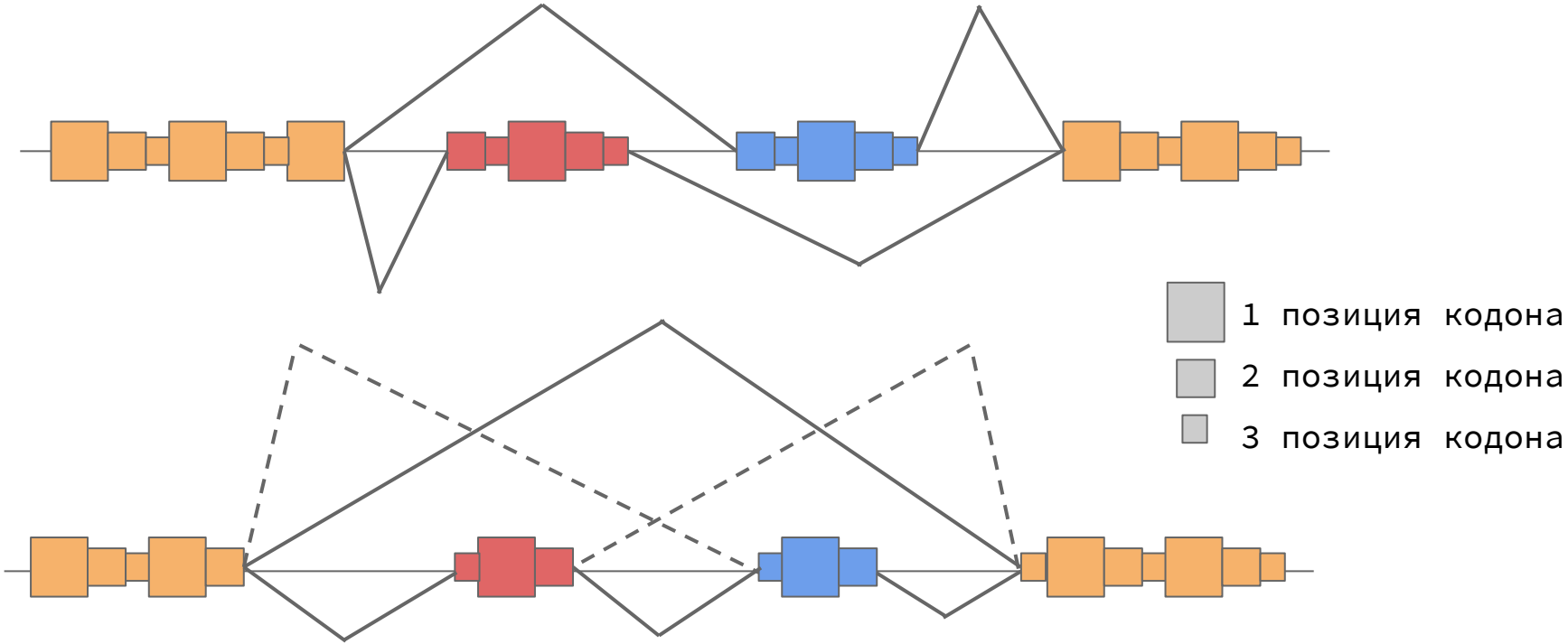
Экзоны-призраки

Даша Латорцева, Диана Марцинова, Кирилл Медведев,

Женя Ходжаева, Зоя Червонцева

Альтернативный сплайсинг и рамка считывания



Взаимоисключающие экзоны: одинаковый остаток от деления длины на 3
Независимые (кассетные) экзоны: длина делится на 3

