NOVEL TRANSCRIPTIONALLY ACTIVE REGIONS (TARS)

Andrea Sboner - 2011.04.15

DISCOVERY OF NOVEL TARS

- Reads overlapping gene annotation set --AceView, GENCODE, UCSC, and repetitive regions are excluded
- Reads from all samples
 pooled together



• Max-gap, min-run algorithm to identify TARs

MIN-GAP, MAX-RUN ALGORITHM: THRESHOLD DEFINITION



Normalized signal track (per million mapped nucleotides)

 $r(g) = \frac{n(g)}{l(g)*M}$ $x(g) = \log_2(r(g) + 1)$

- From exon expression values:
- Determine the median value of each exons:
- $x_c = quantile(\vec{\tilde{x}}, 0.05)$ $\forall \tilde{x}(g) > 0$

 $\tilde{x}(g)$

- Compute the 5th percentile of expressed exons:
- $r_c = 2^{x_c} 1$ Compute the corresponding RPKM value:
- $|\vec{L}| = 0.129Kb$ Consider the median length of GENCODE exons:

 $t_{bgr} = |\vec{L}| \cdot r_c$ • Define threshold (normalized per million mapped nt):

Results



DISTANCE FROM GENES



DIFFERENTIAL EXPRESSION (PRELIMINARY)

- Normalized by the number of mapped nucleotides
- t-test test between neocortex (NCX) and the "rest" (AMY, CBC, HIP, MD, STR)

cutoff	Raw p- values	BH	Bonferroni				
0.05	4180	3242	1067				
0.01	3045	2359	900				





EXAMPLE (ZOOM-OUT)

		r18			
	ᆈᄷᆕᇩᆸ		80 kb		
	DATA FI DATA T DATA T IRACK_ D sample egion mapped	22,020 kb 22,030 kb	22,040 kb 22,050 kb	22,060 kb 22,070 kb	22,080 kb 22,090 kb
				
VFC					
VIC					
ѕтс					
s1C		i i i i i i i i i i i i i i i i i i i			
OFC					
MFC					
м1С					
ІТС					
ІРС					
DFC			- -		
A1C					
АМҮ					
СВС					
нір					
MD					
STR					
all.NOVEL.novel.0.019617					
repeatMasker bol 0		I I INKLO INN KANKA DI INKINEKI INI 3011 DI INI 3001 I DI 310 E DI I NI DI3101300 DI E EK		BED_4088[chr	
repeatingsker_ing19		2_1 MIR_SINE_MIR_1 AluSz_SINE_Alu_1 L2a_LINE_L2_1 AluJr_SINE_Alu_1 L1ME3A_LI	E_L1_1 MLT1B_LTR_ERVL-MaLR_1 AluSx_SINE_Alu_1 AluSq2_SINE_Al	lu_1 AluSx1_SINE_Alu_1 MIRc_SINE_MIR_1 L1MC4_LINE_L1_1 AluSx	1_SINE_Alu_1 L2b_LINEAluJb_SINE_Alu
RefSeq Genes		→ → → → → → → → → → ↓ → ↓ → ↓ → ↓ → ↓ →	$\rightarrow \rightarrow $		
UCSC Genes		$ \qquad \qquad$	BOYJ19		

ANOTHER EXAMPLE

000					IGV - Session: /	/Users/andrea/D	esktop/IGV/igv	session_br	rainSeq.xm	ł						
Human hg19	¢ chr3	¢ chr3:14	7,128,313-147,141,063	Co 🖆 🖣	r 🗇 🗉	x								Ξ		
		p26.1 p25.2	p243 p241 p223	p224 p21	2 p142 p1	141 p13 p123	pl24 all	a12.1	01342	g1331 g214	g22.1 g22.3	g24 g	52 02533	a26.2	g2632 g27.1	g28 g29
								-								
	DATAFLE DATAFLE DATATTRE DATATTRE DATAFLE TRACC_D ID ID ID ID ID ID ID ID ID ID ID ID ID	-	147,130 kb		147,132 ka 		147,134 kb	— 12 kb	1	147,136 ka		147,135 1	•	1	147,149 ka 	•
VFC								-								
V1C			i i													
STC					- X - '									1		
\$1C			2		-	-								-		
OFC			- 1			E		-						- 2		
MFC						F		- L						- 2		
M1C			<u> </u>		-	 								-		
ітс																
IPC			. I			2.13		-						1		
DFC			77 <u>E</u>			- T										
AIC			=			10		-						- 2		
АМҮ						=	- <u></u>							4		
СВС					1				-	-	-				- 7-	
нір			. 1			- 11								-		
MD			tiga 🔤 🕹											4		
STR					=	1.75		-						-		
RefSeq Genes		> → →	· · · 🔳 ·	2474				-								
all.NOVEL.novel.0.019617								_								
UCSC Genes		<u>→</u> →→	860_8388(ch/2).1147138	537(14713)648			BED_0390);	1/31.114713454	esi147134822							
repearMasker_hg19			1	612912					•							
chr3:147.134.986			nes_complexity_complexi	nud.				cos tas pa	Cripes(_5mp)	0.00002				1	T 309M	of 459M

FUTURE STEPS

- Analyze "intronic", "UTR", and intergenic novel TARs
- Classify them with incRNA to determine their "non-coding" potential

