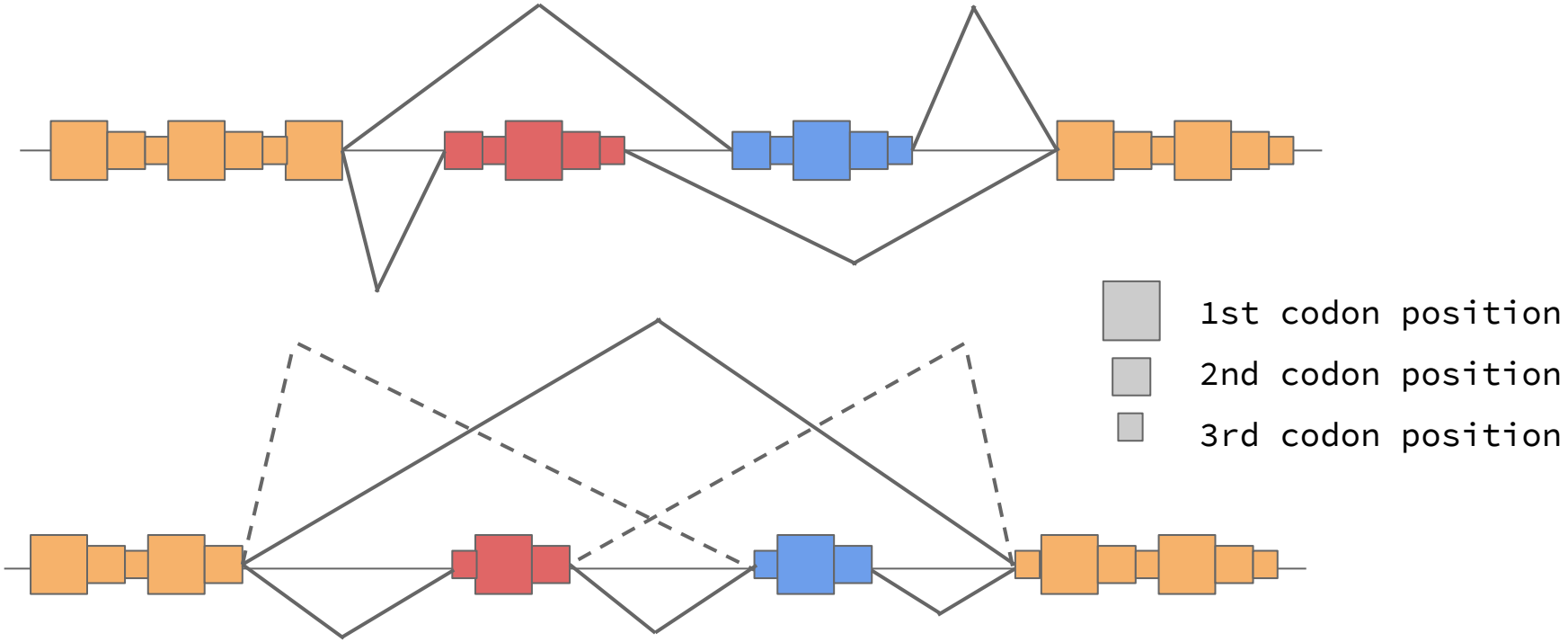
Ghost Exons

Dasha Latortseva, Diana Marcinova, Kirill Medvedev,

Jack Khodzhaeva, Zoe Chervontseva

Alternative splicing and reading frame



Mutually exclusive exons: the same remainder of division by three
Independent (cassette) exons: the length is divisible by three

