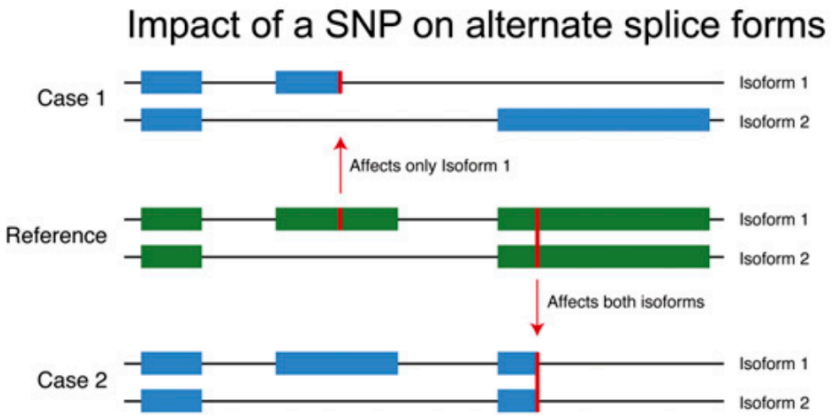
Variant Annotation Tool

Lukas Habegger

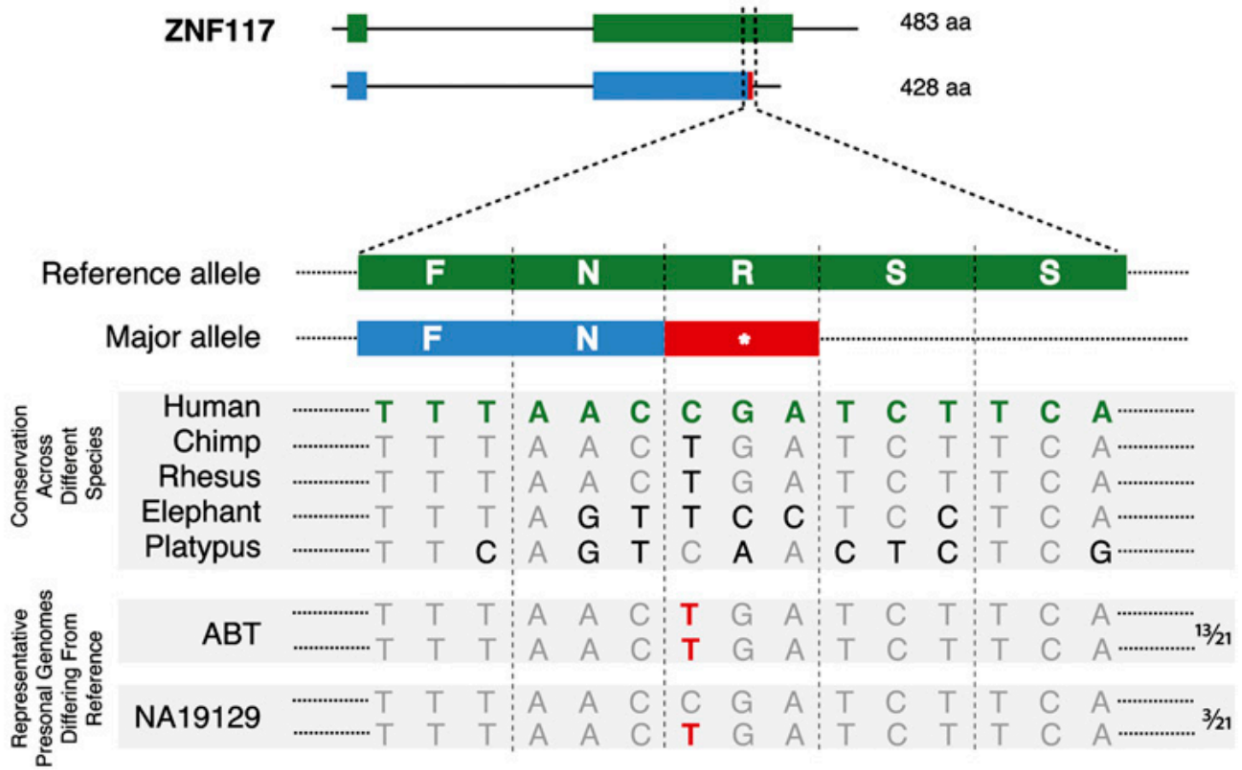
March 31, 2011

Objective

- To annotate genetic variants from personal genomes
  - SNPs
  - Indels
  - SVs
- Efficient algorithm
  - Command line
  - Web-interface
- Visualization of the results

