Анализ дифференциальной экспрессии методом RNA-seq

Анна Клепикова

Лаборатория геномики растений

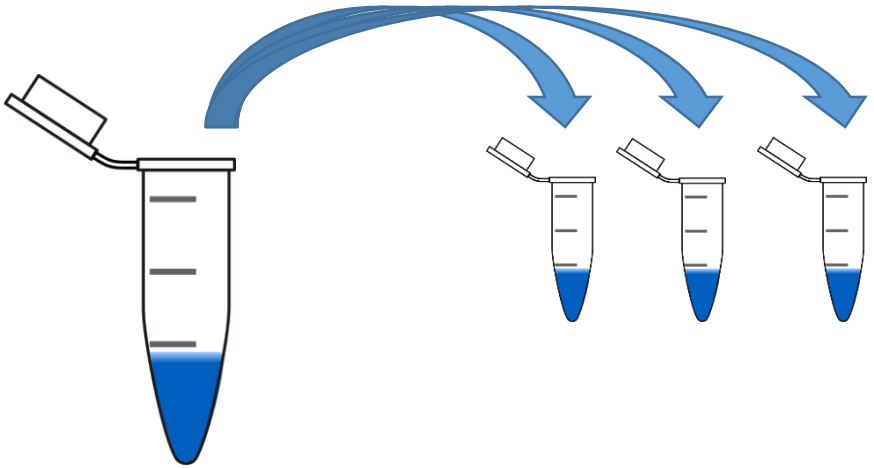
Институт проблем передачи информации РАН

Москва, 2019

Повторности

Биологические

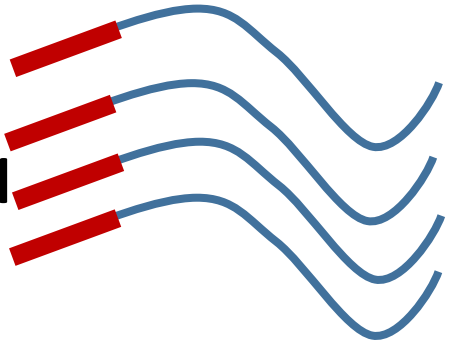
Технические



Пробоподготовка

Выделение РНК

Проверка ее целостности



Отбираем мРНК из образца

Убираем из образца рРНК (деплеция)

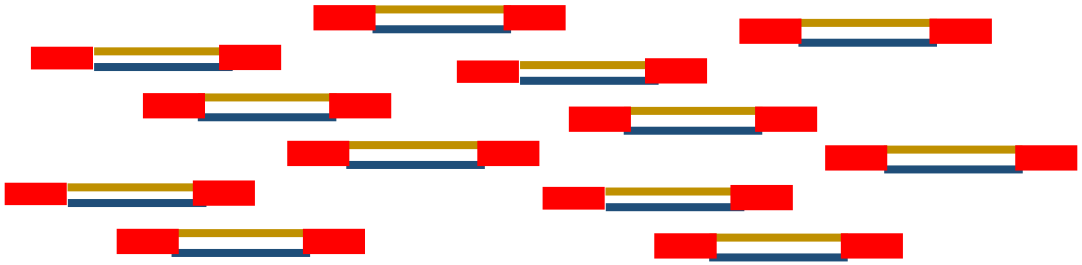
Дробление РНК

Синтез двуцепочечной кДНК

Фрагменты кДНК с адаптерами

Библиотека

Фрагменты кДНК с адаптерами



Секвенирование

Секвенирование

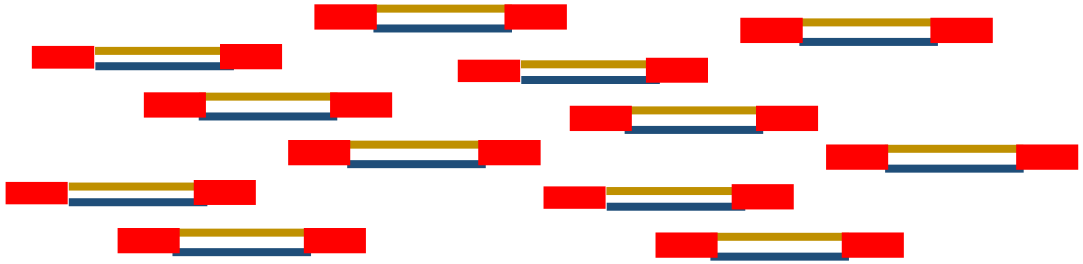
Можно вносить в секвенатор разное количество библиотеки

