«Анализ транскриптомных данных»

Лекция #8.

Контроль за дисперсией

Серёжа Исаев

аспирант MedUni Vienna

Содержание курса

1. Bulk RNA-Seq:
    a. экспериментальные подходы,
    b. выравнивания и псевдовыравнивания,
    c. анализ дифференциальной экспрессии,
    d. функциональный анализ;

1. Single-cell RNA-Seq:
    a. экспериментальные подходы,
    b. **отличия от процессинга bulk RNA -Seq,**
    c. методы снижения размерности,
    d. кластера и траектории,
    e. мультимодальные омики одиночных клеток.

Выделение интересующей нас части дисперсии

Дальше пойдёт разговор о том, как отличить техническую дисперсию от биологической, той, что нас интересует

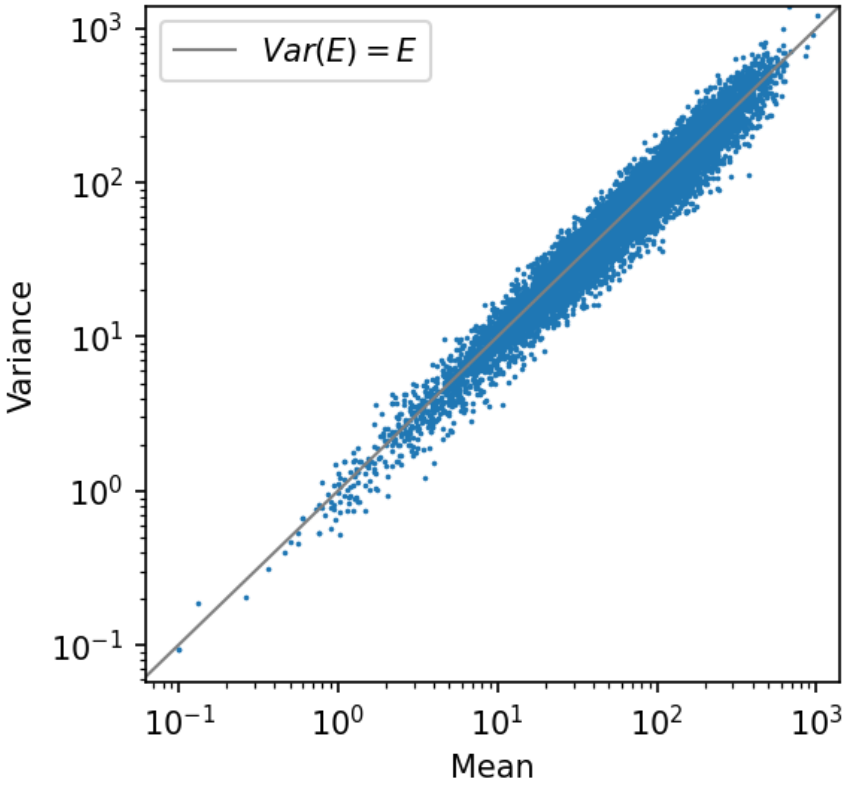
Общее предположение заключается в том, что в целом это техническая вариация похожа для некоторых групп генов (например, для тех, у которых похожа средняя экспрессия), по ним мы можем оценить “ожидаемую” дисперсию и вычислить разницу с этой ожидаемой экспрессией

Эта разница и будет той частью вариации, в которой заключается биологическая разница между клетками

Распределение Пуассона

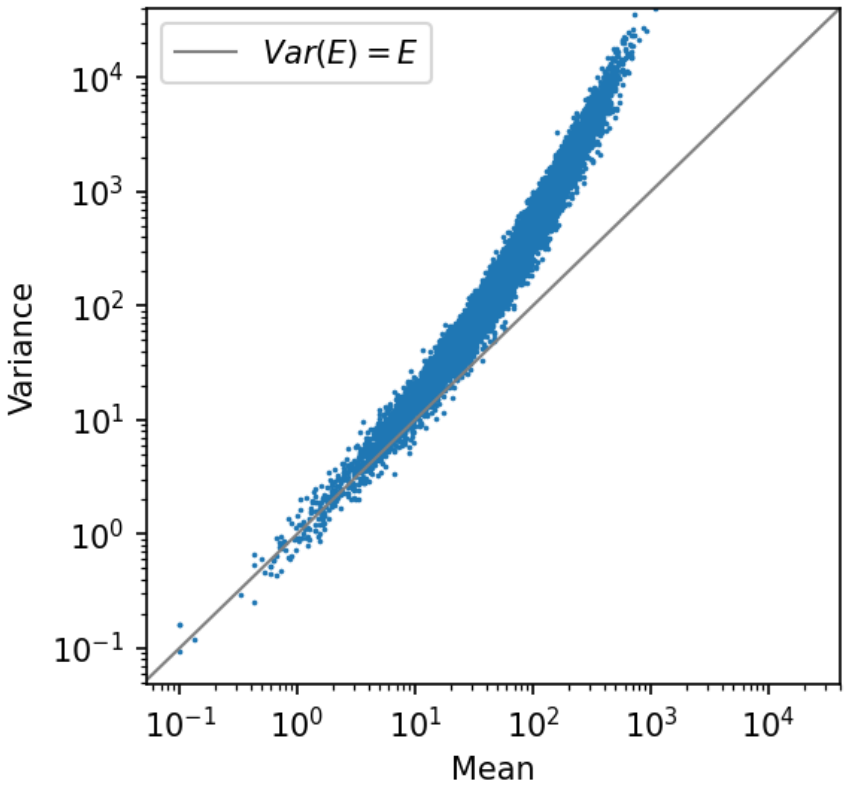
У нас есть большое число молекул РНК в клетке, пропорции РНК каждого типа — это p_i (т. е. вероятность достать конкретно эту молекулу РНК)