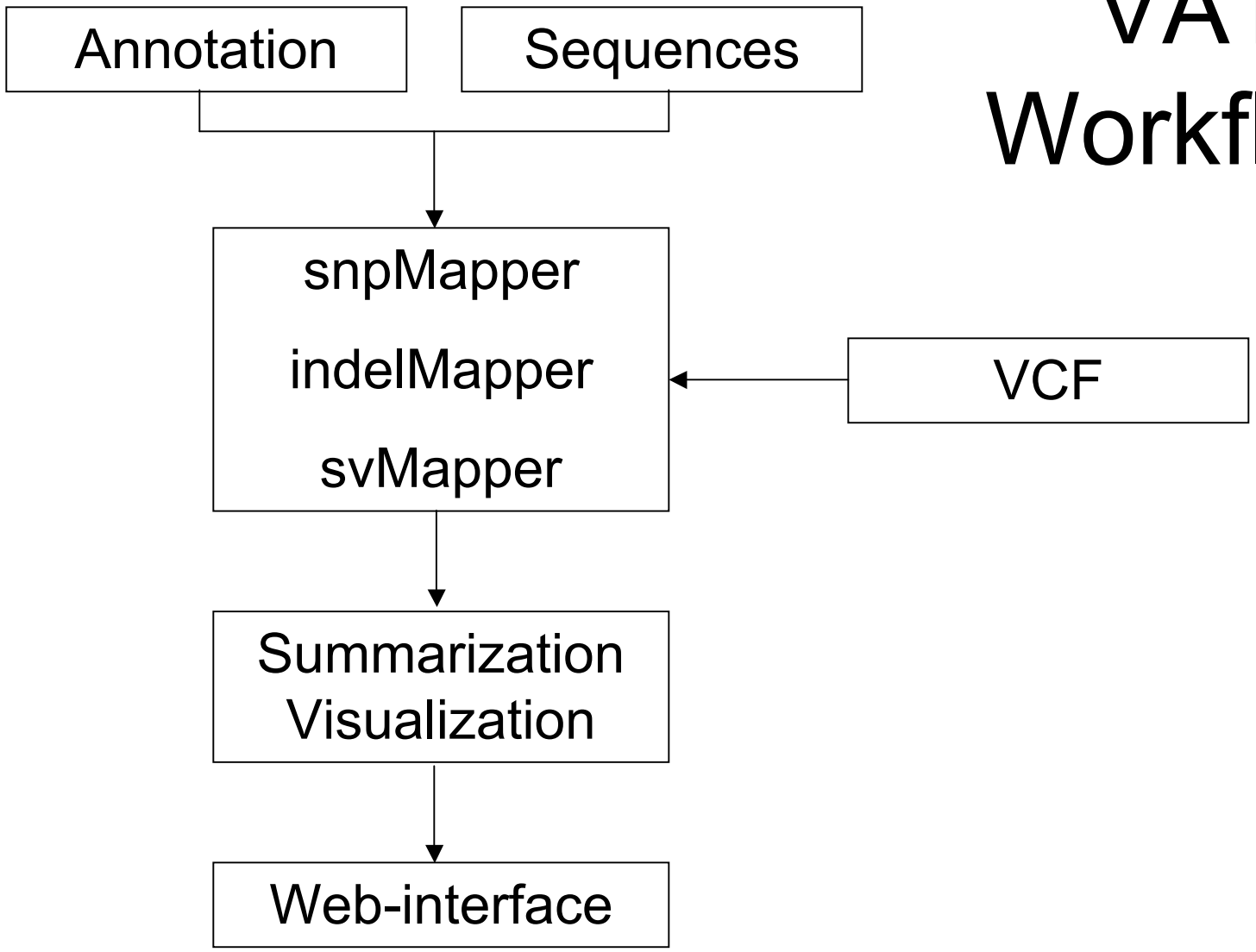


Manifestation of genetic variants



Balasubramanian et al., *Genes Dev.*, 2011

VAT Workflow



Variant Annotation Toool (VAT)	
Download	Documentation
Web Interface	Data Sets

http://archive.gersteinlab.org/proj/VAT/

Results: 1000genomes_lowCoverage

Gene summary based on gencode3b annotation set

Show 25 entries

Search:

Gene ID	Gene name	Number of transcripts	Number of synonymous SNPs	Number of nonsynonymous SNPs	Number of prematureStop SNPs	Number of removedStop SNPs	Number of splice overlaps	Number of frameshift indels	Number of non-frameshift indels	Number of LOF variants	Link
ENSG00000000419	DPM1	6	1	2	0	0	0	0	0	0	Link
ENSG00000000457	SCYL3	4	9	4	0	0	0	0	0	0	Link
ENSG00000000460	C1orf112	4	2	5	0	0	0	0	0	0	Link
ENSG00000000938	FGR	5	1	0	0	0	0	0	0	0	Link
ENSG00000000971	CFH	5	8	11	0	0	0	0	0	0	Link
ENSG00000001036	FUCA2	5	1	5	0	0	0	0	0	0	Link
ENSG00000001084	GCLC	1	3	2	0	0	0	0	0	0	Link
ENSG00000001167	NFYA	2	1	0	0	0	0	0	0	0	Link
ENSG00000001460	C1orf201	10	5	9	1	0	1	0	0	2	Link
ENSG00000001461	NIPAL3	9	3	2	0	0	0	0	0	0	Link
ENSG00000001561	ENPP4	2	2	3	0	0	0	0	0	0	Link
ENSG00000001626	CFTR	5	8	10	0	0	0	0	0	0	Link
ENSG00000001629	ANKIB1	2	3	1	0	0	0	0	0	0	Link
ENSG00000001630	CYP51A1	5	1	1	0	0	0	0	0	0	Link
ENSG00000001631	KRIT1	20	2	2	0	0	0	0	0	0	Link
ENSG00000002016	RAD52	4	0	1	0	0	0	1	0	1	Link
ENSG00000002330	BAD	3	2	1	0	0	0	0	0	0	Link
ENSG00000002726	ABP1	9	8	8	0	0	1	1	0	2	Link
ENSG00000002745	WNT16	3	2	6	0	0	0	1	1	1	Link
ENSG00000002746	HECW1	5	13	9	0	0	0	0	0	0	Link
ENSG00000002822	MAD1L1	15	9	5	0	0	0	0	0	0	Link
ENSG00000002834	LASP1	6	7	4	0	0	0	0	1	0	Link
ENSG00000002933	TMEM176A	5	2	8	0	0	0	0	0	0	Link
ENSG00000003056	M6PR	1	0	0	0	0	0	0	1	0	Link
ENSG00000003137	CYP26B1	2	2	5	0	0	0	1	0	1	Link

Showing 1 to 25 of 18,382 entries

First Previous 1 2 3 4 5 Next Last

[Download compressed VCF file with annotated variants] [View tab-delimited gene summary file]

Sample summary

Show 25 entries Search:

