Using Next Generation Sequencing to Analyze the Transcriptional Output of Cells

David I Smith, Ph.D.
Professor
Director of the Technology Assessment Group
Center for Individualized Medicine
Mayo Clinic

Transcriptional Output of Eucaryotic Cells

- Ribosomal transcripts
- tRNA
- Messenger RNA (mRNAs for expressed genes), usually polyadenylated
- microRNA
- Piwi and Sno RNAs
- lncRNA (long non-coding RNAs)

Non-Coding DNA

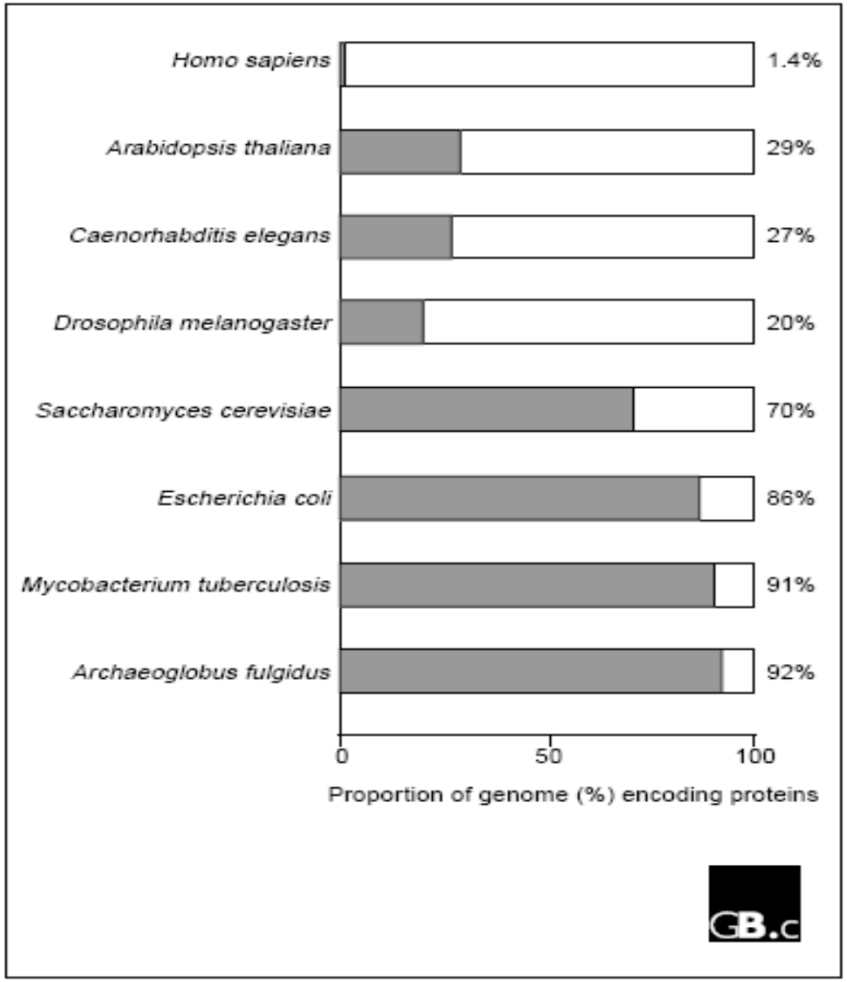
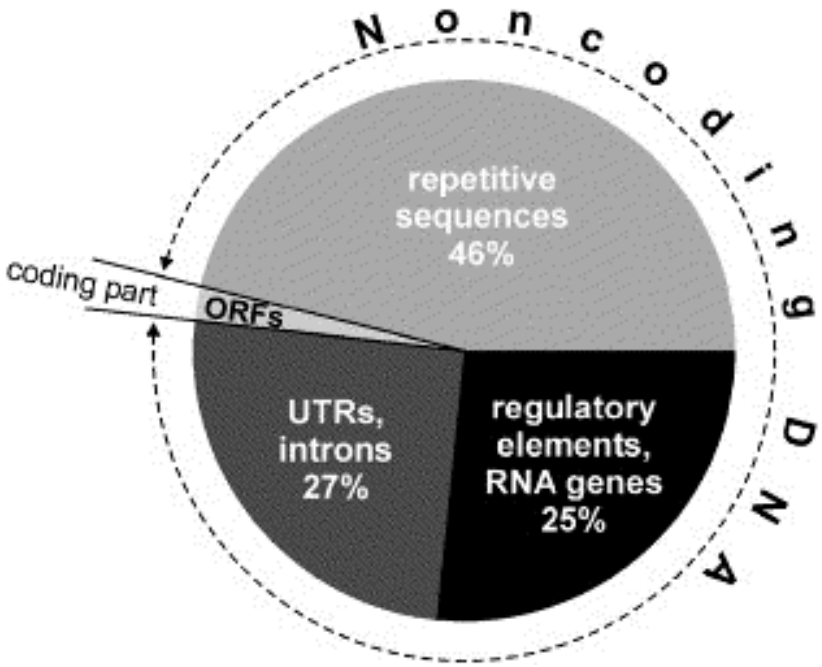


Figure 1
The percentage of protein-coding sequences (gray portions) in several eukaryotic and bacterial genomes.



2% Protein-coding DNA sequence

98% Non-protein coding DNA sequence

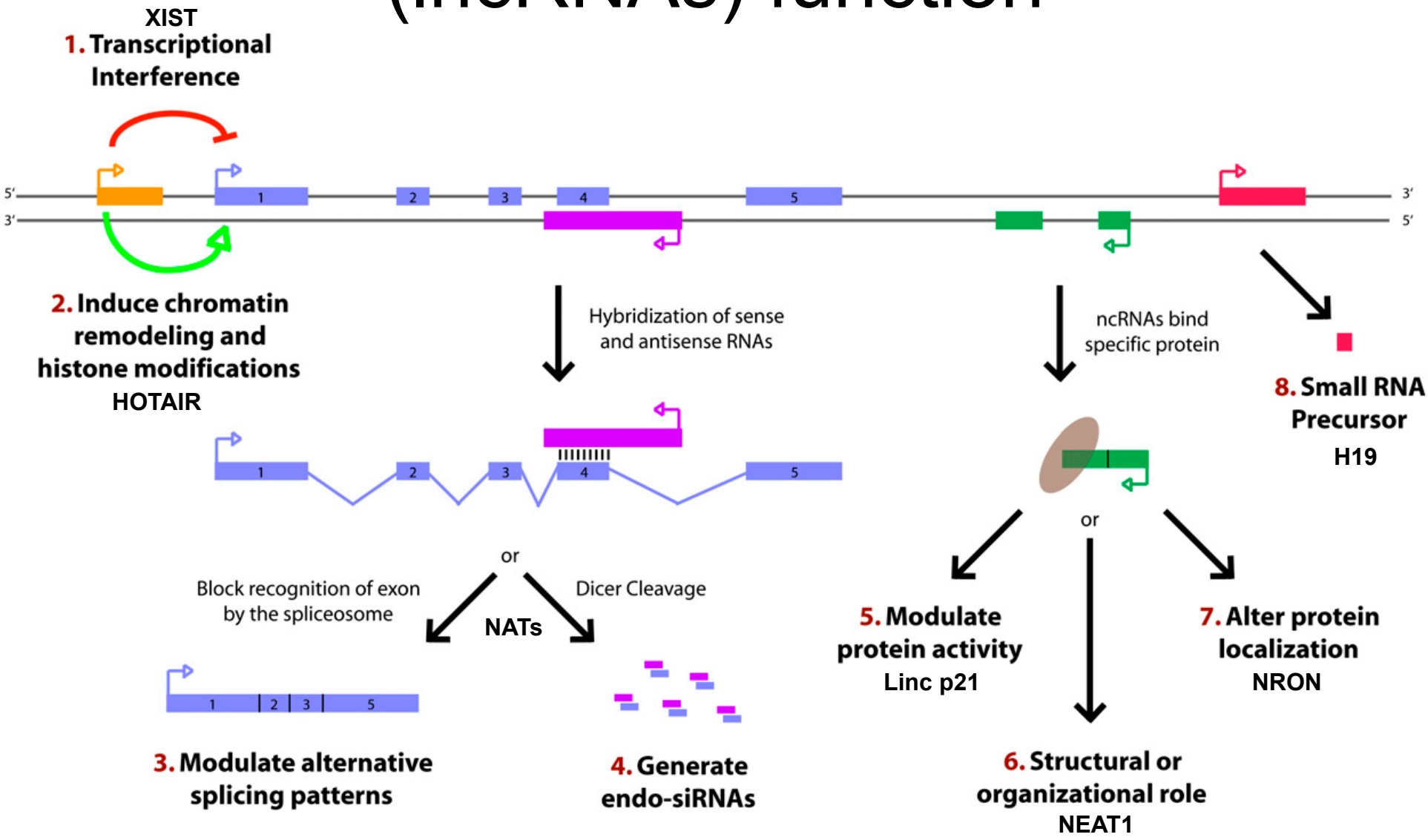
Symanski et al. *Beyond the Proteome:non-coding regulatory RNAs* 2002

Non-Coding Transcripts (NCTs)

- **Transcripts that do not translate into functional protein**
- **Transcripts do not contain any open reading frames**
- **Transcripts found within intronic regions, intergenic regions, and transcribed from coding genes (within exons in antisense direction or overlapping exons and introns)**
- **Some found to be conserved and contain distinct functions**
- **Transcribed in sense and anti-sense**
- **NCTs include all Housekeeping RNAs and Regulatory RNAs**

Housekeeping	**Regulatory**	
	Long	***Small***
tRNAs	**XIST RNA**	**piwiRNAs**
snoRNAs	**H19**	**miRNAs**
rRNAs	**NEAT1**	
	NEAT2	
	HOTAIR	

Paradigms for how long ncRNAs (lncRNAs) function



Wilusz et al. 2009

New Models for Transcription

- Simple operon-based model totally invalid
- Average human gene is not just a single transcript
- Multiple isoforms
- Sense and anti-sense transcripts
- Regulated by transcription, miRNAs, chromatin remodeling and ????
- How to truly study?

Early attempts to characterize transcription

- All of the focus was on the mRNAs
- Assumption was that the protein coding genes were the entire story, hence if you could measure the amount of transcription of each gene you could infer how much of the encoded protein was produced in those cells
- Huge problem with message abundance (some messages are thousands of time more abundantly expressed than others). Most highly expressed transcripts swamp out your sampling when looking for less abundantly expressed transcripts

History of Transcriptional Profiling

- Make a poly A-primed cDNA library and Sanger sequence the clones- Expensive and only good for the most abundant transcripts
- SAGE- serial analysis of gene expression- sequence just the tags (less sequencing and can ID many more genes). Victor Velculescu et al. 1995

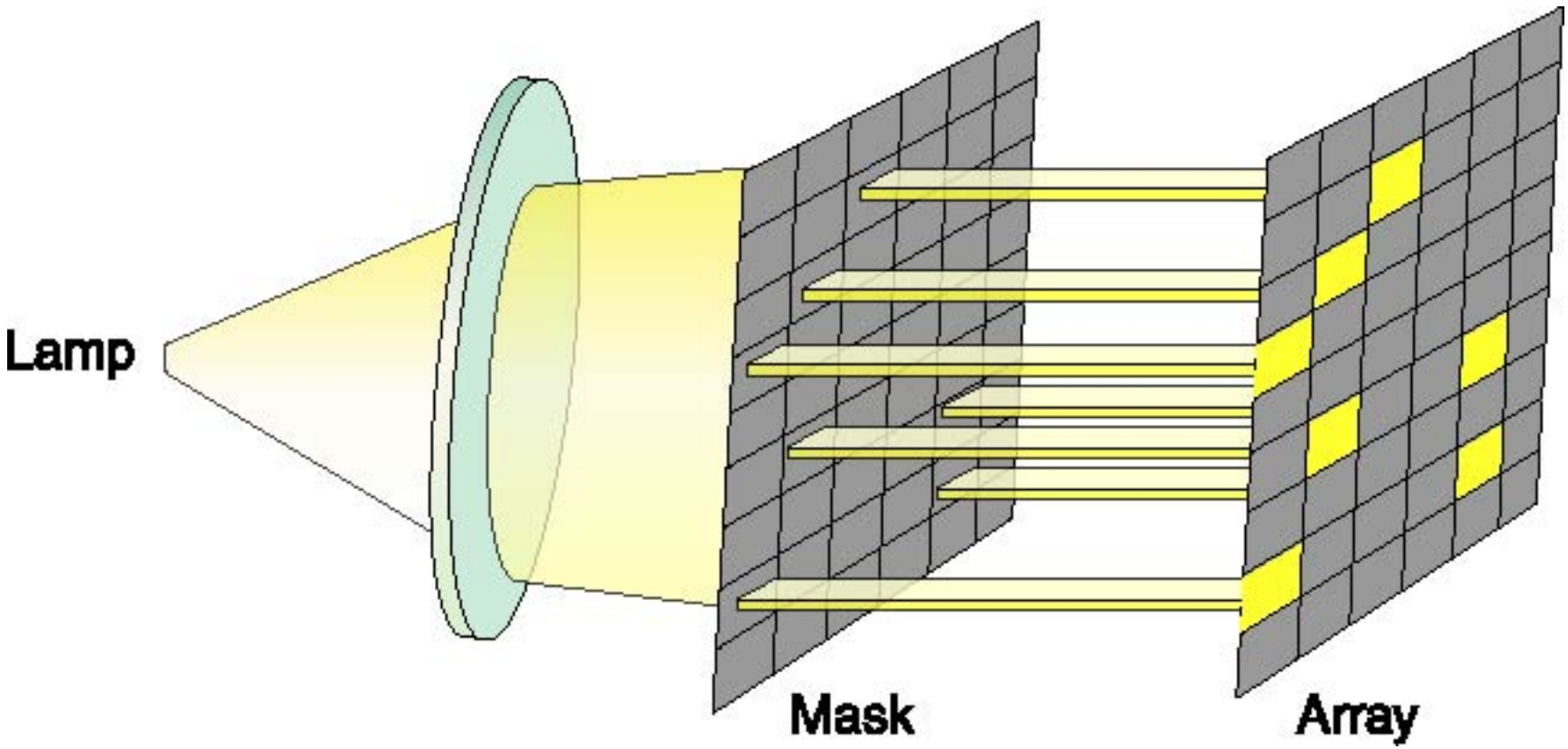

AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
Isolate SAGE tags
Link tags together
Sequence linked tags
Quantitate tags and determine patterns of gene expression
Expression level
Gene product
Normal
Expression level
Gene product
Disease