Мы 30 раз независимо достаём из этой клетки по 10^6 молекул РНК, смотрим на их распределение



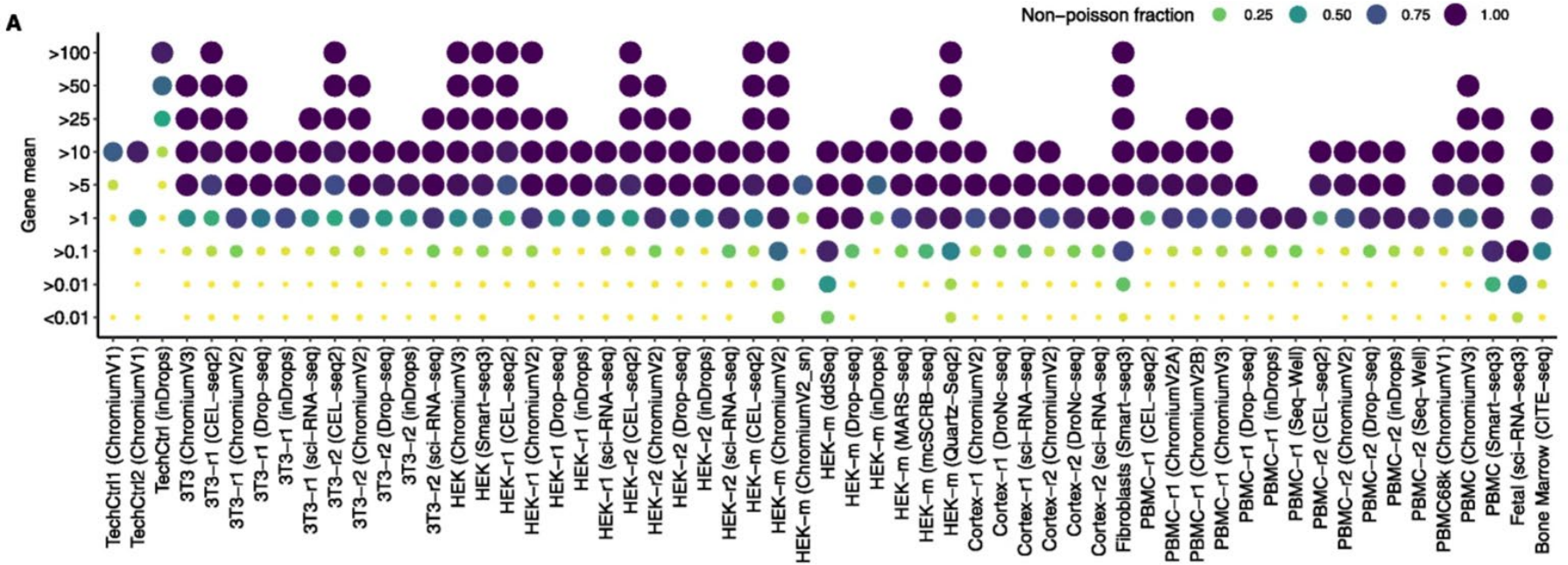
Овердисперсия

Если мы проведём случайную селекцию дважды и между этими шагами добавим случайного шума, пропорционального средней экспрессии гена, то мы получим овердисперсию в данных



