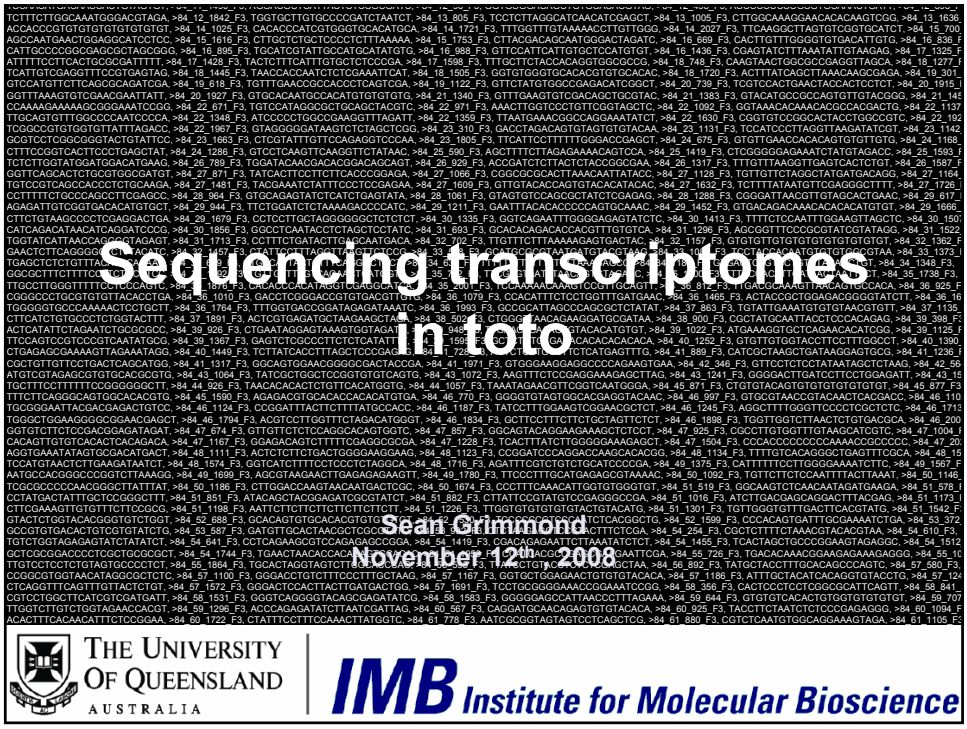
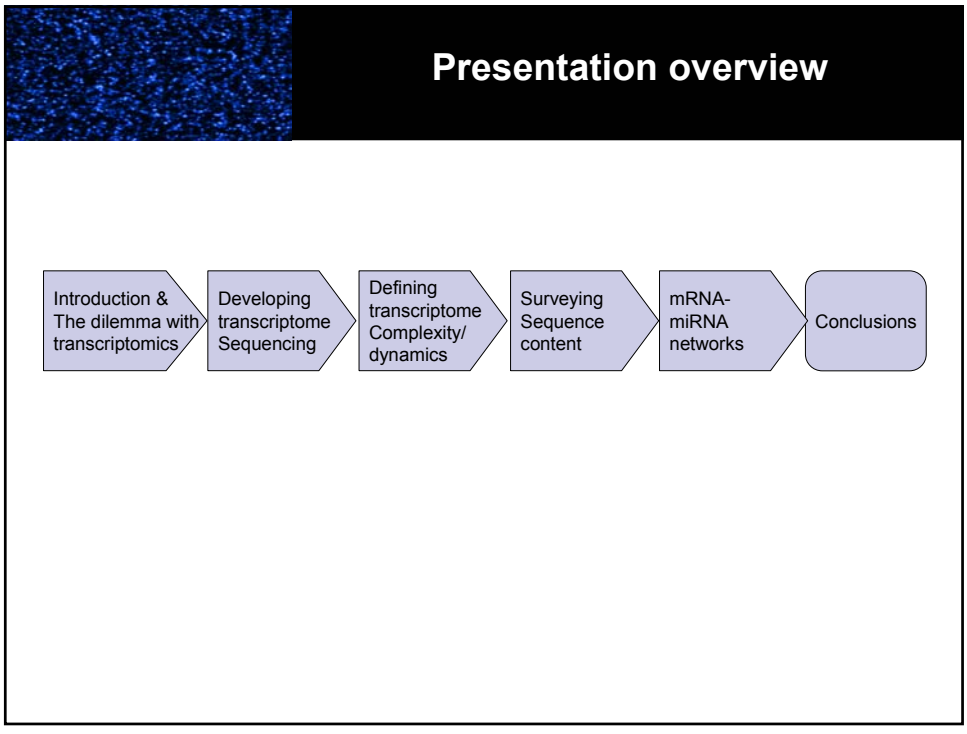

Sequencing transcriptomes in toto
Sean Grimmond
November 12th 2008
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
IMB Institute for Molecular Bioscience


Presentation overview
Introduction & The dilemma with transcriptomics
Developing transcriptome Sequencing
Defining transcriptome Complexity/ dynamics
Surveying Sequence content
mRNA-miRNA networks
Conclusions

Presentation overview

Introduction & The dilemma with transcriptomics

Developing transcriptome Sequencing

Defining transcriptome Complexity/ dynamics

Surveying Sequence content

mRNA-miRNA networks

Conclusions

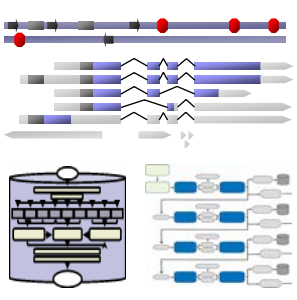
Deriving biological insight from the transcriptome

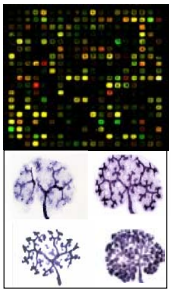
Genome, transcriptome, and proteome annotation

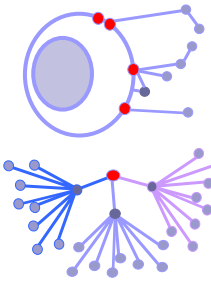
Survey gene expression

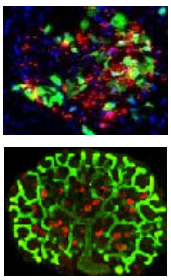
Build networks & candidates

Functional Validation

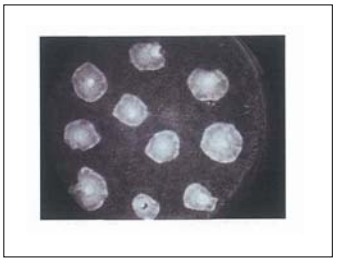


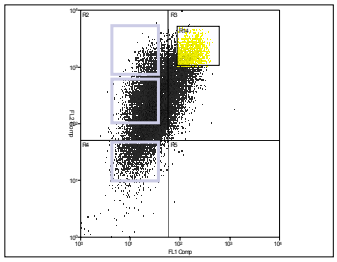


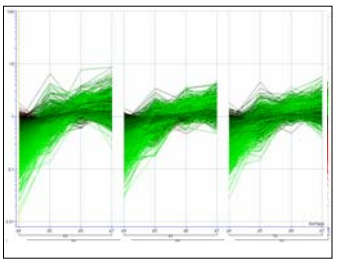


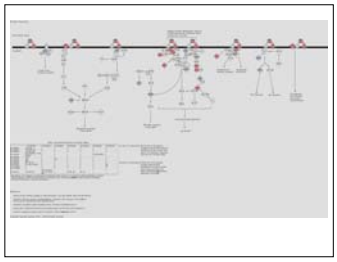


Transcriptome based screening of Human ES cell surface markers



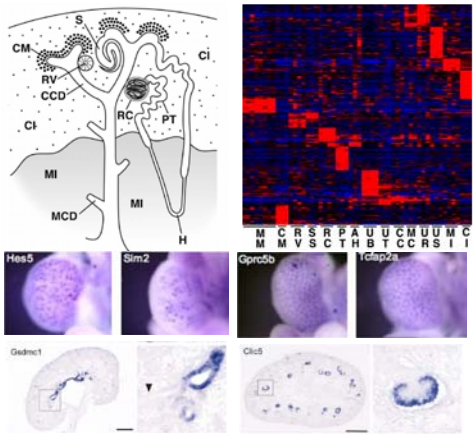


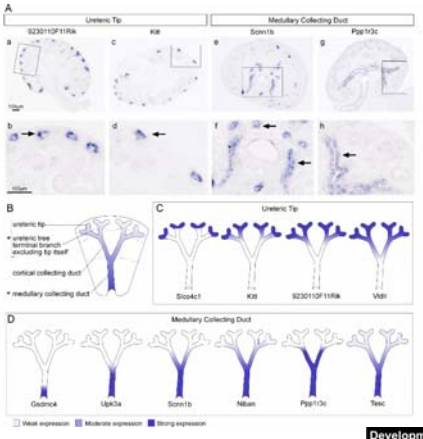




Laslett et al, BMC Dev Biol. 2007

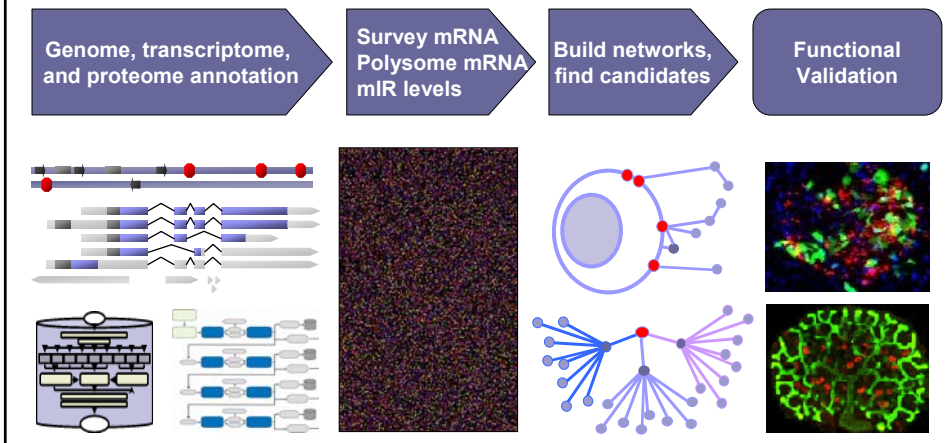
In situ based profiling of Urogenital development