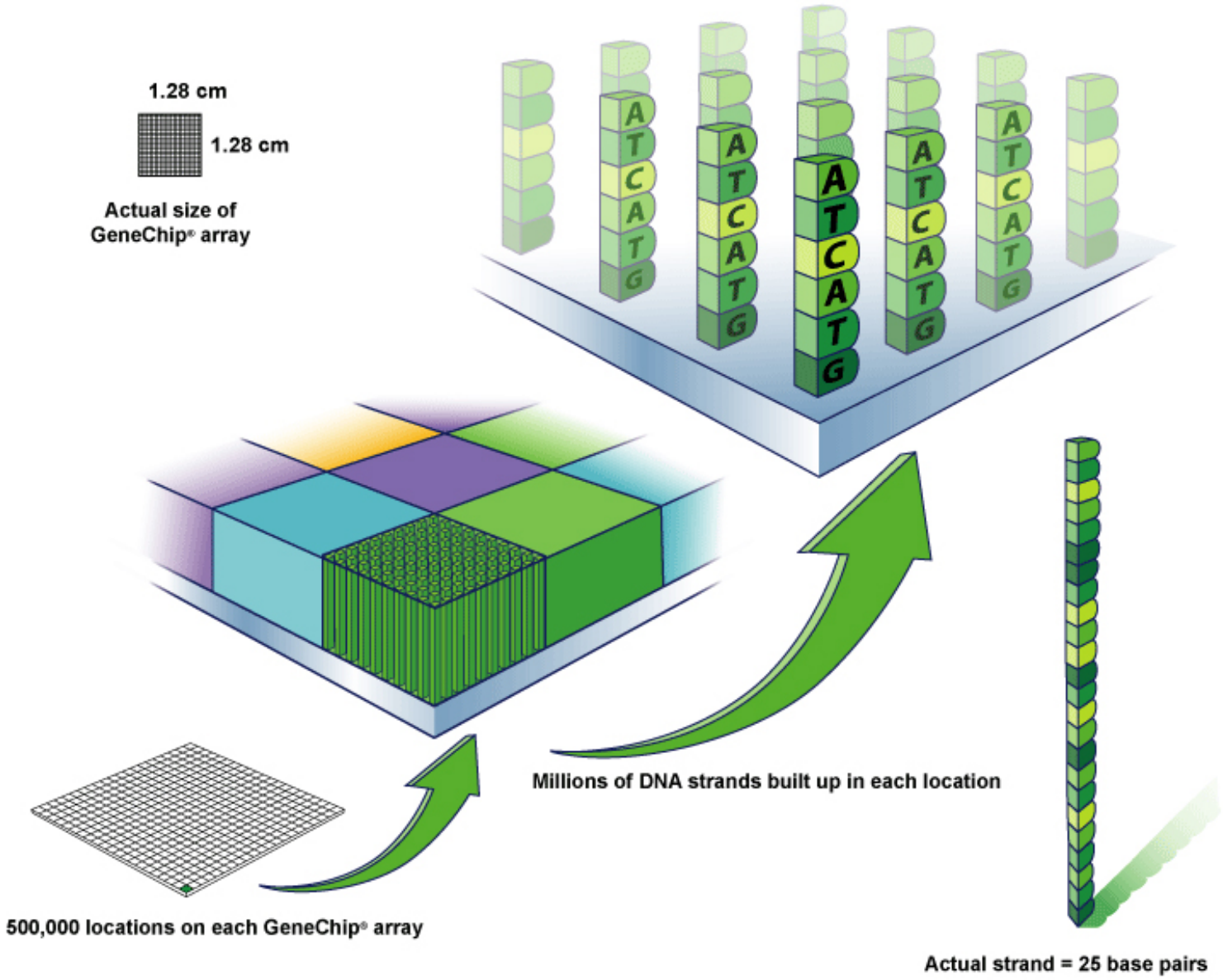
Microarrays

- Immobilize probes onto a microarray
- Originally done with entire cDNA inserts
- Oligonucleotides as probes- for example Affymetrix
- Must know about a specific transcript to make probes for it
- All early arrays were totally focused on protein coding genes

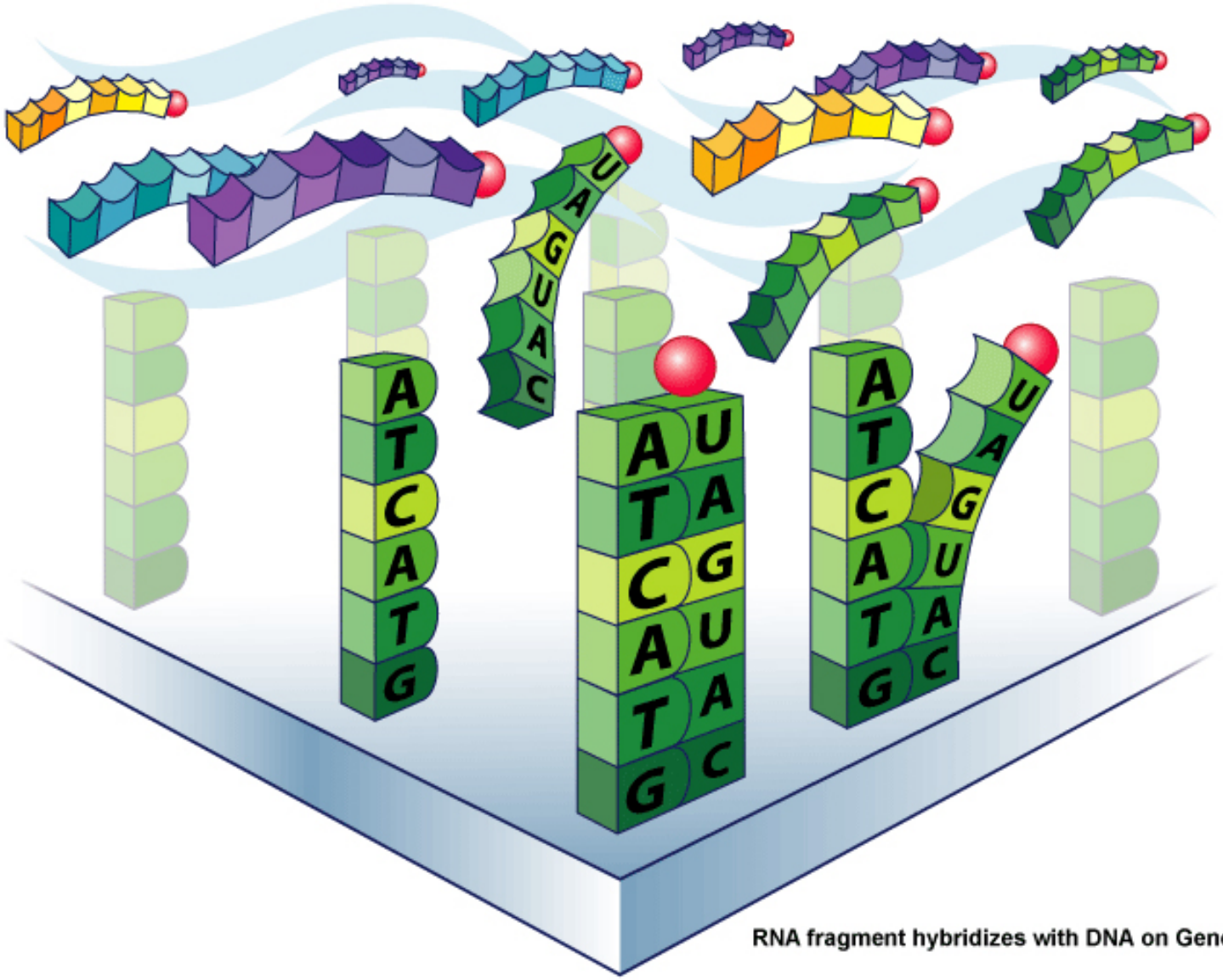
Computers and the human genome

- Progress in the sequencing of genomes came from advances in computing technology (especially high density computer chips)
- As chips were designed with more “features” computers became faster
- Eventually we had computers that were fast enough to deal with and handle all 3 billion base pairs of the human genome sequence
- These technologies spawned all genome sequencing projects
- The semiconductor manufacturing platform can be tuned to biological problems!

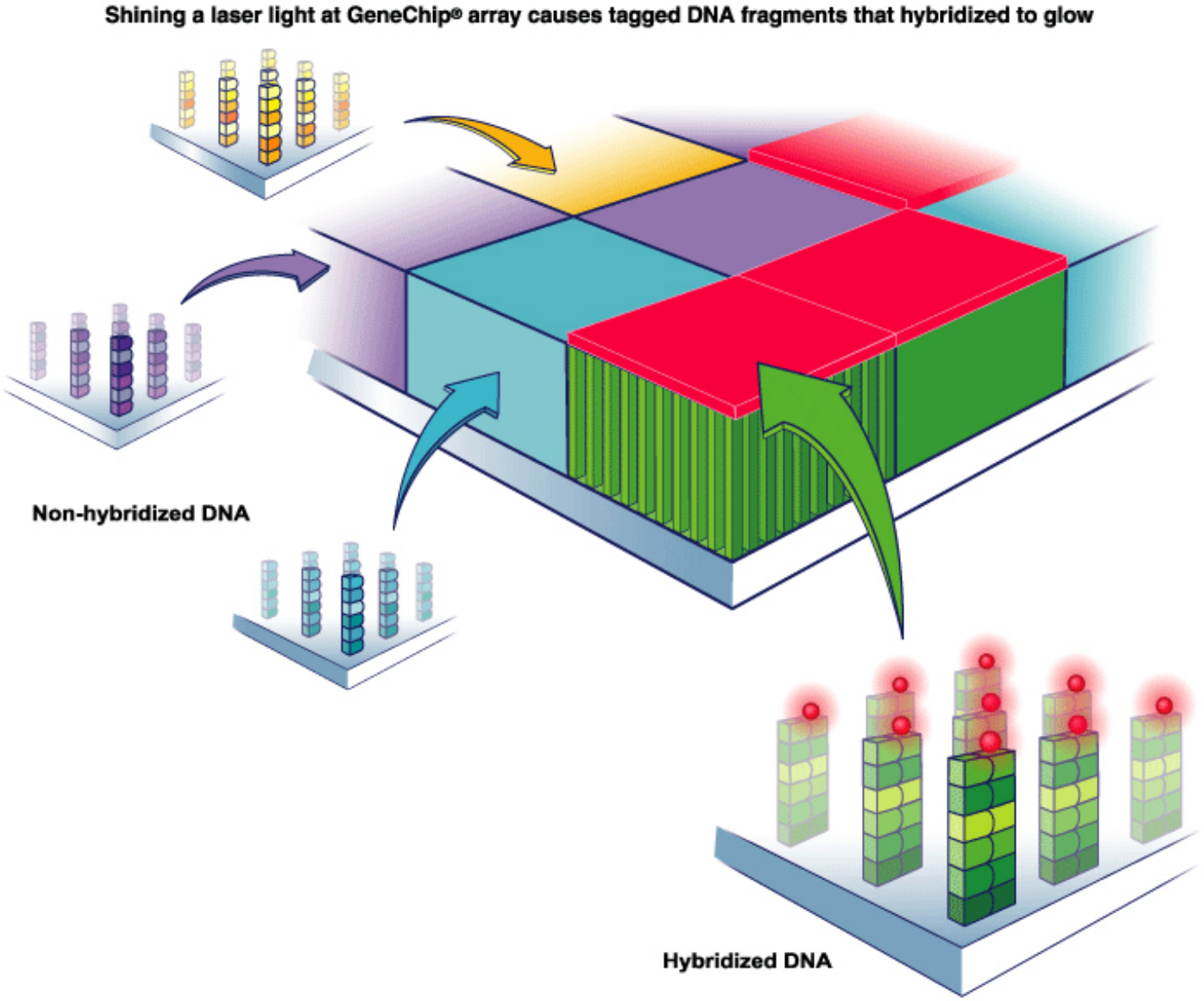

Lamp
Mask
Array


1.28 cm
1.28 cm
Actual size of GeneChip® array
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

RNA fragments with fluorescent tags from sample to be tested



RNA fragment hybridizes with DNA on GeneChip® array


Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

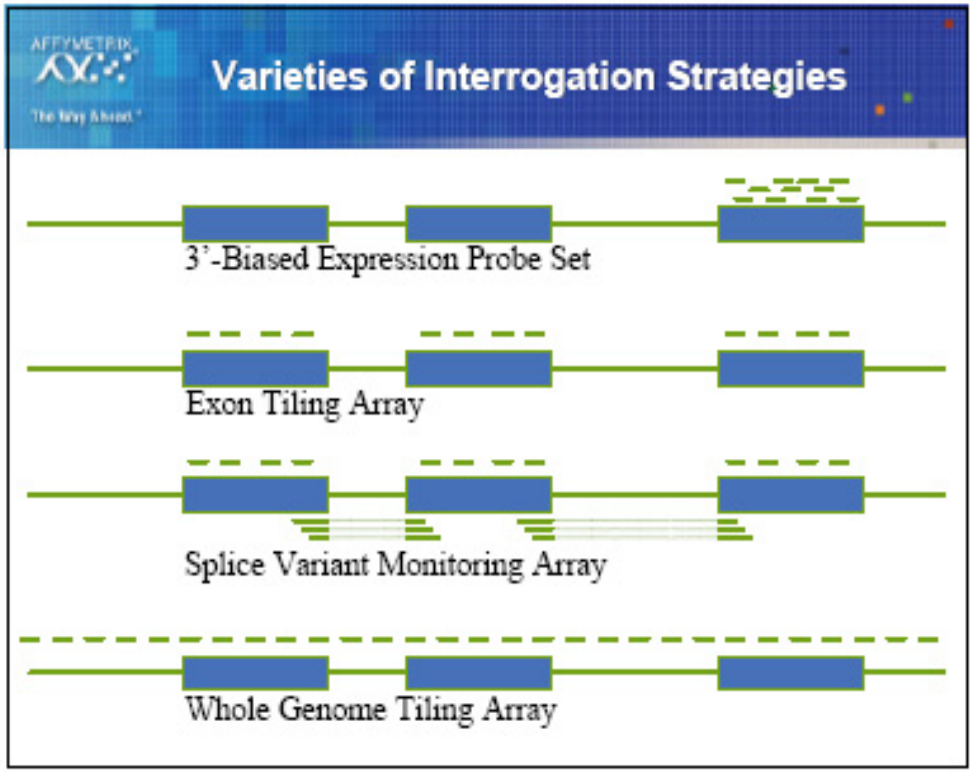
History of Affy arrays

- Several hundred genes
- 5,000 gene array
- U133 Plus 2 array- Multiple probes for each expressed coding transcript
- Probe design- 13 PM (perfect match) probes (25-mers) and 13 MM (mismatch at the 13^{th} base) probes. All derived from the 3' UTR of each transcript


AFFYMETRIX
GeneChip®
Human Genome
U133 Plus 2.0
P/N: 520019
Lot #: SAMPLES
Exp. Date: 00/00/00
For Research Use Only

Number of features on an Affy array

- U133 Plus 2 arrays based upon 500,000 features/array
- Next generation had 6.5 million features/array
- With this many features can do much more than just probe the 5’ end of genes


AFFYMETRIX
"The Way Forward"
Varieties of Interrogation Strategies
3'-Biased Expression Probe Set
Exon Tiling Array
Splice Variant Monitoring Array
Whole Genome Tiling Array

Design of a genome tiling array

Typical design strategy is to select PM,MM probe pairs across non-repetitive regions at a target center-to-center separation which is referred to as the resolution of the array.

Factors considered in probe selection:

- Probe separation
- Probe quality (avoid probes with predicted non-linear intensity vs concentration relationship)
- Probe uniqueness (avoid probes with similarity to multiple genomic locations.

Typically will end up with more 'bad' probes than a conventional 3'-biased array design

Tom Gingeras and Tiling Arrays

- Tom Gingeras worked at Affymetrix. Got the earliest access to genome-wide tiling arrays
- Started with 5-bp tiling arrays (5 bp from the center of one oligonucleotide to the center of the next adjacent oligo)
- Using these tiling arrays they found that non-polyadenylated transcripts were the majority of the transcriptional output of the genome
- Could these non-coding RNAs be important regulators of gene expression?
- All this work inspired the ENCODE (encyclopedia of DNA elements) project

Other Microarray Platforms

- Agilent- HP color printers
- Nimblegen- DLP-based synthesis
- Illumina- long oligonucleotides linked to beads
- All can synthesize much longer probes. Less probes/gene. Can wash at much higher stringencies
- Much greater flexibility to design specific custom arrays

