



2415-14

Winter School on Quantitative Systems Biology

26 November - 7 December, 2011

The Transcriptome: sequencing the effectors of the genome

R. Sachidanandam Mount Sinai Sch. of Medicine New York USA

Transcripts

Ravi Sachidanandam Mount Sinai School of Medicine

mRNA transcription





Alternative Splicing of Dscam Potentially Generates38,016 IsoformsDSCAM: Down syndrome cell adhesion molecule



(A) Schematic representation of *Dscam* gene, mRNA, and protein. The Dscam protein contains both constant and variable domains. The variable domain are encoded by alternative exons. Each block of alternative exons is indicated by a different color. A transcript contains only one alternative exon from each block. The *Dscam* gene encodes 12 alternative exons for the N-terminal half of Ig2 (red), 48 alternative exons for the N-terminal half of Ig3 (blue), and 33 alternative exons for Ig7 (green). There are two alternative transmembrane domains (yellow).

©2011 by National Academy of Sciences

A Model for Dscam-Mediated Interactions

(A) Schematic representation of homophilic binding mediated by the three variable Ig domains. The eight N-terminal Ig domains (circles) are shown. Constant Ig domains are gray, and variable Ig domains are in color. The remainder of the protein is represented by a gray rectangle. Isoforms sharing identical variable Ig domains (represented by matching colors) bind to each other, while isoforms differing in only one variable Ig domain (gray arrowheads) do not. We propose that each variable Ig domain binds to the same variable Ig domain in an opposing molecule. As isoforms sharing any two identical variable Ig domains do not bind or exhibit reduced binding relative to binding between identical isoforms, it is likely that the binding of all three variable Ig domains. (B) Schematic representation of Dscam-mediated interactions between neurites. We propose that differences in levels of Dscam signaling influence the nature of the interactions

The miRNA processing pathway has long been viewed as linear and universal to all mammalian miRNAs. This canonical maturation includes the production of the primary miRNA transcript (pri-miRNA) by RNA polymerase II or III and cleavage of the pri-miRNA by the microprocessor complex Drosha–DGCR8 (Pasha) in the nucleus. The resulting precursor hairpin, the pre-miRNA, is exported from the nucleus by Exportin-5–Ran-GTP. In the cytoplasm, the RNase Dicer in complex with the double-stranded RNA-binding protein TRBP cleaves the pre-miRNA hairpin to its mature length. The functional strand of the matur miRNA is loaded together with Argonaute (Ago2) proteins into the RNA-induced silencing complex (RISC), where it guides RISC to silence target mRNAs through mRNA cleavage, translational repression or deadenylation, whereas the passenger strand (black) is degraded. In this review we discuss the many branches, crossroads and detours in miRNA processing that lead to the conclusion that many different ways exist to generate a mature miRNA.

mRNA post-translational regulation by miRNA

Nature Reviews | Molecular Cell Biology

Structure of Eukaryotic Genes

DNA

Splice-site prediction

Good splice-site prediction implies understanding of its recognition by the spliceosome.

Splice-site predictor/classifier: The Mouse-trap of bioinformatics ?

Predictors generally improve with more data.SpliceRack is the result.