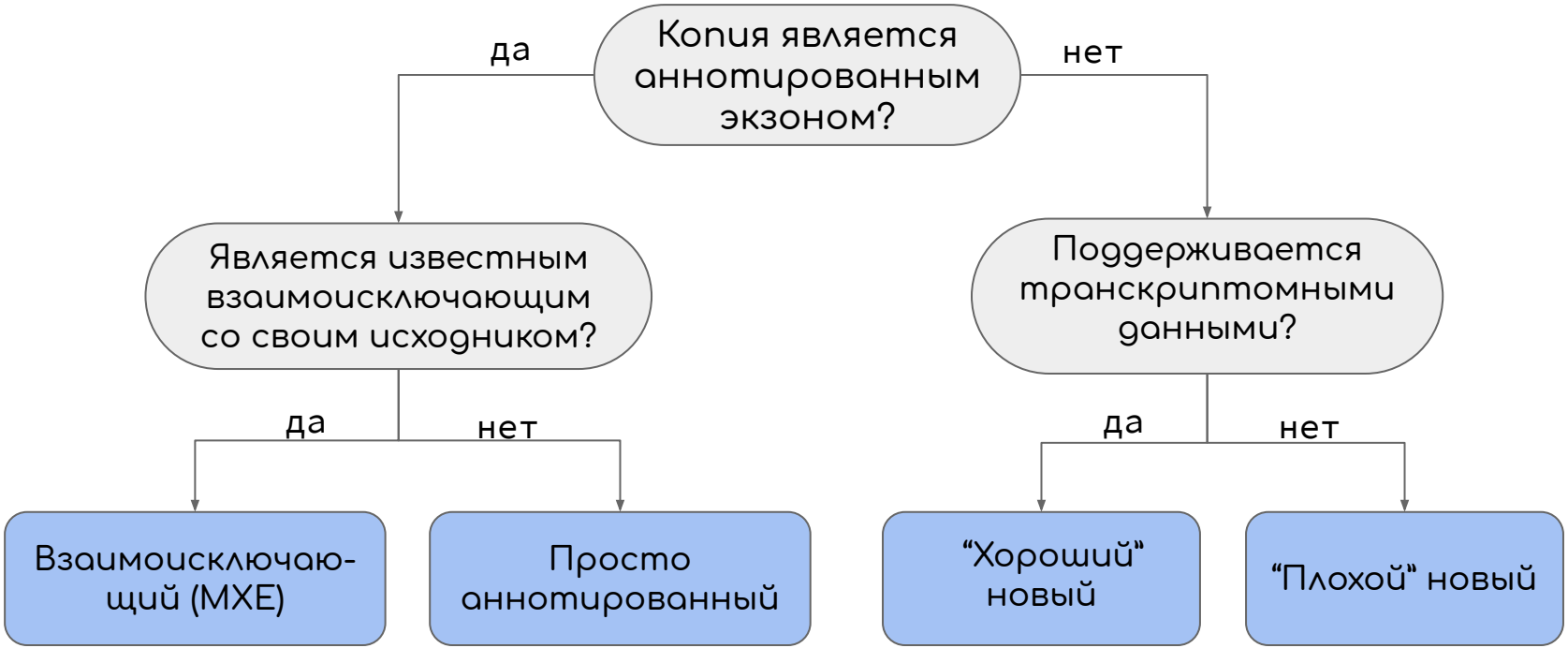
Исходные данные

Наши коллеги взяли все аннотированные экзоны (далее - *исходники*) в геноме человека и нашли все последовательности, похожие на эти экзоны и находящиеся внутри тех же генов (далее - *копии*).