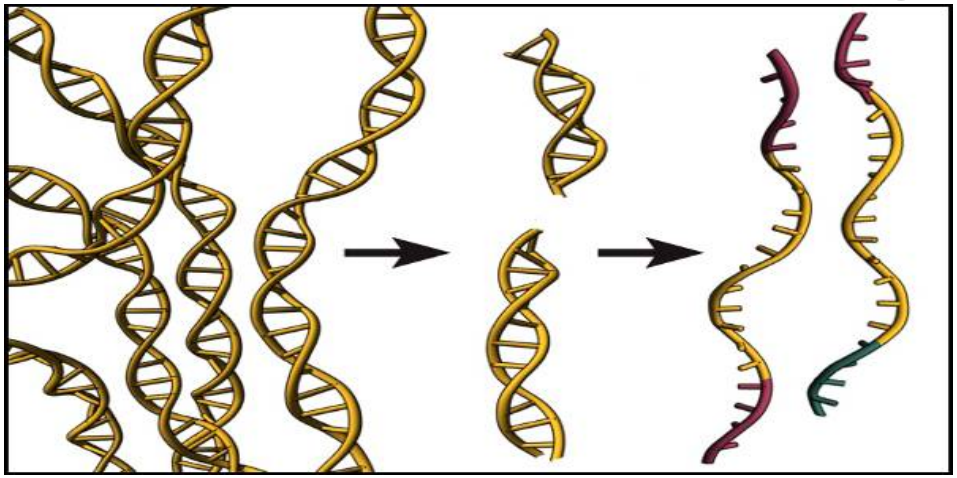
Problems with Microarrays

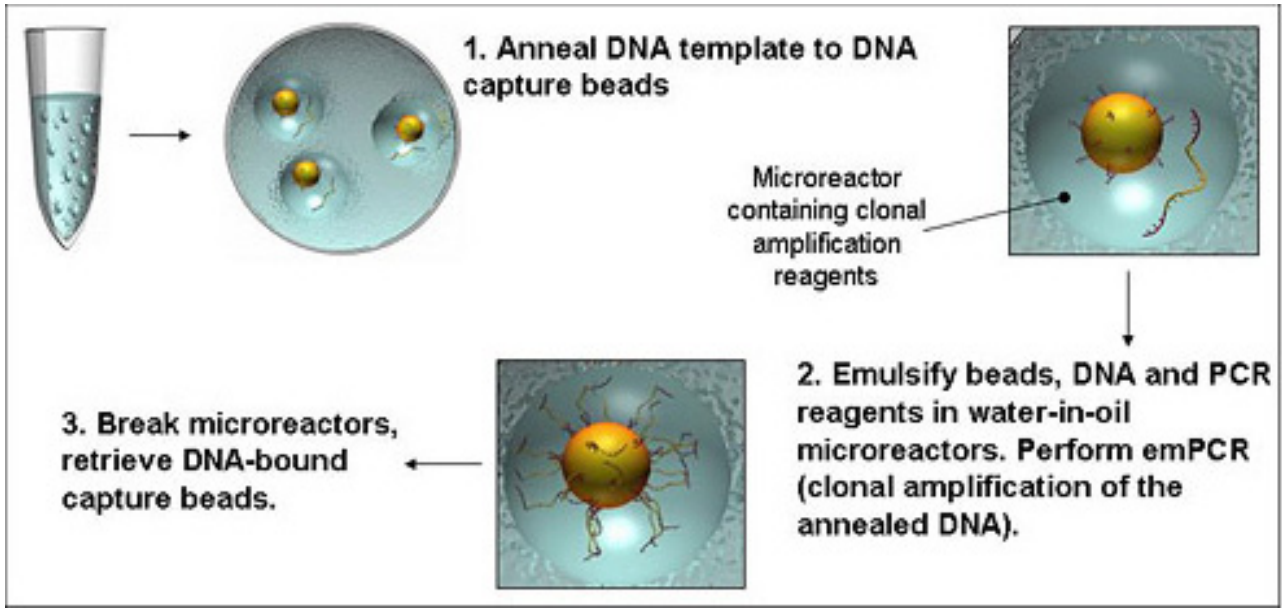
- Lack of sensitivity. Only can measure the expression of the top 50% of expressed genes
- Not really quantitative. More qualitative
- Cross hybridization a real problem
- Are the Gingeras results correct (i.e. that the entire genome is transcriptionally active)?
- What does the concentration of an mRNA species tell you about the proteins encoded by those transcripts?

Next Generation Sequencing

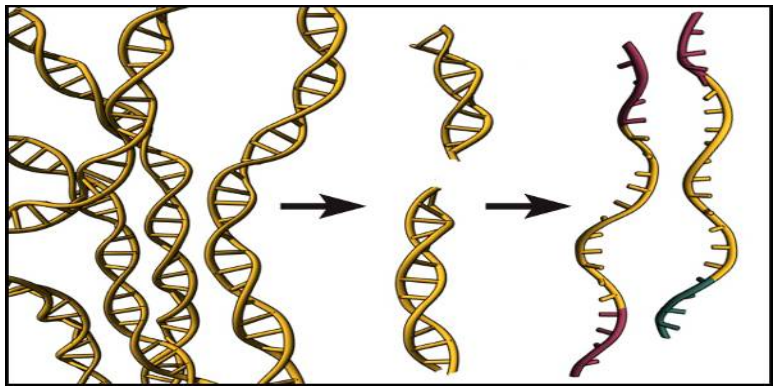
- Based upon massively parallel sequencing
- First commercially available from 454- The Genome Sequencer (GS series)

Clonal Amplification: Emulsion PCR

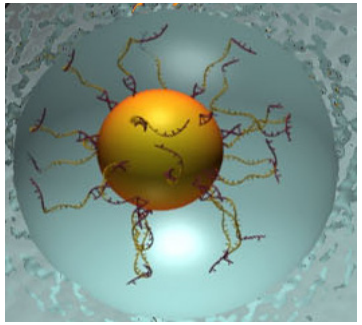




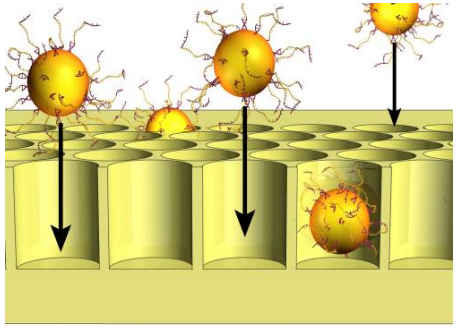
Process Overview - 454



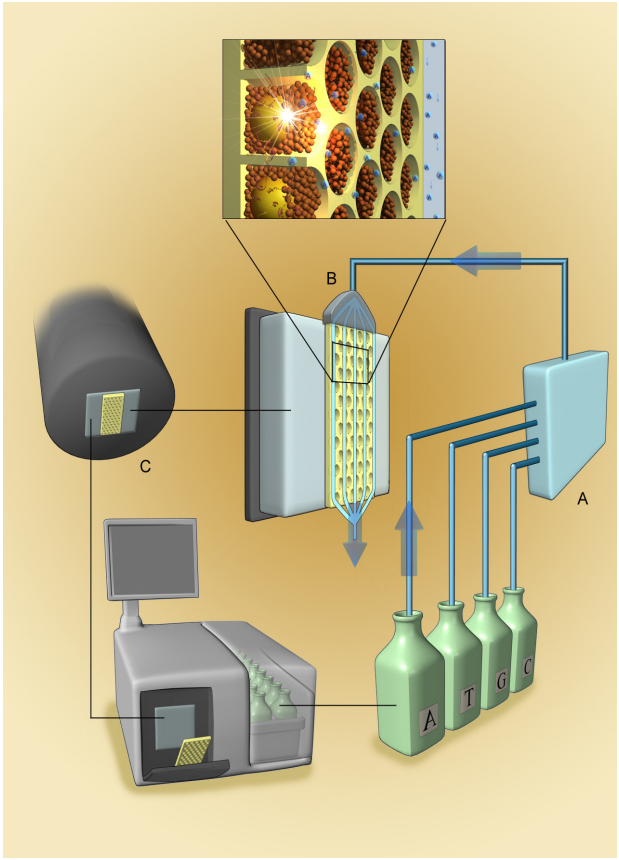
1) Prepare Adapter Ligated ssDNA Library



2) Clonal Amplification on 28 μ beads

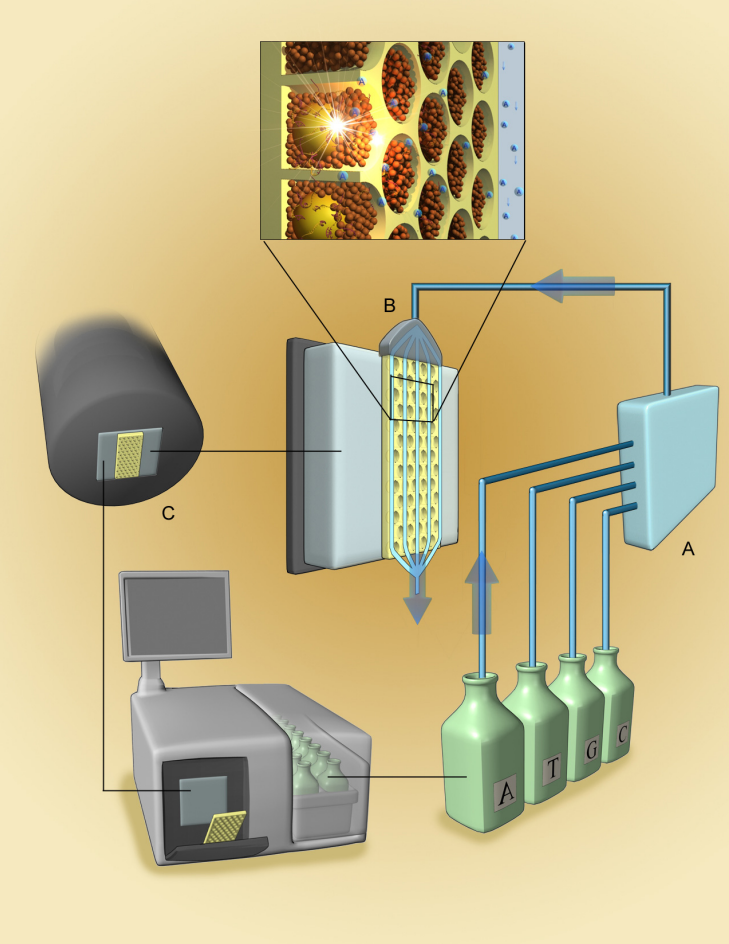


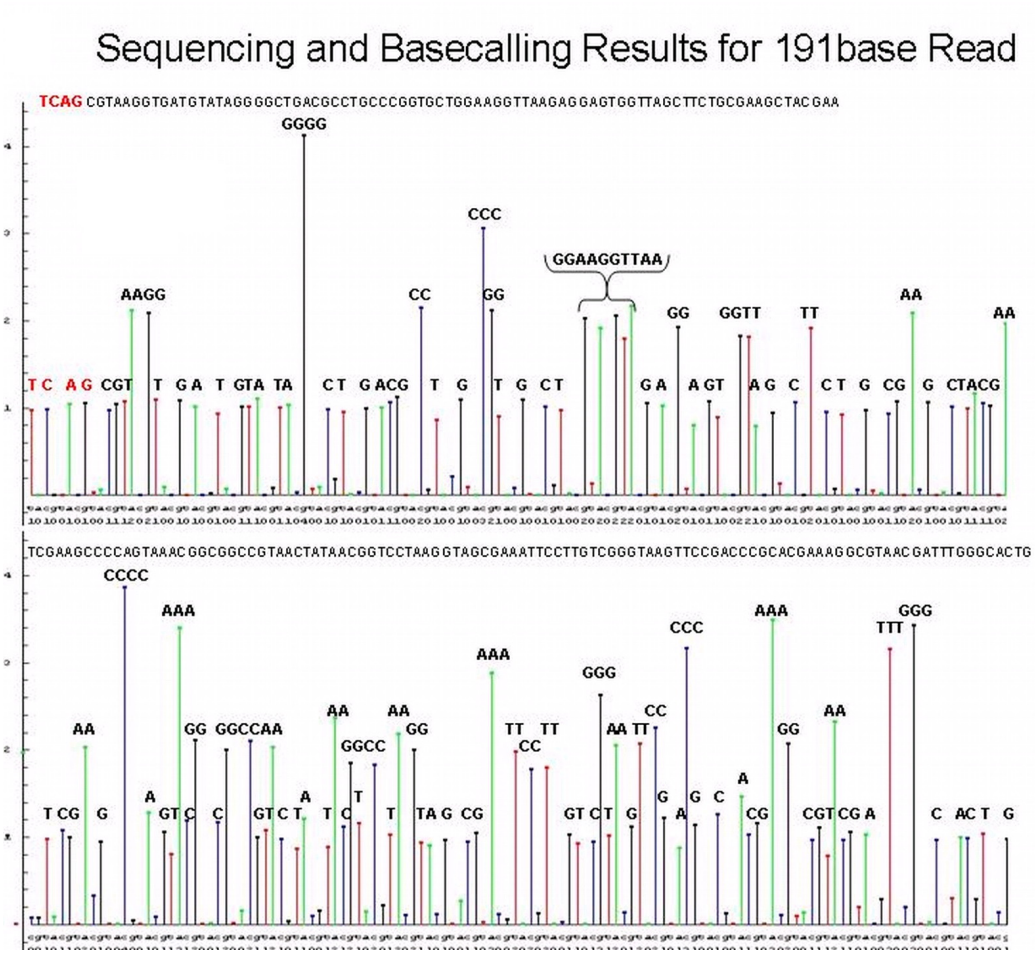
3) Load beads and enzymes in PicoTiter Plate™



4) Perform Sequencing by synthesis on the 454 Instrument

454 Technology - Sequencing Instrument





Strengths and Weaknesses of the 454

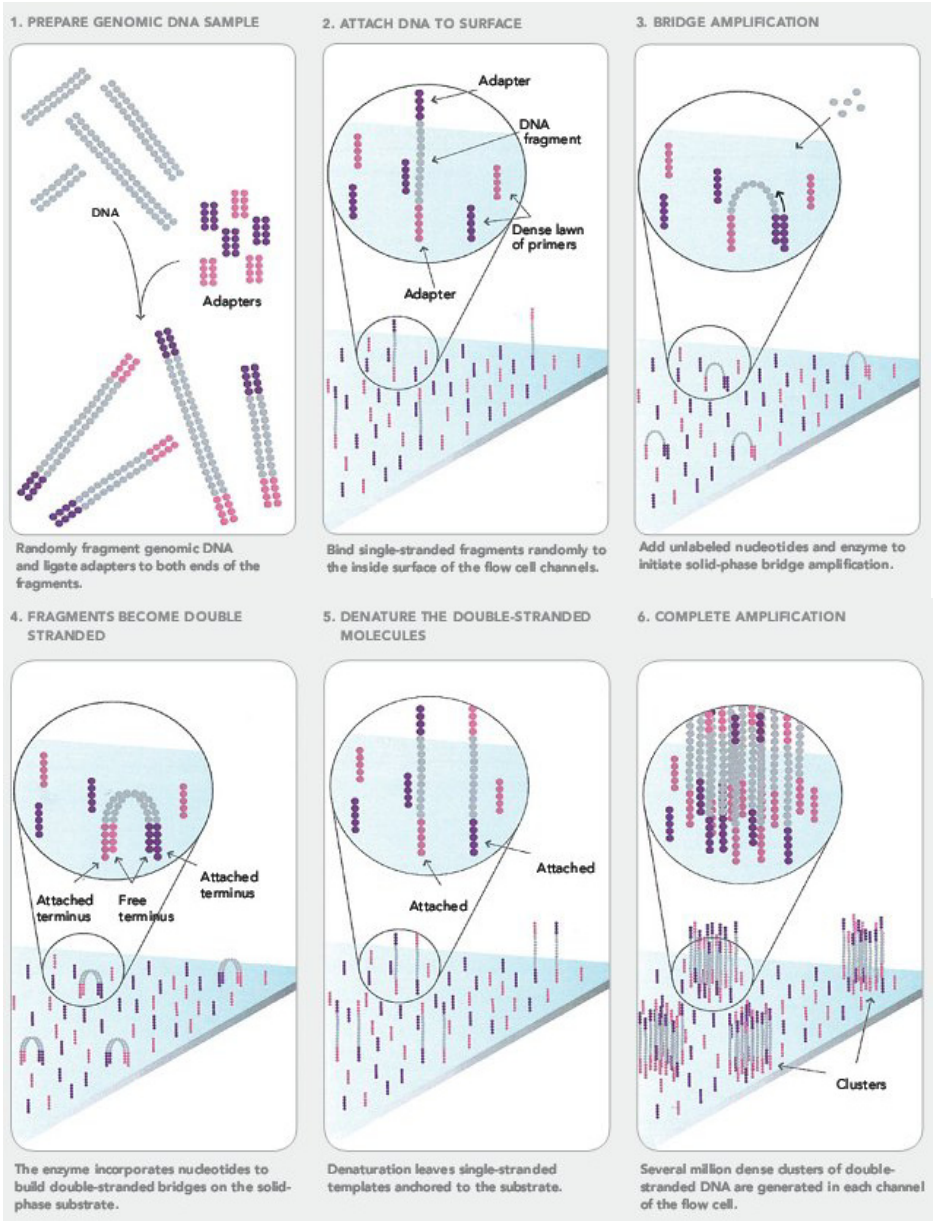
- Long (400 bp+) reads
- Went from 20 Mbs to 500 Mbs output in two years!