HPRT1 Exon-Intron Structure

att	ga tg tcag ta	tgattttac	atgetgatet	tgaccaattt	gaaacagtga	gttaaaatct	ggctgatccg	tactaatcct	aaagaaatat	tctatgaact	attaaatgtt	tccagaa ta
taa	tgagaceece	tetetacaca	aaaagaatta	g t tg tg ca tg	g tgg cg tg ca	cetgtagtee	cagetacttg	ggaggcagag	gcaggagcat	cccttgagcc	taggagtttg	agactgcag
act	cttaatccaa	gtagettggt	aggattttat	ttacctagtg	cctaga tggg	aaattgcctg	gggattccaa	atacctattt	cattaaatta	aagatgtcac	tgattttaag	acttaacac
tgg	tggctcacac	ctg taa teee	agtactttgg	gaagtcaaga	cgggaggctg	gettgaacce	aggagttcaa	gaccageeta	ggcaa tg tag	egagaegeea	tetcaaaata	t taaaaa ta
cga	gtagetggga	ttacaggcgc	ctgctaccat	gcctggctaa	tttttgtatt	tttagttgag	a tgggg tttc	accatgttgg	ccaggetggt	ctagaactct	tgacctctgg	tgatecace
tee	cacctccctt	tecagag tag	cggggaccac	agg tg tg tg c	caccacacct	gactaatttt	tgcattttt	tttg tagaaa	cagggttttg	ccatgttgcc	cagg ttgg tc	tgaaactee
age	agagggaaaa	ttagtetgea	t ta tog to ta	tecagactaa	a tga c tga ta	t taaaa tgaa	attattetta	ggatttgcaa	tettagagaa	aactttttca	tttttattt	tttgagtta
gtt	ttgcaagcag	acagaattca	ttttgccaat	tacgggattt	tccctcagtt	gcagtcaagg	ttcataaaac	tataactett	tatettaat	tagaaatgtt	tttttttg	agacaaggt
cet	anteteaane	aateettete	cottantoto	ccaaagtget	gaga ttacag	atateaacca	ttacacetaa	ccaaaacmat	aacttaaaat	acacacacac	acacacacac	acaaacaca
tat	atttaaggta	tacaacatga	to t ta togga	ta ca ta tagg	togttaaaag	attactocag	tgaagcaaat	taacotatec	ctcaactcac	atagttaccc	atttttttt	tatttaat
	tanatthana	attanatt	htabababab	a ta a ta a th	tattagant	abaabbbbab	ataanata	a bha an ba ba	anto to con a	things hat has	a can a base b b b	
cag	tagetteag	tattatat	gantagan	gtagtegatt	ceccugagg c	atacttttet	tttotagaata	tttagaugua	ggtataceae	tastataata	acactageet	accugaac
gaa	taaaa togaa	accatgaaga	tagga tgaaa	taacttotaa	aactaggtte	aaaag ta ta c	cccctag ctc	aaaataccag	ataatttta	tttattttt	attttttt	tttttgagg
y ca	taaaa uyyaa	acca ugaaga	tagga tgaaa	cara e e ce cera	- agea c	geeegaaaae	ages	aaaa caccay	guggeeeeug	couge course	geececce	ceceegag
aga	actectggge	ttaagegate	ctcccaccte	ageeteccaa	agtgetagga	ttacaggcat	gagecaccat	gcctggcag	aaataccagg	tttttaagta	tcagcactta	ctcttcaat
Cat	gtggetaceg	tatgggacag	tore, tacta	gatgatetgt	aagggetgtg	cttcatcagt	gregttttt	aactgacaaa	aaccuragt	LELELLELLA	gtaatgtgtt	tatttaaaa
taa	aaggacagag	aa tgg t to g	agattatgat	atgaagagaa	aa tg tga t tg	ag tg tgg tag	acttgggggcc	tgcttgaatg	ttgagagaat	gectgttttc	cgataaaaaa	aaaaagtco
cag	ttetttteet	gecer cttat	tatgcatttc	tatacagett	tectectett	tttetatace	a tgc tgcag t	tettattget	acctagaggt	tttc.naatt	cctaggggcg	ga taag tag
cgg	tagaggagag	gtagagcaa	ctctggagga	agettteece	teacetttge	cagteetgtt	atcctagact	taacca taa t	taaaga tgag	ggaggcacus	ag taaaggga	tetagtggg
aag	aaaaaa tara	tagccaggta	tggtgactca	agcetgtaat	ctcggcactt	taggaggccg	aggcaggagg	atogettgag	tccaggagtt	caagaccagc	tgagcaaaa	tag tgaga t
taa	agacat .aac	tggagttgga	ctgtaatacc	aggtatetee	agaaga tggc	actatttaac	agattttata	aataatttga	tg tgag tcac	tg tca tc tga	agettgttgc	cttttcttt