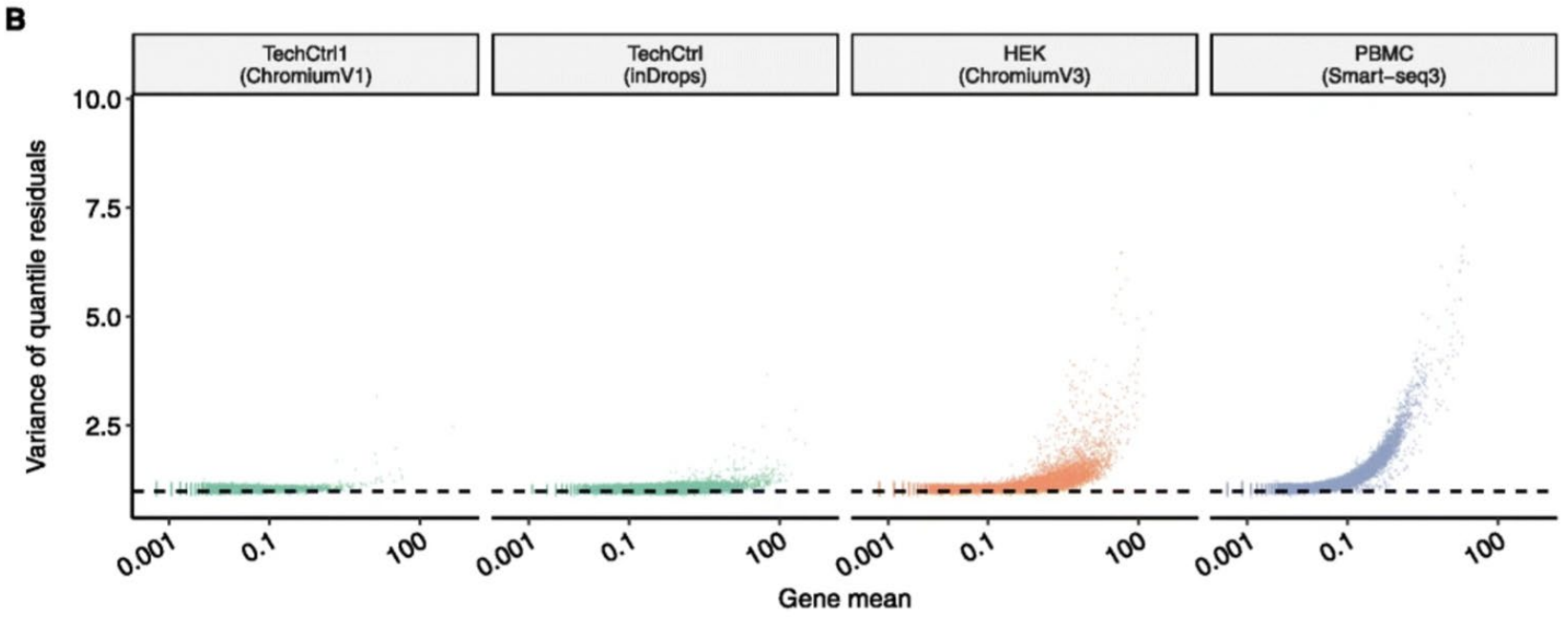
Овердисперсия

При малых средних значениях экспрессии генов овердисперсия не так видна, однако при увеличении покрытия это становится проблемой