- Emulsion PCR
- Homopolymers
- Limited headroom for further increases in sequence output

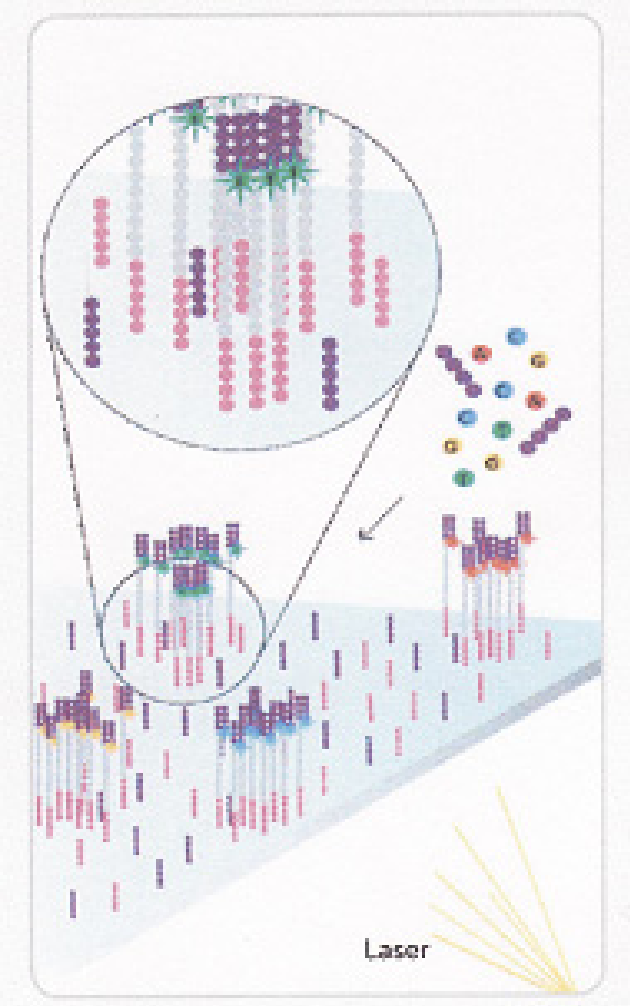
Illumina Genome Analyzer

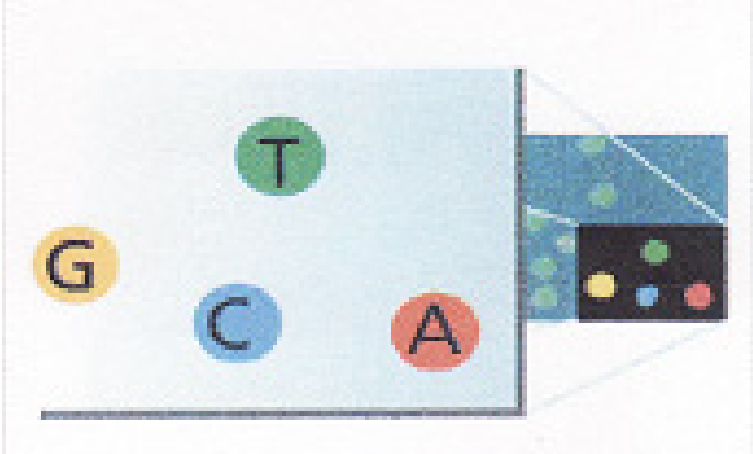
From: Blow, N. et al. *Nature: 2007:* 449, 627-630.

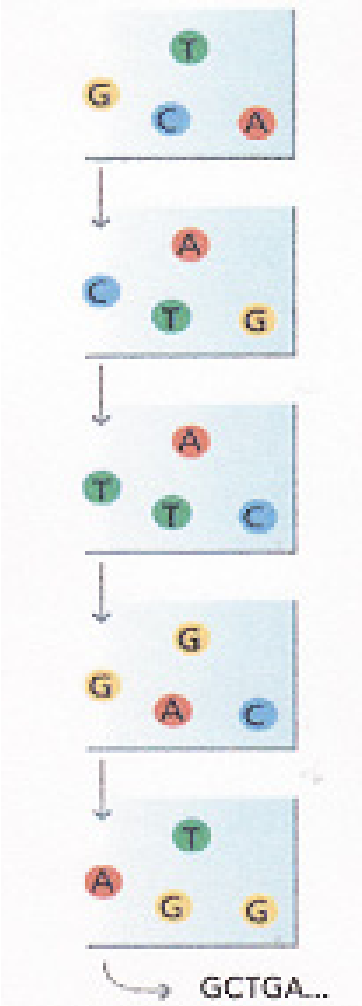
Bridge Amplification



Illumina GA: polymerase-based sequencing with reversible terminators

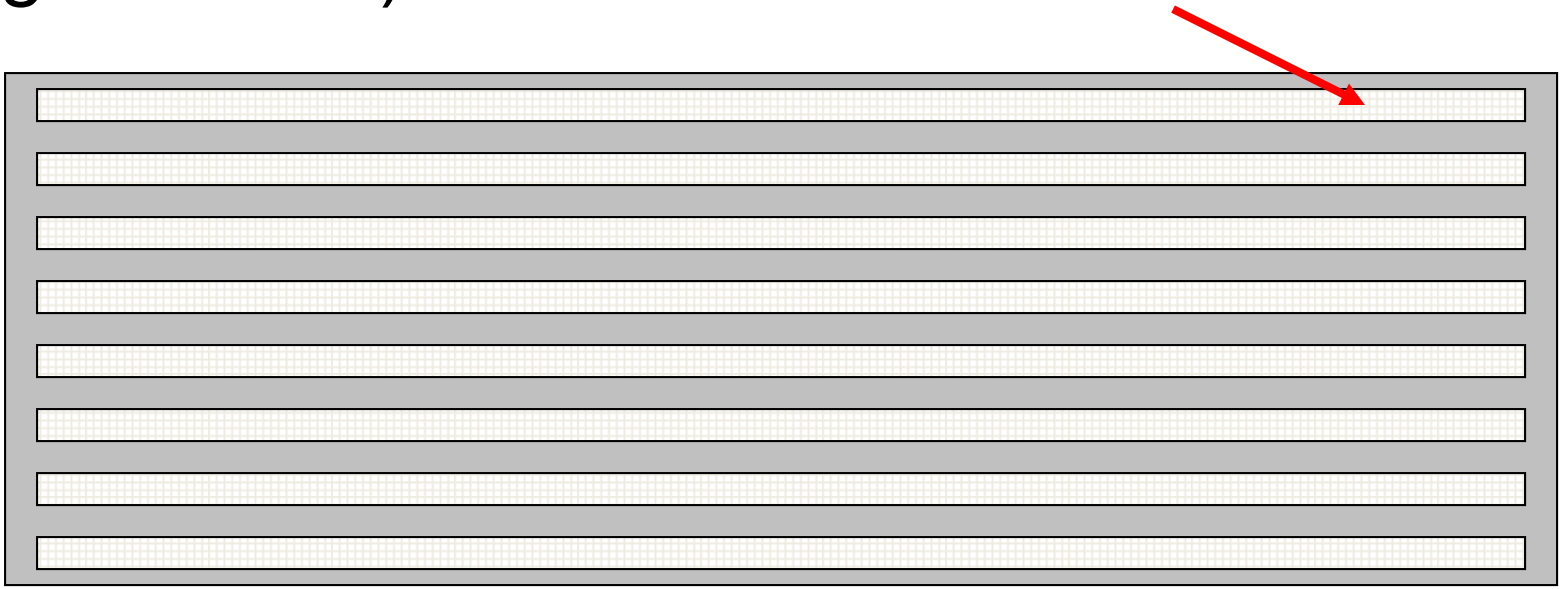






Raw Data is Images

- 8 channels per flow cell
- 300 tiles per channel: First generation
- 20,000 clusters/reads per tile (first generation)



First Generation Next Generation RNAseq

- GA II capable of 6 million reads/lane
- Barely good up to 30 bps
- Do a modified SAGE
- Develop TAGs for transcripts and sequence 6 million/lane
- SAGE on steroids!!
- Not fully exploiting the power of Next Gen

HiSeq 2000

Evolution of Instrument Performance

400

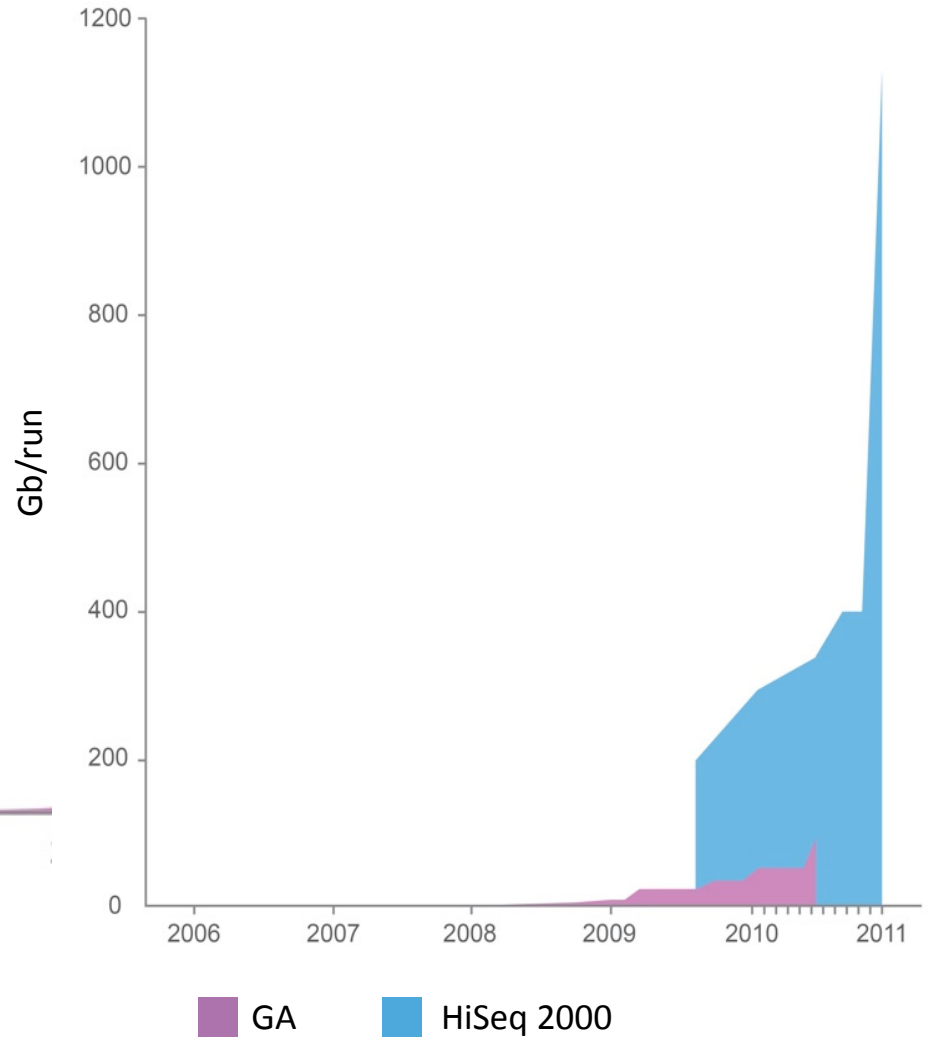
Sequencing Run Parameters

Run format: 2x150 bp

Output full run	**1.13Tb**
Output per day	**81Gb**

Data passing filter
88.9%

2008



Massively Parallel Sequencing as the Solution!

- Even with the first generation Genome Analyzer could look at 6 million sequences/lane. HiSeq 2000 is now up to 300 million reads/lane
- Orders of magnitude better than SAGE
- Much more sensitive than microarrays
- How many reads to characterize a transcriptome?

Source of the RNA

- Fresh frozen versus FFPE
- Advantage of FFPE- Many more samples. Clinical Follow-up
- Disadvantage- RNA is quite beat up
- Three solutions- (1) DNS protocol, (2) Genomic Health propriety protocol, or (3) Use fresh or fresh-frozen

What to Sequence?

- If you sequence a library made from total RNA more than 95% of the transcripts will be ribosomal
- Two solutions: (1) poly A+ selection; (2) selective removal of ribosomal sequences (RiboMinus or RiboZero)

Poly A+ Selection

- First strand synthesis done on oligo-dT attached to magnetic beads
- Strengths- Very effective at removing ribosomal sequences. Less overall sequencing required.
- Disadvantages- RNA quality an issue. Degraded RNA makes it difficult to sequence the 5’ ends of transcripts. Only selects for polyadenylated transcripts (many non-coding transcripts are not polyadenylated). None of the miRNAs are polyadenylated

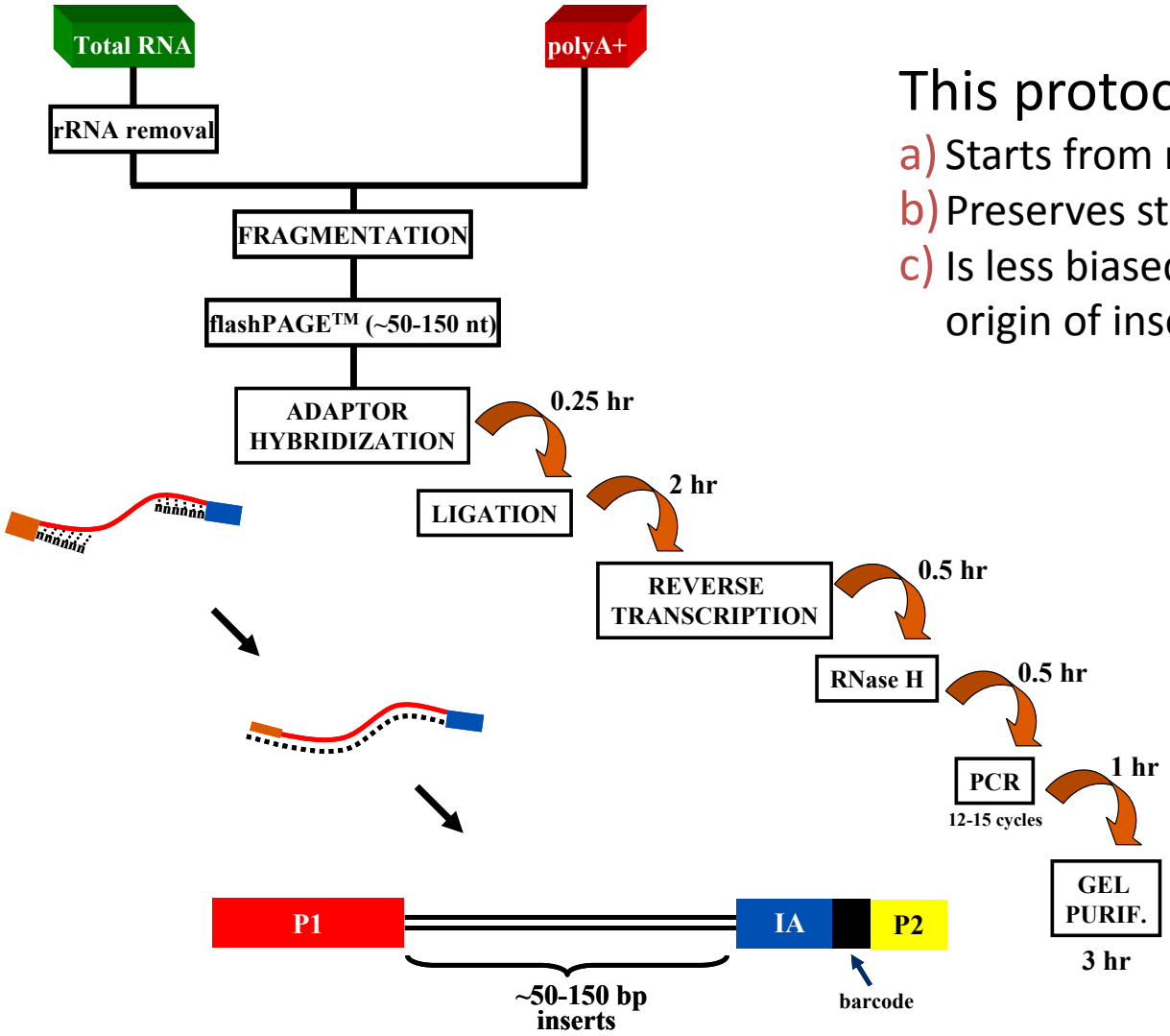
RiboMinus or RiboZero Ribosomal Removal

- Advantage- Can sequence all (not just polyadenylated) non-ribosomal transcripts
- Disadvantage- need to sequence more than poly A+ selection for the same coverage
- RNA degradation decreases the efficiency of either RiboZero of RiboMinus to remove ribosomal sequences

Directionality?

- Standard library construction does not preserve the strandedness of each sequenced transcript
- Is this important? Depends.
- Protocols are available to generate libraries that do preserve strandedness

The WT kit from Ambion



This protocol:

a) Starts from ribo-cleared total or polyA.
b) Preserves stranded-ness.
c) Is less biased with respect to positional origin of inserts within transcripts.