taa	a tog a to ta t	a ta tagaage	ataacaaaaa	tagaagcaca	taaaag tgaa	aagtetcata	aacgccattg	tcactactca	tgtaattgct	gttacaaatt	tott aa to	t tgaa taaa
ttt	at ttttaaa	taaaaaaag	agagacaggg	tettgetgtg	tttctcgggc	tggccttgaa	ctcctggggt	caagegateg	tettgeetet	geeteeetgg	gattack ggc	a tgaagcca
cat	a tta tragge	ttaagttttg	ccan ta toon	toggagaaca	g tagaa teac	attatttag	tatttattte	traga taga t	ataattttac	accttataac	ettetettet	ataaattat
tee	tctggaatt	tattttata	tatoctotoa	ggtaggagacc	atacttttat	tttttcccaa	atogottact	agttggccaa	acatcattta	ttgaataatt	catetttte	ctactgact
ate	cagcacagce	aaattatogt	cattrtcacc	accaactaca	a taga ta t ta	agcatttecc	attgaatete	ctgtaagggt	tttattogat	tetataataa	cag taaaa to	ggagectaa
++-	agaaataaat	a tan anna	a trans to sale t	attatata ta	treesttr	an a baran ba	tasaasttt	thtoganata	ttoopoort		a tag aga ti	ages at the
tat	agaaa teact	g uga cugaag		TGACCAGTCA	ACAGGGGACA T	ΔΔΔΔΩΤΔΔΤ ΤΟ	CTCCACAT CA		TAACTCC AAAC	caaa cagta	atggeegact	aggactitt
cta	a ta ta tag cc	actotaatt	aggatestre	TUNCCHUTCH		AAAAOTAAT IQ	JUTUUAUAT UA	TOTOTOAR OTT		adage	acctugaaad	agaagaaaa
c ug	ce ce uj ujgg	acto	gggacc	aaaaaacacc	gaugggggaa	aayatayeee	c caaaaaaaaa	aaaaaaaaca	aaccuatyty	ag cc ig ig	agguagace	acatagett
agt	gagetetta	caacagtttc	tggtaaatca	ctcaataaat	tcagacatac	tattattta	agaaatetea	aagagttttc	ttgtacctta	aaatteteet	ag tg tgaa c	attggtttt
get	CILCLECEE	gaagetgeea	gecatigett	acttacacta	tgecaaa ta t	aaaggcatta	ateteataaa	agtttcacaa	caateetgtg	agggagacga	ta tececa ct	ttacaaa to
tca	ggg tca tg tc	aaactaatgt	cctcctcage	atctttggaa	aacttcagag	gagaaa tgag	ctttgcccct	cctgttcatt	tca ta ta cca	ctgttagacc	tgtcct.ccc	tttcagcat
gta	aagt cattt	attatggaaa	aatcaataag	tataacgagt	gaaagttatt	tettggtggt	aagattatgg	gattatttga	actttctgtt	tcattgtatt	ttatitattt	atttattt
tet	ccaactictg	a to to agg ta	tecacetgee	tcagcetece	aaagtaccgg	gattacgggt	gtgagccacc	etgeetggee	teatttgte	ttttggggggt	atticttgtgt	gcaga ta ta
att	cttttgttt	ttagacteta	g tg tc tg tac	tegttgtace	atgetgggat	tcatttgaac	aattgcatgg	ctttttagt	gtattattaa	atttgcagtt	cacttagaat	ttactggga
gca	tgagagaaat	a, ta tgaa tt	gettgecaca	agttatgggc	tageettact	tcattctgta	cttggacctg	tttaggette	taagagatet	tacetectae	aataaactgc	tttgagaca
caa	gg tgagg taa	aaaatutaaa	agttetaate	tttcttgcaa	accag tgga t	cttttg tgcc	ttactctqqt	aaacactgtc	ttagaagaat	a ta .agaaca	ttaaaa tott	aa tocta ta
aat	gtagaaagcg	aag tgaggg .	t ta tgg tgag	aggaagcatt	gg ta tca tg t	tttagtgtag	tecaagaata	tggacacate	cagaaaa to	agatcaagtt	tagectaatg	agaaaa ta t
aaa	cagg tggg tc	aaattctgtt	tttannattt	ccattatgat	gaaaatttca	gtattacagg	cttccaaatc	ccagcaga tg	ggconettgt	ttaaaggaga	gtttga ta ta	a taaagca t
taa	gecaggagaga	gaaatteete	treatcartt	the tea tag	acttatatte	taaaggag tg	agattogttt	tttataga	ctacttagta	atttatttt	accaa taa to	gaa tog ta t
tta	ttatgtgcta	aagtatttg	tatettagea	cogagagget	daycaybtto	ataggattag	and tad	actaagggaa	acctttactt	cetttagete	agtggttctc	aaaa to too