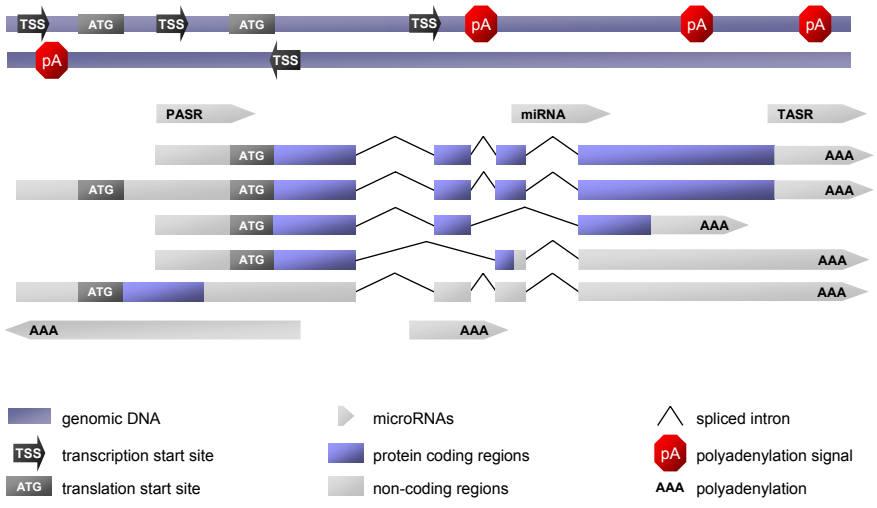


Brunskill et al, Developmental Cell, 2008

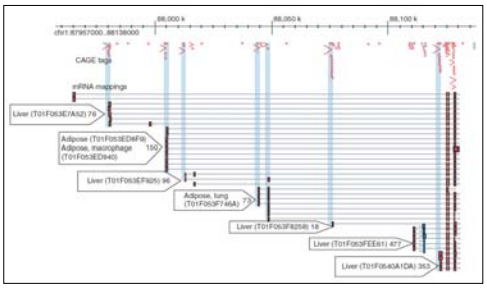
Deriving biological insight from the transcriptome



Representative Gene Vs Transcriptional complexity



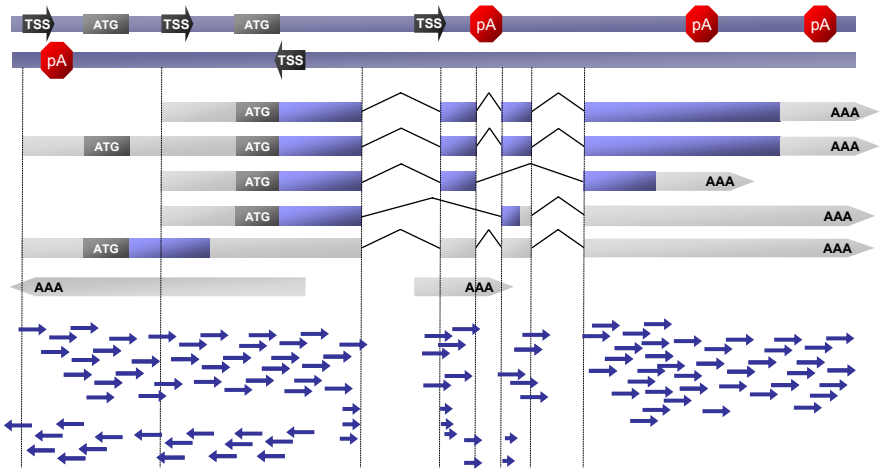
Creating transcriptional frameworks against the genome:



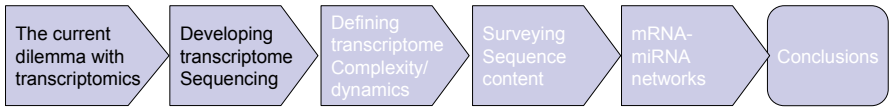
Exhaustive surveying of expression space:
~30,000 genes, >100,000 transcripts, > 65,000 ORFs
Multiple promoters, Novel layers of control…

Carninci et al, Nat Genet, 2005.

Using shotgun sequencing to survey transcriptomes?



Presentation overview



Sequencing by ligation

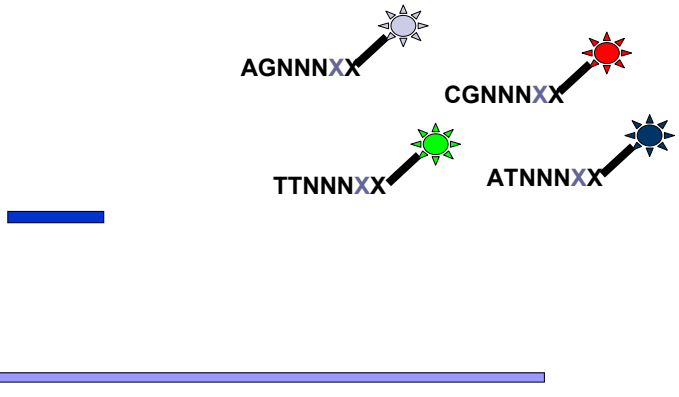
Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome

Jay Shendure,[1*†] Gregory J. Porreca,[1*†] Nikos B. Reppas,[1] Xiaoxia Lin,[1] John P. McCutcheon,[2,3] Abraham M. Rosenbaum,[1] Michael D. Wang,[1] Kun Zhang,[1] Robi D. Mitra,[2] George M. Church[1]

We describe a DNA sequencing technology in which a commonly available, inexpensive epifluorescence microscope is converted to rapid nonelectrophoretic DNA sequencing automation. We apply this technology to resequence an evolved strain of *Escherichia coli* at less than one error per million consensus bases. A cell-free, mate-paired library provided single DNA molecules that were amplified in parallel to 1-micrometer beads by emulsion polymerase chain reaction. Millions of beads were immobilized in a polyacrylamide gel and subjected to automated cycles of sequencing by ligation and four-color imaging. Cost per base was roughly one-ninth as much as that of conventional sequencing. Our protocols were implemented with off-the-shelf instrumentation and reagents.

Shendure et al (2005) Science 309:1728

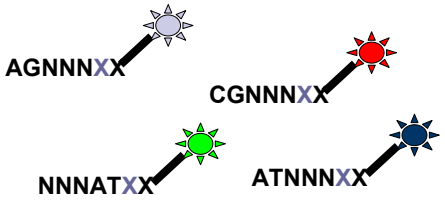
SOLiD sequencing



Sequencing by Supported Oligo Ligation and Detection (SOLiD).

1. Anneal sequencing primer to amplicon.
2. Anneal and ligate first dinucleotide detection probe. From pool of 16.
3. Wash and capture image of slide.
4. Cleave after dinucleotide detector and wash.
5. Repeat cycle 5-7 times (25-35 base read).

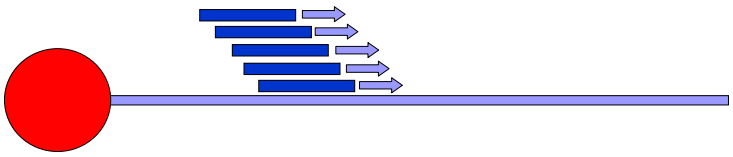
SOLiD sequencing



Sequencing by Supported Oligo Ligation and Detection (SOLiD).

1. Anneal sequencing primer to amplicon.
2. Anneal and ligate first dinucleotide detection probe. From pool of 16.
3. Wash and capture image of slide.
4. Cleave after dinucleotide detector and wash.
5. Repeat cycle 5-7 times (25-35 base read).

SOLiD sequencing



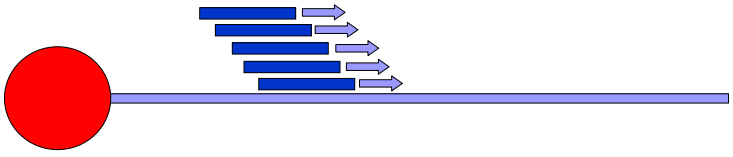
```
PrimerA:   T  T 03 04 05  A  T 08 09 10  C  C 13 14 15  G  G 18 19 20 ..
PrimerB:   T 02 03 04  C  A 07 08 09  G  C 12 13 14  T  G 17 18 19  A ..
PrimerC:  01 02 03  A  C 06 07 08  A  G 11 12 13  A  T 16 17 18  T  A ..
PrimerD:  01 02  G  A 05 06 07  A  A 10 11 12  A  A 15 16 17  C  T 20 ..
PrimerE:  01  T  G 04 05 06  T  A 09 10 11 12 13 14 15 16  G  C 19 20 ..

Sequence   T  T  G  A  C  A  T  A  A  G  C  C  A  A  T  G  G  C  T  A
```

- No homo-polymer problems, no phase problems
- Errors are random
- Each base is _read_twice_
- Independent measures of each base >> raw accuracy (99.5% for single runs)
- Consensus accuracy is 99.97% at 10x (based on human BAC sequencing)

Output = 80-160,000,000 reads = 6.0Gb-18.0 Gb per run (3-9Gb per slide)
Latest output is > 200,000,000 reads per slide at 50mers per run.