Глубина секвенирования библиотеки

Число чтений (ридов, reads):
Миллионы
Десятки миллионов
Сотни миллионов

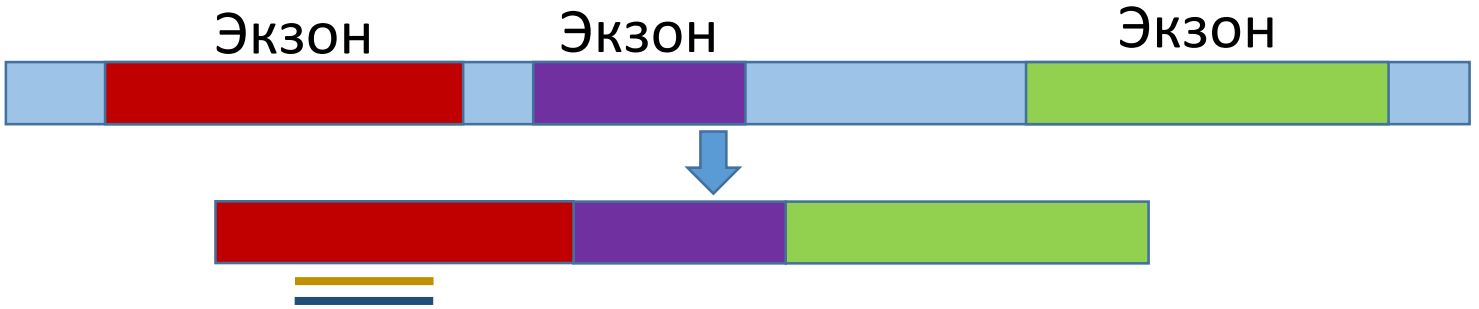
Препроцессинг данных

- Тримминг
- Картирование



ATTTTGGAATGCTGATCGTAAAGTC

- Подсчет числа чтений, приходящихся на данный ген



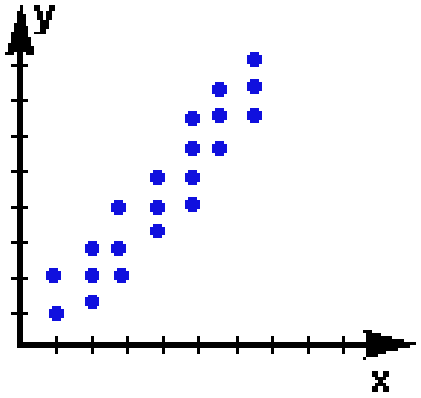
- Общее число чтений – на весь ген
- Число чтений, приходящееся на экзон
- Число чтений, приходящееся на сайт сплайсинга

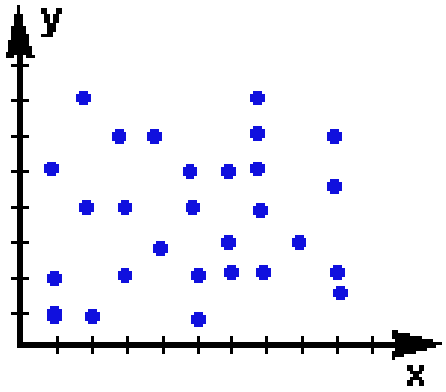
Read counts – число чтений для всех генов