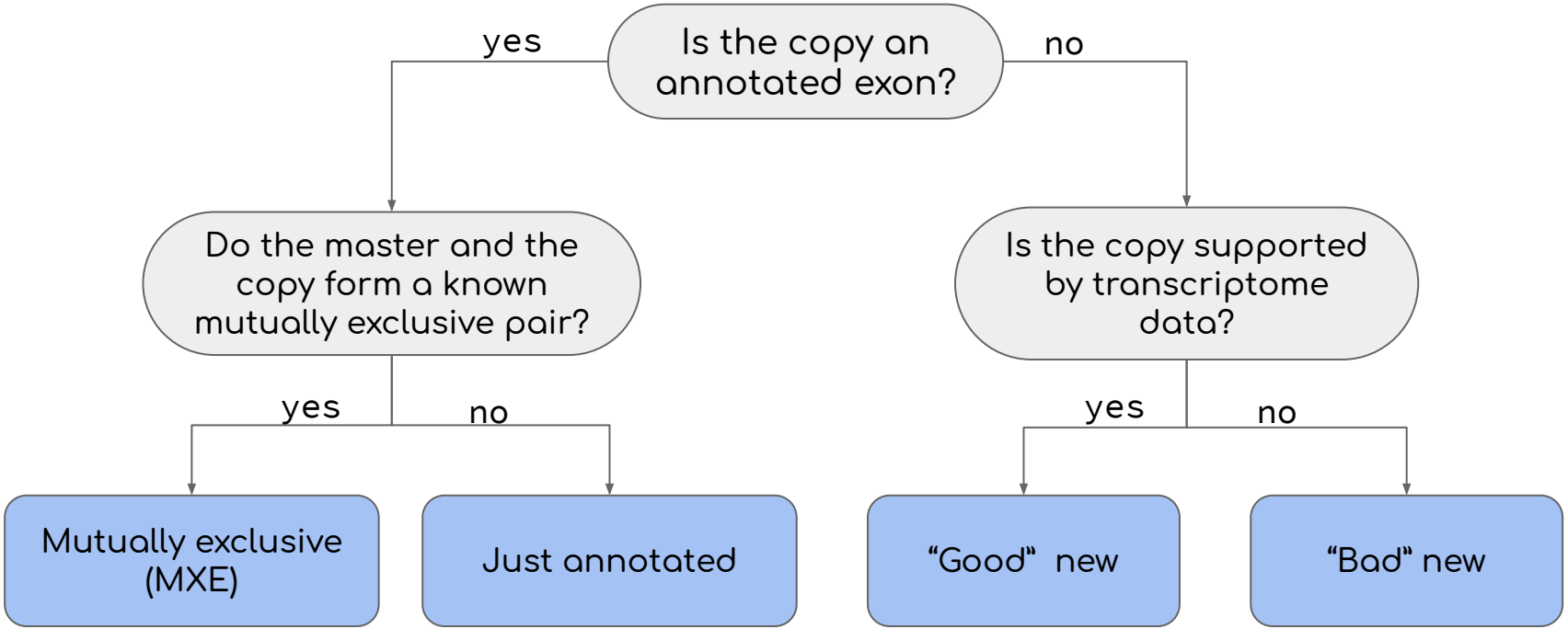
Initial data

Our colleagues collected all annotated exons in the human genome (further - *masters*) and found all sequences similar to these exons in the same genes (further - *copies*).

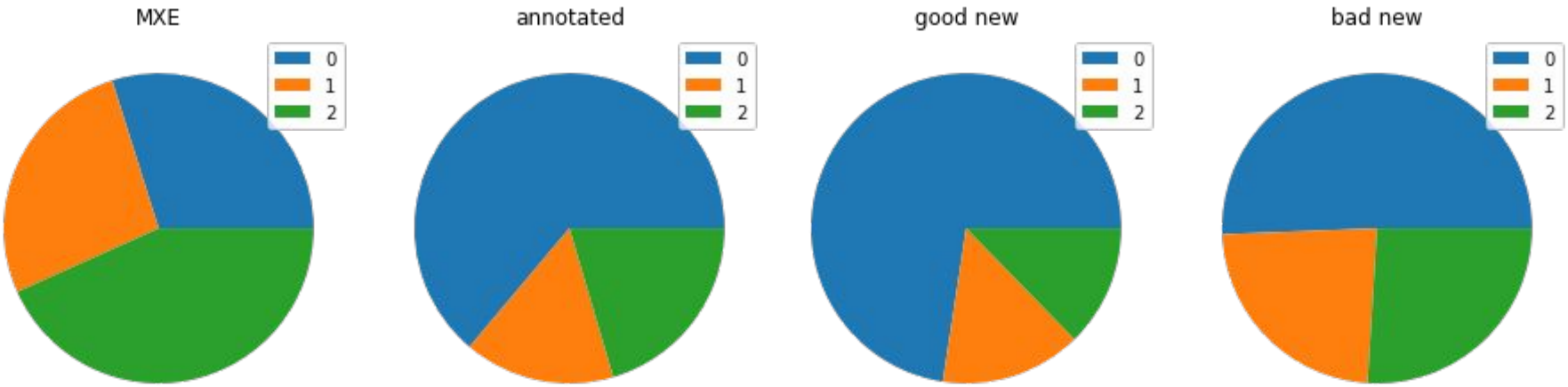
This yielded a table of **~30 thousands of master-copy pairs** having identity of 55-95%. We calculated and explored various properties of these pairs.