SOLiD sequencing



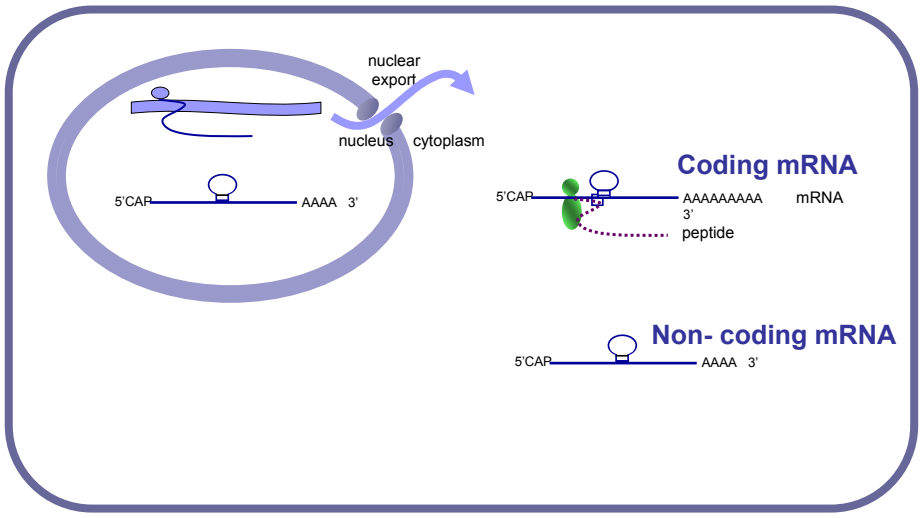
```
PrimerA:   T  T 03 04 05  A  T 08 09 10  C  C 13 14 15  G  G 18 19 20 ..
PrimerB:   T 02 03 04  C  A 07 08 09  G  C 12 13 14  T  G 17 18 19  A ..
PrimerC:  01 02 03  A  C 06 07 08  A  G 11 12 13  A  T 16 17 18  T  A ..
PrimerD:  01 02  G  A 05 06 07  A  A 10 11 12  A  A 15 16 17  C  T 20 ..
PrimerE:  01  T  G 04 05 06  T  A 09 10 11 12 13 14 15 16  G  C 19 20 ..

Sequence   T  T  G  A  C  A  T  A  A  G  C  C  A  A  T  G  G  C  T  A
```

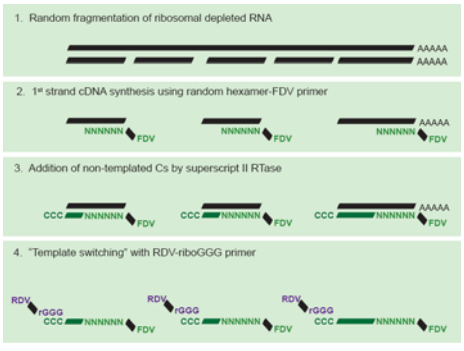
Scale:

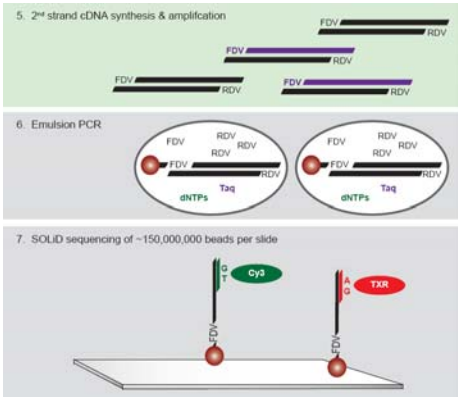
- Output = 80-160,000,000 reads = 6.0Gb-18.0 Gb per run (3-9Gb per slide)
- Latest output is > 200,000,000 reads per slide at 50mers per run.

Poly A transcriptome



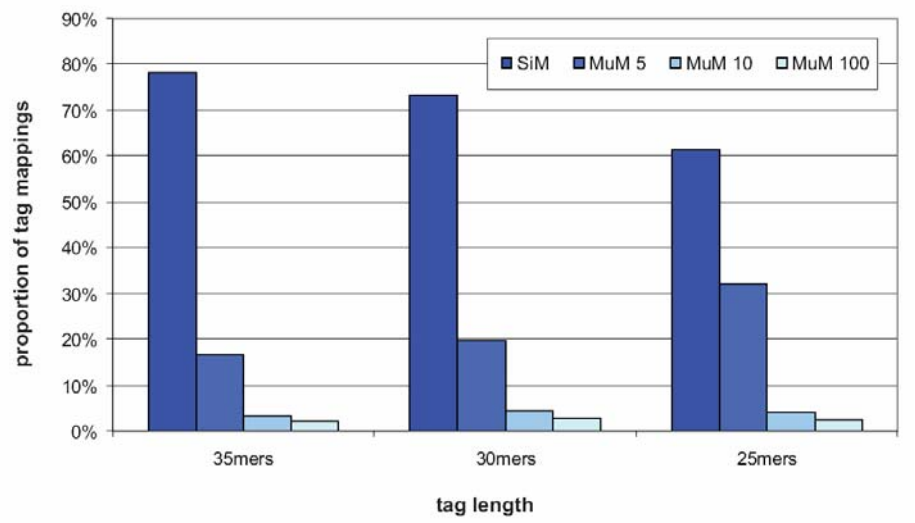
Shotgun transcriptome libraries:



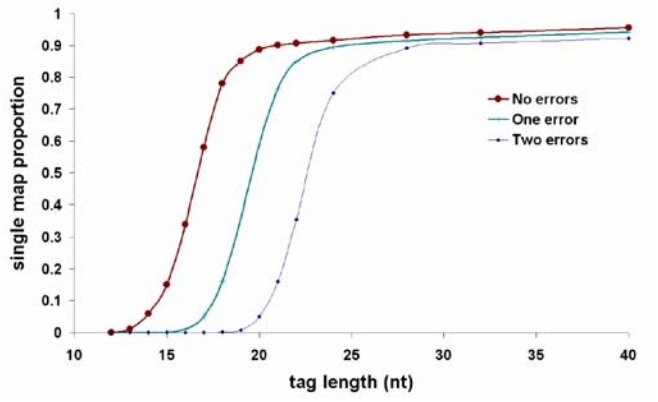


Cloonan et al, Nature Methods, 5: 613-9 (2008)

Accuracy of mapping at 25-35bp

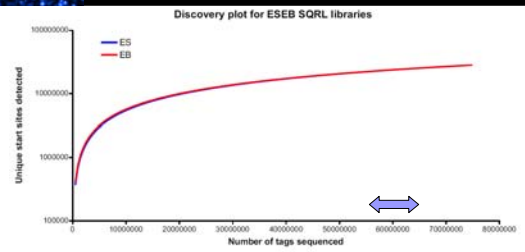


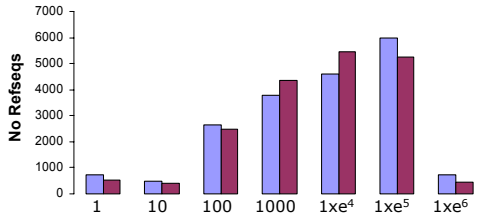
Unique mapping of short tags (impact of length & error):



Faulkner et al Genomics 91 (2008) 281–288

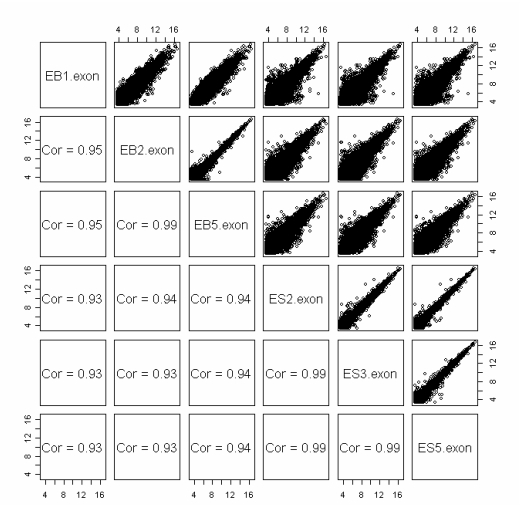
Depth of sequencing required for SQRL of mRNA





Tag count per Refseq (3Gb sequence)

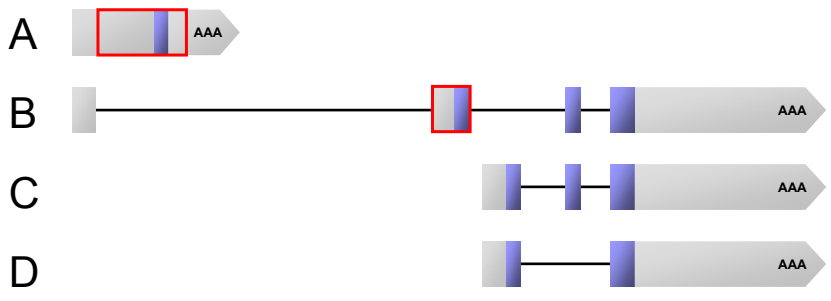
Biological and technical replication of SQRL



Presentation overview

The current dilemma with transcriptomics → Developing transcriptome Sequencing → Defining transcriptome Complexity/ dynamics → Surveying Sequence content → mRNA-miRNA networks → Conclusions

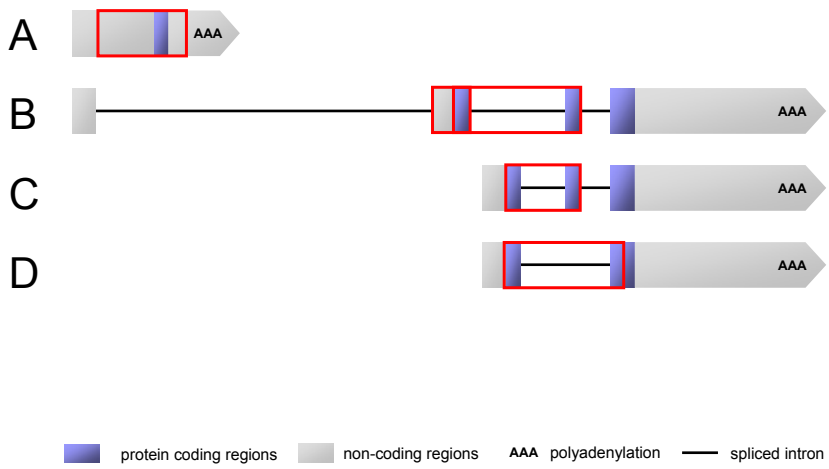
Defining transcript Specific expression



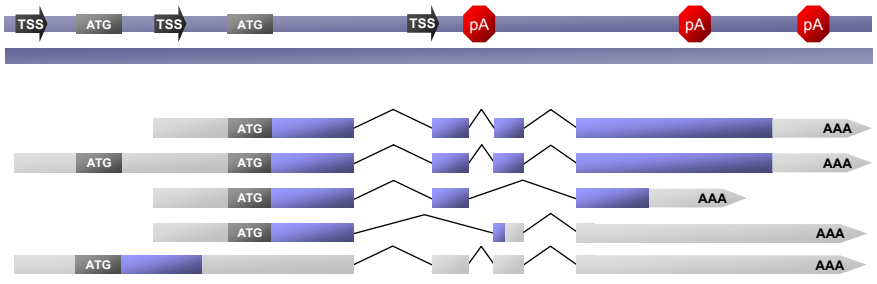
Transcripts defined by Aceview (September 2007 release)

protein coding regions | non-coding regions | AAA polyadenylation | spliced intron