Figure from Scott Kuersten

Setting up an RNAseq Experiment

- What is the source material?
- Tissue culture is probably the best source
- Clinical specimens pose a number of problems
- What are you trying to determine? Helps to define how many samples to run and how much transcriptome sequence to derive from each sample

RNAseq at its' best

- Take your favorite tissue culture cells.
- Stress them. Knock down your favorite gene. Add some chemical.
- Measure transcription before and after
- Sequence all important transcripts
- No need for prior knowledge of the transcriptional output of your cells of interest

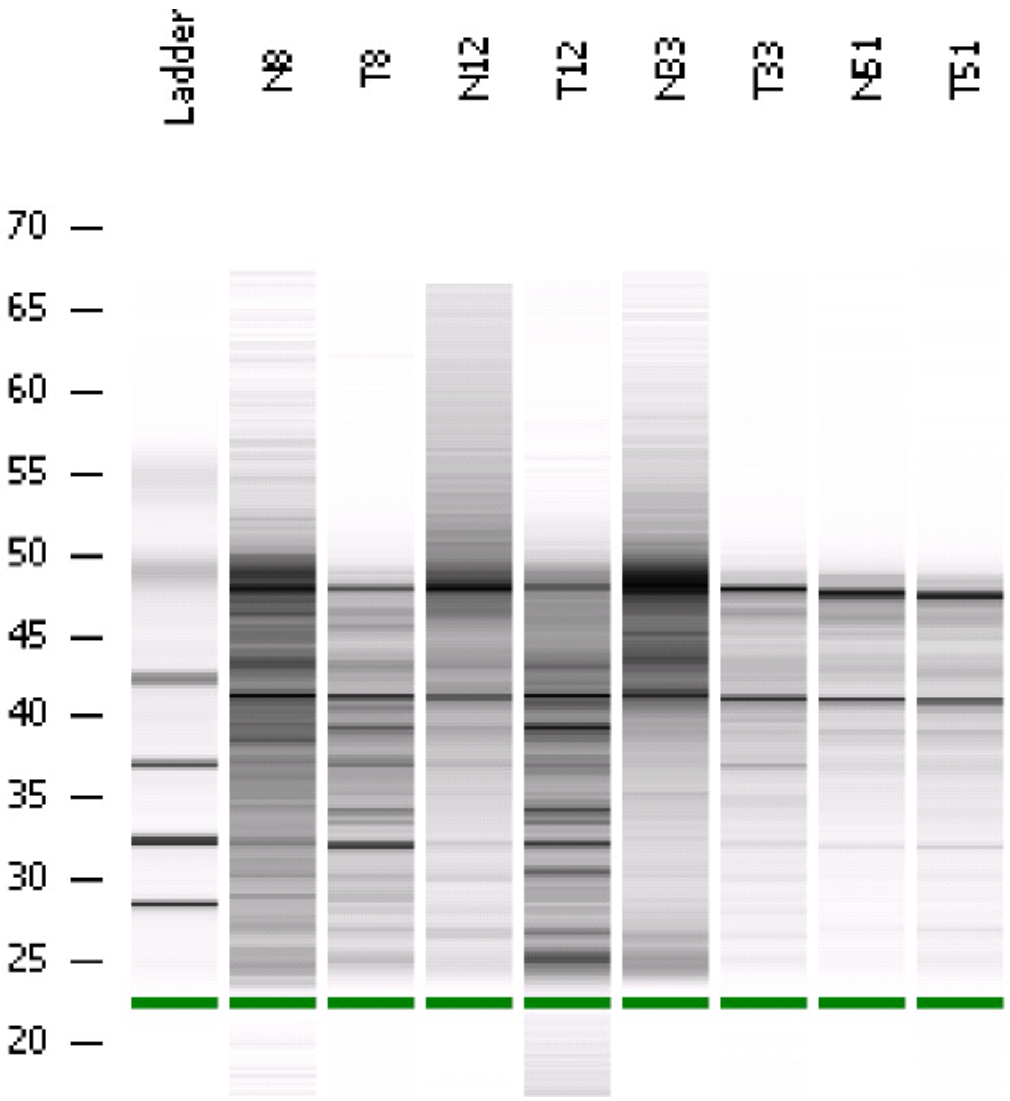
RNAseq and Cancer

- Compare gene expression in tumor as compared to matched normal tissue- essential for RNAseq!
- How many samples to run? How many do you have? How much money do you have?
- What are you going to compare? Normal to tumor? Good outcome to poor outcome? Different risk factors?
- Make sure tumor is >80% tumor! Otherwise think about using Laser Capture Microdissection

RNAseq and Cancer

- How many RNAseq reads are enough?
- Bare minimum 75-100 million, but you could also do 300 million plus (depends upon what you want to see). How much heterogeneity in your cancer?
- How many tumor-normal pairs to run?
- Why do this experiment at all if it will be soon available from CGAP?

RNA degradation



What to do with all the data?

- Current generation HiSeq 2000 can generate 300 million reads/lane
- How to handle this imaging data and convert it into sequence?
- How to align sequence to the transcriptome to figure out what you have?
- How to analyze the resulting transcriptome output and make sense out of it?

Commercial versus In-house Solutions

- Many different commercial vendors selling packages for dealing with Next Gen data- Geospiza, NextGene, and many others
- Possible in-house pipelines for data analysis
- What to do? Which is best?
- What was your original plan for how you were going to analyze this data?

So who is doing Next Gen sequencing and data analysis right?

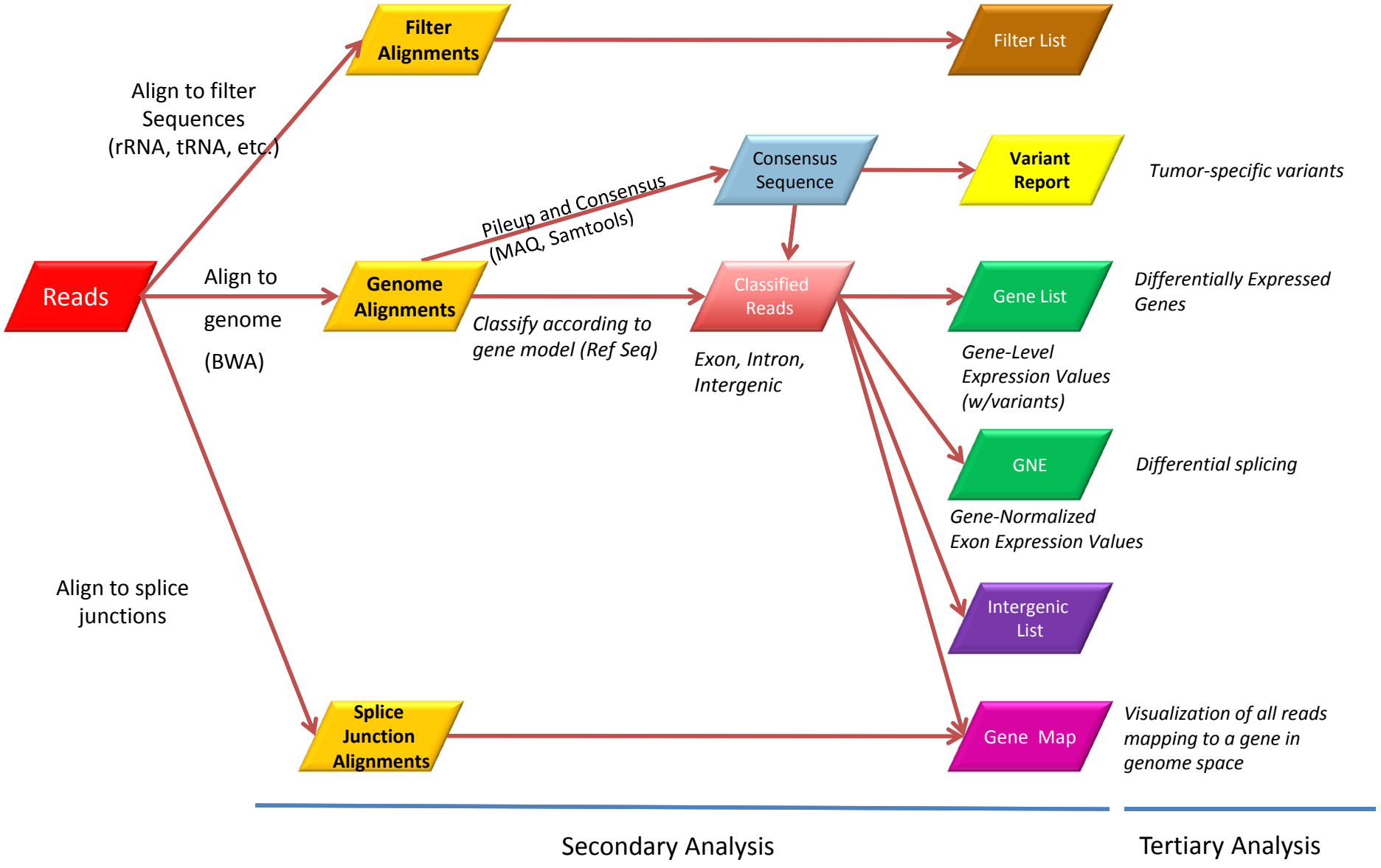
- Broad
- Wash U
- BGI
- Baylor

- The key is investing significantly in data analysis!

So what about me?

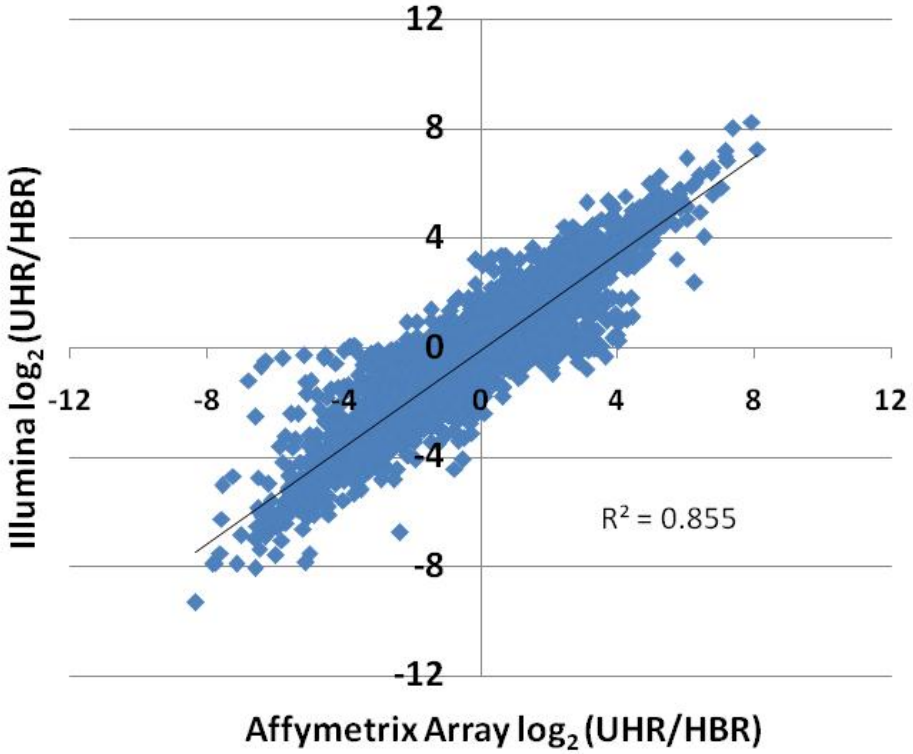
- Mayo Clinic Bioinformatics Core has been developing pipelines for different Next Gen datasets
- Active collaboration with Todd Smith/Eric Olson at Geospiza. Phase II SBIR grant to Geospiza
- Jian Ma- UIUC as part of the Mayo/UIUC Partnership

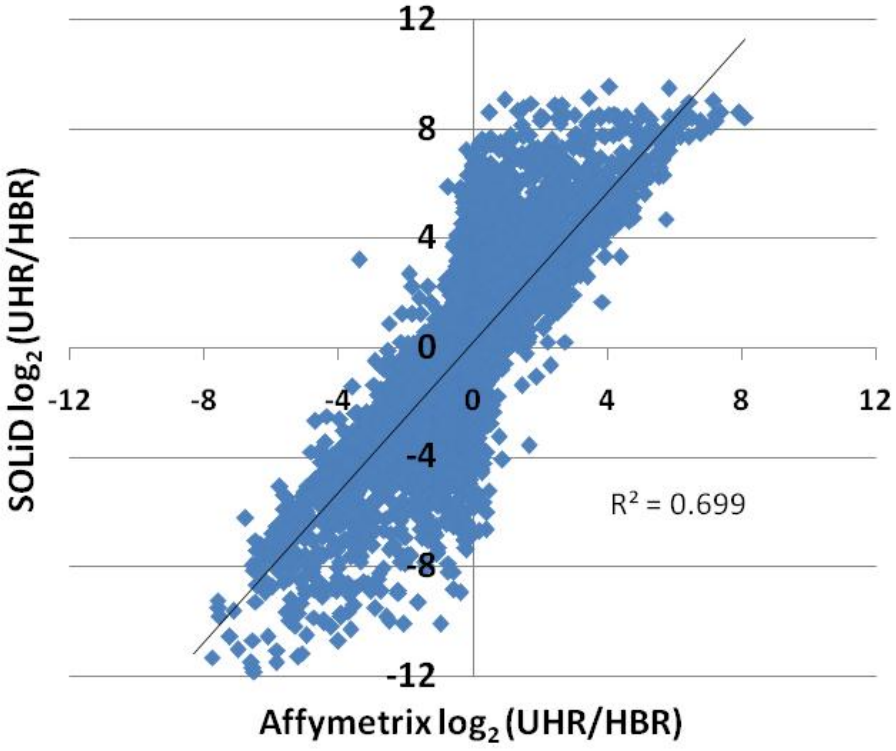
Data Analysis Workflow



Patient number	Total Reads		Aligned to filter		Aligned to genome		Aligned to RefSeq Exon	
	T	N	T	N	T	N	T	N
1	53M	67M	8M 16%	9M 14%	42M 80%	45M 68%	29M 55%	32M 48%
2	71M	72M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
3	58M	62M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
4	67M	59M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
5	61M	62M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
6	75M	71M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
7	61M	57M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
8	61M	53M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
9	64M	65M	15M 21%	16M 22%	52M 73%	52M 72%	33M 47%	32M 45%
10	65M	61M	3M 5%	5M 8%	53M 82%	48M 79%	35M 59%	32M 53%

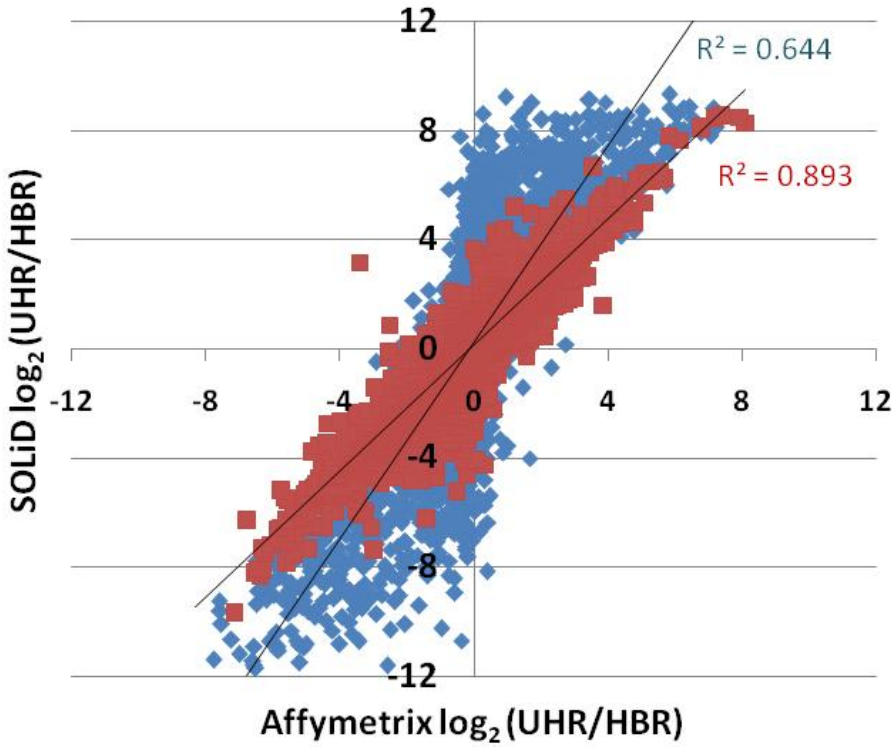
Comparison with microarrays: UHR/HBR





- SOLiD reports over a wider dynamic range.
- Correlation is good, but what do all the off-diagonal measurements represent?

