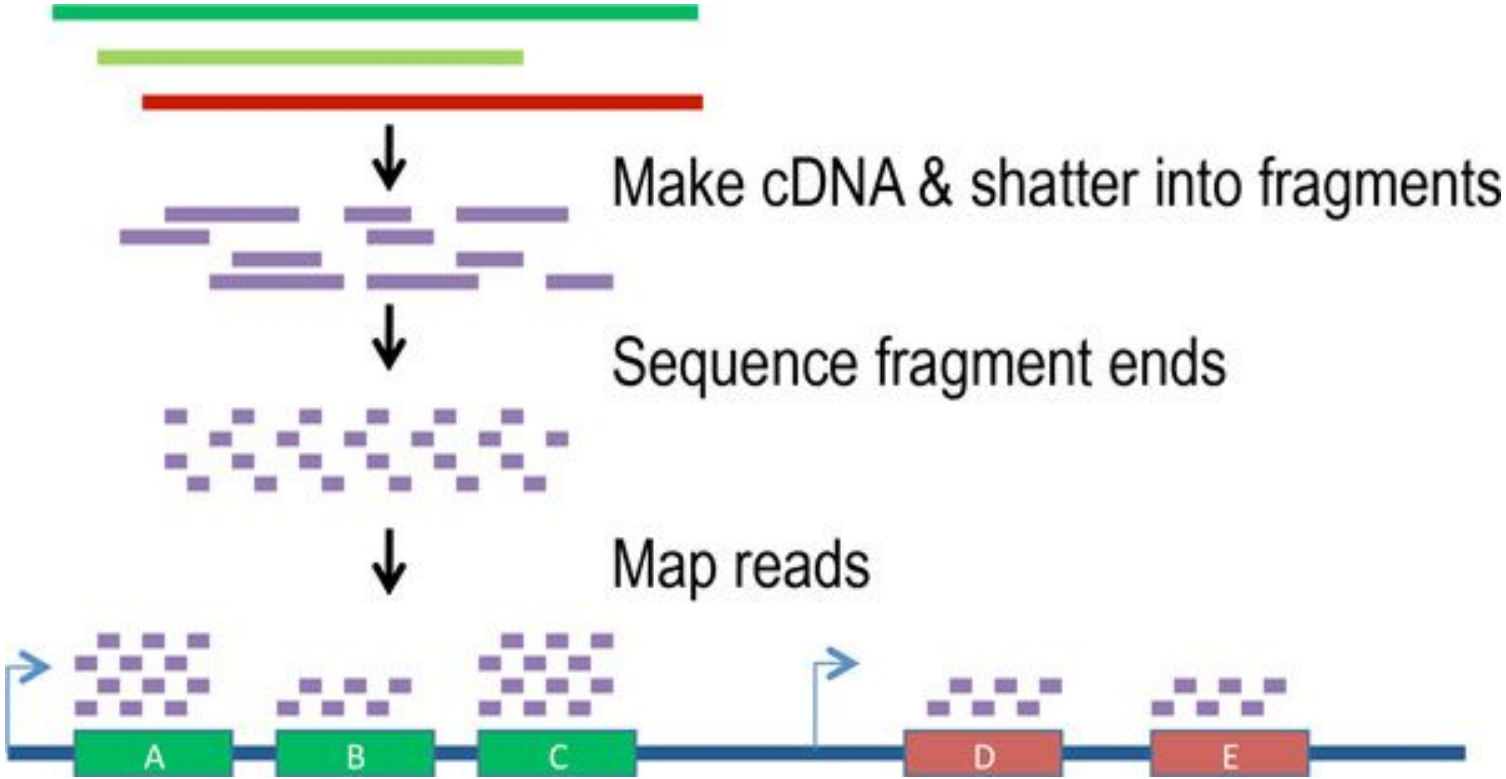
# Создание пайплайна для добавления результатов RNAseq в программу GeneQuery

Студенты: Лаврентий Данилов, Родион Сахабеев

Научный руководитель: Александр Предеус (Институт биоинформатики)

# RNAseq


Make cDNA & shatter into fragments
Sequence fragment ends
Map reads
A
B
C
D
E

# Система GeneQuery

http://genome.ifmo.ru/genequery/searcher/

1. Экспрессия генов характеризуется уровнем мРНК
2. Современные методы позволяют измерить одновременно все мРНК
3. Экспрессия сильно меняется не только между разными тканями, но и между разными индивидуумами, после обработки лекарствами
4. GeneQuery позволяет получить кластера ко-экспрессирующихся генов, которые характеризуют фенотипы