“Diagnostic” features



Defining transcript Specific expression

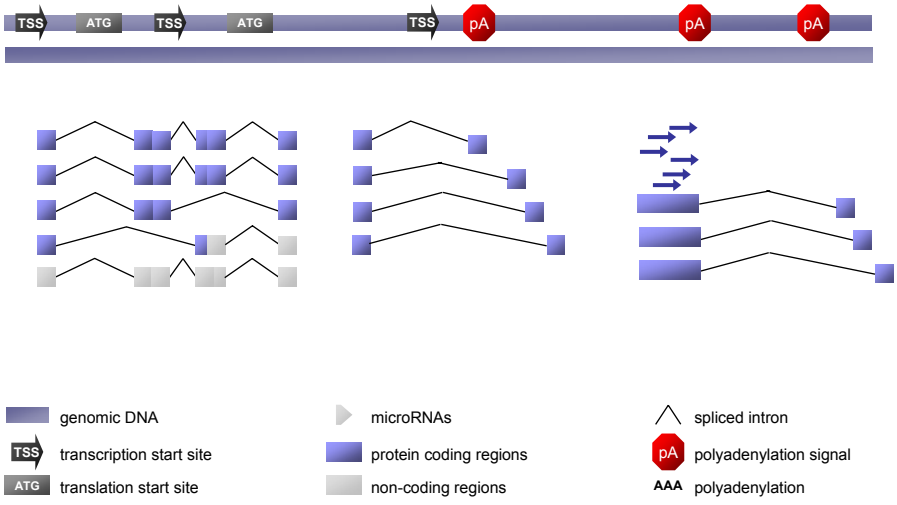


92.6% known transcripts have diagnostic features (covers 99.8% of loci)

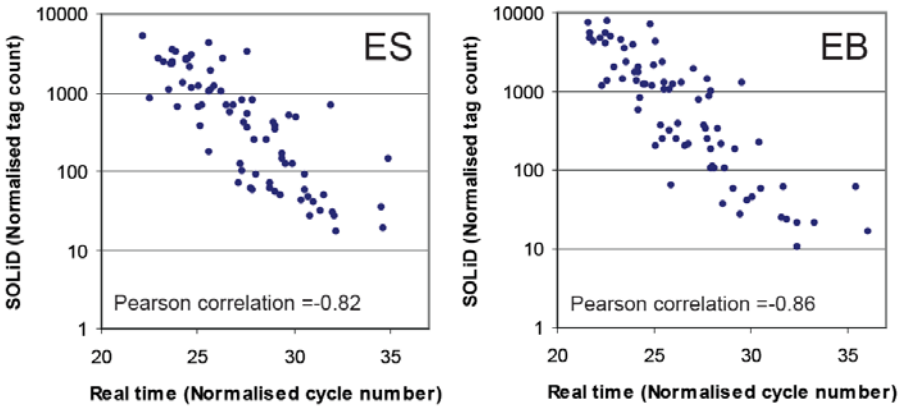
217127 diagnostic features covering 160156 individual transcripts from 65254 loci

genomic DNA
TSS transcription start site
ATG translation start site
microRNAs
protein coding regions
non-coding regions
spliced intron
pA polyadenylation signal
AAA polyadenylation

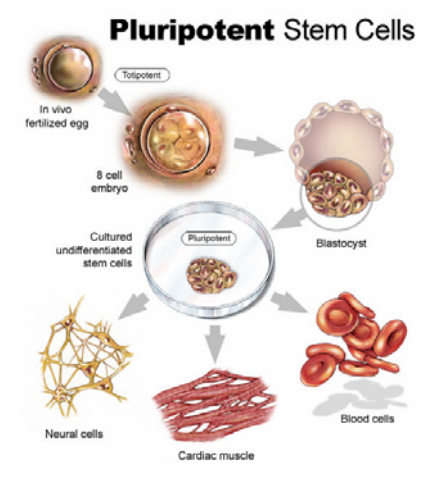
Surveying exon usage



SQRL is quantitative correlation with qRT-PCR



Key questions in ES cell biology



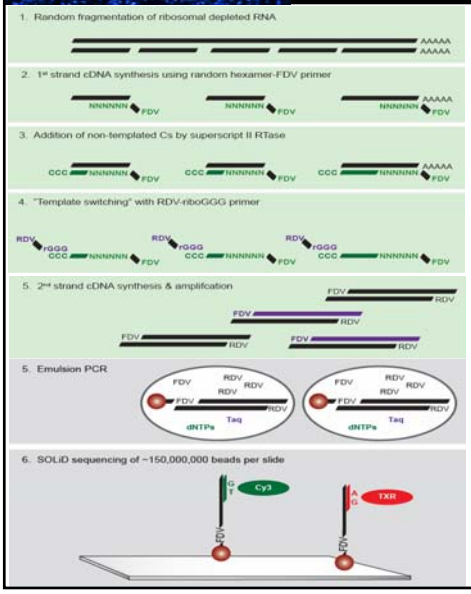
What are the signals required to maintain stem cell Pluripotency?

What are the molecular programs controlling lineage determination?

What role does RNA play in ES cell maintenance and differentation?

http://www.stemcellresearchfoundation.org/

Using shotgun sequencing to survey transcriptomes



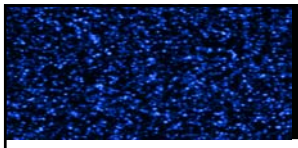
Starting material: Ribo-depleted polyA and polysome associated RNA from mouse or human ES cells & EBs.

-Sequence ~100,000,000 mappable reads per library using SOLiD.

-Reads are 25, 30 or 35bp in length

-Make 3 libraries per biological state

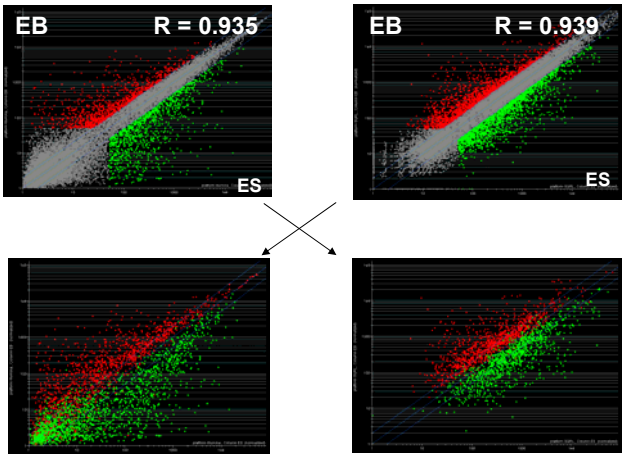
-All tags are mapped to the genome first and then to a database of all known exon junctions.



RNAseq is sensitive & quantitative

(8,800) Bead array

RNAseq (13,000)



Red: >2x up regulated in EB, **Green**: <2x down regulated in ES, **Grey**: Marginal detection (<.95 detection score for Illumina or 50tags for SQRL)

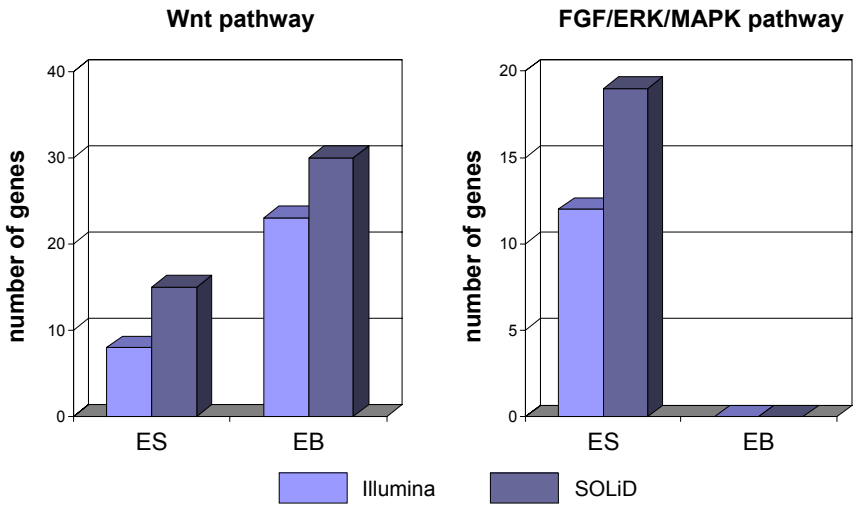
Up-regulated in EB library

Refseq UCSC ID	Bstat	Fold change	Common name	Refseq ID
16427	11.24	245.51	Hbb-bh1	NM_008219
2964	9.91	16.76	Car4	NM_007607
14226	9.47	43.17	Asb4	NM_023048
21211	9.11	28.00	Hand2	NM_010402
8843	8.77	20.70	Dkk1	NM_010051
6847	8.53	7.94	St3gal6	NM_018784
11406	8.40	42.28	Car14	NM_011797
19444	7.51	13.57	Hmgb3	NM_008253
18996	7.42	7.22	Pthr1	NM_011199
10966	7.35	7.65	Cldn11	NM_008770
2620	6.81	6.41	Tmem88	NM_025915
2957	6.69	22.21	Lhx1	NM_008498
16428	6.62	8.00	Hbb-y	NM_008221
2677	6.60	7.64	Alox15	NM_009660
11726	6.55	5.51	Slc39a8	NM_026228
21817	6.50	10.40	Evx1	NM_007966
10049	6.46	9.75	Lmo2	NM_008505
18640	6.40	5.32	Dapk2	NM_010019
2140	6.36	9.51	1500041B1	NM_029861
19723	6.30	5.86	Nxf7	NM_130888
13035	6.30	5.88	C1qdc2	NM_026125
11132	6.23	20.82	Tdo2	NM_019911
12231	6.20	43.36	Cer1	NM_009887
3097	6.15	5.24	Skap1	NM_001033186
6586	6.14	6.38	Klhl6	NM_183390
3257	6.13	7.34	Ramp2	NM_019444
11243	5.83	6.17	Efna1	NM_010107
7976	5.83	4.00	AK220484	NM_001083628