Feature_ID	A1_R.A_11	A1_R.A_12	A2_R_21	A2_R_22	A3_S.R_31	A3_S.R_32	A6_S.H_61	A6_S.H_62
AT1G01010	175	262	670	680	262	343	100	156
AT1G01020	180	200	435	505	279	251	396	462
AT1G01030	37	33	66	80	19	16	2	0
AT1G01040	749	980	1364	1375	1110	1409	1163	1105
AT1G01046	0	0	0	0	0	0	0	0
AT1G01050	2744	2609	2118	2109	1982	1724	1628	1409
AT1G01060	90	125	108	81	82	89	18	10
AT1G01070	387	565	333	292	371	325	237	226
AT1G01073	0	0	0	0	0	0	0	0
AT1G01080	53	94	133	150	38	43	1873	1720
AT1G01090	7017	5594	4845	4678	5806	6869	4613	4834
AT1G01100	9115	7703	7460	9110	6900	7985	3139	3710
AT1G01110	328	407	379	441	411	556	44	103
AT1G01115	0	0	0	0	0	0	0	0
AT1G01120	333	510	1187	1344	418	490	2764	3357
AT1G01130	79	52	99	85	39	45	333	215
AT1G01140	51	53	53	53	54	66	1800	1794
AT1G01150	0	0	5	10	3	2	0	0
AT1G01160	498	493	823	813	584	632	661	586
AT1G01170	10	18	42	36	20	13	1046	1261
AT1G01180	99	173	165	143	122	105	180	262

Повторности

Должны быть похожи друг на друга





Квадрат коэффициента корреляции Пирсона > 0,92

Что на самом деле считаем

Экспрессия гена
=
число мРНК этого гена в образце

Число чтений
пропорционально
числу мРНК этого гена в образце

$$y = k*x + b$$

Нормализация на размер библиотеки

Размер библиотеки		Библиотека 1	Библиотека 2
		20M	10M
Ген 1	Число чтений	1000	500
	мРНК	10	10

Размер библиотеки 1 / размер библиотеки 2

20M / 10M = 2

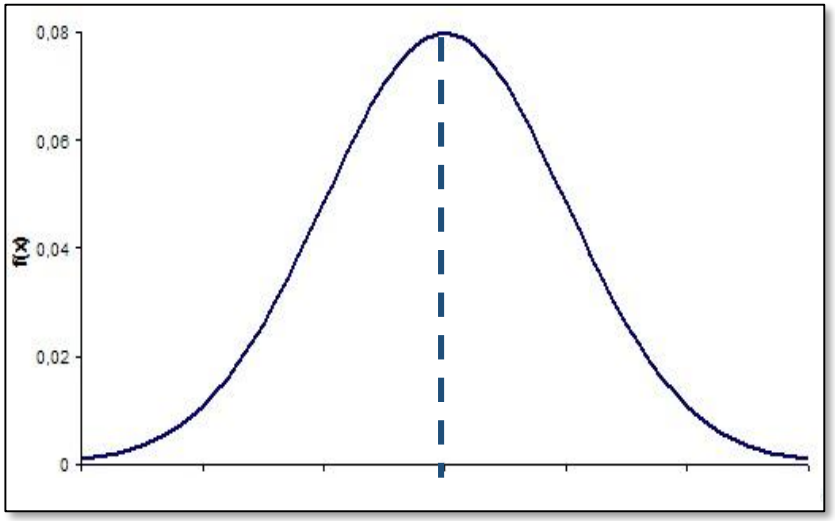
	Ген 1	Ген 2
Длина	1000	5000
Число чтений	100	500
мРНК	10	10

Есть более сложные способы нормализации

Статистика

- Случайная величина: число чтений, приходящихся на ген i в образце j.