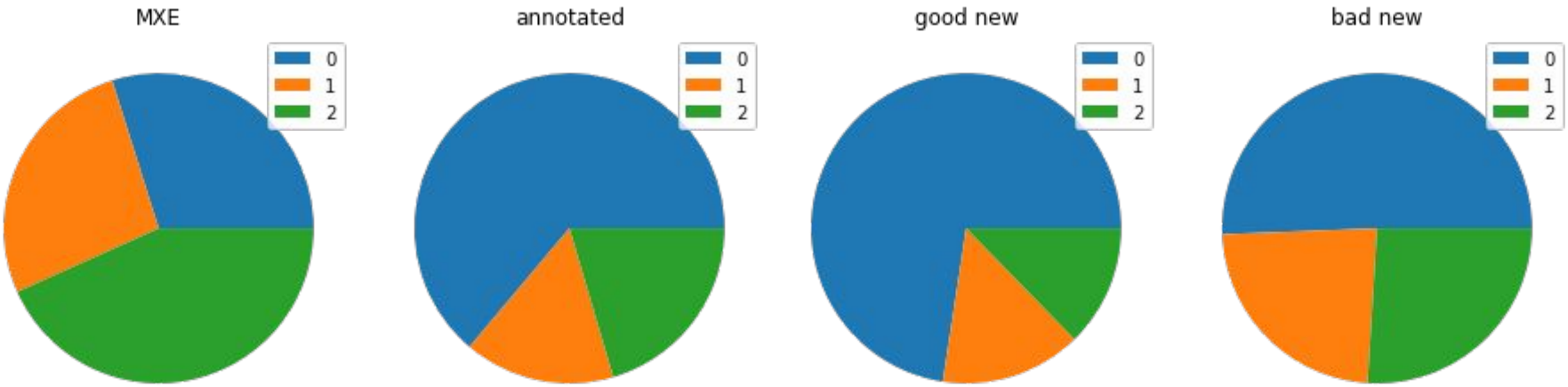
Есть **таблица из ~30 тыс. пар** исходник-копия с identity 55-95%. Для них мы посчитали много разных характеристик (см. далее).

Disclaimer: Кто из каждой пары появился раньше в эволюции, мы не знаем.

Деление копий на 4 класса

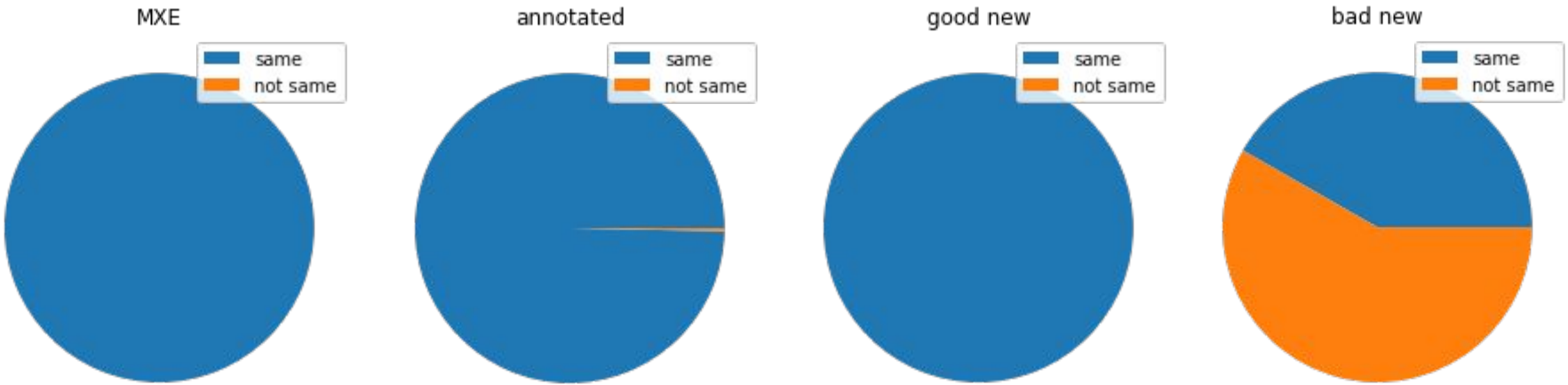


Хорошие экзоны предпочитают иметь длину, кратную трем - кроме взаимоисключающих, им все равно