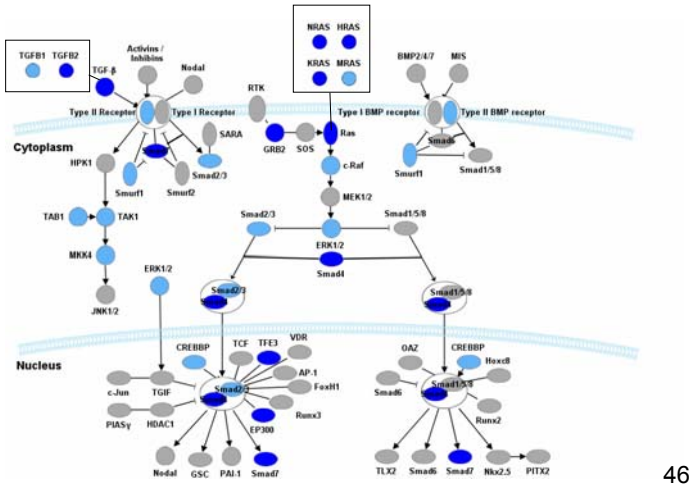
Up-regulated in ESlibrary

Refseq UCSC_ID	Bstat	Fold change	Gene symbol	Refseq ID
577	10.26	21.43	Tcfcp2l1	NM_023755
5456	9.77	41.83	Tdh	NM_021480
10018	9.54	13.69	Gm1967	NM_001033452
15887	8.95	12.79	EG435970	NM_001034893
985	8.33	19.32	Ifi202b	NM_008327
10644	8.18	7.66	Mybl2	NM_008652
6807	8.07	22.29	Dppa2	NM_028615
21838	8.01	15.36	Fgf4	NM_010202
319	7.92	8.96	Mreg	NM_001005423
15897	7.83	7.27	4933405K0	NM_028913
20848	7.82	20.54	Krt42	NM_212483
726	7.46	12.93	Nr5a2	NM_030676
5367	7.44	10.89	Tgm1	NM_019984
2133	7.42	14.45	Cobl	NM_172496
12114	7.38	23.83	Klf4	NM_010637
14257	7.34	7.26	Cav1	NM_007616
11017	7.26	8.65	Phf17	NM_172303
1041	7.23	7.13	Lefty2	NM_177099
1597	7.22	6.78	6330514A0	NM_183152
13160	7.21	10.56	Cnpy1	NM_175651
12236	7.11	6.54	Bnc2	NM_172870
1043	7.09	23.63	Lefty1	NM_010094
15996	7.08	12.92	Hsd17b14	NM_025330
7789	7.00	12.23	1700061G1	NM_030141
21186	6.86	8.41	E130014J0	NM_001040400
17531	6.85	10.96	Clgn	NM_009904
16769	6.80	29.50	Mylpf	NM_016754
7565	6.79	7.14	Zfp57	NM_001013745

Better sensitivity improves functional annotations



Increased transcriptional complexity of TGF-β signalling in ES

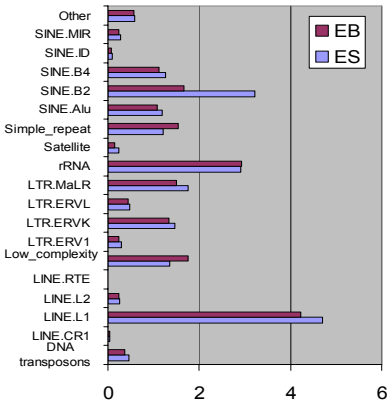


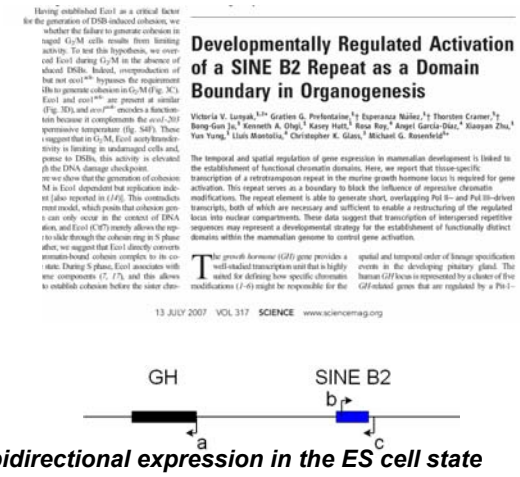
46 nodes in the pathway
25 loci transcriptionally active in ESCs
41 variant mRNAs expressed in ESCs

Transcriptome Discovery: in ES cells and EBs

Category	ES	EB
Known exons	74.2%	75.5%
Predicted exons	5.6%	7.8%
Known regions	7.9%	5.6%
Predicted regions	4.8%	3.9%
Conserved regions	2.8%	3.2%
Other regions	5.6%	7.8%

Widespread and dynamic Repeat expression:



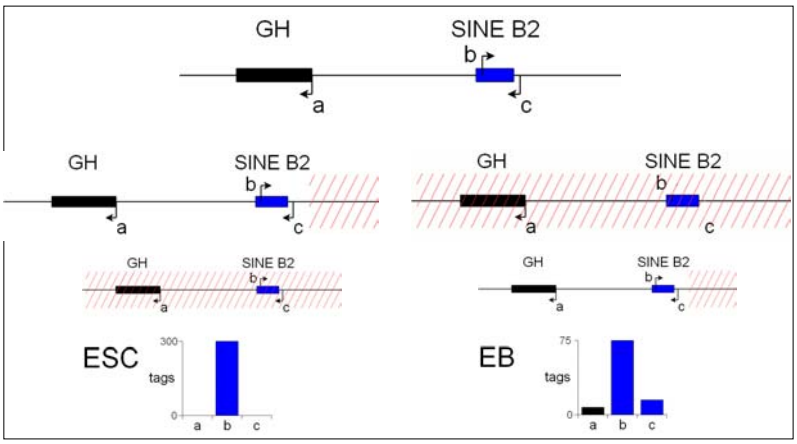


Developmentally Regulated Activation of a SINE B2 Repeat as a Domain Boundary in Organogenesis

13 JULY 2007 VOL 317 SCIENCE www.sciencemag.org

>300 Repeat elements display bidirectional expression in the ES cell state

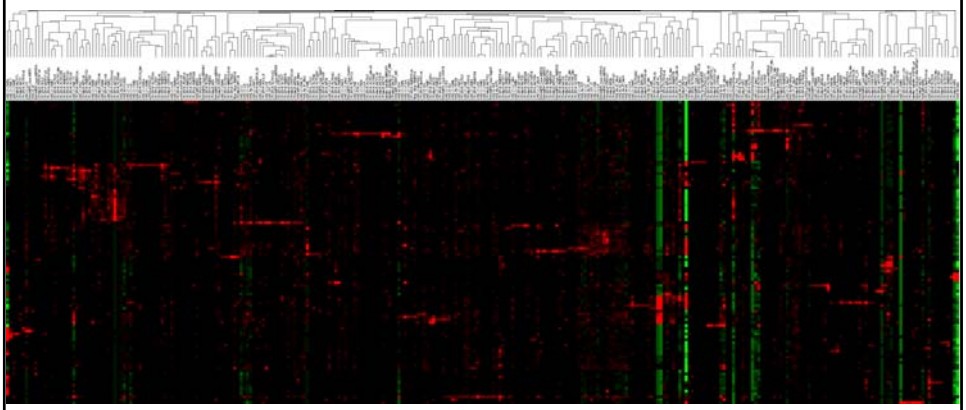
TEs can act as locus boundary elements:



>300 Repeat elements display bidirectional expression in the ES cell state

Lunyak, V. V., et al., *Science* (2007).

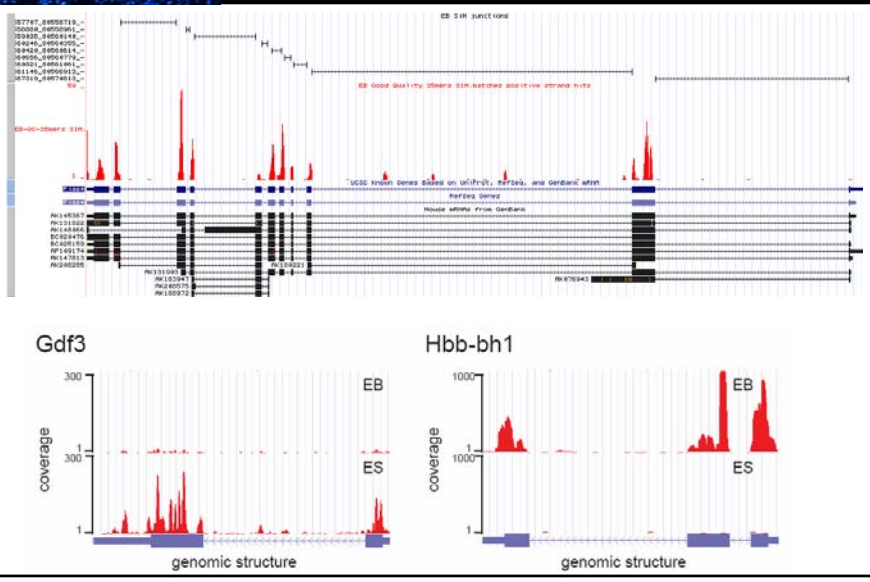
Repeat element expression across 50 mouse tissues