-	accatesact	togtattet	cotona ta ct	tttatttoo	trateactta	tenetancag	cototototo	tototototo	tetetetete		to to to topo	ttateacto
tac	ctattetoga	cattttatat	aaa tagaa to	atacaatato	tageetttta	tatetagett	ctctcactta	atattttcaa	gattcattca	tattatagaa	tatateterea	ctcatttcc
gac	agacgttagg	thatttecac	tattactec	ttatttctcg	tacctgaaat	gteettatte	cetecettet	tateccatot	ttaagtcatt	taagacccag	ctcaaacgtc	acctecaca
				- to be be be be		- too too too too too too	- the back and					
ggt	gugegeeace	augeeugget	aa cug ug ug c	gugugugugu	g ta tg ta tg t	acguatacatac	gugugugugu	g ta ta ta ta ta t	a ca	acatatata	a ta ta ta ta ta c	a ca ca ca ca
auc	ccatteraat	ca caaaa ca c	ccugaactea	yaaaaaayyy	La Ly Cuyaa C	accuacytac	ccacaaaayc	actaacatte	ujcatati	yceedyace	cuattettet	ujayaaa cu
ttg	gcagaattta	gtteettgtg	attgtaggac	tgagggcccg	ttttctcact	ggetgetgge	caggggttgc	teccagatat	ttaaaggete	a tgccc tagc	ccatgacagt	ctcacaaca
atc	cctaaaggag	gcaggaattt	tgagagccat		CACACTOCCA	CCCAGAAA C	ACTITICATE TO					tttatteca
age	cacetatate	ggagag cogg	gaggagagag	GATATAATT	GACACIGGCA F	AACAAIGCA G	ACTITICTI IC	CITGOTCA GG	CAGTATAA TUU	AAGATG GICA	AGGICG CAAG	yca cy ta ta
att	taaagctatg	caatgtette	tttttgaaa	ggatataatt	gacactggca	aaacaa tgca	gactttgctt	teettggtea	gg cag ta taa	tccaaagatg	g tcaagg tcg	caagg ta tg
tat	taacatttgg	tttttcagca	tgctaattat	atcagtttgt	cctgaatage	atggcagagg	atttgggcc	cccttgcaaa	attaagaata	aggattecaa	ag cggg tgag	gaag tga ta
tca	acaactetgg	ggtggcatta	ttattcccac	ttttcagata	aggttactga	ggca taggga	attgtccaaa	ggtacagagc	tagtccgcta	tagaga tgag	atttgaaccc	agggaacct
cct	tttgaaaaaa	teacggtate	tgtcgagcat	ctttgaatca	gagtaageet	tctag tgag t	ca ta tg tcag	cagtttgact	gtatgggett	ttetaa ta te	cagttcaagt	g ttta tcag
gac	tttaateete	tgggtattct	tttgttgttc	tttcctggta	ta tg c tg tgg	aa t tgaga ta	gactggttcg	tgagcgagag	atttgtgtt	gccacaggta	ggacatgete	aaacaatac
tct	ccctgctgag	aattagtttt	ggetteettg	gagg tga ta t	egectetgtt	gag ta taag t	ggcctactgt	gatcacacca	ctgcactcca	gcctggg tga	cagag tgaga	ccctgtete
taa	aggggaa tga	ttattaacat	ctttttctca	gggaaactat	a tgag tcaag	gaga taa ta t	atttgaaaat	ctttttaact	gcaaageget	gtttcactgt	tggttataat	g tga t tga t
get	accactattt	catttogaa	cccaaagaaa	cettetacee	a ttagcagte	atteteett	cteccagece	etggcaacta	ctaa totact	ttetacagaa	ag teeg ta ca	gatttatat
ctt	t ta tgaa taa	tgttgatttg	aatgtttgtg	tacaagtatg	aatacctgtt	ttcaggtctc	ttgag ta ta t	agttgctagg	tca ta tag ta	actetatet	taacattta	aggaattgo
ggc	catttacata	tatettetta	agaacggtta	cccatttaca	g ta tggaaaa	tgettcagat	gcaactetag	tcatgcctta	gaga tggage	tttattaaac	attcagatet	ctaggca ta
ato	cettagaate	teegaettag	atcaataatt	atctacttag	getgeacatt	ggaatcacet	gagagttaaa	aaaccaggat	aacctetace	tatateteat	ctccagcaat	totatata
	eee uggug ce	and a grand a g	g cong ogg cc	accurcuag	geogenearce	ggaa course c	gagag caaaa	anderenggar		ag ag ac caa c		and and and and