# GeneQuery[α]

**Database species:** ○ Homo Sapiens ◉ Mus Musculus ○ Rattus Norvegicus

**Query species:** ○ Homo Sapiens ◉ Mus Musculus ○ Rattus Norvegicus

**Gene list** (separated by newline/whitespace/tab)

```
Col5a1
Tgm2
Gpc1
Phkg1
Efna1
Ampd3
Tktl1
Pnrc1
Plaur
Glrx
Maff
Serpine1
Cited2
Gapdh
Errfi1
Ets1
Aldoa
Tmem45a
Pdk1
Pgam2
Atf3
```

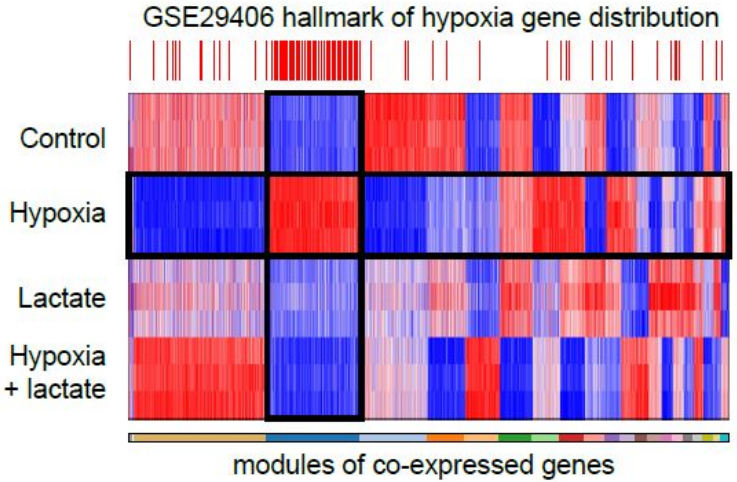
Search | Run example

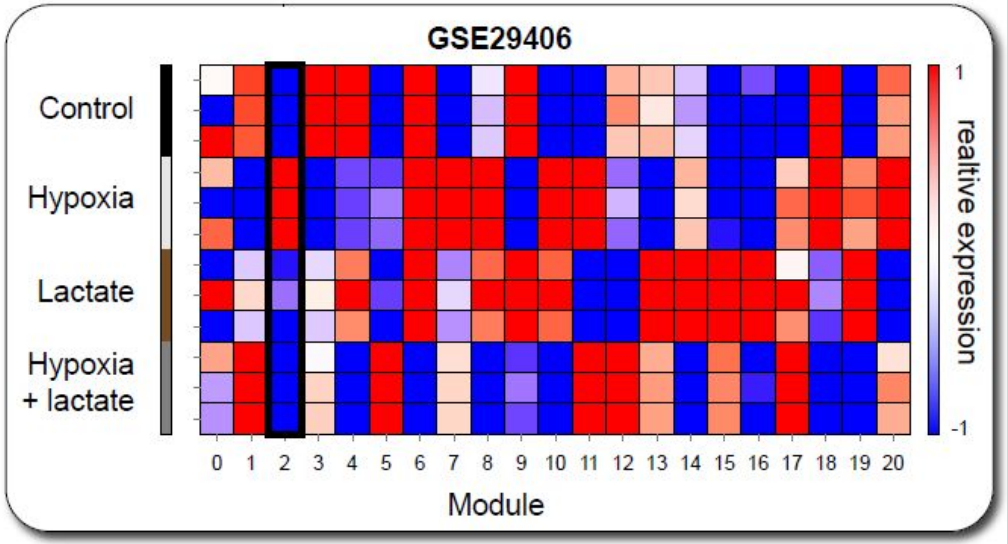
| | | | |
|---|---|---|---|
| **Modules** | 415 | **Detected gene format** | SYMBOL |
| **Detected groups** | 12 | **Genes entered** | 185 |
| **Min $\log_{10}$(adj.p$_{value}$)** | -36.14 | **Unique entrez IDs** | 185 |
| | | | show gene conversion table |
| | | **Apply orthology** | no |

Export result to CSV | ☐ Group results

| # | Experiment title |
|---|---|
| 1 | Gene expression in MEFs in response to treatment with dipyridyl and trichostatin A |
| 2 | Two transactivation mechanisms are responsible for the bulk of HIF-1alpha-responsive gene expression |
| 3 | Gene expression in hypoxic MEFs having only p300 and CBP with deleted CH1 domains |
| 4 | Whole gene expression data from osteocyte-like cell line MLO-Y4 under large gradient high magnetic field (LG-HMF) |