Comparison with microarrays: UHR/HBR



Low abundance (≤ 100 counts)

High abundance (> 100 counts)

Many of the off-diagonal measurements are for low-abundance transcripts. Is this true dGEx detected only by SOLiD or a technical artifact?

Pairwise Analysis: Human Whole Transcriptome

[Reports: Ontology | KEGG | Chromosome | Interactive Plots] [Results: Export | Save]

	Group 1	Group 2
Conditions:	Normal	Tumor
Experiments:	80426, 80429, 80431, 80432, 80444, 80449, 80451, 80452, 80455, 80460, 80462	80425, 80428, 80430, 80433, 80442, 80448, 80450, 80454, 80458, 80459, 80461
Significance:	1.5, t-test, Benjamini and Hochberg	
Normalization:		
Quality Cutoff:	50	
Data Transformation:	Log Transformed	

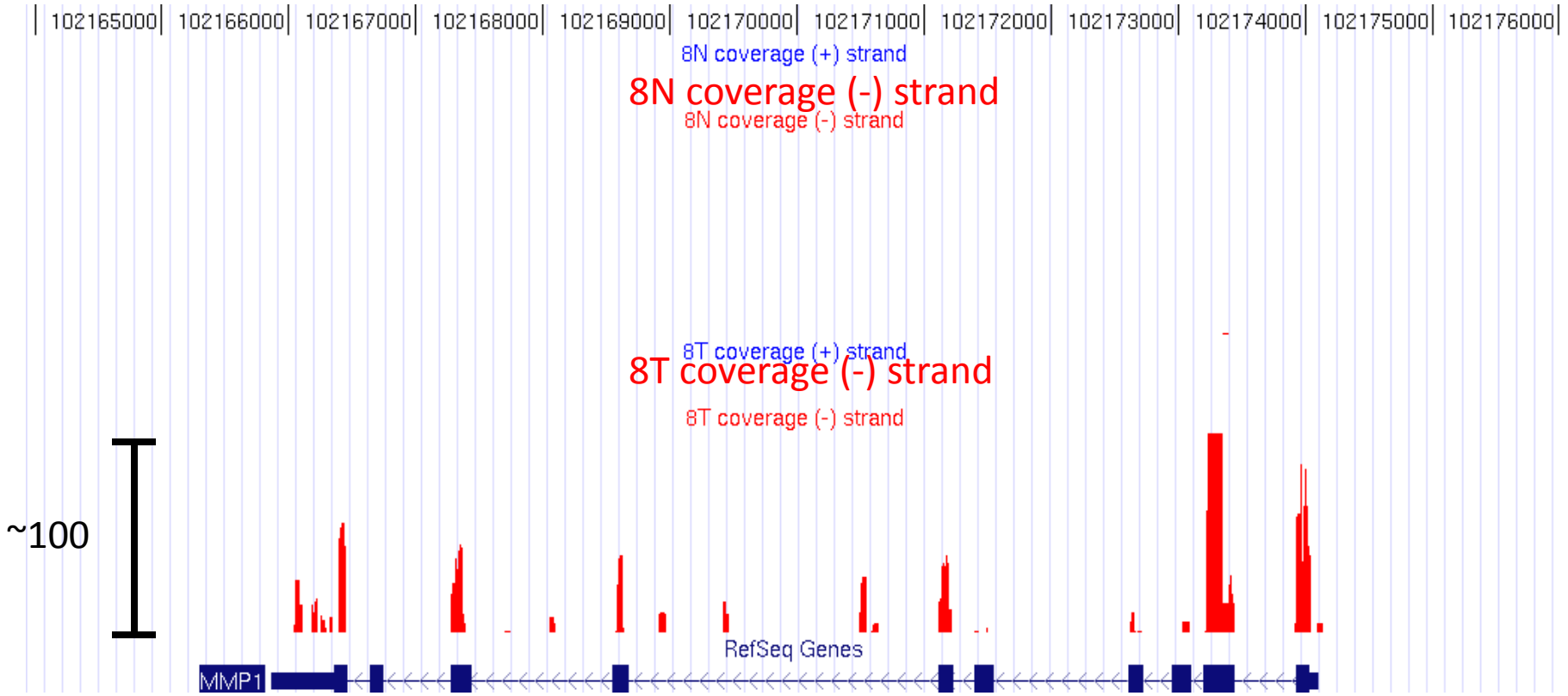
Show: 20 | p Cutoff: 0.05 | adjusted p | Threshold: 1.5 | Search | (2144 results found) [1 - 20] [21 - 40]

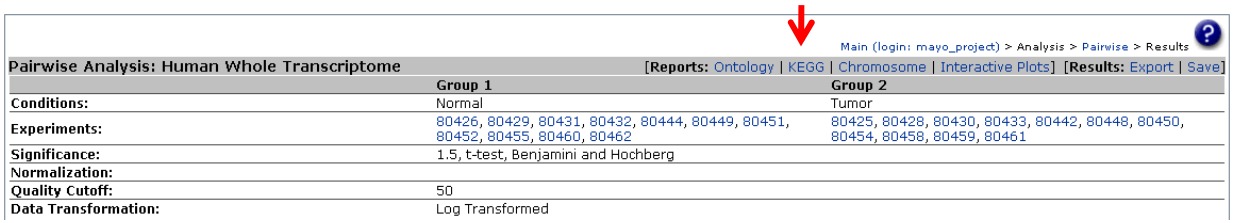
No.	Ratio	p-value	adj. p	Identifier	Gene Name
1	▼ 77.85	1.06e-05	0.00290	ADH1B	Alcohol dehydrogenase 1B (class I), beta polypeptide
2	▲ 74.04	3.45e-07	0.00060	HOXC10	Homeobox C10
3	▼ 73.72	0.00229	0.02848	CRISP3	Cysteine-rich secretory protein 3
4	▼ 67.79	0.00025	0.01056	MYOC	Myocilin, trabecular meshwork inducible glucocorticoid response
5	▼ 62.15	0.00152	0.02362	MUC7	Mucin 7, secreted
6	▼ 52.96	0.00641	0.04945	C20orf114	Chromosome 20 open reading frame 114
7	▼ 52.16	0.00400	0.03778	MUC5B	Mucin 5B, oligomeric mucus/gel-forming
8	▲ 51.84	8.21e-09	4.58e-05	HOXC8	Homeobox C8
9	▼ 51.43	0.00198	0.02680	PIP	Prolactin-induced protein
10	▲ 51.10	1.61e-09	1.35e-05	HOXC11	Homeobox C11
11	▲ 42.36	0.00011	0.00736	HOXB13	Homeobox B13
12	▼ 41.84	0.00037	0.01209	HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)
13	▼ 41.56	1.90e-07	0.00044	SERPINA5	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), membe
14	▼ 41.40	3.40e-05	0.00435	PEBP4	Phosphatidylethanolamine-binding protein 4
15	▲ 41.38	1.93e-05	0.00352	ZIC2	Zic family member 2 (odd-paired homolog, Drosophila)
16	▲ 40.55	0.00014	0.00839	HOXD13	Homeobox D13
17	▲ 39.64	1.43e-05	0.00327	HOXB9	Homeobox B9
18	▼ 38.95	0.00030	0.01102	C7	Complement component 7
19	▼ 37.46	0.00011	0.00769	HSPB6	Heat shock protein, alpha-crystallin-related, B6
20	▲ 33.87	1.31e-05	0.00314	HMGA2	High mobility group AT-hook 2

Show: 20 | p Cutoff: 0.05 | adjusted p | Threshold: 1.5 | Search | (2144 results found) [1 - 20] [21 - 40]

Transcripts that are UP in tumors: MMP1

RefSeq Id	Gene	$\log_2$ (T8 / N8)	$\log_2$ (T12 / N12)	$\log_2$ (T33 / N33)	$\log_2$ (T51 / N51)
NM_002421	MMP1	4.91	7.38	4.59	1.13





Main (login: mayo_project) > Analysis > Pairwise > Results

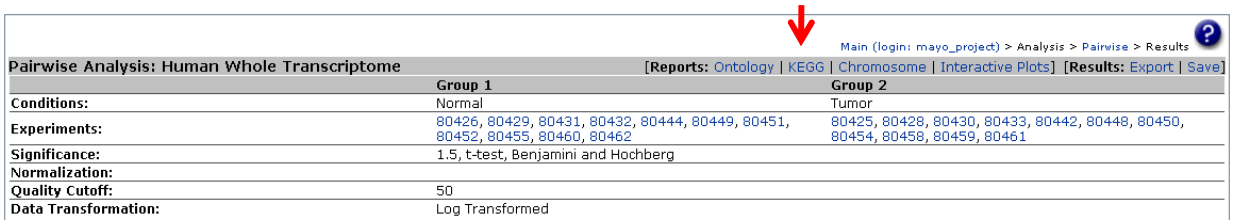
Pairwise Analysis: Human Whole Transcriptome [**Reports:** Ontology | KEGG | Chromosome | Interactive Plots] [**Results:** Export | Save]

	Group 1	Group 2
Conditions:	Normal	Tumor
Experiments:	80426, 80429, 80431, 80432, 80444, 80449, 80451, 80452, 80455, 80460, 80462	80425, 80428, 80430, 80433, 80442, 80448, 80450, 80454, 80458, 80459, 80461
Significance:	1.5, t-test, Benjamini and Hochberg	
Normalization:		
Quality Cutoff:	50	
Data Transformation:	Log Transformed	

Group 1: Normal
Group 2: Tumor

Export Report

Pathway	Genes	KEGG	Totals List			Totals Gene Set	z-score	
Cell cycle			55	55	0	125	13.20	-1.91
DNA replication			21	21	0	36	9.95	-1.02
Spliceosome			39	39	0	127	8.16	-1.92
Pyrimidine metabolism			27	26	1	98	5.73	-1.06
Mismatch repair			10	10	0	23	5.52	-0.81
Proteasome			15	15	0	47	5.21	-1.16
Homologous recombination			10	10	0	28	4.70	-0.90
Oocyte meiosis			28	25	3	113	4.56	-0.07
One carbon pool by folate			10	7	3	17	4.42	3.75
p53 signaling pathway			16	16	0	68	3.92	-1.40
Aminoacyl-tRNA biosynthesis			11	11	0	41	3.75	-1.08
Nucleotide excision repair			11	11	0	44	3.48	-1.12
Base excision repair			10	9	1	34	3.35	0.06
RNA degradation			13	13	0	59	3.25	-1.30
Pancreatic cancer			15	14	1	70	2.97	-0.69
Small cell lung cancer			16	16	0	84	2.96	-1.56
Purine metabolism			31	26	5	159	2.93	0.29
Lysine degradation			11	10	1	46	2.80	-0.25
RNA polymerase			7	7	0	29	2.66	-0.91
Non-homologous end-joining			4	4	0	13	2.59	-0.61
Basal transcription factors			8	8	0	36	2.58	-1.02
Systemic lupus erythematosus			24	22	2	138	2.56	-0.96
Progesterone-mediated oocyte maturation			16	15	1	86	2.48	-0.91
Drug metabolism - cytochrome P450			7	1	6	72	-2.38	2.90
Malaria			5	0	5	51	-2.34	3.08
Metabolism of xenobiotics by cytochrome P450			6	1	5	70	-2.34	2.25
Retinol metabolism			6	1	5	64	-2.20	2.47
Bladder cancer			8	8	0	42	2.08	-1.10



Main (login: mayo_project) > Analysis > Pairwise > Results

Pairwise Analysis: Human Whole Transcriptome [**Reports:** Ontology | KEGG | Chromosome | Interactive Plots] [**Results:** Export | Save]

	Group 1	Group 2
Conditions:	Normal	Tumor
Experiments:	80426, 80429, 80431, 80432, 80444, 80449, 80451, 80452, 80455, 80460, 80462	80425, 80428, 80430, 80433, 80442, 80448, 80450, 80454, 80458, 80459, 80461
Significance:	1.5, t-test, Benjamini and Hochberg	
Normalization:		
Quality Cutoff:	50	
Data Transformation:	Log Transformed	

Group 1: Normal
Group 2: Tumor

Export Report

Pathway	Genes	KEGG	Totals List			Gene Set	z-score	
Proximal tubule bicarbonate reclamation			8	2	6	23	-0.15	6.83
Tyrosine metabolism			9	2	7	41	-1.03	5.60
Fatty acid metabolism			9	3	6	42	-0.55	4.57
ABC transporters			7	1	6	44	-1.66	4.41
Starch and sucrose metabolism			7	1	6	52	-1.89	3.87
One carbon pool by folate			10	7	3	17	4.42	3.75
Synthesis and degradation of ketone bodies			3	1	2	9	0.15	3.56
Glycine, serine and threonine metabolism			8	4	4	31	0.62	3.45
Propanoate metabolism			8	4	4	33	0.49	3.29
Malaria			5	0	5	51	-2.34	3.08
Drug metabolism - cytochrome P450			7	1	6	72	-2.38	2.90
Pyruvate metabolism			8	4	4	40	0.08	2.80
Circadian rhythm - mammal			3	1	2	13	-0.24	2.78
Aldosterone-regulated sodium reabsorption			10	6	4	42	1.03	2.68
Valine, leucine and isoleucine degradation			7	3	4	44	-0.63	2.57
Terpenoid backbone biosynthesis			4	2	2	15	0.49	2.50
Glycolysis / Gluconeogenesis			11	6	5	64	-0.06	2.47
Retinol metabolism			6	1	5	64	-2.20	2.47
ECM-receptor interaction			13	7	6	84	-0.40	2.46
Complement and coagulation cascades			8	3	5	68	-1.46	2.32
Renin-angiotensin system			3	1	2	17	-0.52	2.26
Metabolism of xenobiotics by cytochrome P450			6	1	5	70	-2.34	2.25
Butanoate metabolism			5	2	3	35	-0.78	2.10
Fatty acid biosynthesis			2	1	1	6	0.59	2.08
Vitamin B6 metabolism			2	1	1	6	0.59	2.08
Cardiac muscle contraction			9	4	5	76	-1.30	2.04

Pairwise Analysis: Human Whole Transcriptome

[**Reports:** Ontology | KEGG | Chromosome | Interactive Plots] [**Results:** Export | Save]

	Group 1	Group 2
Conditions:	Normal	Tumor
Experiments:	80432	80433
Significance:	None	
Splice Index:	Max	
Normalization:		
Quality Cutoff:	500	
Data Transformation:	None	

Show: 20 Threshold: None Search (7324 results found) [1 - 20] [21 - 40]

No.	Ratio	Splice Index	Identifier	Gene Name
1	▼ 1.23	0.09411	ASPH	Aspartate beta-hydroxylase
2	▼ 8.29	0.09992	TPM3	Tropomyosin 3
3	▼ 2.60	0.10628	MAP4	Microtubule-associated protein 4
4	▼ 34.05	0.10727	TPM1	Tropomyosin 1 (alpha)
5	▼ 5.58	0.11305	OBSL1	Obscurin-like 1
6	▼ 1.93	0.11569	AARSD1	Alanyl-tRNA synthetase domain containing 1
7	▲ 1.21	0.11776	100132299	100132299
8	▼ 78.13	0.12083	PDE4DIP	Phosphodiesterase 4D interacting protein
9	▼ 11.47	0.12152	TMOD1	Tropomodulin 1
10	▼ 1.14	0.12225	C21orf91	Chromosome 21 open reading frame 91
11	▼ 1.72	0.12225	PDLIM5	PDZ and LIM domain 5
12	▲ 1.10	0.12381	LOC197350	Hypothetical protein LOC197350
13	▲ 1.11	0.12465	RBM16	RNA binding motif protein 16
14	▼ 3.82	0.12554	ABCA2	ATP-binding cassette, sub-family A (ABC1), member 2
15	▲ 1.92	0.12647	PALLD	Palladin, cytoskeletal associated protein
16	▼ 1.11	0.12746	RGS3	Regulator of G-protein signaling 3
17	▲ 1.07	0.12851	SLC37A1	Solute carrier family 37 (glycerol-3-phosphate transporter), member 1
18	▲ 1.08	0.12963	BMP6	Bone morphogenetic protein 6
19	▲ 1.14	0.12963	CACNB3	Calcium channel, voltage-dependent, beta 3 subunit
20	▼ 8.18	0.13082	ISLR	Immunoglobulin superfamily containing leucine-rich repeat

Show: 20 Threshold: None Search (7324 results found) [1 - 20] [21 - 40]