Sample	Group	Number of synonymous SNPs	Number of nonsynonymous SNPs	Number of prematureStop SNPs	Number of removedStop SNPs	Number of splice overlaps	Number of frameshift indels	Number of non-frameshift indels	Number of LOF variants
NA06985	CEU	10730	10257	103	37	167	385	293	655
NA06986	CEU	11265	11006	110	32	182	407	302	699
NA06994	CEU	10934	10605	128	35	178	368	297	674
NA07000	CEU	11010	10701	119	35	188	402	303	709
NA07037	CEU	11136	10845	124	35	175	408	311	707
NA07051	CEU	10840	10653	105	36	183	411	317	699
NA07346	CEU	10886	10633	124	38	185	410	311	719
NA07347	CEU	11126	10698	106	36	174	407	297	687
NA07357	CEU	11315	11030	127	35	198	411	333	736
NA10847	CEU	10778	10590	111	37	189	396	286	696
NA10851	CEU	11285	11306	134	41	177	411	301	722
NA11829	CEU	11068	10673	121	38	172	398	324	691
NA11830	CEU	11042	10552	115	32	177	391	319	683
NA11831	CEU	10784	10433	112	41	176	399	340	687
NA11832	CEU	11101	10584	116	35	179	397	306	692
NA11840	CEU	11055	10688	134	32	175	388	292	697
NA11881	CEU	10930	10525	118	31	185	408	311	711
NA11894	CEU	11091	10680	121	37	180	400	312	701
NA11918	CEU	10740	10432	114	34	164	390	296	668
NA11919	CEU	10982	10561	126	36	181	395	296	702
NA11920	CEU	11267	11018	126	39	206	418	332	750
NA11931	CEU	11084	10743	127	34	180	392	308	699
NA11992	CEU	11142	10800	117	37	190	402	329	709
NA11993	CEU	10848	10519	121	38	176	402	299	699
NA11994	CEU	10877	10350	116	34	171	436	297	723

Showing 1 to 25 of 179 entries First Previous 1 2 3 4 5 Next Last

[View tab-delimited sample summary file]

1000genomes_lowCoverage: gene summary for FUZ [ENSG00000010361]

External links:

[UCSC genome browser] [Ensembl genome browser] [Gene Cards]

Transcript summary

Transcript name	Transcript ID	Chromosome	Strand	Start	End	Number of exons	Transcript length
FUZ-203	ENST00000421740	chr19	-	55006214	55008175	6	705
FUZ-204	ENST00000445575	chr19	-	55002152	55008175	13	1161
FUZ-202	ENST00000377092	chr19	-	55002222	55008175	10	1146
FUZ-201	ENST00000313777	chr19	-	55002222	55008175	11	1254

Graphical representation of genetic variants

LEGEND FOR VARIATION TYPES:

spliceOverlap synonymous nonsynonymous prematureStop removedStop insertion deletion substitution

Detailed summary of variants

Chromosome	Position	Reference allele	Alternate allele	Identifier	Type	Fraction of transcripts affected	Transcripts	Transcript details	Alternate allele frequencies			Genotypes
									CEU	CHBJPT	YRI	
chr19	55006228	AAGAG	A	.	deletion	1/4	ENST00000421740	705_692	0.025	0.000	0.000	Link
chr19	55002239	G	T	.	nonsynonymous	2/4	ENST00000313777 ENST00000377092	1254_1238_413_A->D 1146_1130_377_A->D	0.000	0.000	0.051	Link
chr19	55002278	G	A	rs12610577	nonsynonymous	2/4	ENST00000313777 ENST00000377092	1254_1199_400_T->I 1146_1091_364_T->I	0.000	0.117	0.085	Link
chr19	55003512	G	T	.	nonsynonymous	3/4	ENST00000445575 ENST00000313777 ENST00000377092	1161_1004_335_T->N 1254_1004_335_T->N 1146_896_299_T->N	0.008	0.000	0.000	Link
chr19	55003860	G	A	rs11557714	synonymous	3/4	ENST00000445575 ENST00000313777 ENST00000377092	1161_819_273_D->D 1254_819_273_D->D 1146_711_237_D->D	0.033	0.000	0.000	Link
chr19	55004465	C	T	rs2305921	synonymous	3/4	ENST00000445575 ENST00000313777 ENST00000377092	1161_672_224_L->L 1254_672_224_L->L 1146_564_188_L->L	0.125	0.133	0.136	Link
chr19	55006232	G	A	.	nonsynonymous	1/4	ENST00000421740	705_688_230_L->F	0.008	0.000	0.000	Link
chr19	55006282	A	G	.	spliceOverlap	1/4	ENST00000421740	705	0.050	0.000	0.000	Link
chr19	55006283	G	A	.	spliceOverlap	1/4	ENST00000421740	705	0.100	0.058	0.000	Link
chr19	55006519	G	A	rs35499921	synonymous	4/4	ENST00000445575 ENST00000313777 ENST00000377092 ENST00000421740	1161_405_135_I->I 1254_405_135_I->I 1146_297_99_I->I 705_405_135_I->I	0.000	0.000	0.025	Link
chr19	55008076	C	A	rs35138412	nonsynonymous	4/4	ENST00000445575 ENST00000313777 ENST00000377092 ENST00000421740	1161_100_34_A->S 1254_100_34_A->S 1146_100_34_A->S 705_100_34_A->S	0.000	0.000	0.051	Link