# А зачем все это надо...



# Планы

1. Разобраться с тем, что было сделано до нас

2. Добавить результаты RNAseq экспериментов в базу GeneQuery

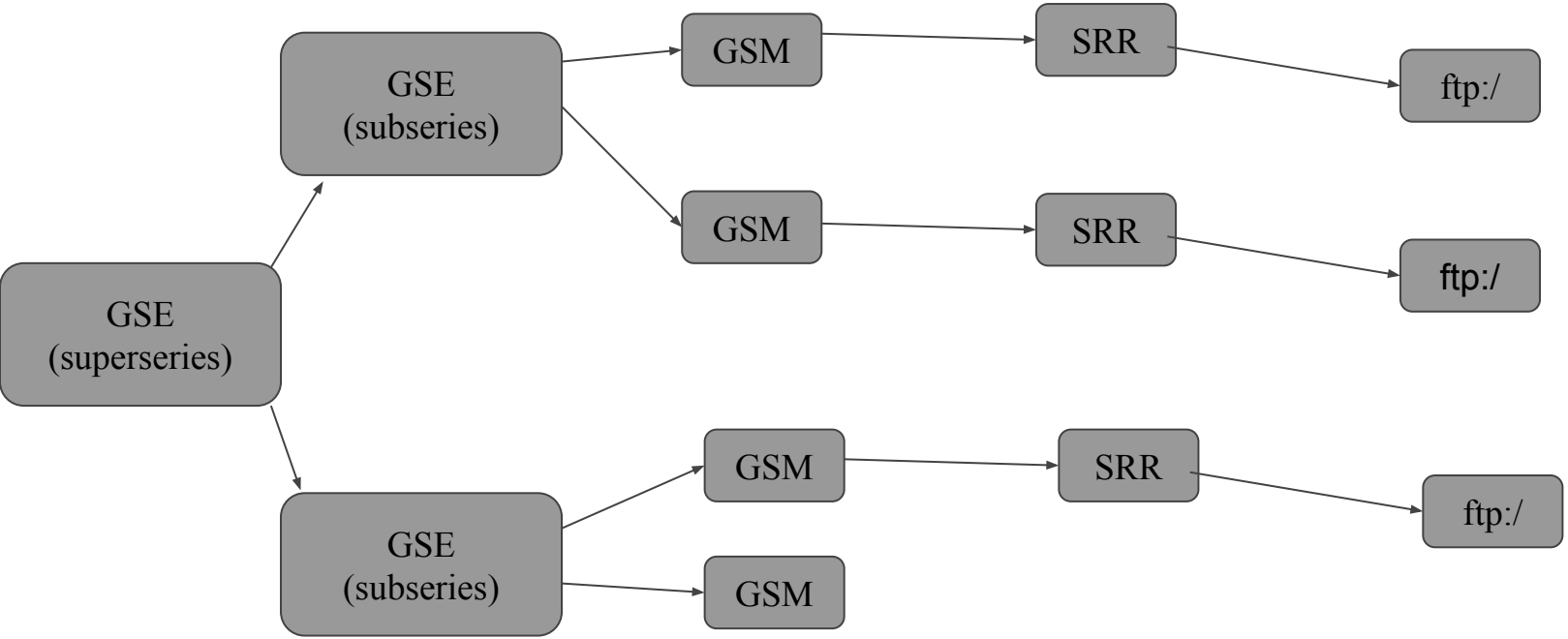
3. Разобраться с предлагаемыми новыми методами кластеризации, выбрать оптимальный

4. Применить новый метод кластеризации к собранным экспериментам, сравнить результаты

5. Добавить визуализацию кластеризации и систему

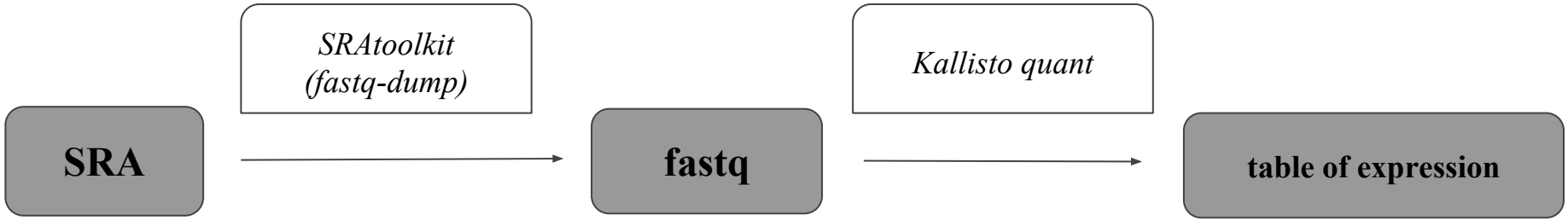
фидбэка для отбраковки некачественных экспериментов