Disclaimer: we do not know which exon in a pair emerged first in evolution.

Copies classification: 4 classes

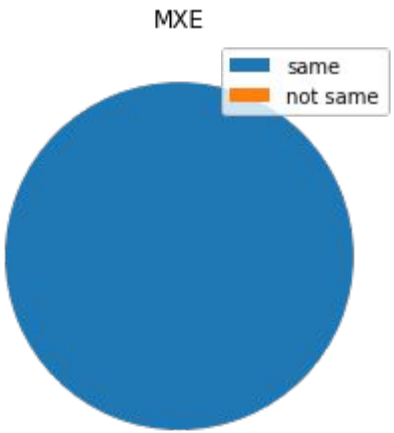


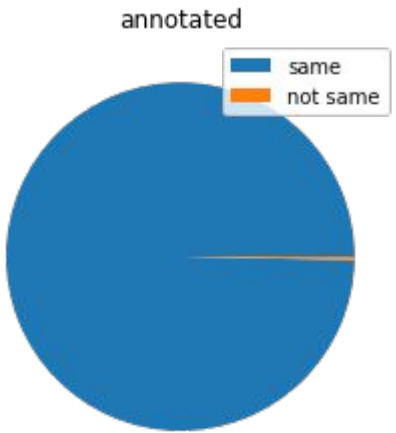
Good exons tend to have length divisible by three - except for mutually exclusive exons that do not care

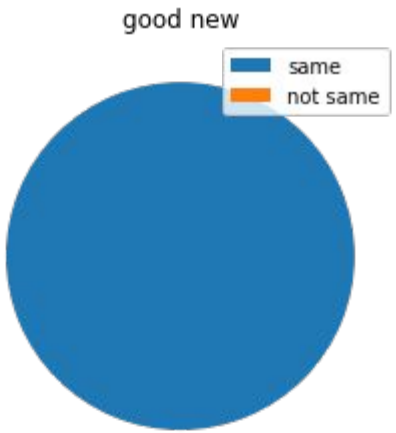


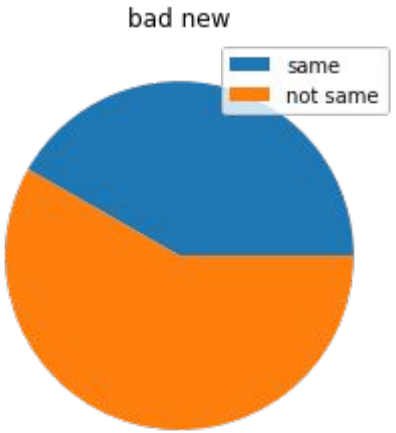
Colours denote exons by reminders of division by three (see the legend). Preference of zero reminder may be explained by maintaining the coding frame regardless of the exon's inclusion or exclusion.