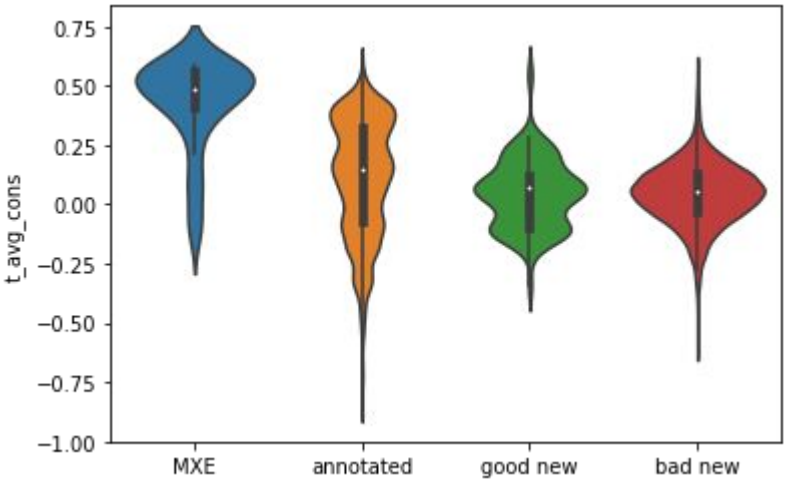
Цветами показаны группы экзонов, длины которых при делении на 3 дают соответствующие остатки. Предпочтение остатка ноль может объясняться сохранением рамки считывания при включении/исключении экзона.

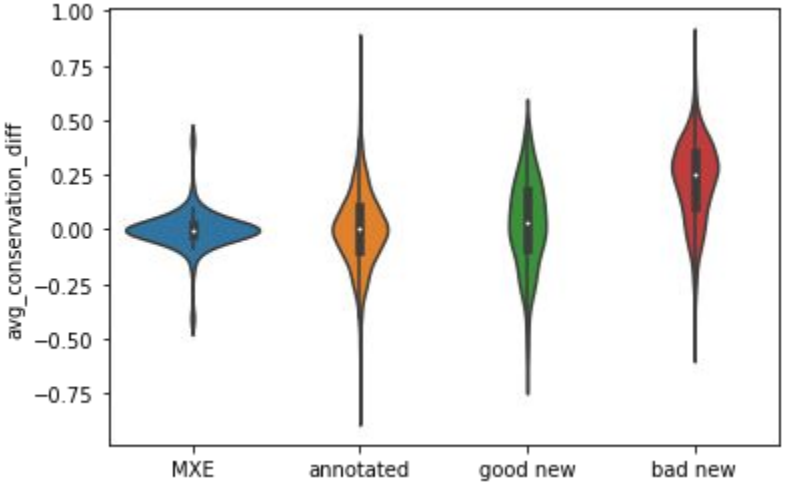
Хорошие копии предпочитают иметь такой же остаток деления длины на три, что и исходный экзон



Это может позволять копии использоваться вместо исходника.
Оранжевые в annotated - возможно, это регуляторные копии, которые специально сбивают рамку.

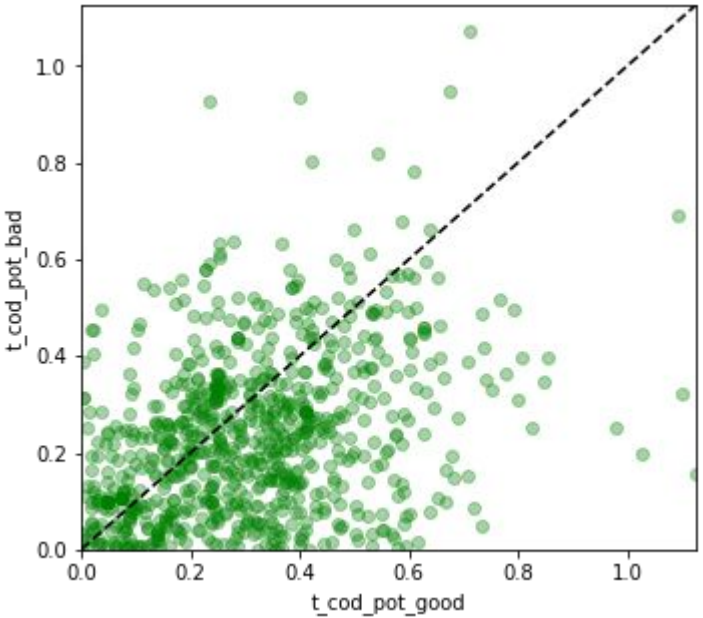
Взаимоисключающие экзоны консервативны, и плохие копии менее консервативны, чем исходники