Variant summary

Chromosome	Position	Reference allele	Alternate allele
chr19	55004465	C	T

Genotype information

CEU	CHBJPT	YRI
RefCount = 105, AltCount = 15	RefCount = 104, AltCount = 16	RefCount = 102, AltCount = 16
NA06985: 0|0	NA18526: 0|0	NA18486: 0|0
NA06986: 0|0	NA18532: 0|0	NA18489: 0|0
NA06994: 0|0	NA18537: 0|0	NA18498: 0|0
NA07000: 0|0	NA18542: 0|1	NA18499: 0|1
NA07037: 0|0	NA18545: 0|0	NA18501: 0|0
NA07051: 1|0	NA18547: 0|0	NA18502: 0|1
NA07346: 0|0	NA18550: 0|0	NA18504: 0|0
NA07347: 0|0	NA18552: 0|1	NA18505: 0|0
NA07357: 0|0	NA18555: 0|0	NA18507: 0|1
NA10847: 0|0	NA18558: 0|0	NA18508: 0|0
NA10851: 0|0	NA18561: 0|0	NA18510: 0|0
NA11829: 0|0	NA18562: 0|1	NA18511: 0|0
NA11830: 0|0	NA18563: 0|0	NA18516: 0|0
NA11831: 0|0	NA18564: 0|0	NA18517: 0|0
NA11832: 0|0	NA18566: 0|0	NA18519: 0|1
NA11840: 0|0	NA18570: 0|1	NA18520: 1|0
NA11881: 0|1	NA18571: 0|1	NA18522: 0|0
NA11894: 0|0	NA18572: 0|0	NA18523: 0|0
NA11918: 0|0	NA18573: 0|0	NA18853: 0|1
NA11919: 1|0	NA18576: 0|0	NA18856: 0|0
NA11920: 0|0	NA18577: 0|0	NA18858: 0|1
NA11931: 0|0	NA18579: 1|0	NA18861: 0|0
NA11992: 1|0	NA18582: 0|0	NA18870: 0|1
NA11993: 0|1	NA18592: 0|0	NA18871: 0|0
NA11994: 0|1	NA18593: 0|0	NA18907: 0|0
NA11995: 0|0	NA18603: 0|1	NA18909: 0|0
NA12003: 0|0	NA18605: 1|0	NA18912: 0|0
NA12004: 0|0	NA18608: 0|0	NA18916: 1|1
NA12005: 0|0	NA18609: 0|0	NA19093: 0|0
NA12006: 0|1	NA18638: 0|1	NA19098: 0|0
NA12043: 0|0	NA18940: 0|1	NA19099: 0|0
NA12044: 0|0	NA18942: 0|0	NA19102: 0|1
NA12045: 1|0	NA18943: 0|0	NA19108: 0|0
NA12144: 0|0	NA18944: 0|0	NA19114: 1|0
NA12154: 0|0	NA18945: 0|0	NA19116: 0|1
NA12155: 0|0	NA18947: 0|0	NA19119: 0|0
NA12156: 0|0	NA18948: 0|0	NA19129: 1|0
NA12234: 0|0	NA18949: 0|0	NA19131: 0|0
NA12249: 0|1	NA18951: 0|0	NA19137: 0|0
NA12287: 0|0	NA18952: 0|0	NA19138: 0|0
NA12414: 0|0	NA18953: 0|0	NA19141: 0|0
NA12489: 0|0	NA18956: 1|0	NA19143: 1|0
NA12716: 0|1	NA18959: 0|0	NA19144: 0|0
NA12717: 0|0	NA18960: 0|0	NA19147: 0|0
NA12749: 0|1	NA18961: 1|0	NA19152: 0|0
NA12750: 0|0	NA18964: 0|0	NA19153: 0|0
NA12751: 0|0	NA18965: 0|0	NA19159: 0|0
NA12760: 0|0	NA18967: 1|0	NA19160: 1|0
NA12761: 0|0	NA18968: 0|0	NA19171: 0|0
NA12762: 0|0	NA18969: 0|1	NA19172: 0|0
NA12763: 1|0	NA18970: 0|0	NA19190: 0|0
NA12776: 0|0	NA18971: 0|0	NA19200: 0|0
NA12812: 0|0	NA18972: 1|0	NA19201: 0|0
NA12813: 1|1	NA18973: 1|0	NA19204: 0|0
NA12814: 0|0	NA18974: 0|0	NA19207: 0|0
NA12815: 0|0	NA18975: 0|0	NA19209: 0|0
NA12828: 0|0	NA18976: 0|0	NA19210: 0|0
NA12872: 0|0	NA18980: 0|0	NA19225: 0|0
NA12873: 1|0	NA18981: 0|0	NA19257: 0|0
NA12874: 0|0	NA19005: 0|0	