Good copies tend to have the same remainder modulo three as the masters



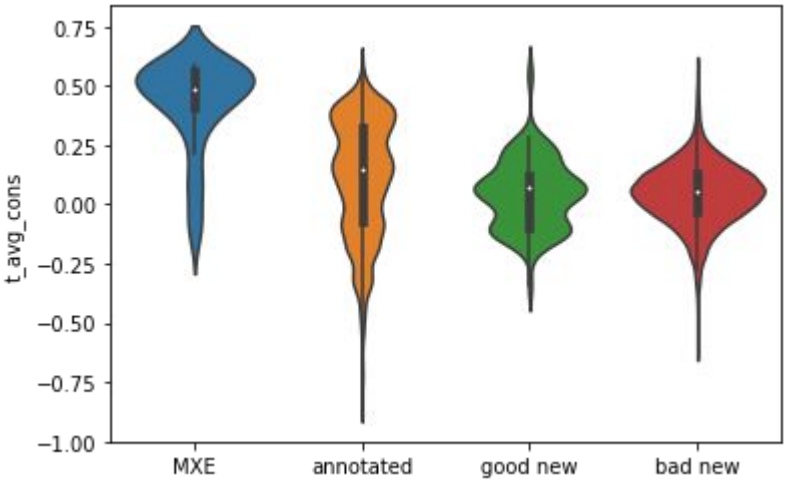


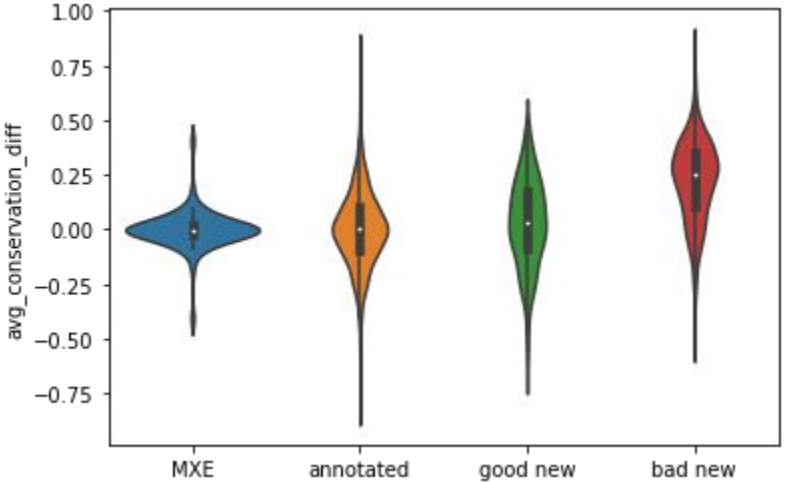




This could help the copy to be used instead of the master.

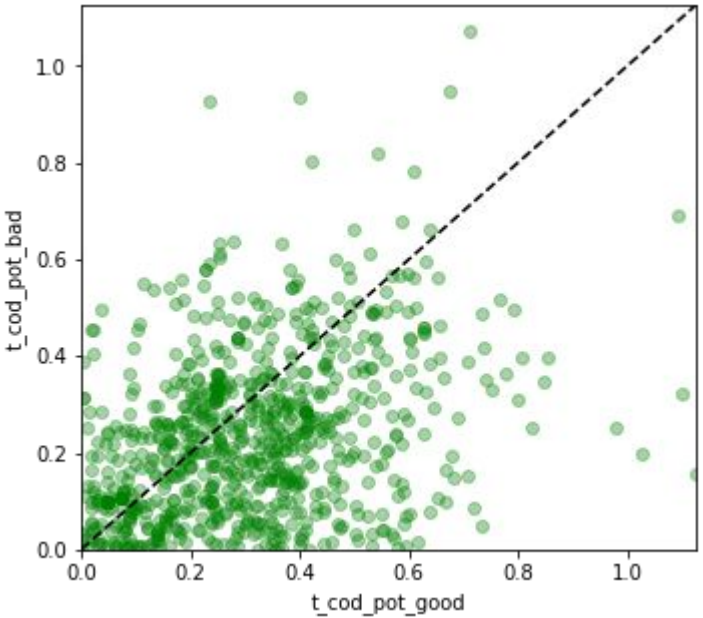
Mutually exclusive exons are conserved.
Bad copies are less conserved than masters