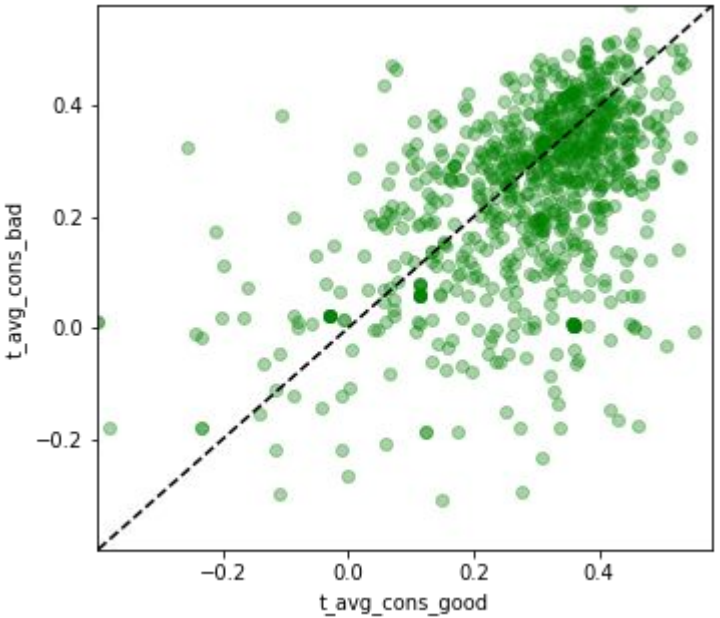




Слева - распределения консервативности копий, справа - распределения того, насколько исходник консервативнее копии. Консервативность была посчитана как среднее по экзону значение трека phyloP для 46 позвоночных.

Из двух копий одного экзона хорошая копия чаще имеет больший кодирующий потенциал и большую консервативность





... это можно заключить из того, что ниже диагонали точек больше, чем выше диагонали. Каждая точка - пара копий одного исходника. По оси x значение для хорошей, по y - для плохой.

Некоторые распространённые дуплицирующиеся домены и их аннотации

*Содержит большое количество цистеиновых остатков

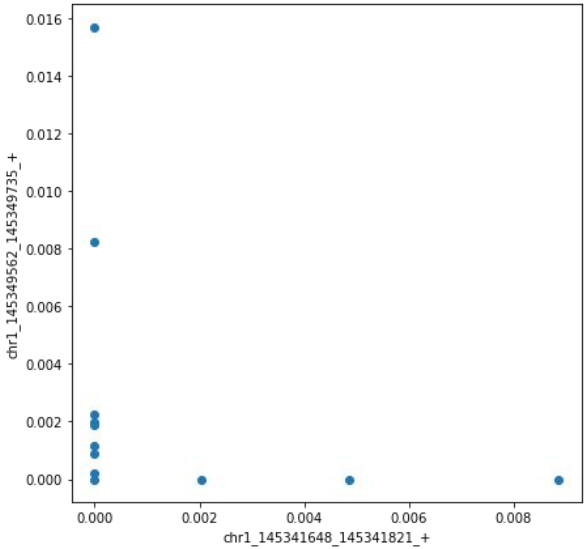
SRCR*	Binds to ligands (involved in immune response)
FXa_inhibition	A short domain of coagulation enzyme factor Xa
Ldl_recept_a	Cell surface receptors
hEGF*	hEGF involved in growth and proliferation of cells, in proteins of neurogulin and selectins
TILa*	Occurs along side the TIL PF01826 domain and is likely to be a distantly related relative
Myb_DNA-bind_4	Greatly expanded in plants and related to transposons
TIL*	Trypsin Inhibitor, found in many extracellular proteins
fn1*	Fibronectin type I domain involved in fibrin-binding
fn2*	Fibronectin type II domain, collagen-binding
EGF_3	Includes the C-terminal domain of the malaria parasite MSP1 protein
GATA*	Binds to DNA. Two GATA zinc fingers are found in the GATA transcription factors
C8*	Found in disease-related proteins including von Willebrand factor, Alpha tectorin, Zonadhesin and Mucin
Sushi	Found in variety of complement and adhesion proteins