Potential functional role of pseudogenes

March 31, 2011

Gerstein Lab

Xiu Huang

Suganthi Balsubramanian

Lukas Habegger

Alex Abyzov

Mark Gerstein

ENCODE pseudogene sub-group

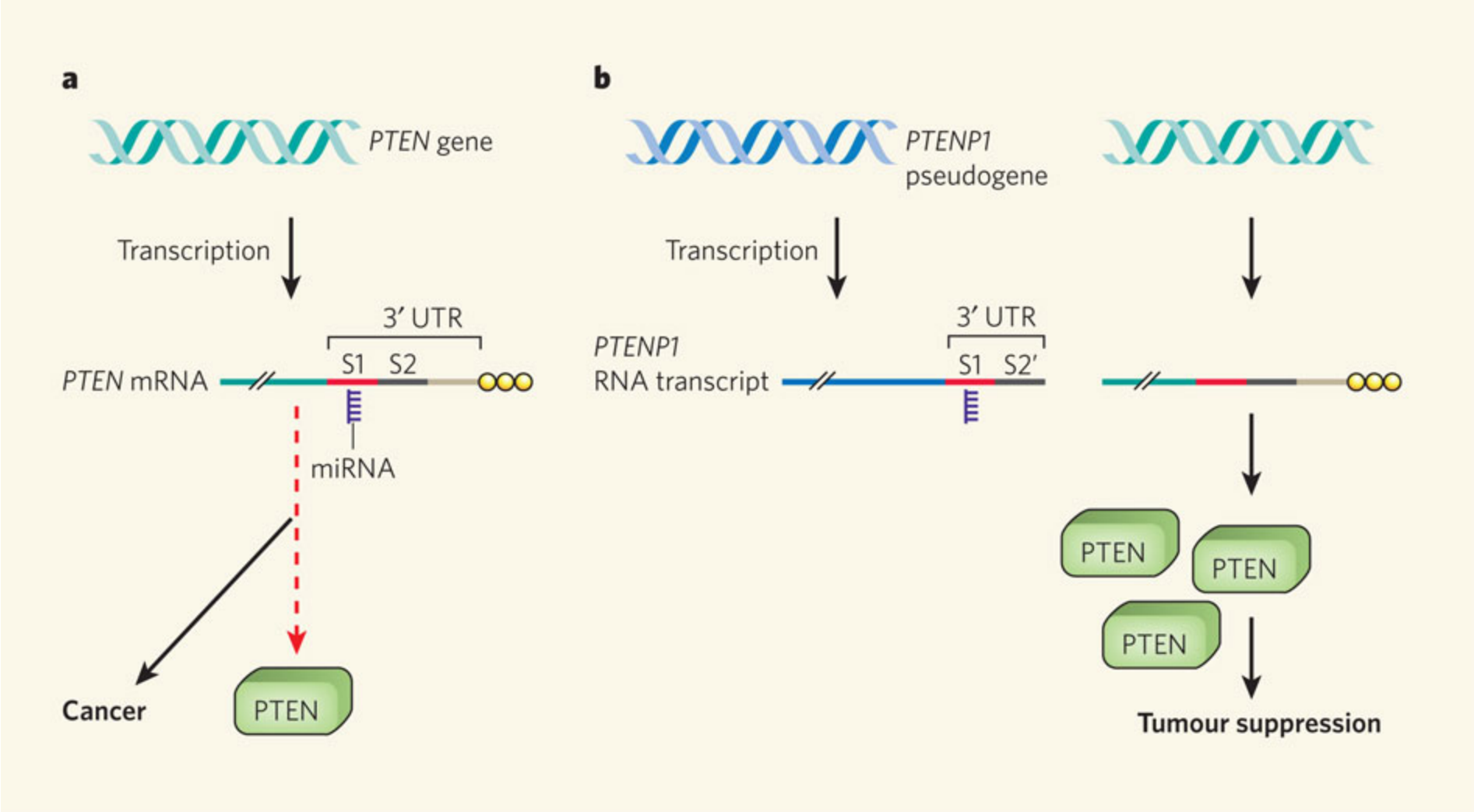
Adam Frankish

Jennifer Harrow

Rachel Harte

Mark Diekhans

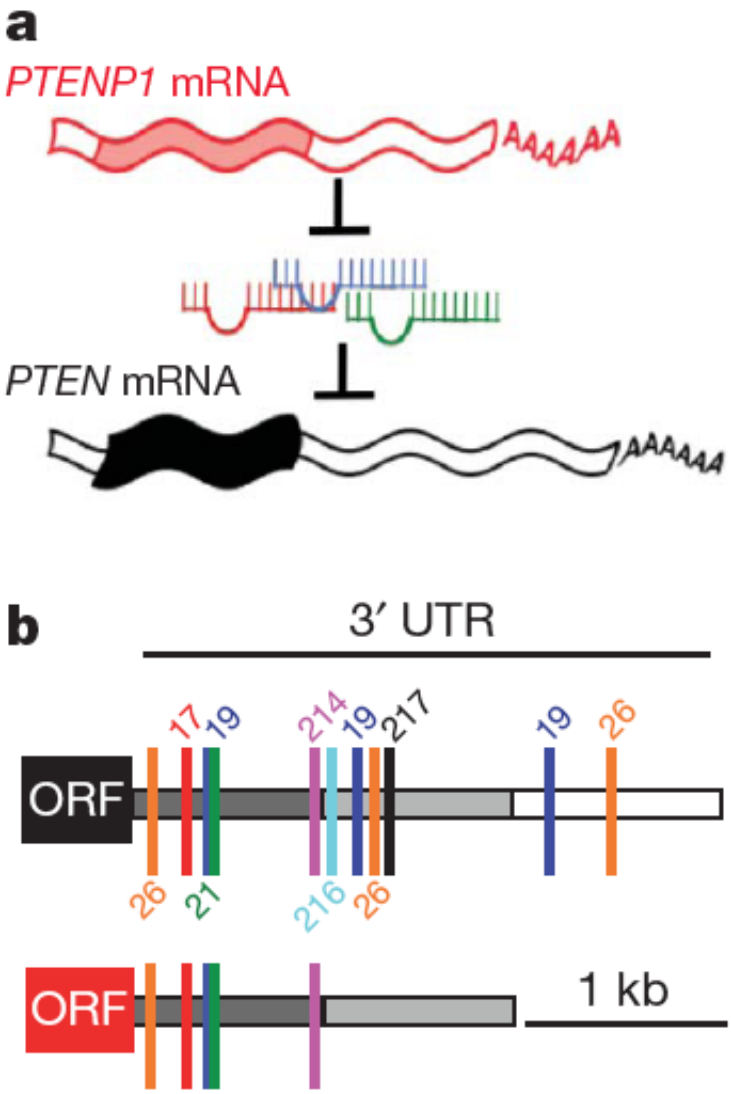
Pseudogene acts as a decoy



A coding-independent function of gene and pseudogene mRNAs regulates tumour biology

