



**RNA-Seq statistical analysis, visualization, and
systems biology: Neural Differentiation of hESCs
using paired-end sequencing technology**

Matt Newman

484-918-0515

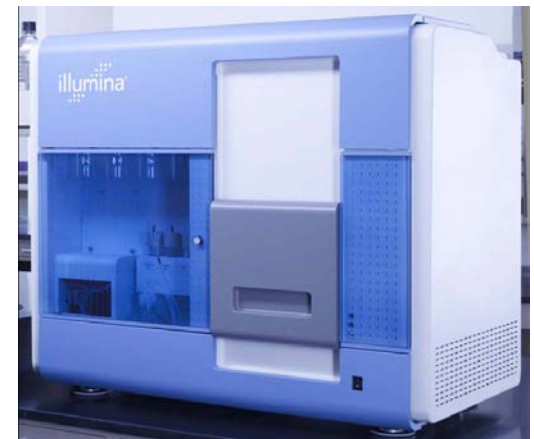
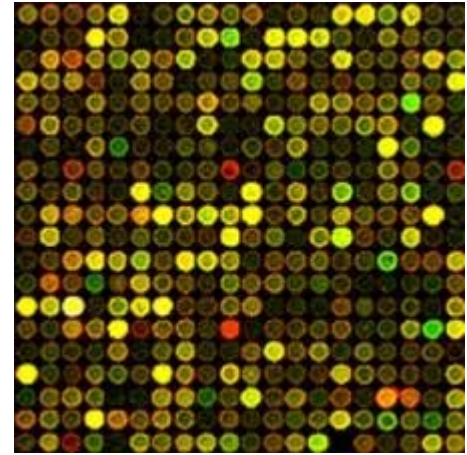
matt.newman@omicsoft.com

Omicsoft Corporation

7-30-10

Transitions

- Microarrays for Gene Expression
- Next Generation Sequencing
 - Gene Expression
 - Exon-exon junctions
 - Gene Fusion
 - Novel Genes
 - Mutation Analysis
 - More.....



Challenge

- Can the bioinformatician easily analyze Next Generation Sequencing Data?
 - Quality control of data
 - Speed of analysis
 - Size of data
 - Types of analysis
 - Type of Computer

Benefits

- Explore areas not previously explored
 - Gene Fusion
 - Alternative Transcription
 - Mutations linked to expression changes
- New platform for more accurate expression analysis

Dataset Background

- 193 million reads
- 36 bp/read
- Paired end technology used
- Illumina Genome Analyzer

hESC Differentiation

- 4 stages of neural differentiation
 - hESC
 - N1 (early initiation)
 - N2 (neural progenitor)
 - N3 (early glial-like)

Hypotheses

- We can take RNA-Seq data from the public domain, perform QC and alignment, then use the results to explore the following:
 - Neuronal differentiation can be detected using systems biology and next generation sequencing technology at the early stages of differentiation
 - Next generation sequencing can be used to explore the possibility of alternative splicing in neuronal development.

Sequence Read Archive

SRP002079 GSE20301: Dynamic transcriptomes during neural differentiation of human embryonic stem cells

Study Type: Transcriptome Analysis
Submission: [SRA012181](#) by INDIVIDUAL on 2010-03-08 21:26:29
Abstract: n/a
Description: Summary: Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short long, and paired-end sequencing In order to examine the fundamental mechanisms governing neural differentiation, we analyzed the transcriptome changes that occur during the differentiation of human embryonic stem cells (hESCs) into the neural lineage. Undifferentiated hESCs as well as cells at three stages of early neural differentiation, N1 (early initiation), N2 (neural progenitor), and N3 (early glial-like) were analyzed using a combination of single read, paired-end read, and long read RNA sequencing. The results revealed enormous complexity in gene transcription and splicing dynamics during neural cell differentiation. We found previously unannotated transcripts and spliced isoforms specific for each stage of differentiation. Interestingly, splicing isoform diversity is highest in undifferentiated hESCs and decreases upon differentiation, a phenomenon we call "isoform specialization." During neural differentiation, we observed differential expression of many types of genes including those involved in key signaling pathways, and a large number of extracellular receptors exhibit stage-specific regulation. These results provide a valuable resource for studying neural differentiation and reveal insights into the mechanisms underlying in vitro neural differentiation of hESCs, such as neural fate specification, NPC identity maintenance and the transition from a predominantly