Main (login: mayo_project) > Analysis > Pairwise > Results > Gene Summary

» Gene Summary: Aspartate beta-hydroxylase

• By Group

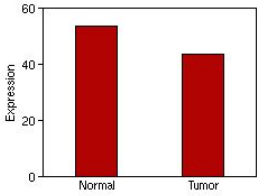
Group	Condition	N	Mean	SEM	SEM/Mean	Quality Mean
1	Normal	1	53.4728	-	-	2714.000
2	Tumor	1	43.4700	-	-	2190.000

• By Target

Group	Sample	Expression	Quality
1	14 N YP	53.4728	2714
2	13 T YP	43.4700	2190

• Exon Usage

View Density Plots
View Exon Data



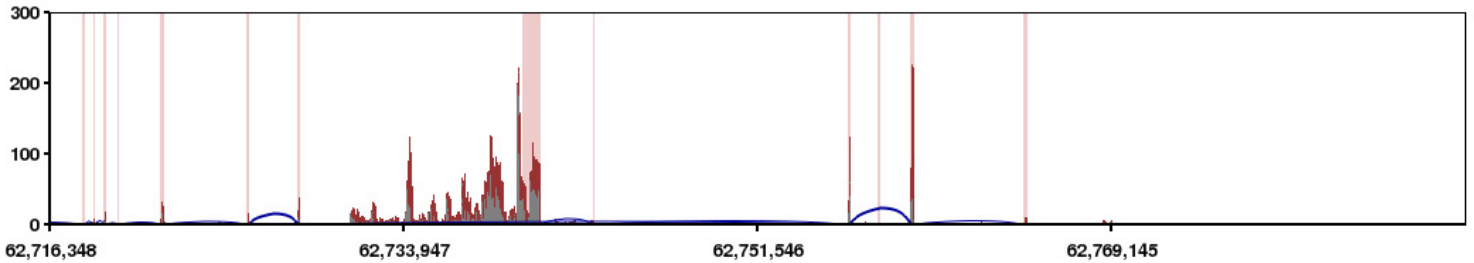
Tumor down-regulated 1.23 fold

» One-Click Gene Summary™

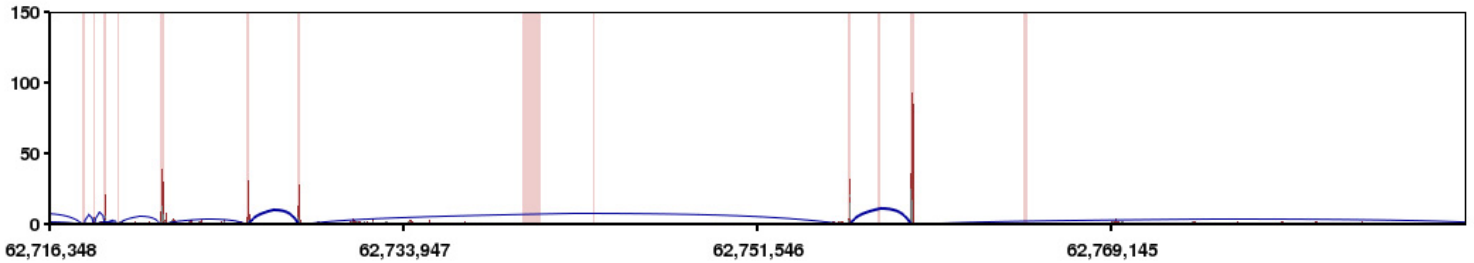
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x Chrom.

Unique Only ☐ Hide Variants ☑

Sample 1



Sample 2



Main (login: mayo_project) > Analysis > Pairwise > Results > Gene Summary

» Gene Summary: Tropomyosin 3

• By Group

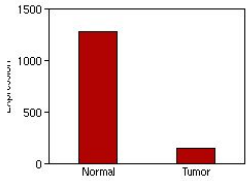
Group	Condition	N	Mean	SEM	SEM/Mean	Quality Mean
1	Normal	1	1277.240	-	-	64826.000
2	Tumor	1	154.070	-	-	7762.000

• By Target

Group	Sample	Expression	Quality
1	14 N YP	1277.240	64826
2	13 T YP	154.070	7762

• Exon Usage

View Density Plots
View Exon Data

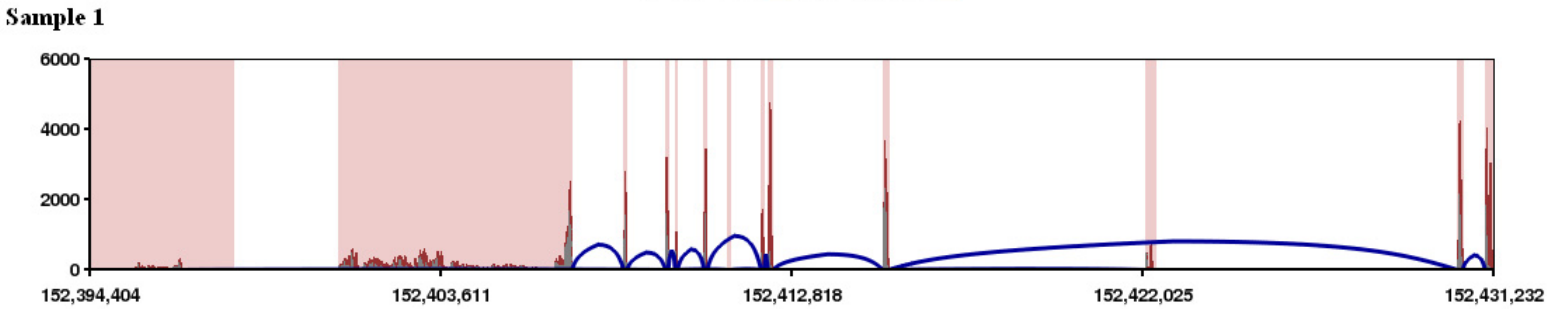


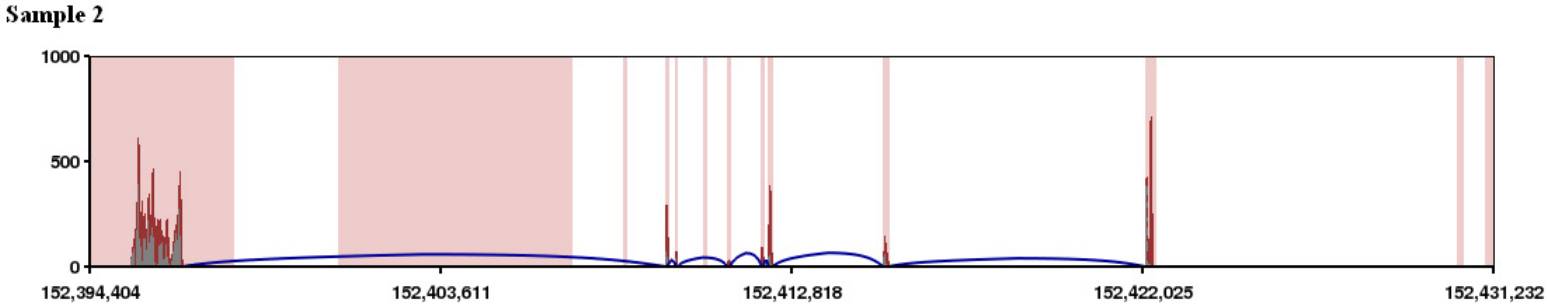
Tumor down-regulated 8.29 fold

» One-Click Gene Summ

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x Chrom.

Unique Only ☐ Hide Variants ☑





Main (login: mayo_project) > Analysis > Pairwise > Results > Gene Summary

» Gene Summary: Microtubule-associated protein 4

• By Group

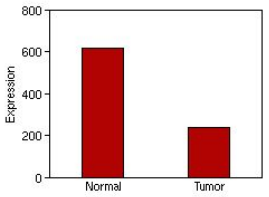
Group	Condition	N	Mean	SEM	SEM/Mean	Quality Mean
1	Normal	1	619.706	-	-	31453.000
2	Tumor	1	238.033	-	-	11992.000