Poliseno et al, Nature,V465, June 2010, 1033-1040

Conservation of miRNA binding sites



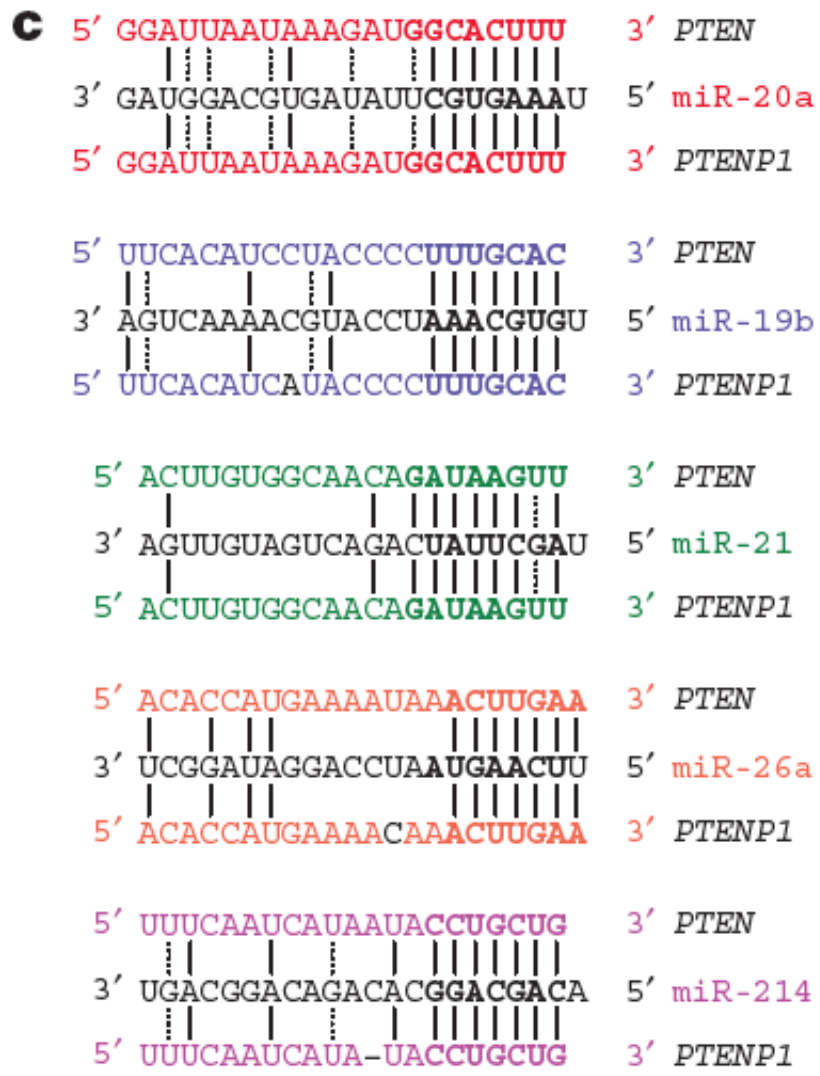


Table 3. Conservation of validated miRNA binding sites in cancer-related target genes.

wt genes	corresponding pseudogene(s)	validated miRNA families	conservation of the binding site between *wt* and pseudo
CCND3	*CCND3P*	*miR-16* [1]	no*
CDK4	*CDK4PS*	*miR-34* [2]	yes
DNMT3A	*DNMT3AP1*	*miR-29* [3]	no
		miR-143 [4]	no
E2F3	*E2F3P1*	*miR-17* [5]	yes
		miR-34 [6]	no
c-MYC	*MYCL3*	*let-7* [7]	no*
		miR-145 [8]	no*
OCT4	*OCT4-pg1,2,3,4,5,6*	*miR-145* [9]	yes
KRAS	*KRAS1P*	*let-7* [10]	yes
		miR-143 [11]	yes
PTEN	*PTENP1*	*miR-17* [12]	yes
		miR-19 [13,14]	yes
		miR-21 [15]	yes
		miR-26 [16]	yes
		miR-214 [17]	yes
		miR-216 [18]	no
		miR-217 [18]	no
FOXO3	*FOXO3B*	*miR-182* [19]	yes

Are pseudogenes decoys?

- Are 3'-UTR of pseudogenes more conserved than the rest of the pseudogene?
- Do they contain conserved miRNA binding sites?
- If pseudogene expression regulates parent gene, do we also see 5'-UTR conservation?

Analysis of pseudogenes

Added 1kb to 12,381 pseudogenes
Aligned against parent cDNA (bl2seq)