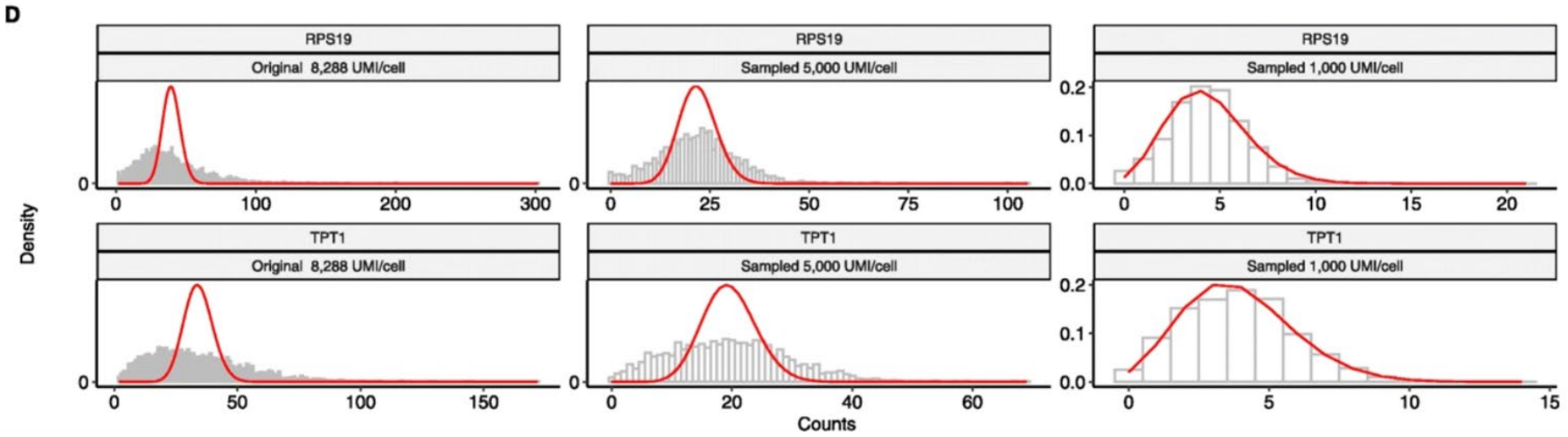
Овердисперсия

При малых средних значениях экспрессии генов овердисперсия не так видна, однако при увеличении покрытия это становится проблемой



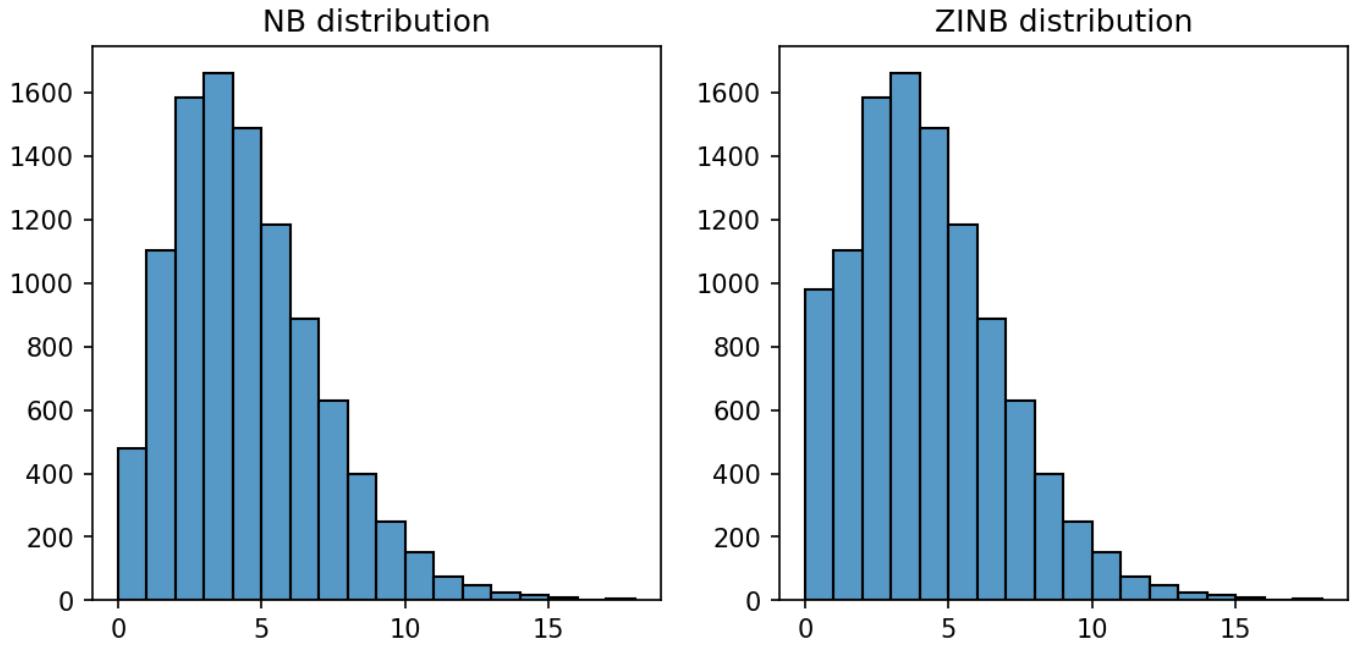
Овердисперсия

При малых средних значениях экспрессии генов овердисперсия не так видна, однако при увеличении покрытия это становится проблемой



Zero-inflated NB -распределение (ZINB)

В scRNA-Seq наблюдается проблема дропаутов — того, что экспрессию какого-то гена в какой-то клетке мы не видим, хотя она там в принципе должна быть



Zero-inflated NB -распределение (ZINB)

Но, по всей видимости, объяснение этому кроется в истинной биологической гетерогенности исследуемых образцов, а не в особенности распределения

