Tibshirani.

Statistical significance for genomewide studies.

Proceedings of the National Academy of Sciences, 100(16):9440–9445, 2003.

Nizar Touleimat and Jörg Tost.

Complete pipeline for infinium® human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation.

Epigenomics, 4(3):325–341, 2012.