Pseudo 5' Splice Sites and 3' Splice Sites

Authentic 5' Splice Sites and 3' Splice Sites

Hore's floated a floated a

SpliceRack

Largest and more thoroughly curated collection of splice sites.

Allows exploration of splice sites in 5 genomes (H. sapiens, M. musculus, D. melanogaster, C. elegans, A. thaliana).

Uses new methods of classifying splice sites into the four categories.

Identified rare non-canonical sites, conserved in several species.

Offers a platform for further genomic exploration.

Sheth et. al., NAR, 2006, http://katahdin.cshl.org:9331/SpliceRack

Characterizing the Splice sites

Position Weight Matrix

		A	С	G	Т
-3		0.334	0.362	0.185	0.119
-2		0.637	0.107	0.115	0.141
-1		0.099	0.027	0.805	0.069
1		0.000	0.000	1.000	0.000
2		0.000	0.000	0.000	1.000
3	and the second	0.597	0.027	0.350	0.026
4		0.699	0.071	0.119	0.111
5		0.089	0.054	0.782	0.075
6		0.181	0.149	0.193	0.477
7		0.296	0.194	0.295	0.215
8		0.226	0.250	0.237	0.287
9		0.224	0.262	0.242	0.272
10		0.227	0.238	0.255	0.280

Occurrence of G at -1

U2: GT_AG 5'ss: Consensus Motif \rightarrow <u>CAG</u> <u>GTAAGT</u>

5' is snRNA-mRNA binding, while 3' is protein-mRNA binding, less sequence-specific.

One reason for prediction failure

The weight matrix reflects UI-snRNP 5' splice-site binding

Exon Intron Ul snRNA GUC CAUUCA rank 10 síte CAG GTAAGT

Exon Intron Ul snRNA GUC CAUUCA rank 1 síte CAGGTGAGG

5

Summary I

Collections of splice sites give useful information about the splicing machinery.

5' donor sites show ancient origins, while 3' acceptor sites show shuffling between U2 and U12 types.

Weight matrices can explain many features of splice sites, but not all.

Next Level of splice site characterization

2-pt correlation is a natural extension.

Only feasible at the 5'ss GT-AG-U2 type

TAAGTAAGI -3-2-1 1 2 3 4 5 6 Observed frequency = O(-3T, -1A) = 2537Expected frequency = E(-3T, -1A) $= p(-3T) \times p(-1A) \times N(5'ss)$ = 0.1192 × 0.0995 × 183682 = 2178.11

Correlation = (Obs. freq. - Exp. freq)/Exp. Freq = 0.164

Color-Scheme for Correlations

Two nucleotide correlation in H. sapiens

H.sapiens

D.melanogaster

C.elegans

A.thaliana

Cross-species dinucleotide correlations

A.thaliana

C.elegans

Are the correlations relevant?

Second-order effect, probably visible only in weak splice sites.

Context is important in splicing, experiments difficult to perform. But Nature provides us with two examples

Mutations – sporadic changes that cause disorders.