Correspondence | Published: 14 January 2020

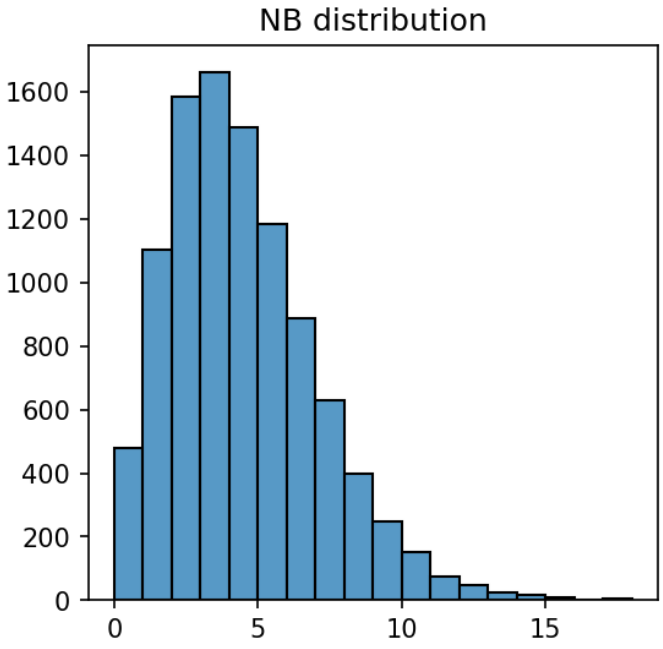
Droplet scRNA-seq is not zero-inflated

Valentine Svensson

Nature Biotechnology **38**, 147–150 (2020) | Cite this article

12k Accesses | **89** Citations | **89** Altmetric | Metrics

Поиск среднего и дисперсии



Для оценки параметров отрицательного биномиального распределения чаще всего используют maximum likelihood estimators (MLE)

Почему нельзя просто посчитать выборочную дисперсию? Из-за того, что у нас присутствует некоторый скрытый фактор — глубина секвенирования клетки

Модель Пуассона для распределения каунтов

Сначала запишем аналитические решения для распределения Пуассона:

$$X_{cg} \sim \text{Poisson}(\mu_{cg})$$

$$\mu_{cg} = n_c p_g.$$

$$\hat{\mu}_{cg} = \frac{\sum_j X_{cj} \cdot \sum_i X_{ig}}{\sum_{ij} X_{ij}}$$

Analytical Pearson Residuals

Для начала кратко вспомним, как связаны дисперсия и среднее в отрицательном биномиальном распределении:

$$\mathbb{E}[X] = \frac{r(1-p)}{p},$$

$$\mathrm{Var}[X] = \frac{r(1-p)}{p^2} = \frac{r(1-p)(p+(1-p))}{p^2} = \frac{r(1-p)p + r(1-p)^2}{p^2} =$$

$$= \frac{r(1-p)}{p} + \frac{r(1-p)^2}{p^2} = \mathbb{E}[X] + \frac{1}{r}\frac{r^2(1-p)^2}{p^2} = \mathbb{E}[X] + \frac{1}{r}\mathbb{E}[X]^2,$$

Analytical Pearson Residuals

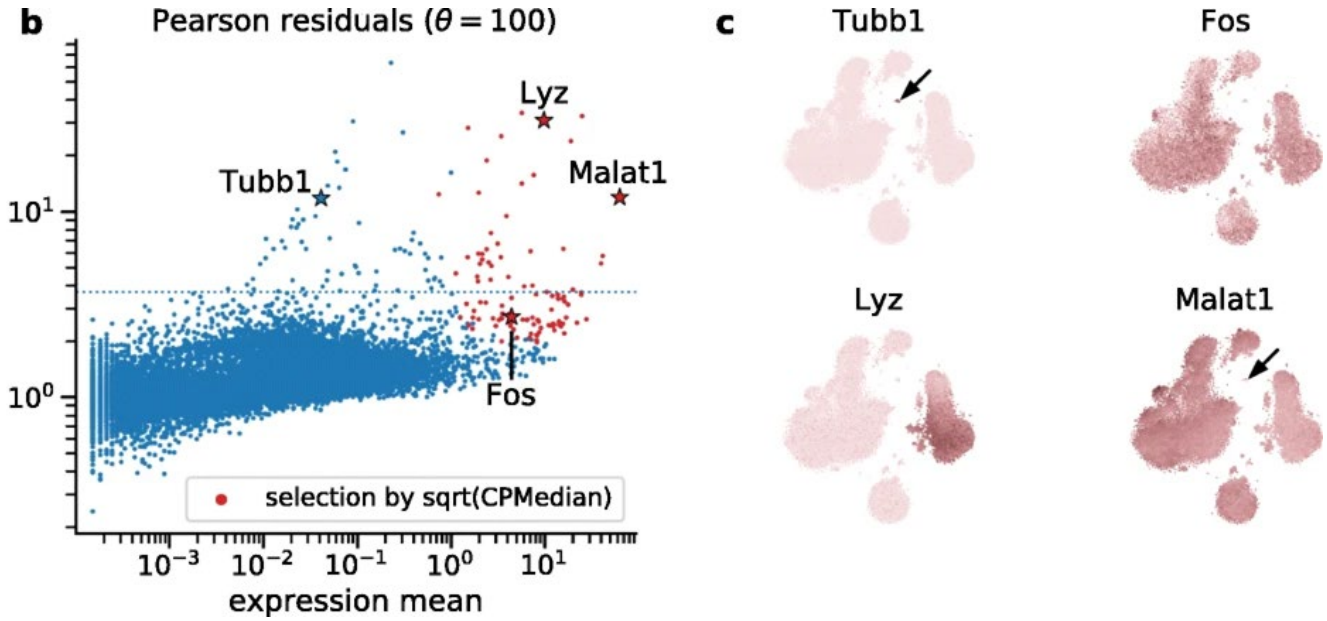
Таким образом, если мы воспользуемся оценкой среднего из распределения Пуассона и зафиксируем параметр 1/r (его называют θ), то мы сможем найти оценочную дисперсию