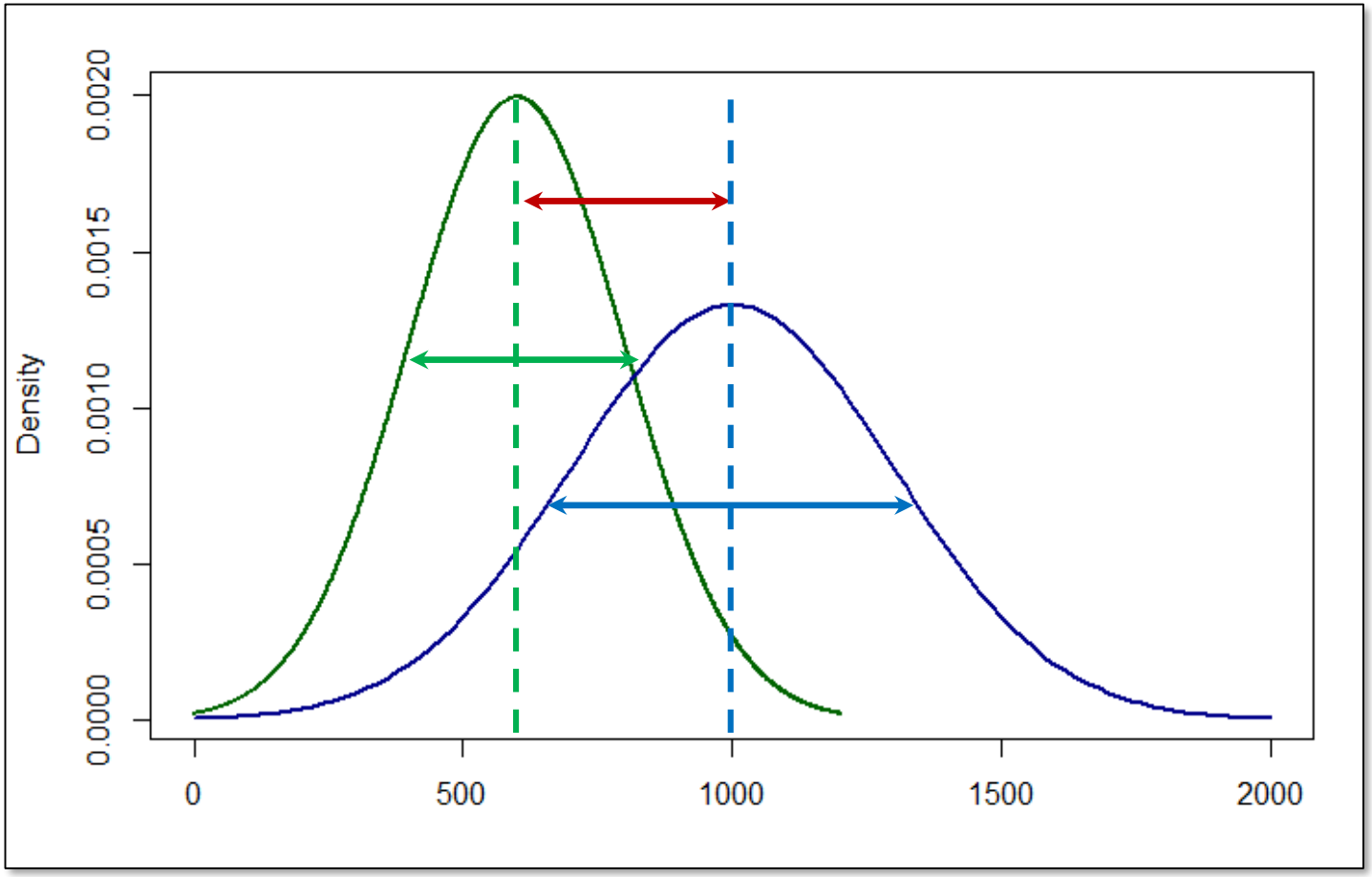
Feature_ID	A1_R.A_11	A1_R.A_12	A2_R_21	A2_R_22
AT1G01010	175	262	670	680
AT1G01020	180	200	435	505
AT1G01030	37	33	66	80



Параметры распределения:
Среднее значение
Дисперсия

Статистика

Измерение в случайной величины (числа чтений) в одних условиях и в других условиях



Статистический критерий (тест)

Случайная величина: число ридов, приходящихся на ген i в образце j

Нужно узнать, если различия между числом чтений на ген в образце 1 и образце 2

Используется статистический критерий – величина, посчитанная по параметрам распределения, и дающая ответ на вопрос – есть ли разница между ними.