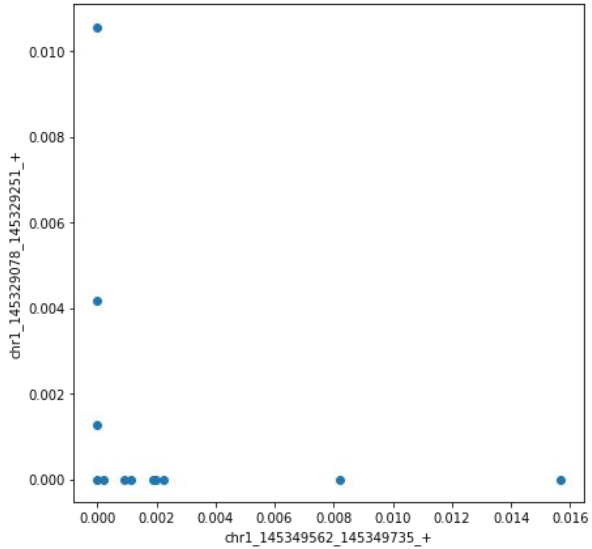
Очевидные (и не очень!) корреляции между параметрами

**Полная расшифровка названий параметров - на слайде 14

	ePI	qjunc_l	qjunc_r	tjunc_l	tjunc_r	q_cod_pot	q_avg_cons	t_cod_pot	t_avg_cons	q5ss_score	q3ss_score	t5ss_score	t3ss_score	correlation	cod_pot_dif
ePI		-0.311	-0.295	-0.323	-0.316	-0.27	-0.481	-0.266	-0.476	-0.118	-0.039	-0.115	-0.02	-0.315	0.005
qjunc_l	-0.311		0.897	0.734	0.729	0.211	0.511	0.235	0.472	0.155	-0.102	0.134	-0.132	0.177	-0.019
qjunc_r	-0.295	0.897		0.724	0.734	0.237	0.505	0.235	0.459	0.147	-0.164	0.14	-0.141	0.181	-0.002
tjunc_l	-0.323	0.734	0.724		0.908	0.229	0.479	0.223	0.513	0.14	-0.1	0.163	-0.087	0.189	0.007
tjunc_r	-0.316	0.729	0.734	0.908		0.227	0.469	0.249	0.507	0.15	-0.113	0.152	-0.143	0.2	-0.005
q_cod_pot	-0.27	0.211	0.237	0.229	0.227		0.327	0.341	0.377	0.093	0.096	0.093	0.117	0.238	0.543
q_avg_cons	-0.481	0.511	0.505	0.479	0.469	0.327		0.383	0.725	0.135	0.187	0.17	0.121	0.271	-0.061
t_cod_pot	-0.266	0.235	0.235	0.223	0.249	0.341	0.383		0.319	0.087	0.126	0.084	0.083	0.24	-0.527
t_avg_cons	-0.476	0.472	0.459	0.513	0.507	0.377	0.725	0.319		0.164	0.127	0.129	0.177	0.258	0.068
q5ss_score	-0.118	0.155	0.147	0.14	0.15	0.093	0.135	0.087	0.164		0.024	0.364	0.07	0.064	0.034
q3ss_score	-0.039	-0.102	-0.164	-0.1	-0.113	0.096	0.187	0.126	0.127	0.024		0.075	0.411	-0.004	-0.031
t5ss_score	-0.115	0.134	0.14	0.163	0.152	0.093	0.17	0.084	0.129	0.364	0.075		0.021	0.052	-0.01
t3ss_score	-0.02	-0.132	-0.141	-0.087	-0.143	0.117	0.121	0.083	0.177	0.07	0.411	0.021		-0.018	0.045
correlation	-0.315	0.177	0.181	0.189	0.2	0.238	0.271	0.24	0.258	0.064	-0.004	0.052	-0.018		-0.004
cod_pot_dif	0.005	-0.019	-0.002	0.007	-0.005	0.543	-0.061	-0.527	0.068	0.034	-0.031	-0.01	0.045	-0.004	