Исходя из этого мы можем посчитать остатки Пирсона и оценить “биологическую” часть дисперсии, которую мы можем оставить для дальнейших вычислений

$$Z_{cg} = \frac{X_{cg} - \hat{\mu}_{cg}}{\sqrt{\hat{\mu}_{cg} + \hat{\mu}_{cg}^2/\theta}}$$

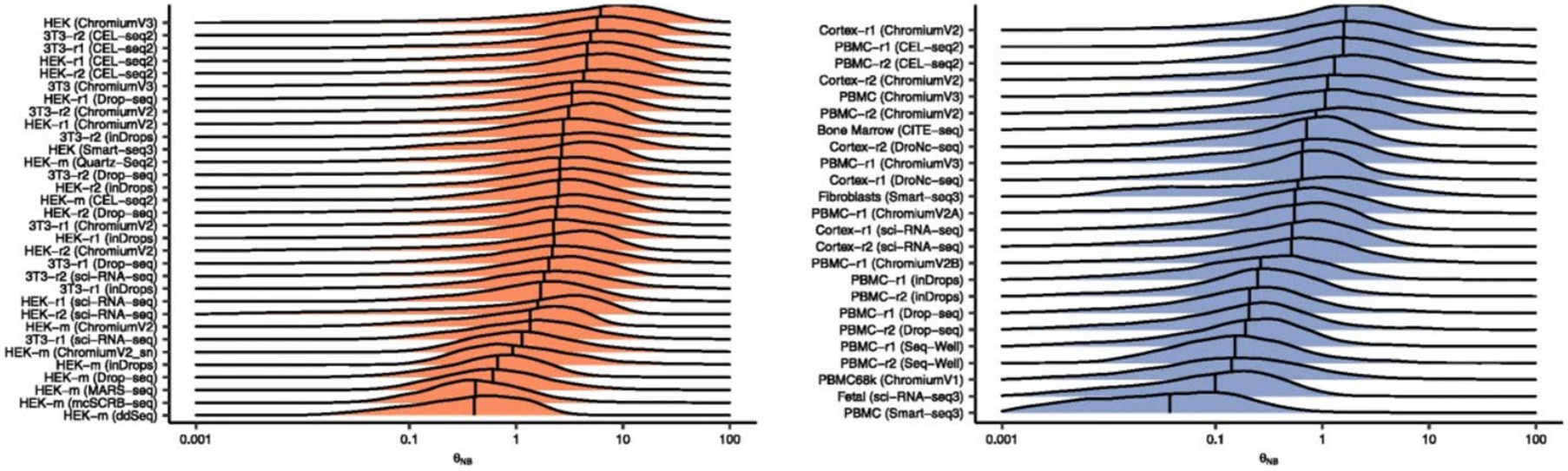
Analytical Pearson Residuals

По всей видимости, такая простая трансформация позволяет выделить действительно биологически овердисперсные гены и улучшить анализ датасетов