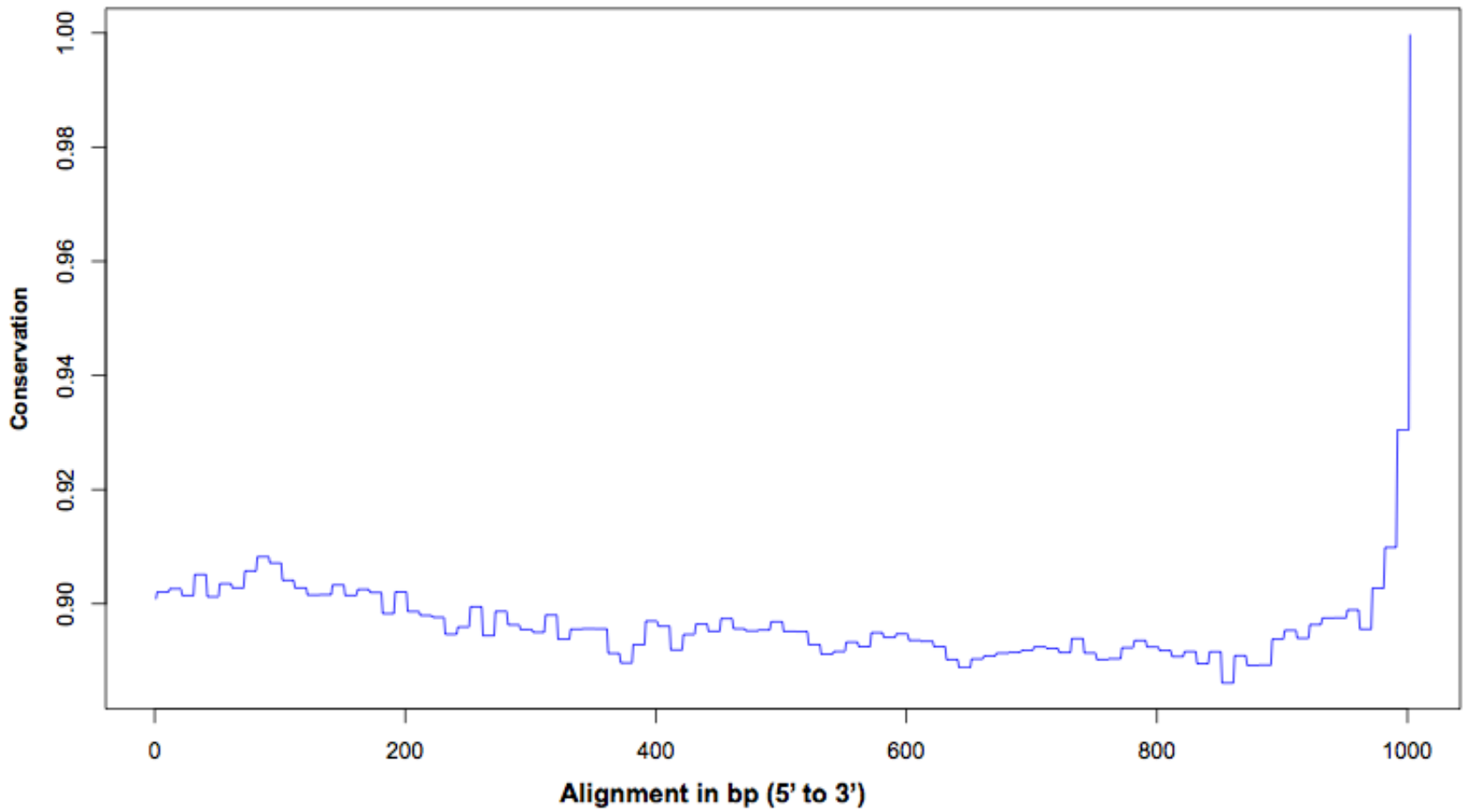
↓

Removed alignments that did not include 3'-UTR regions

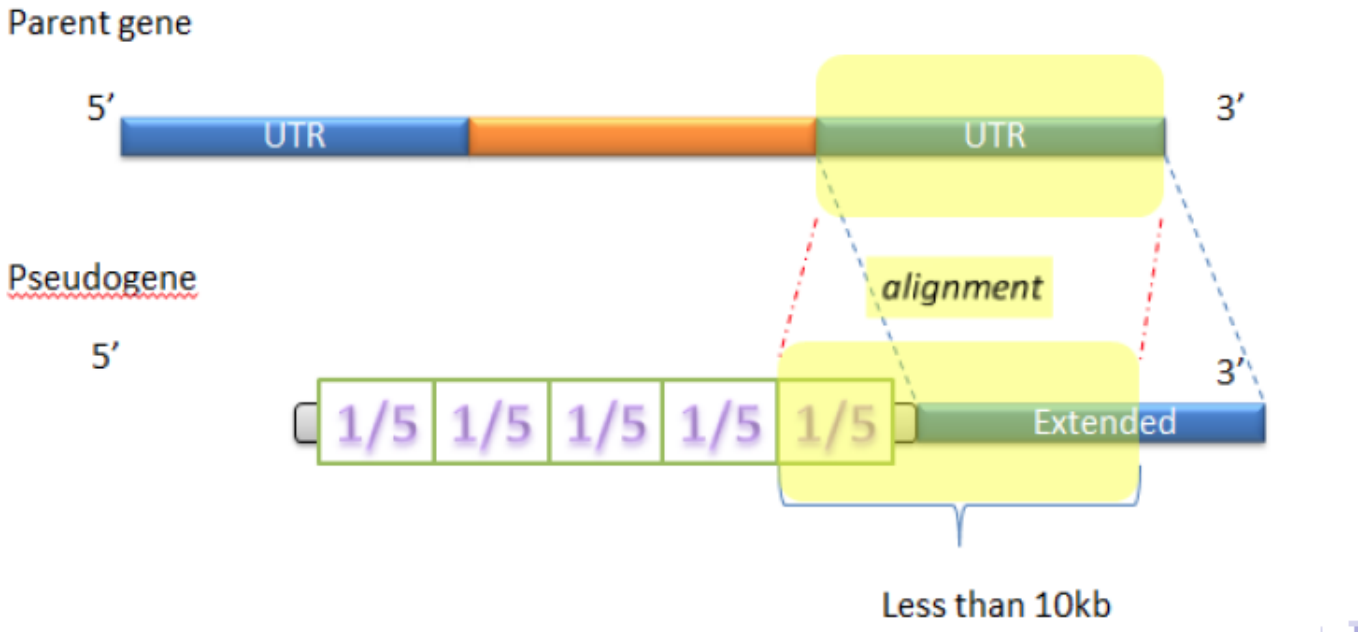
↓

Obtained alignments for 7636 pseudogenes

Parent-pseudogene alignments

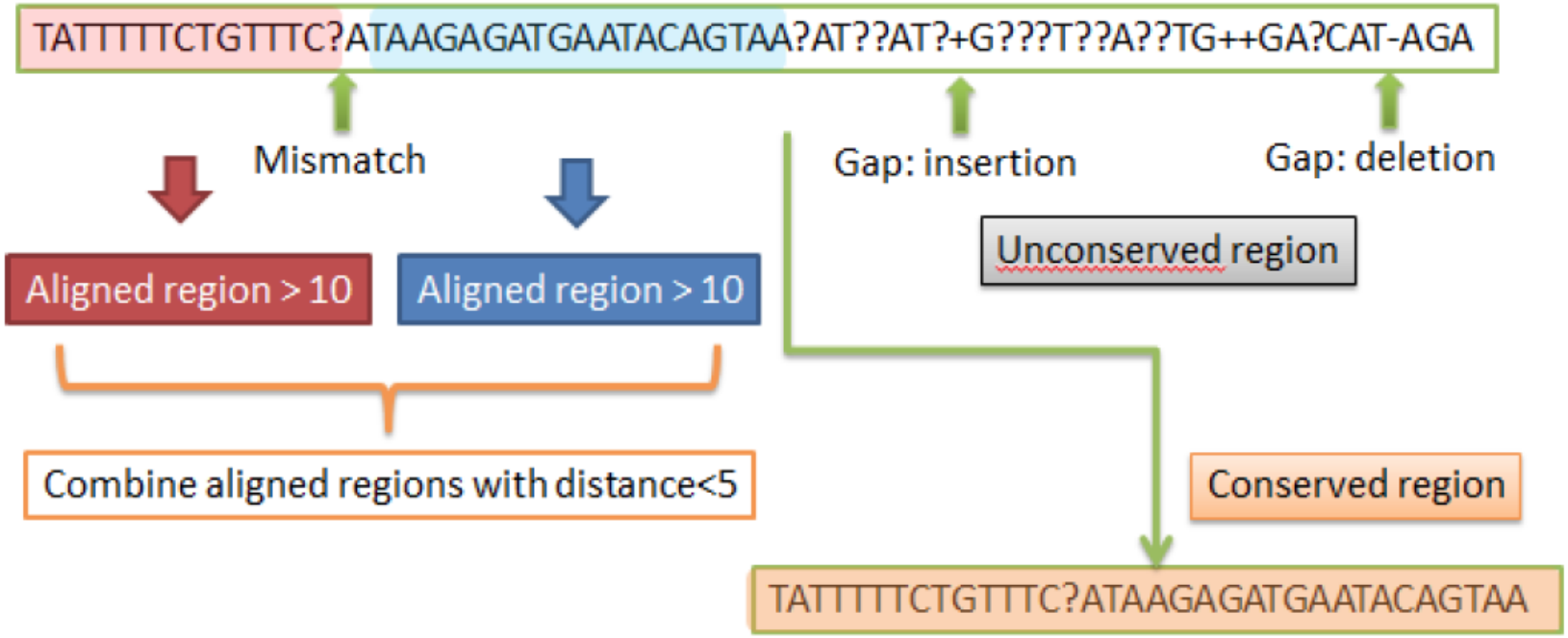


Global alignment (R package)

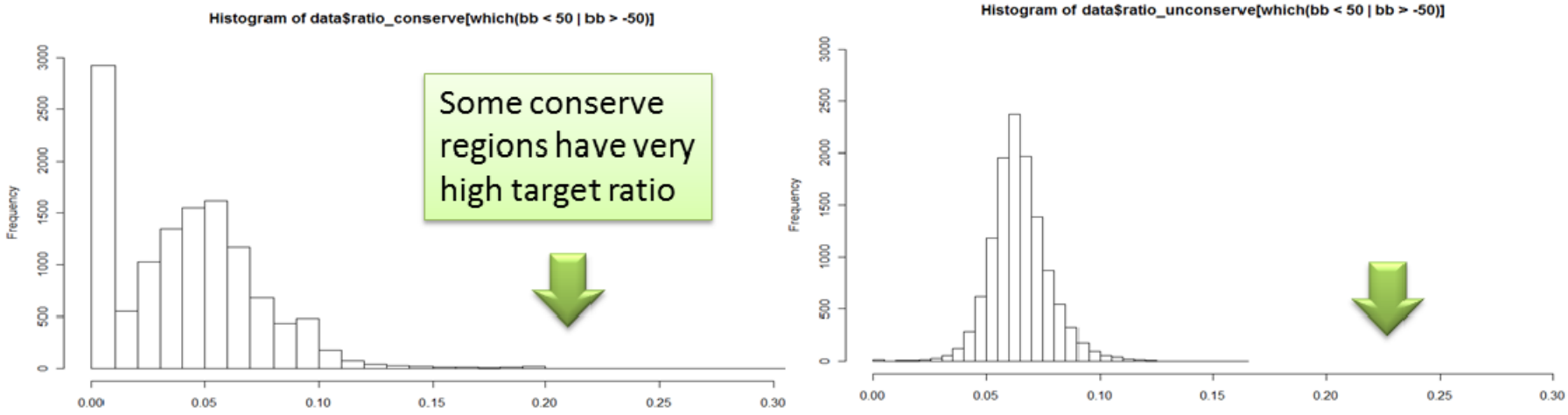


Preliminary results on miRNA binding sites analysis

- Process the result of the alignment by finding aligned regions with length > 10nt.



- Compare miRNA target site number/sequence length ratio in sequences with |conserve_region_length-unconserve_region_length|<50nt.



PTENP1


ENCODE Gencode Manual Gene Annotations (level 1+2) (Oct 2009)
PTENP1
ENCODE Gencode Automated Gene Annotations (level 3) (Oct 2009)
PTENP1
RefSeq Genes
PTENP1
Human mRNAs from GenBank
Human mRNAs
ln(x+1) 4
Transcription
0
ENCODE Transcription Levels Assayed by RNA-seq on 6 Cell Lines
50
Enhanced H3K4Me1
0
ENCODE Enhancer- and Promoter-Associated Histone Mark (H3K4Me1) on 8 Cell Lines
50
Enhanced H3K27Ac
0
ENCODE Enhancer- and Promoter-Associated Histone Mark (H3K27Ac) on 8 Cell Lines
100
Promoter H3K4Me3
0
ENCODE Promoter-Associated Histone Mark (H3K4Me3) on 9 Cell Lines
ENCODE Digital DNaseI Hypersensitivity Clusters
2
9
11
5
Txn Factor ChIP
ENCODE Transcription Factor ChIP-seq
3
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
-0.5
Multiz Alignments of 44 Vertebrates
Chimp
Orangutan
Rhesus
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
Segmental Dups
Duplications of >1000 Bases of Non-RepeatMasked Sequence

Future directions

- A better method to identify miRNA binding sites with higher specificity.

- Comparative genome analysis (preliminary results indicate most conservation if at all seen is primate-specific).

- Look for differential enrichment of miRNA binding sites in transcribed versus nontranscribed pseudogene (RNASEQ, histone marks, identification of TF-binding from ChipSeq), duplicated versus processed pseudogenes.