Статистический критерий (тест)

Точный критерий Фишера

$$F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \sim F(m-1, n-1)$$

P-value (p-значение)

Уровень изменения – fold change
Экспрессия гена в образце 1 / экспрессия гена в образце 2

Пороги

P-value < 0,05

Поправка на множественное тестирование < 0,05

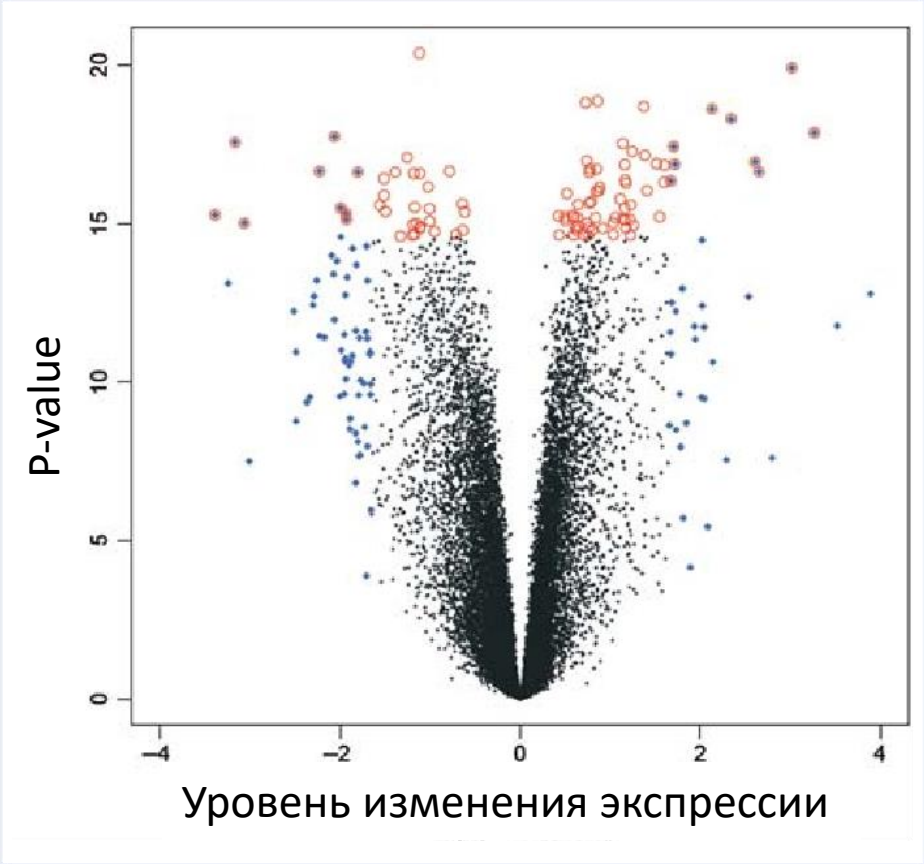
Уровень изменения – fold change > 2

Дифференциально экспрессирующиеся гены

Как изобразить результаты

MA-plot

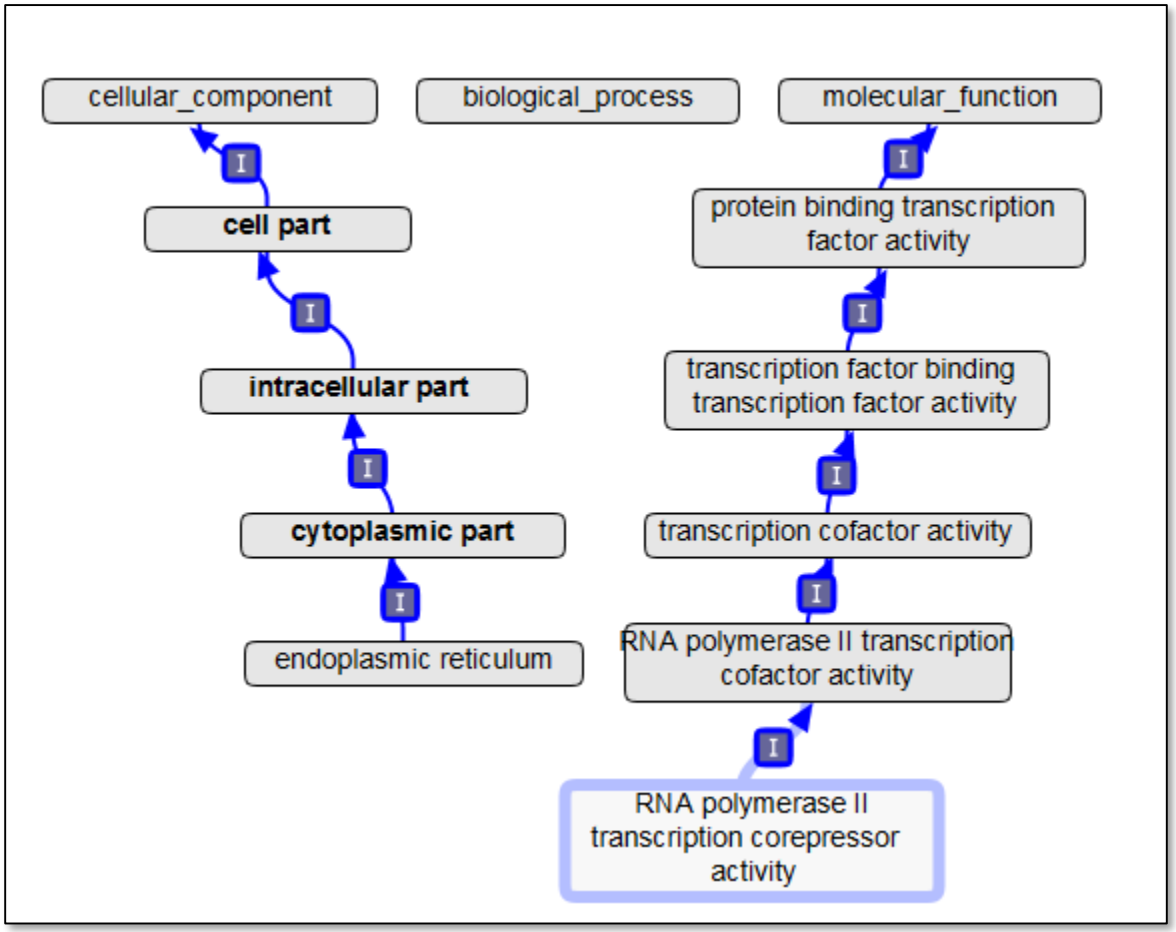
Volcano plot



Программы для поиска дифференциально экспрессии

- DESeq (DESeq2)
- EdgeR
- baySeq
- DEGSeq
- PoissonSeq
- Cuffdiff
- NOISeq

Gene Ontology – Генная Онтология



Gene Ontology – Генная Онтология

- Энричмент (обогащение) терминами GO

	Список DE генов	Список всех генов организма
Термин GO:0005783	100	200
Остальные термины	900	20000

1/10	1/100

Category	Term	Fold Enrichment
GOTERM_BP_FAT	GO:0046686~response to cadmium ion	13.5
GOTERM_BP_FAT	GO:0010038~response to metal ion	13.2
GOTERM_BP_FAT	GO:0009628~response to abiotic stimulus	11.9
GOTERM_BP_FAT	GO:0010035~response to inorganic substance	12.4
GOTERM_BP_FAT	GO:0006970~response to osmotic stress	12.3
GOTERM_BP_FAT	GO:0009651~response to salt stress	12.3

Альтернативный сплайсинг