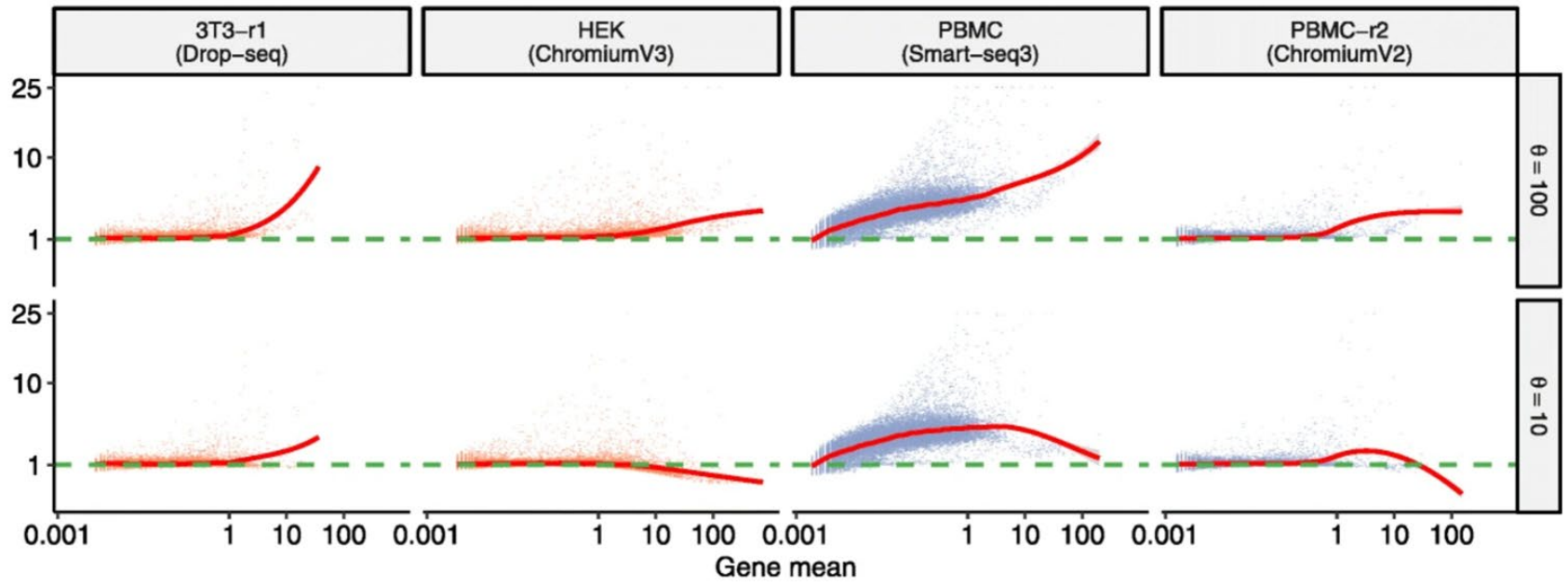
Более аккуратная оценка параметров

Однако такая оценка, по всей видимости, слишком грубая, так как параметр θ отличается от эксперимента к эксперименту



Более аккуратная оценка параметров

Однако такая оценка, по всей видимости, слишком грубая, так как параметр θ отличается от эксперимента к эксперименту и, более того, от гена к гену



SCTransform

$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m$$

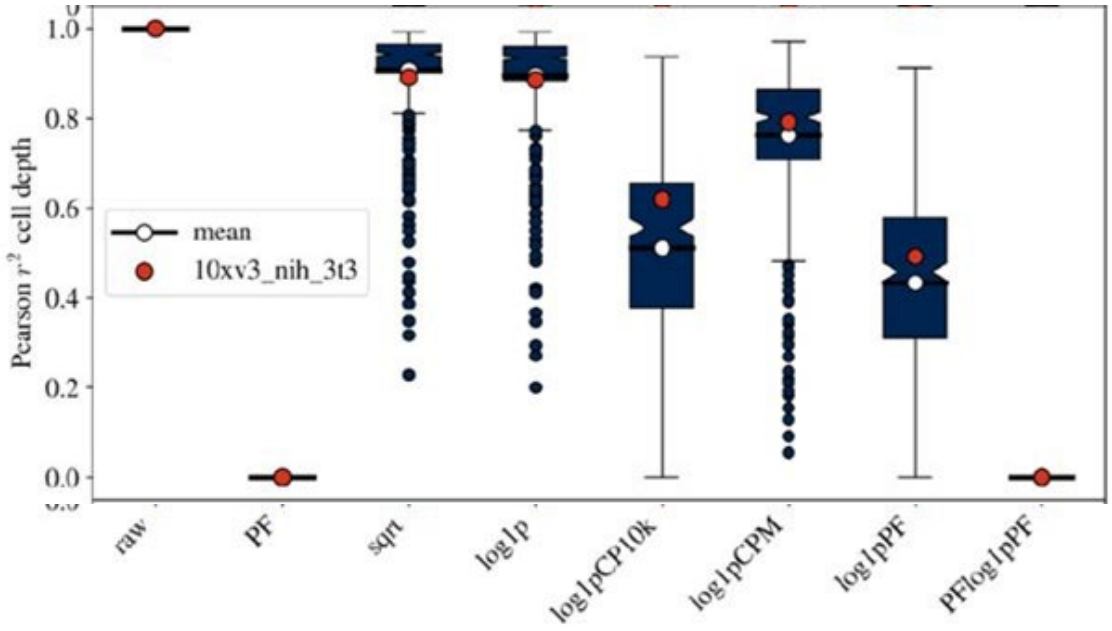
$$\begin{aligned} z_{ij} &= \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}}, \\ \mu_{ij} &= \exp\left(\beta_{0_i} + \beta_{1_i} \log_{10} m_j\right) \\ \sigma_{ij} &= \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}}, \end{aligned}$$