 SNPs - Single Nucleotide Polymorphisms, natural differences at single positions in the DNA between people.

+3 A-->G mutation in disease

In collaboration with Brage S. Andresen's lab

Research Unit for Molecular Medicine and Institute of Human Genetics, Aarhus University, Aarhus, Denmark

SBCADD: Autosomal recessive disorder of L-Isoleucine catabolism

Madsen et al., HG, 2006

U1 snRNA GUC CAYYCA CAG/GTAAGT -3-2-1/123456 GGG/GTACAT GGG/GTACAT GGG/GTGCAT

<u>4A 4C 4G 4T 5A 5C 5G 5T 6A 6C 6G 6T</u>

3A				
3C				
3G				
3T				

Use of splice-site database and tools

- We do better than many machine-learning based approaches.
- We cannot ab-initio predict good splice sites
- BUT, we can tell when a splice-site change will be deleterious, and have explained several raregenetic disorders using our splice-site database.

Muscular Dystrophy

Genetic Disorder leading to muscle wasting
Defects in Dystrophin, helps connect cytoskeleton of muscle fibre to the surrounding ECM.
on X, 2.6MB gene, 97 exons, 16 hours to transcribe, mRNA is 14kb, protein 3500

Spinal Muscular Atrophy (SMA)

SMA is caused by the homozygous loss of the survival of motor neuron 1, telomeric (SMN1) gene; either by deletion or rarely by mutation. Leads to death of affected patient, in childhood.

Humans have a paralogue gene called SMN2, also on chromosome 5, which differs from SMN1 by 11 nucleotides but has an identical coding sequence. One of the nucleotide changes between SMN1 and SMN2 genes is a C-to-T transition within exon 7, and although it is a synonymous change, it weakens the 3' splice site, resulting in skipping of exon 7. Some good transcripts are made. Can you make SMN2 take over the role of SMN1 ?

Mechanism of action of an antisense drug that modulates SMN2 splicing.

Rigo F et al. J Cell Biol 2012;199:21-25

What is RNAi ?

- RNA interference.
- Mechanism for gene regulation.
- PTGS (non-coding genes, miRNAs).
- •TGS (epigenetic modifications).
- Silencing of transposons, stem-cell differentiation etc. Other components of the pathway (e.g. piRNAs)

Non-coding silencing genes

miRNA Targeting rules

GeoSeq-Algorithm

used to detect spurious miRNA in miRBASE

Geoseq: a tool for dissecting deep-sequencing datasets. BMC Bioinformatics.2010 Oct 12;11:506. PubMed PMID: 20939882;

Protocol to prepare small RNAs for deep-sequencing

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

Different adaptor sequences give different results from the same sample

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2. 42

Mount Sinai School of Medicine

.....

A gel-shift to show 3'-adapter ligations to a synthetic construct depend on sequence of the adapter

Mount Sinai School of Medicine

Modeling the ligation efficiency and experimental verification of the model

 $m_{in}^k = F_{ij} * M^k * E_{mn}$

A-293T_fNN	B-293T_eNN
AG-GA.103.fNN_eCT - AG-GA.103.fNN_eCT - CA-AG.93.fNN_eCT - CA-AG.93.fNN_eCT - TA-AG.18a.fNN CA-AG.17.fNN_eCT - TA-GT.92a.fNN_eCT - TA-GT.92a.fNN_eCT - TA-GG.16.fNN_eCT - TA-GG.20a.fNN_eCT - TA-AG.20a.fNN_eCT - TA-AG.20a.fNN_eCT - TA-AG.20a.fNN_eCT -	AG-GA.103.fTC_eNN - AG-GA.103eNN - CA-AG.93eNN - CA-AG.93eNN - TA-AG.18a.fTC_eNN - CA-AG.17.fTC_eNN - CA-AG.17.fTC_eNN - TA-GT.92a.fTC_eNN - TA-GT.92aeNN - TA-GG.16.fTC_eNN - TA-CG.16eNN - TA-AG.20a.fTC_eNN -
A A COLOR A CO	10.10 10.00 10 10 10 10 10 10 10 10 10 10 10 10 1

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

Most variability comes from the 2 nucleotides at the adapter's ligating ends

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011 Nov; 39(21):e141. Epub 2011 Sep 2.