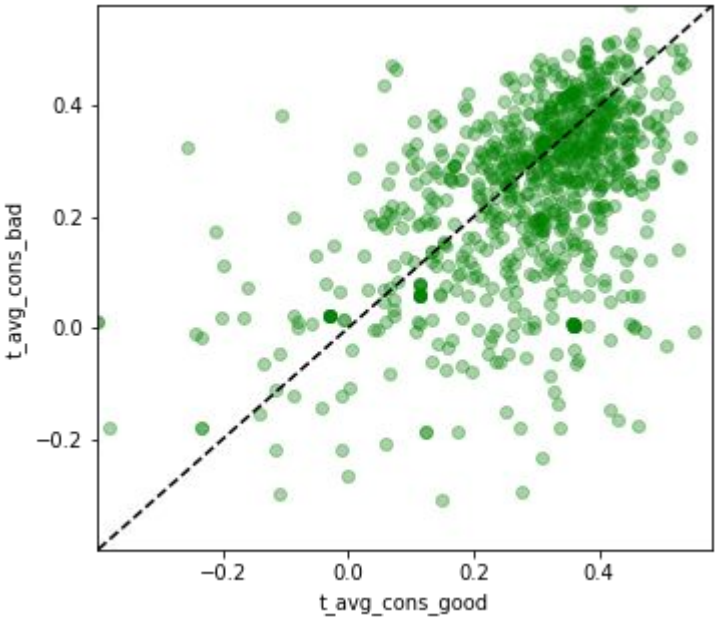




Left - conservativity of copies. Right - the difference in conservativity between masters and copies. Conservativity is defined as the average (over exon length) value of the phyloP track.

Among two copies of the same master, the good one tends to have larger coding potential and be more conserved





... indeed, there are more dots below the diagonal than above. Each dot represents a pair of copies for one original. The X axis is the analyzed value for the good copy ; the Y axis is the value of the bad one.

Some frequently duplicated protein domains

*Contains lot's of conserved cysteines

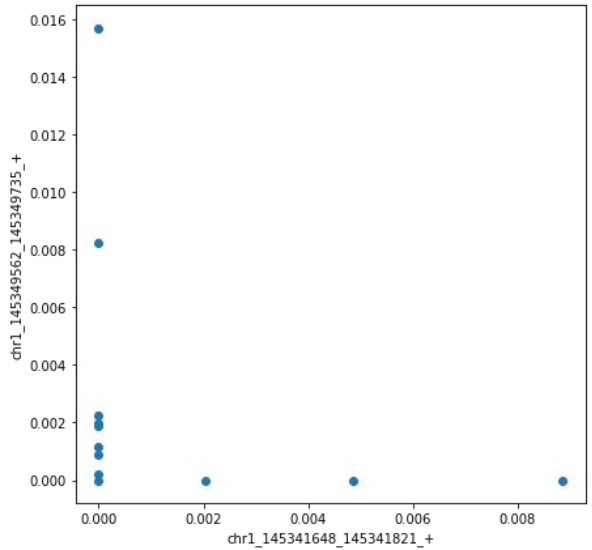
SRCR*	Binds to ligands (involved in immune response)
FXa_inhibition	A short domain of coagulation enzyme factor Xa
Ldl_recept_a	Cell surface receptors
hEGF*	hEGF involved in growth and proliferation of cells, in proteins of neurogulin and selectins
TILa*	Occurs along side the TIL PF01826 domain and is 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

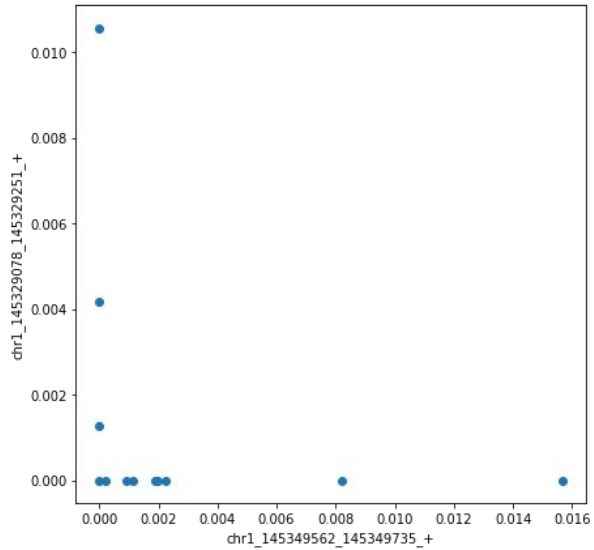
Obvious (and not so!) correlations among the parameters