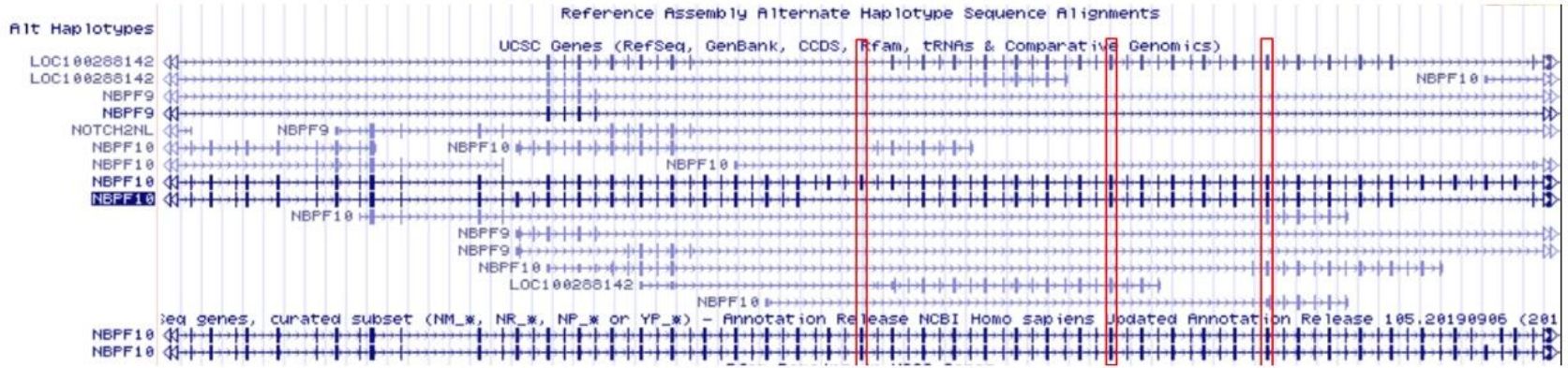
Экспрессии почти всех пар исходник-копия положительно коррелируют между собой в 16 тканях. Одно из немногих исключений - ген NBPF10





Каждая точка - одна из тканей. По осям отложены экспрессии для исходника и копии. Видно, что экспрессия одного в каждой из пар исключает экспрессию другого.

Ген NBPF10 состоит из многочисленных дупликаций и имеет затейливый сплайсинг. Связан с болезнями мозга. Появился у приматов



Красными прямоугольниками отмечены анти-коррелирующие экзоны.
Первый анти-коррелирует со вторым и с третьим.

To take home:

- В геноме человека есть много дуплицированных экзонов
- Часть из них включаются в транскрипты, часть нет
- Рабочие копии отличаются от нерабочих (призраков) рядом свойств: остаток от деления длины на 3, сохранение кодирующего потенциала, сохранение консервативности

Благодарности

Тимофею Иванову – за данные по дупликациям

Дмитрию Первушину – за постановку задачи

Михаилу Сергеевичу Гельфанду – за обсуждения

**Расшифровка названий параметров

ePI - степень сходства исходника и копии
qjunc_l - число ридов, подтверждающих наличие левого сайта сплайсинга исходника по данным РНК-секвенирования
qjunc_r - число ридов, подтверждающих наличие правого сайта сплайсинга исходника по данным РНК-секвенирования
tjunc_l - число ридов, подтверждающих наличие левого сайта сплайсинга копии по данным РНК-секвенирования
tjunc_r - число ридов, подтверждающих наличие левого сайта сплайсинга копии по данным РНК-секвенирования
q_cod_pot - кодирующий потенциал исходника
q_avg_cons - средняя консервативность исходника
t_cod_pot - кодирующий потенциал копии
t_avg_cons - средняя консервативность копии
q5ss_score - сила сайта сплайсинга на правом конце исходника
q3ss_score - сила сайта сплайсинга на левом конце исходника
t5ss_score - сила сайта сплайсинга на правом конце копии
t3ss_score - сила сайта сплайсинга на левом конце копии
correlation - корреляция между экспрессиями исходника и копии в 16-ти тканях
cod_pot_dif - разница в кодирующих потенциалах исходника и копии