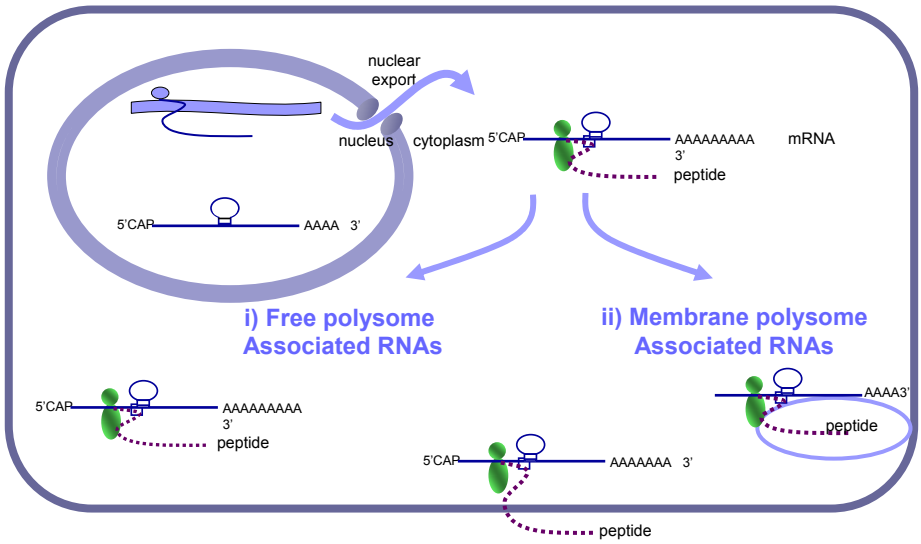
TE can act as alternate promoters

- TEs known to provide alternative promoters to nearby genes (Speek et al, Genomics, 2008).
- CAGE delineated >100,000 TE promoters
  - -<100kb upstream of a RefSeq
  - -same strand as a RefSeq
  - -not overlapping a RefSeq
- >700 confirmed as alt. promoters by ESTs
- Confirmation of novel promoters RT-PCR, cDNA sequencing and RACE

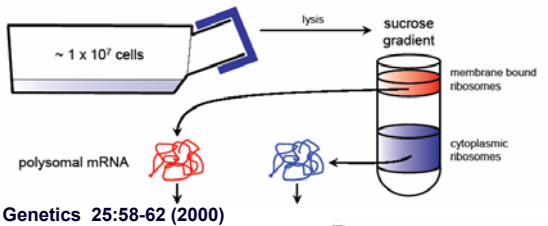
Viewing exon and junction expression in a genomic context:



The Polysome-associated transcriptome:

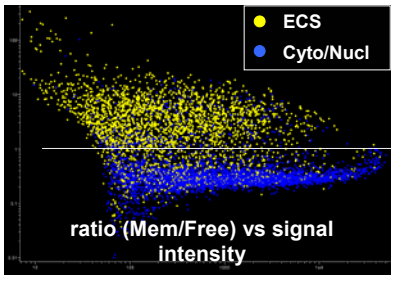


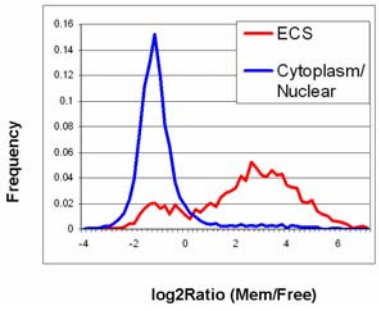
hES PolyA Vs Polysome RNA sequencing:



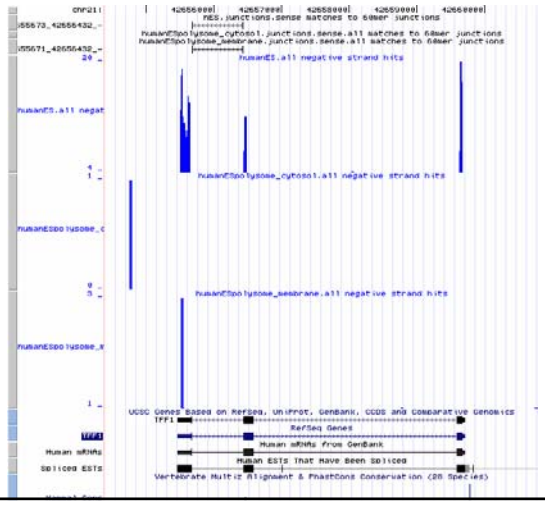
Diehn *et al.*, Nature Genetics 25:58-62 (2000)

B



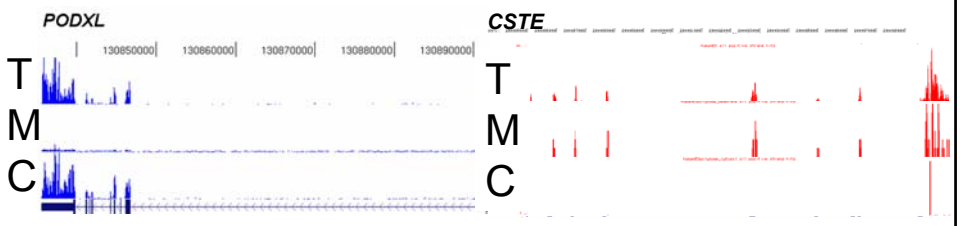


Expression Vs translation



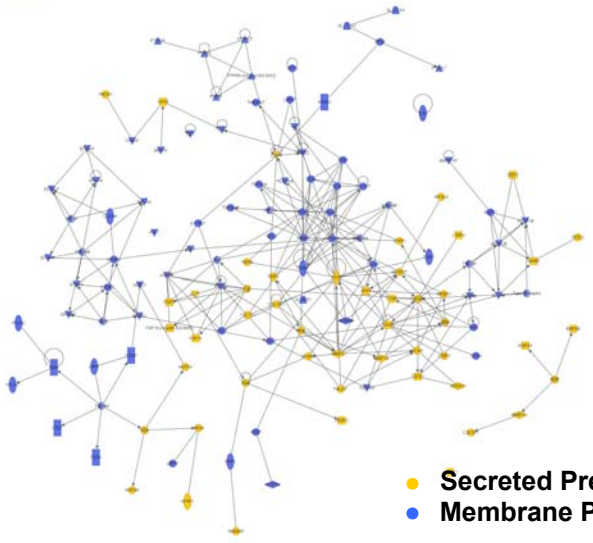
Tag count
Total: 1115
Cytosol: 0
Membrane: 3

hES PolyA Vs Polysome RNA sequencing:

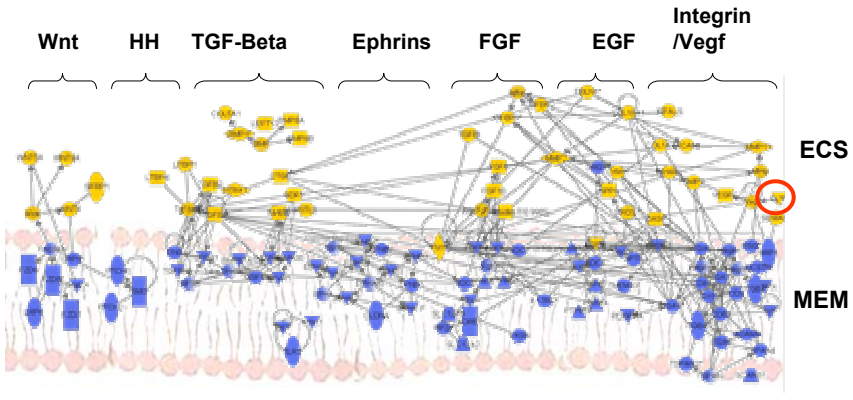


	Total	Cytosol	Membrane
Expressed Refseq genes	12,497	9,387	6,700
All aceview diagnostic variants	53,190	26,005	11,060
All aceview diagnostic loci	30,281	17,851	9,160
Ratio (variants/loci)	1.76	1.46	1.21

Network analysis of ECS signalling:



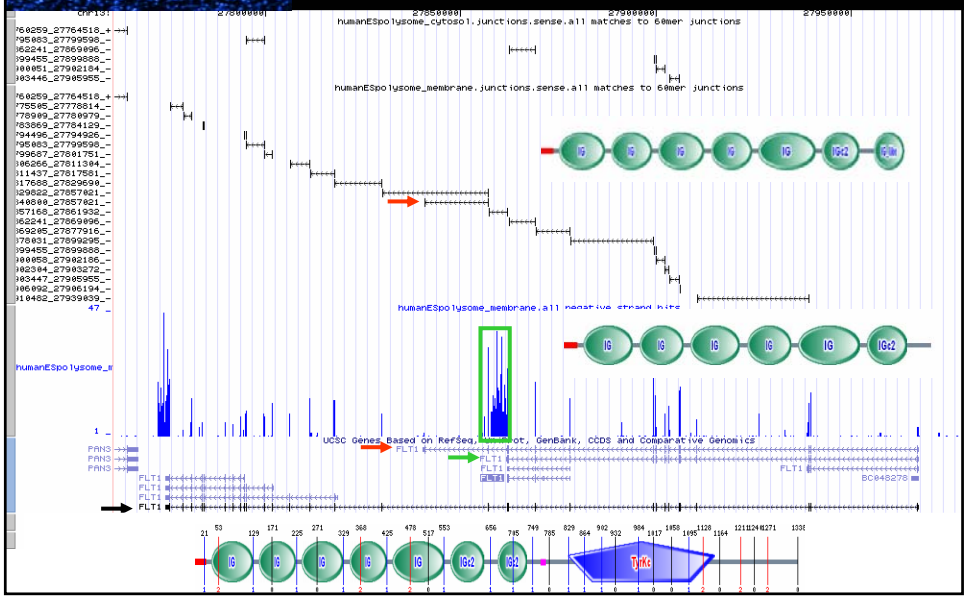
Modelling the extracellular space in ES cells:



Ingenuity Pathways Analysis ™

Secreted Predictions
Membrane Predictions

Transcription output of VEGFR1 In hES:



Presentation overview

The current dilemma with transcriptomics

Developing transcriptome Sequencing

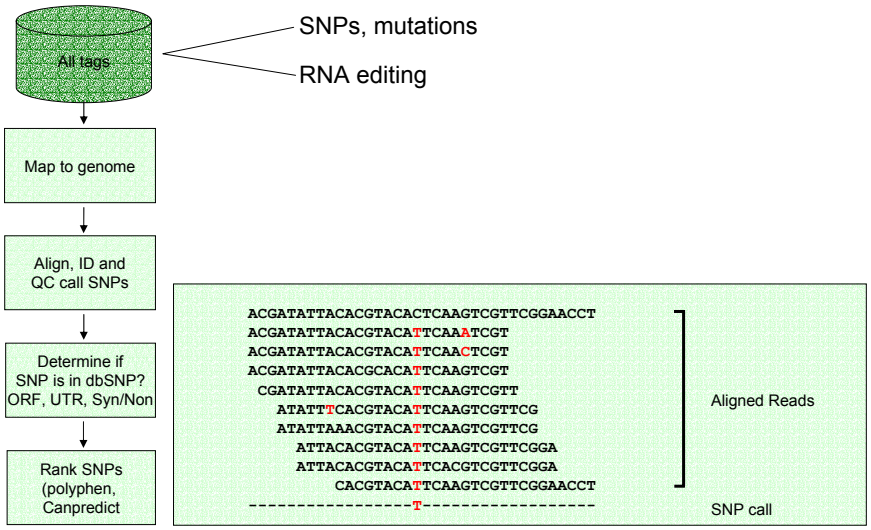
Defining transcriptome Complexity/ dynamics