# Структура данных



# Этапы обработки данных

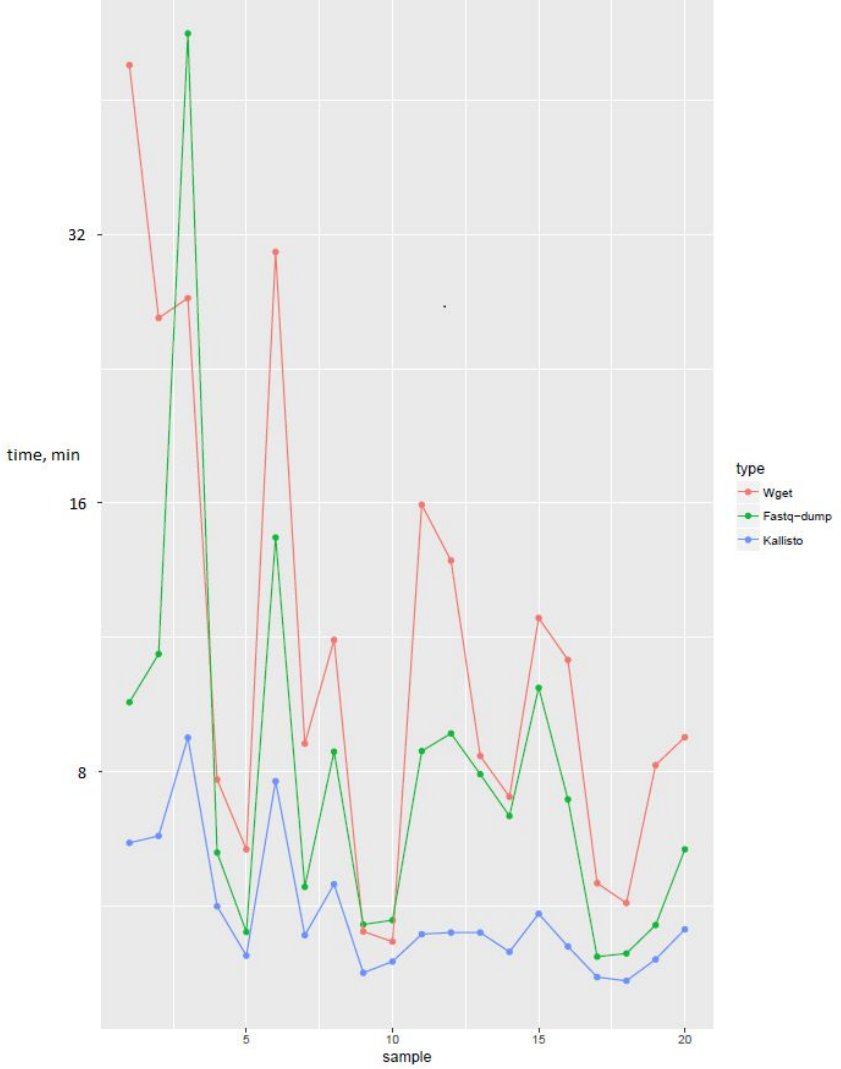
1. Создание последовательностей транскриптомов при помощи утилиты **rsem-prepare-reference** для мыши, крысы и человека
2. Создание индексов kallisto для данных транскриптонов **(kallisto index)**
3. Квантификация RNA-seq экспериментов при помощи индексов kallisto **(kallisto quant)**




4. Процесс квантификации экспериментов доведен до автоматизма

# Скорость работы

- Wget дольше всех
- Fastq-dump средне
- Kallisto быстрее всех



# Варианты оптимизации пайплайна

- Распараллеливание загрузки так, чтобы wget закачивал в 4 потока
- Использование модифицированного fastq-dump (parallel-fastq-dump)
- Реализация скрипта на python при помощи модуля multyprocessing

# Обработка результатов

- Подсчет уровня экспрессии каждого гена в конкретном эксперименте (суммирование транскриптов)

- Статистически обработать результаты при помощи анализа взвешенных сетей коэкспрессии генов (WGCNA) и выявление дифференциально экспрессированных генов

# А что хочется сделать дальше...

- Оптимизация ошибок кластеризации разных типов экспериментов в системе GeneQuery (для RNAseq, Chipseq и DNAseq)

- Добавление визуализации кластеризации и системы фидбэка для отбраковки некачественных экспериментов

# Спасибо за внимание!