The SRA runs have been pre-filtered by NCBI to remove contaminating human sequence. To obtain the unfiltered version of these data, please navigate to [dbGaP authorized access system](#)

[Download fastq for entire study](#)

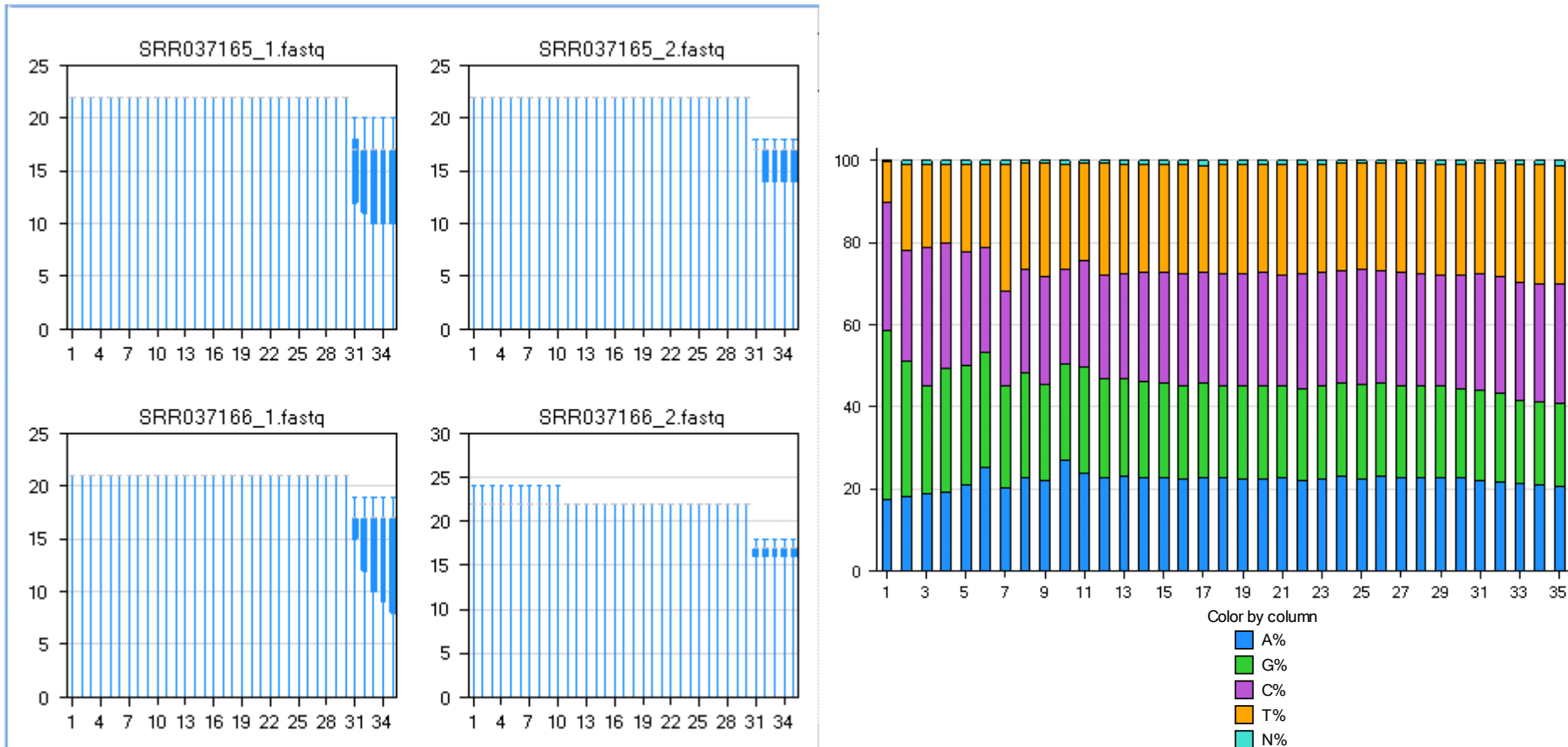
Experiments

[Show RUNs for each experiment](#)