• By Target

Group	Sample	Expression	Quality
1	14 N YP	619.706	31453
2	13 T YP	238.033	11992

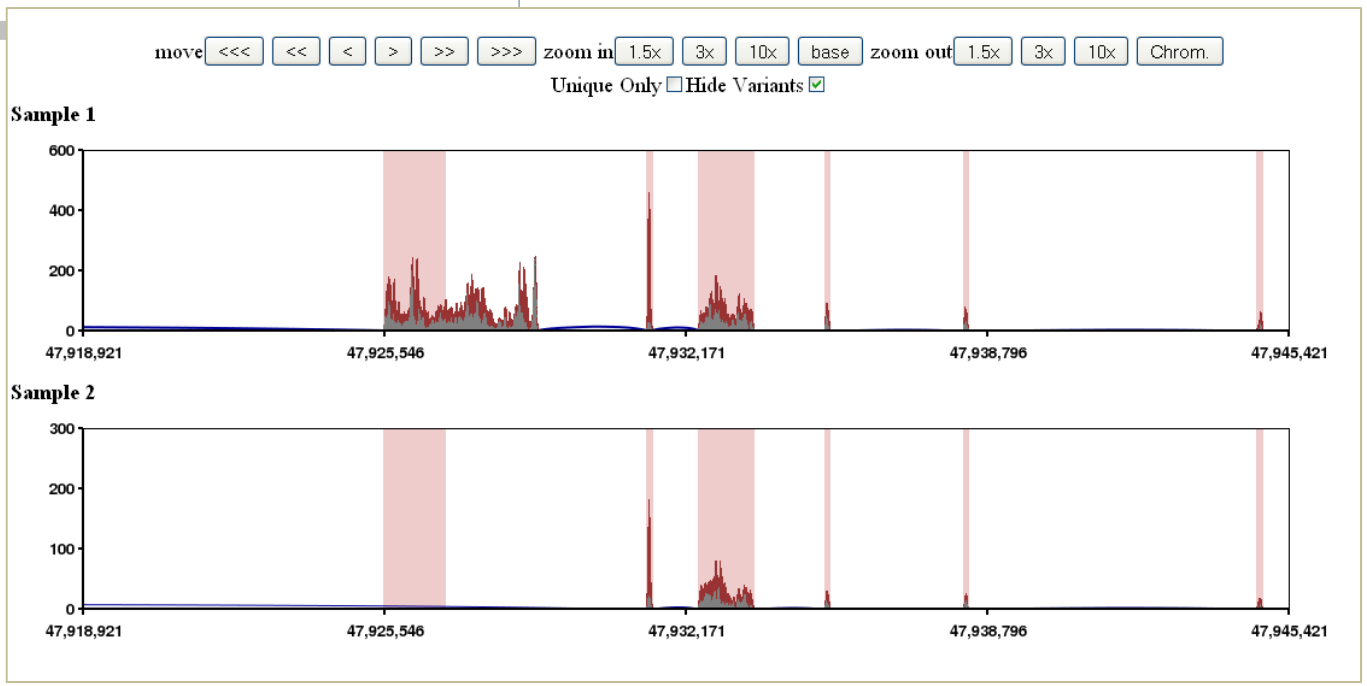
• Exon Usage

View Density Plots
View Exon Data

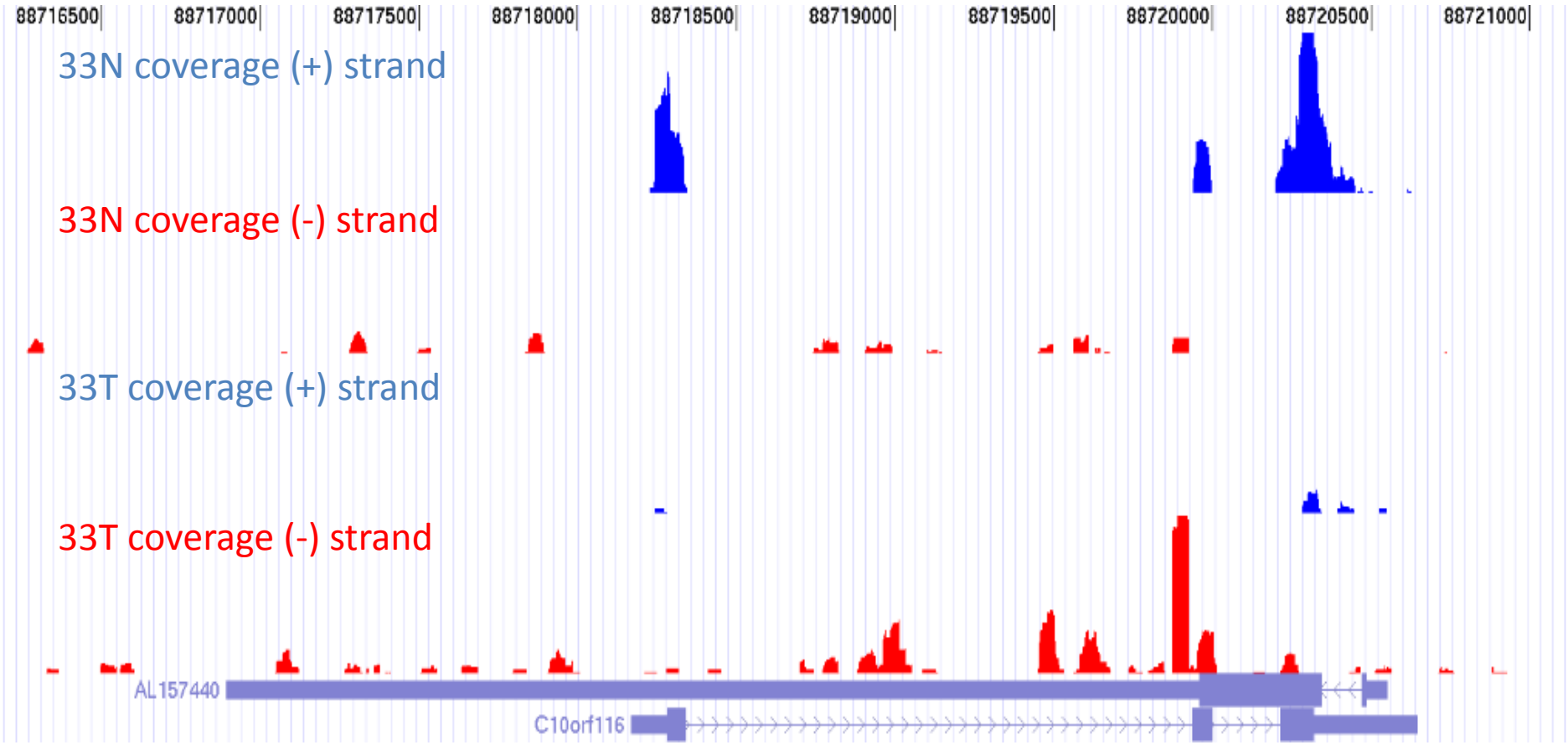


Tumor down-regulated 2.60 fold compared to Normal

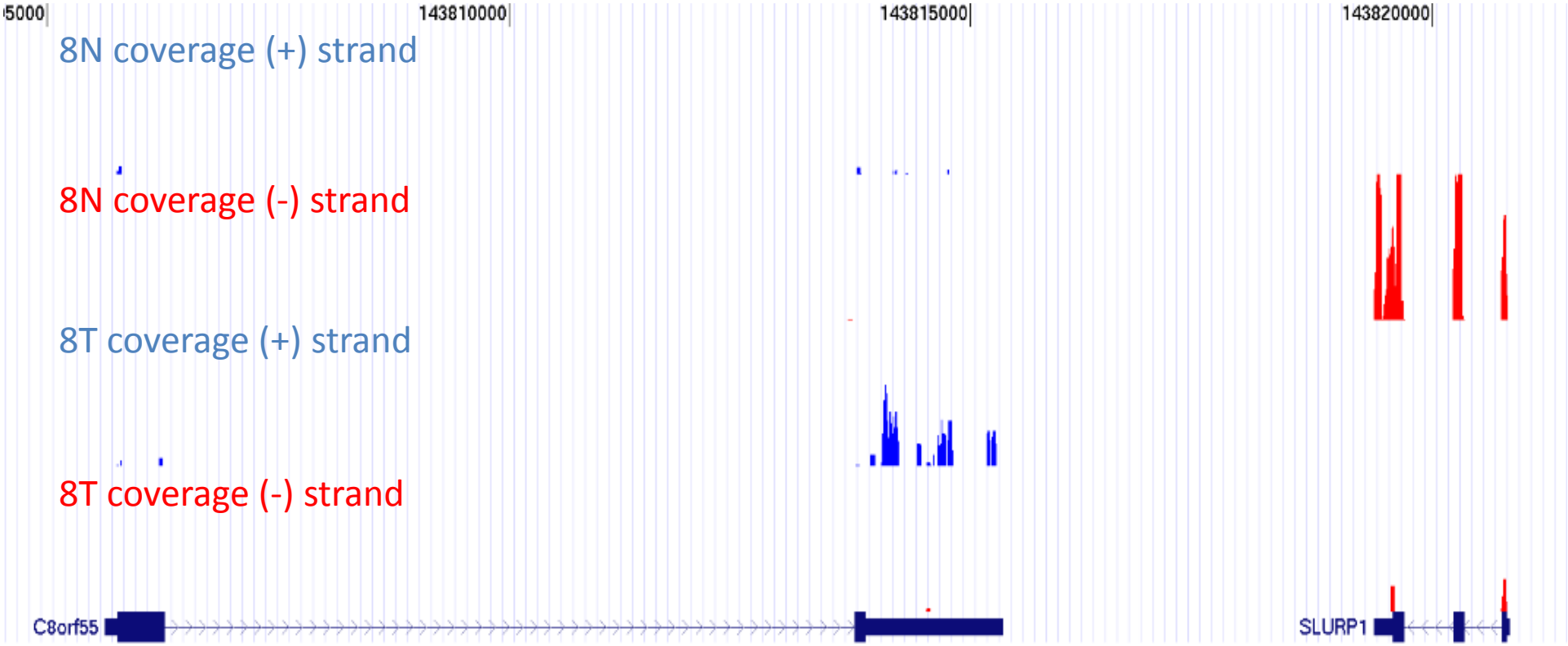
» One-Click Gene Summary™



Antisense transcripts: AL157440 & C10orf116

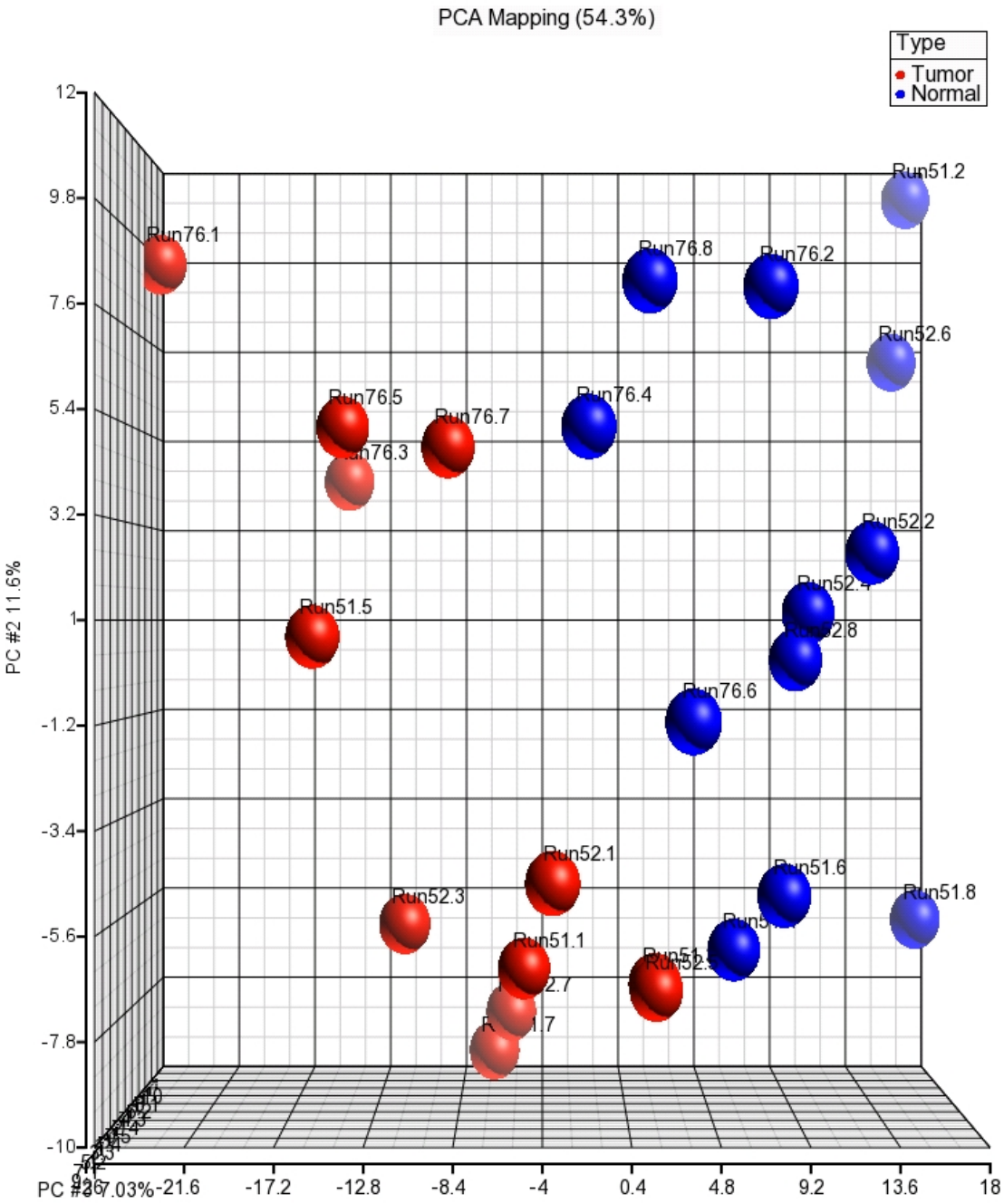


Antisense transcripts: C8orf55 & SLURP1



Long Non-coding Transcripts

- 2 databases of ncRNA were evaluated: one had 2,500 and the other had 400,000 ncRNAs
- Use the ncRNA sequences as a reference to map the fastq sequence reads
- Obtain read counts as a measure of expression of each ncRNA
- 2-19% of total reads maps to these ncRNAs
- Normalize the count per sample reads
- Differential expression analysis for tumor versus normal

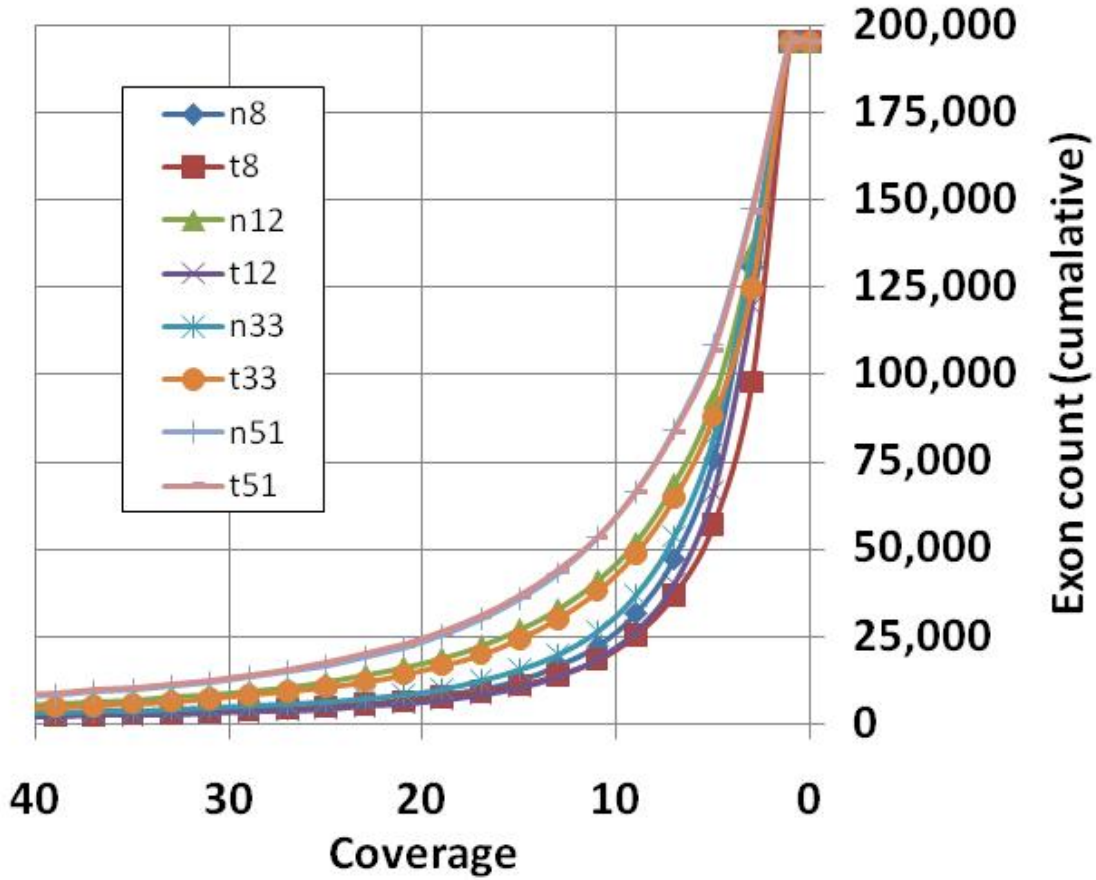

PCA Mapping (54.3%)
Type
Tumor
Normal
PC #2 11.6%
PC #3 7.03%
12
9.8
7.6
5.4
3.2
1
-1.2
-3.4
-5.6
-7.8
-10
-21.6
-17.2
-12.8
-8.4
-4
0.4
4.8
9.2
13.6
18
Run51.2
Run76.1
Run76.8
Run76.2
Run52.6
Run76.5
Run76.7
Run76.4
Run52.2
Run51.5
Run76.6
Run52.1
Run51.6
Run51.8
Run52.3
Run51.1

What other information is available from RNAseq?

- RNAseq is actually sequencing the transcripts
- For more abundantly expressed genes can determine if there are mutations in your transcribed sequences
- Allele-specific expression changes can also be detected

Coverage of exons

5K-25K exons may have sufficient coverage for mutation detection/ allele-specific gene expression analysis.



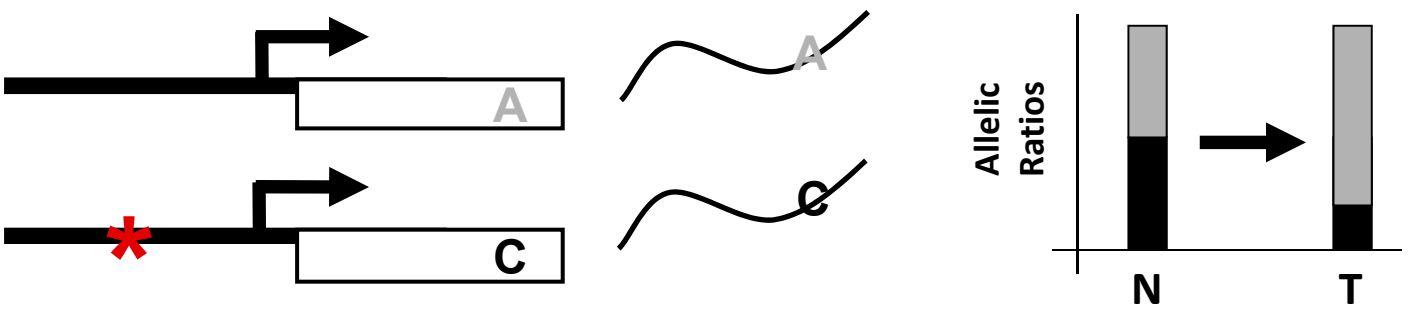
Searching for Cancer-Specific Mutations

- Examine the more abundantly expressed genes and compare sequence in tumor to matched normal tissue
- Gene must be expressed well in both tumor and normal
- Can still do this for the top 10% of expressed genes (more if you sequence deeper!)

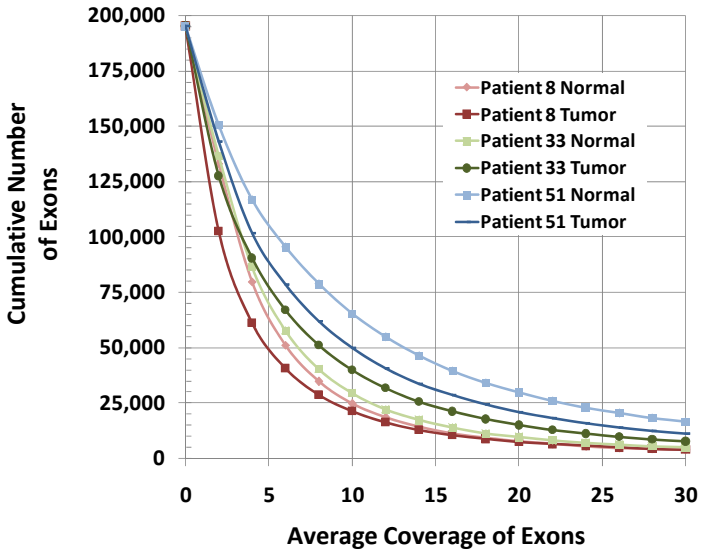
Allele-Specific Expression

- An important mechanism in cancer development (deletions) is detected as loss of heterozygosity of polymorphic markers
- This can result in loss of expression of just one allele
- Can this have an effect beyond just a 50% reduction in expression?
- Yes- Allele-specific expression

a



b



c

GO Category Description	Count in Category	Count in Overlap	P-value
cell adhesion	189	9	3.5E-03
organ development	294	12	1.4E-03
epidermis development	55	7	9.9E-05
ectoderm development	60	8	1.6E-05
intermediate filament cytoskeleton	25	8	1.7E-08
plasma membrane	578	16	1.4E-03

d

	Allelic Imbalance?			Gene Expression Tumor vs Normal			Allelic Imbalance Details			Cancer gene?	Adhesion/ECM?	
	8	33	51	8	33	51	8	33	51			
CD44	●	●	●	-0.9	0.5	1.3	Syn, I	K->R	I/NSS	○	●	Cd44 molecule
DSC3	●	●	●	0.6	0.5	1.1	3	3, K->R, S->T	3 (3)	○	●	desmocollin 3 isoform Dsc3a preproprotein
DST	●	●	●	-1.3	0.1	0.0	L->F	Syn, I/NSS	G->R	○	●	dystonin isoform 1e precursor
MALAT1	●	●	●	0.3	-0.6	0.3	NC	NC	NC	○	○	metastasis associated lung adenocarcinoma 1
PERP	●	●	●	0.7	0.0	1.7	3 (2)	Syn, R->P	3	○	●	PERP, TP53 apoptosis effector
ALDH3A1	●	●	○	-0.9	1.4	-2.4	3, Syn	D->G		○	○	aldehyde dehydrogenase 3A1
CCND1	●	●	○	3.7	-0.1	1.2	3 (2)	3		●	○	cyclin D1
CTNND1	●	●	○	1.2	0.0	1.0	I, 3	I, 3		○	●	catenin, delta 1 isoform 1A
DSP	●	●	○	-0.1	-0.3	1.6	I->F, Syn	Syn		○	●	desmoplakin isoform I
FAT2	○	●	●	0.1	0.3	1.9		Syn	S->L, I, I->M,	○	●	FAT tumor suppressor homolog 2
GJB2	○	●	●	0.6	-0.7	1.2		W->C	I, 3	○	●	gap junction protein, beta 2

RNAseq is complementary to other technologies

- Compare expression to methylation (which can either be done on arrays or with some form of methylation sequencing)
- Compare the exome sequence searching for mutations in genes to gene expression
- Attempt to integrate all of these into a cohesive model (for example of cancer development)

RNAseq is also becoming affordable!

- Current generation HiSeq 2000 can generate 300 million reads per lane of a flow cell.
- If 75 million reads is sufficient can bar code and run 4 samples/lane. Cost per sample (with library prep) is then $500 per sample
- If you need 150 million reads the cost per sample is $900 per sample
- As sequence output further increases the cost of RNAseq will further decrease

RNAseq as part of clinical practice

- Several institutions and companies are already exploring using RNAseq on cancer specimens to better inform clinical decisions (University of Michigan, Genomic Health)
- Can determine important changes in transcription with much greater granularity than microarrays.
- Can also determine non-human transcripts (such as viral transcripts)
- Information is very complementary to exome sequencing
- Will this become a standard part of cancer care very soon?

CONCLUSIONS

- RNAseq is a powerful tool to analyze the transcriptional output of cells
- Think carefully to design the proper RNAseq experiment before you waste your money and time
- Decide on the number of samples, which samples, what type of RNAseq library and how many sequences/sample
- Can determine message abundance, transcript isoforms produced, allele-specific expression and even mutations in more abundantly expressed transcripts
- Will quickly be replacing microarrays for measuring transcription