Оценка параметров распределения для каждого из генов производится с использованием модели связи числа каунтов гена с глубиной секвенирования клетки

Параметр θ оценивается для каждого гена отдельно, более того, после этого происходит регуляризация этого параметра с учётом θ генов с похожей средней экспрессией

Эффект от глубины секвенирования клетки

Однако, по всей видимости, не всё так радужно, и эффект от глубины секвенирования клетки остаётся даже после SCTransform



Иные способы контроля дисперсии

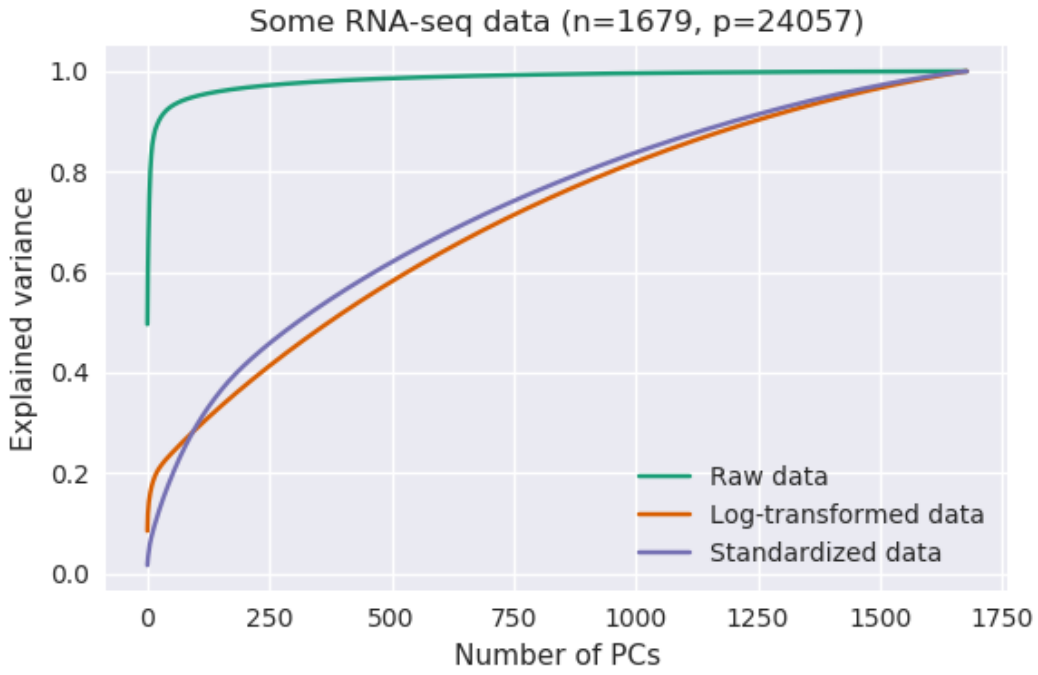
Более простые способы контроля дисперсии включают в себя следующие варианты:

1. scale(log1pCP10k),
2. log1pCP10k,
3. log1pPF,
4. PFlog1pPF

Каждый из этих способов сначала проводит линейную нормализацию на глубину секвенирования библиотеки, а потом трансформирует эти значения либо просто при помощи логарифмирования, либо при помощи логарифмирования и шкалирования

Ещё раз: зачем нам это всё?

Когда мы будем применять в дальнейшем методы кластеризации или снижения размерности, они будут чувствительны в основном к генам, экспрессия которых имеет высокие абсолютные значения дисперсии



Преобразования данных: шаг за шагом

SCTransform:

1. Регрессия на глубину секвенирования,
2. Определение значений распределения для гена,
3. Нахождение остатков регрессии,
4. …

Остальные методы:

1. Нормализация на глубину библиотеки,
2. Логарифмирование (делает наши распределения близкими к нормальному),
3. Шкалирование (дополнительно контролирует дисперсию),
4. …