**See slide 14 for the complete list of parameters meanings

	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	

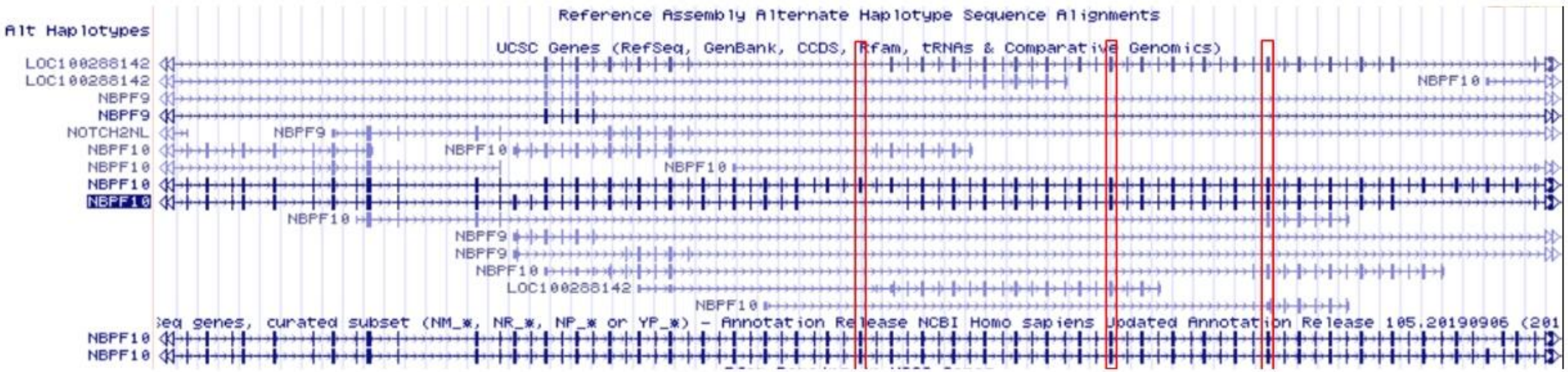
Expressions of almost all master-copy pairs are positively correlated in 16 tissues. One of rare exceptions is the gene NBPF10





Each dot represents a tissue. The X axis shows expression of the master, the Y axis shows expression of the copy. If the master is expressed, the copy is not, and vice versa.

The gene NBPF10 consists of multiple duplications and has a complicated splicing pattern. It is associated to brain disease and emerged in evolution in the primate clade



The anti-correlated exons are shown with red rectangles. The first one anti-correlates with the second one and with the third one.

To take home:

- ➔ Human genes often have duplicated exons.
- ➔ Some of them are incorporated in transcripts, some are not.
- ➔ Good (functional) copies are different from bad (non-functional) ones: they more often have length divisible by 3, have stronger coding potential and are more conserved.

Acknowledgements

Tim Ivanov - for the data on duplications

Dmitri Pervouchine - for the problem statement

Mikhail Gelfand - for fruitful discussions

**Definitions

ePI - the level of identity of the master and the copy
qjunc_l - the number of reads supporting the 5'-splice site of the master
qjunc_r - the number of reads supporting the 3'-splice site of the master
tjunc_l - the number of reads supporting the 5'-splice site of the copy
tjunc_r - the number of reads supporting the 3'-splice site of the copy
q_cod_pot - coding potential of the master
q_avg_cons - average conservativity of the master
t_cod_pot - coding potential of the copy
t_avg_cons - average copy conservativity of the copy
q5ss_score - the copy's 5'-splice site strength
q3ss_score - the copy's 3'-splice site strength
t5ss_score - the original's 5'-splice site strength
t3ss_score - the original's 3'-splice site strength
correlation - correlation between the expression levels of the master and the copy (in 16 tissues)
cod_pot_dif - difference of the coding potential of the master and the copy