Final improved protocol for unbiased small RNA sequencing

Use of random-ends (N's) on the adapter averages biases

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

Measurement of dme-mir-6 from early drosophila embryos in
various experiments over the years.

Publication	3' adapter	5' adapter	Platform	3'-adapter ligation	5'-adapter ligation	Result
Elbashir <i>et al.</i> , G&D, 2001[27]	UUUaaccg (RNA/dna hybrid)	tcactAAA (RNA/dna hybrid)	Sanger	T4 RNA- ligase	T4 RNA ligase	miR-6-3p high , miR-6-1,2,3-5p low
Aravin <i>et al.</i> Dev. Cell, 2003 [28]	UUUaaccg (RNA/dna)	tcactAAA (RNA/dna)	Sanger	T4 RNA- ligase	T4 RNA ligase	miR-6-3p high , miR-6-1,2,3-5p low
Ruby <i>et al.</i> Genome Res., 2007 [29]	CTGTAGG(DNA)	UGAAA (RNA)	454	T4 RNA- ligase	T4 RNA ligase	miR-6-3p high, miR-6-1,2,3-5p low
Wang and Liu Frontiers in Gen., 2011[30]	CTGTAGG(DNA) adenylated 5' dideoxy	3' CGCAUC (RNA)	GAII- (Solexa)	Truncated T4 RNA- ligase	T4 RNA ligase	miR-6-3p low, miR-6-1,2-5p low, miR-6-3-5p high

★ mod-ENCODE data is inconsistent with other data

Mount Sinai School of Me

	NAME	SEQUENCE			
	dme-miR-6-1-5p (star-1)	agggaauaguugcugugcugua			
	dme-miR-6-2-5p (star-2)	agggaacuucugcugcugauaua agggaacgguugcugaugaugua uaucacaguggcuguucuuuuu			
	dme-miR-6-3-5p (star-3)				
	dme-miR-6-3p (mature)				
	Three instances of dme-r	niR-6 in drosophila genome			
edic	ine from the	same cluster			

47

unpublished

Correct distribution of miRNAs is restored using our new protocol with randomized adapter ends

Curing hepatitis-C

Hep-C virus binds to mature miR-122 and uses that to enhance its replication Blocking this binding is one way to halt infection and effect a cure

Inhomogenous coverage of exons

How do you assign a number to the expression level of a gene from such inhomogenous data ?

Gene Ontology (GO)

From Gene Ontology, a portion of the biological process ontology describing DNA metabolism. Note that a node may have more than one parent. For example, *DNA ligation* has three: *DNA-dependent DNA replication*, *DNA repair* and *DNA recombination*. *This is a directed acyclic graph (DAG)*.

•GO organizes the genes in hierarchies, according to function.

•Changes in expression levels are used to identify branches or pathways in GO that are affected in the experiment

Gene Ontology : http://www.geneontology.org/

GO Enrichment:

Gene Ontology allows the genes to be mapped to processes via gene products. For a given list of genes (selected via any class comparison like SAM etc.), the GO processes that are enriched in it can be identified. There are standard tools like DAVID that perform these functions. A hypergeometric test can be used to test for GO term enrichment.

Bayesian Networks.

Causality relations can be derived from data and modeled using bayesian networks.

Heat Maps

•A matrix with genes in the rows and experiments in the columns

•Colors represent values relative to mean

•Visually striking patterns can be detected, especially by organizing rows and columns, using clustering methods (shown in the next slide)

Heat maps with clustering of rows and columns

Patterns in the colors show groups of genes/experiments that are co-regulated

Questions ?

Contact me at <u>ravi.mssm@gmail.com</u> for criticisms, comments and questions.

I have tried to give a quick overview in lectures 1 and 2 of the broad ideas in genomics/ bioinformatics.

I have attached to this presentation, slides containing notes, for some material I worked out on the board, as well as some extra population genetics material that I did not have time to get into.