Surveying Sequence content

mRNA-miRNA networks

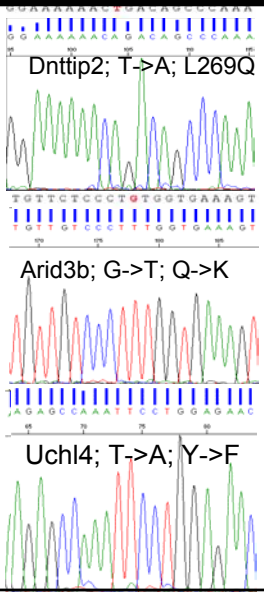
Conclusions

Screening sequences for substitutions (expressed SNPs, mutations, RNA editing)

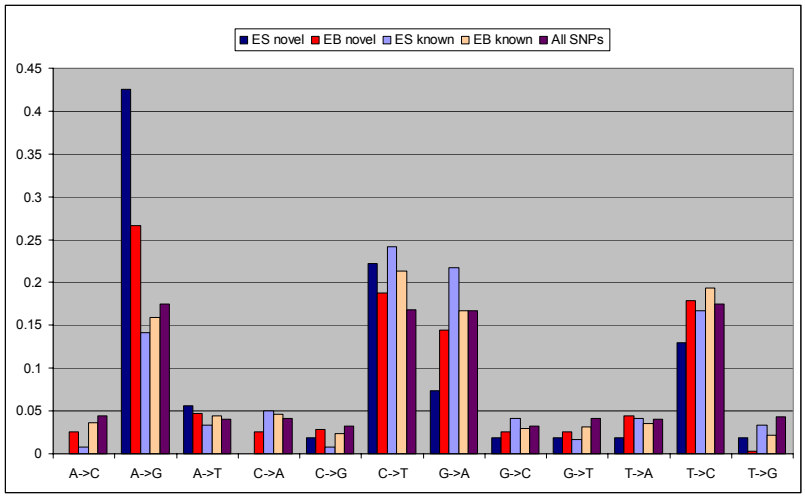


Validation of sequence variations

Chromosome	Position	Strand	Gene UTRme	Predicted SNP	Amino acid change	Validated
chr1	174437452	+	Tagln2	C->T	*UTR*	Yes
chr1	34501748	+	Imp4	C->T	*UTR*	Yes
chr11	84777618	+	Car4	C->T	A->V	Yes
chr14	22616785	+	Samd8	A->G	*UTR*	Yes
chr18	38460154	+	Rnf14	C->T	A->V	Yes
chr19	41992404	+	Pgam1	C->T	*UTR*	Yes
chr2	119453391	+	Nusap1	A->G	K->R	Yes
chr3	121978860	+	Dnttip2	T->A	L->Q	Yes
chr3	93330080	+	S100a11	C->T	*UTR*	Yes
chr4	15903284	+	Nbn	C->G	Y->F	Yes
chr5	138548192	+	Zkscan1	A->G	*UTR*	Yes
chr5	138548273	+	Zkscan1	A->G	*UTR*	Yes
chr7	38752418	-	C80913	T->C	N->D	Yes
chr7	50209670	-	EG668668	C->G	G->R	Yes
chr7	5076892	-	Rasl2-9	T->A	E->D	No
chr9	123371057	+	Lars2	T->G	*UTR*	Yes
chr9	57681529	-	Arid3b	G->T	Q->K	Yes
chr9	64083463	+	Uchl4	A->T	D->E	No



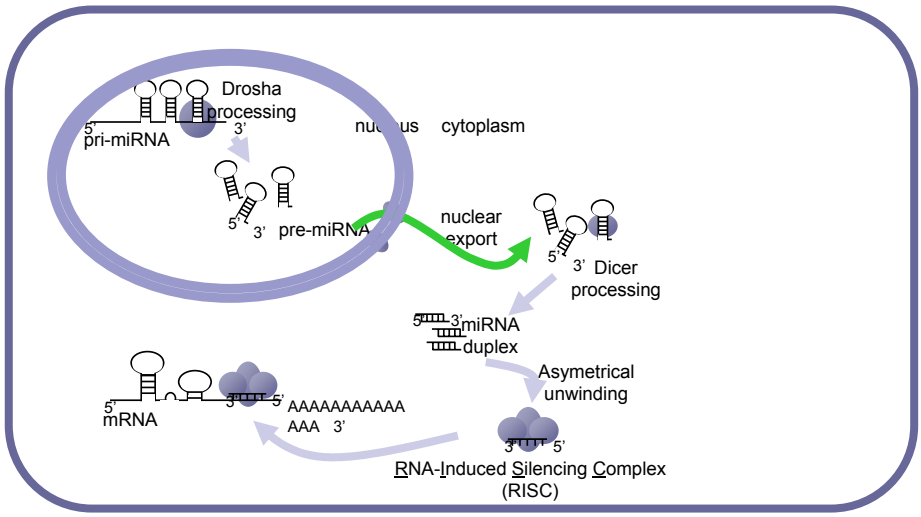
SNP detection in ES: Evidence of RNA editing?



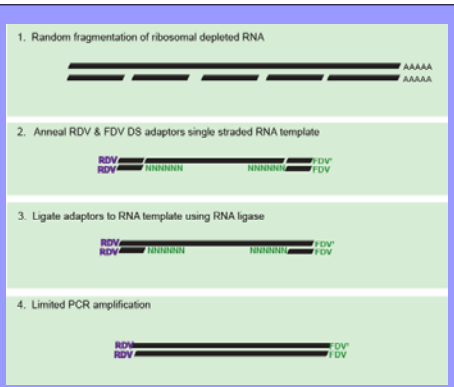
Presentation overview

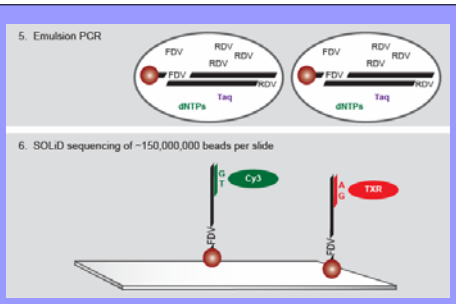
The current dilemma with transcriptomics → Developing transcriptome Sequencing → Defining transcriptome Complexity/ dynamics → Surveying Sequence content → mRNA-miRNA networks → Conclusions

miRNA suppression of mRNA targets



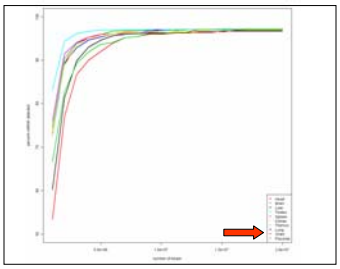
miRNA sequencing in mammalian ESCs & EBs

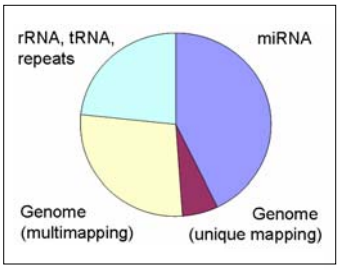


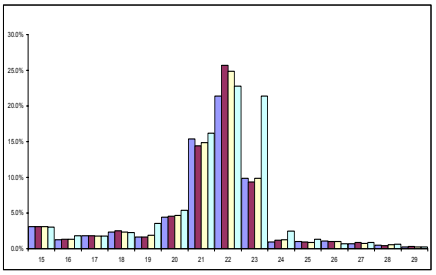


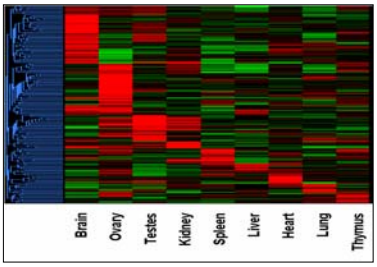
Sequence ~10,000,000 15-35bp reads using SOLiD
Map against miRBase, then the genome
Cluster miRbase matches and summarise isomiR complexity

miRNA sequencing in mammalian ESCs & EBs



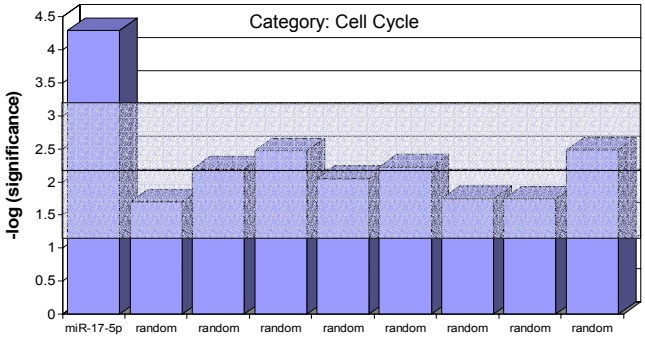






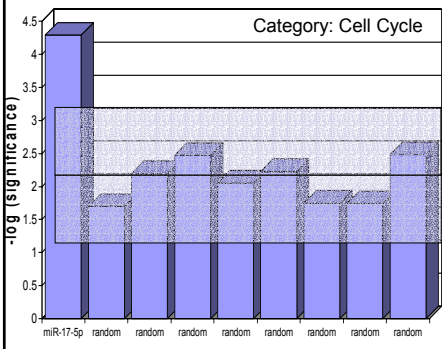
Studying miRNA mediated control of gene networks:

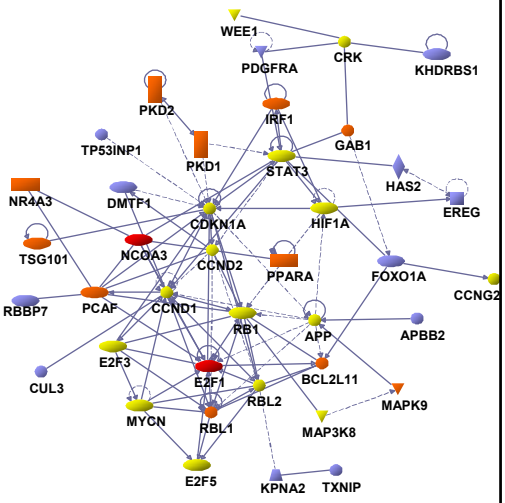
1. Define the miR targets of the polycistronic miRs (PicTar)
2. Determine functional relationship of targets (IPA)
3. Benchmark against similarly size random sets of targets



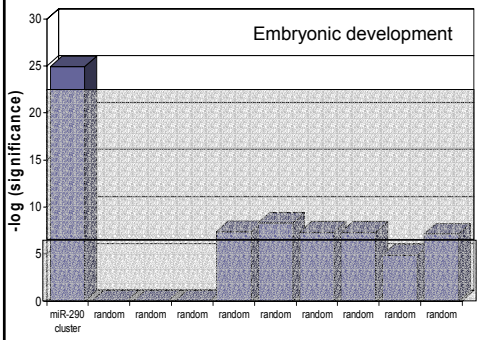
Cloonan et al, Genome Biology, 9: R127 (2008)

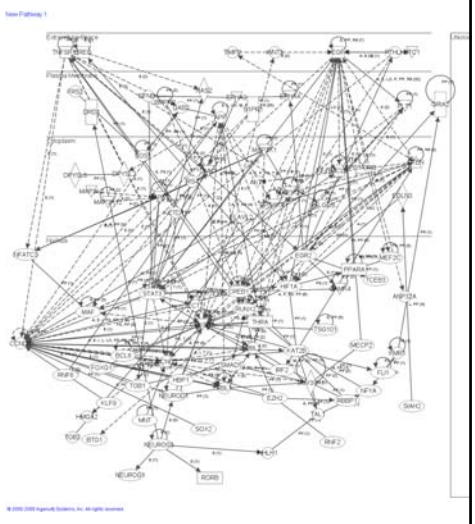
Studying miRNA mediated control of gene networks



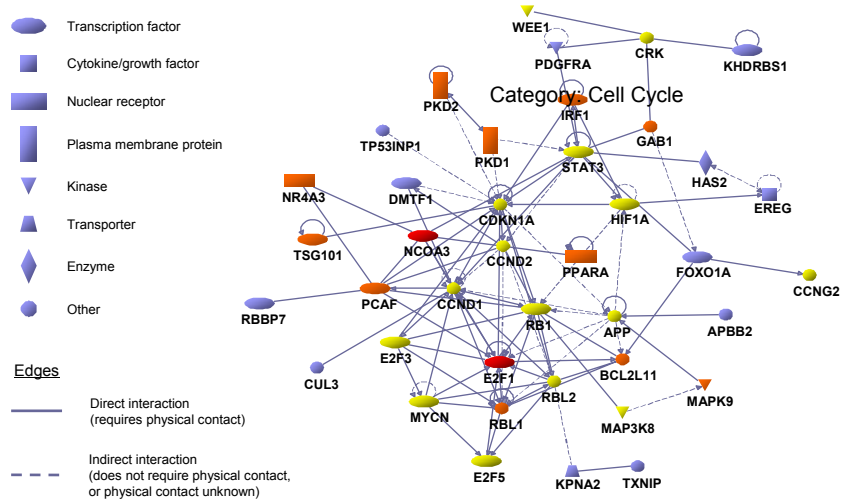


Studying miR-290 mediated control of gene networks:



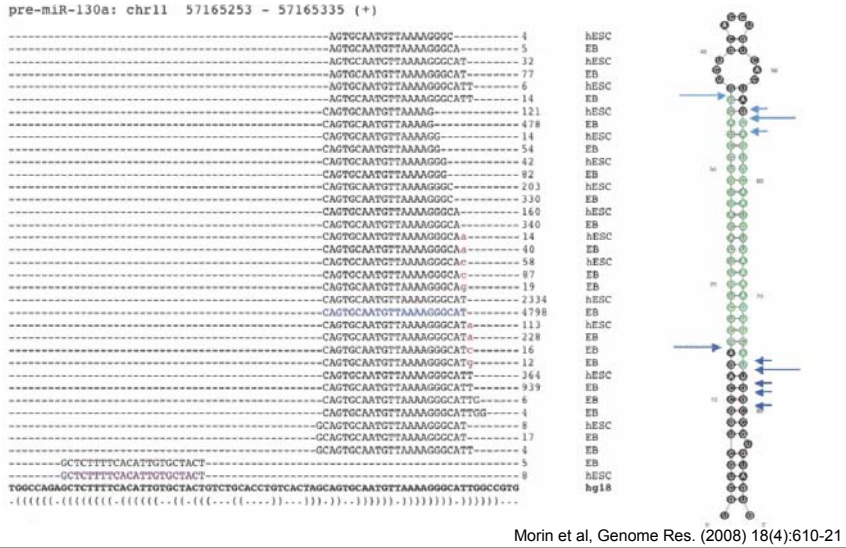


Studying miRNA mediated control of gene networks



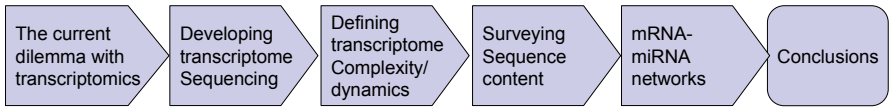
Cloonan et al, Genome Biology, 9: R127 (2008)

miRNA expression: Complexity from IsomiRs

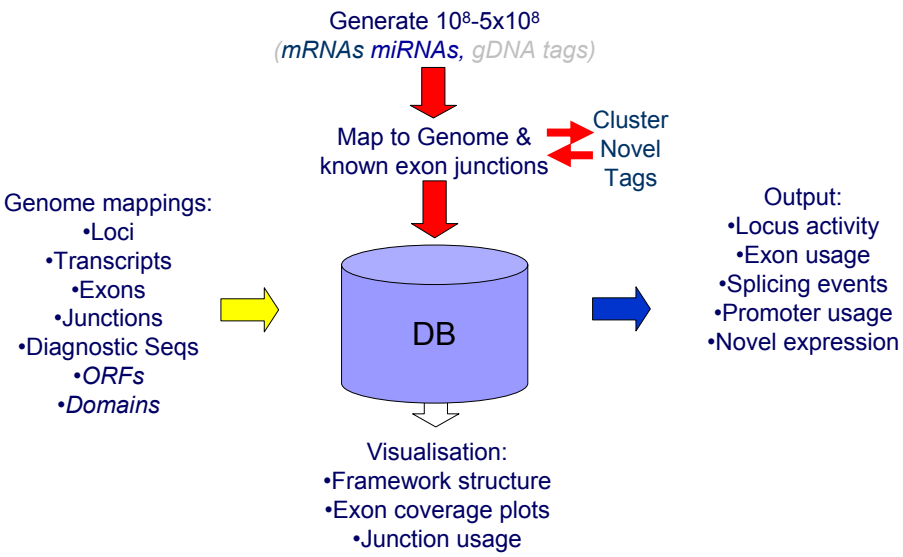


Morin et al, Genome Res. (2008) 18(4):610-21

Presentation overview



Summary:



Conclusions:

RNAseq is a powerful tool for studying RNA abundance, dynamics and complexity with unprecedented sensitivity

Transcriptome discovery is identifying roles for novel expression

Polysome associated RNAseq gives novel insights into those transcripts that are actively translated.

SQRL provides a complete view of small RNA complexity and abundance.

It is also now possible to screen the transcriptome for expressed SNPs, mutations (mis-sense, and indels), allelic specific expression and RNA editing.

Challenges:

-Sequence content (more reads, longer reads, more accurate sequencing).

-Ambiguity of mapping. “Black holes” in the transcriptome.

-Transcript length affect shotgun tag sensitivity.

-Matching to a reference is easy, de novo assembly is hard.

-Diagnostic sequences only infer transcription complexity

-Novel combinations of junction usage (ie alt splicing event between exon 1& 3 cannot be conclusively tied to alternate 3’utr usage.)

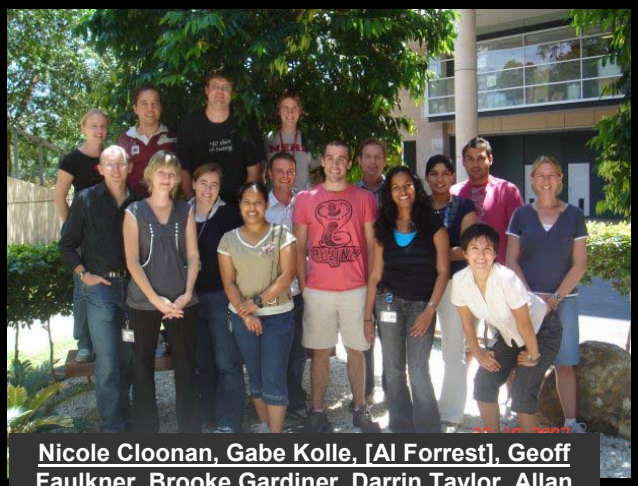
SOLiD:
Kevin McKernan
Clarence Lee
Heather Peckham
Stephen McLaughlin

SREK:
Scott Kuersten
Jian Gu
Catalin Barbacioru

IMB array facility:
Tina Maguire
Keerthana Krishna

Stem Cell Research:
Andrew Perkins
Steve Bruce
Andrew Laslett
George Zhou

Repeat analysis:
Piero Carninci
Shintaro Katayama
Valerio Orlando



Nicole Cloonan, Gabe Kolle, [Al Forrest], Geoff Faulkner, Brooke Gardiner, Darrin Taylor, Allan Robertson, Shivangi Wani, Graham Bethel.

The University Of Queensland AUSTRALIA
Australian Government Australian Research Council
Australian Government National Health and Medical Research Council
Australian Stem Cell Centre