↑ Accession	Spots	Bases
Total: 20	252.7M	11.9G
SRX017367	4.7M	332.4M
SRX017368	9.7M	701.4M
SRX017369	4.4M	153.6M
SRX017370	4.3M	302.4M
SRX017371	7.3M	482.2M
SRX017372	34.1M	921.6M
SRX017373	42.3M	1.5G
SRX017374	1.8M	128.9M
SRX017375	14.0M	1.0G
SRX017376	7.0M	245.2M

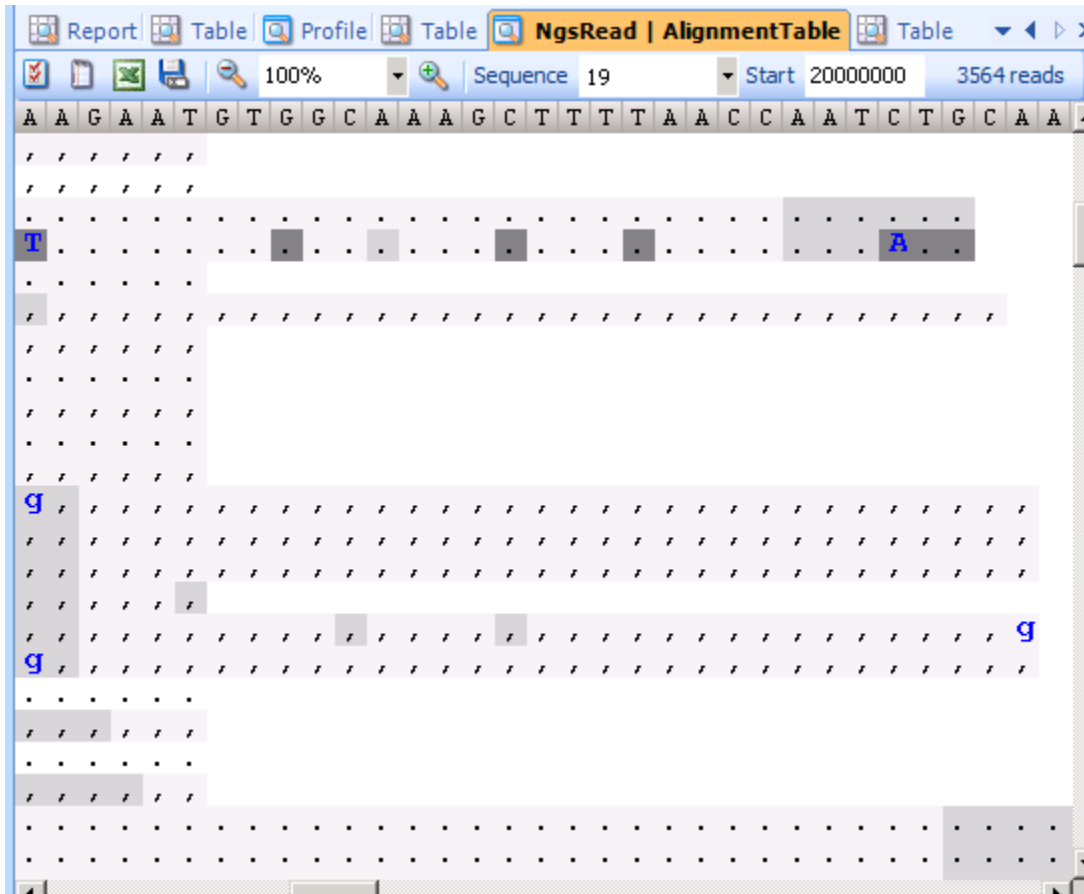
- Public database containing raw sequence reads including next generation platforms

Quality Control of Reads



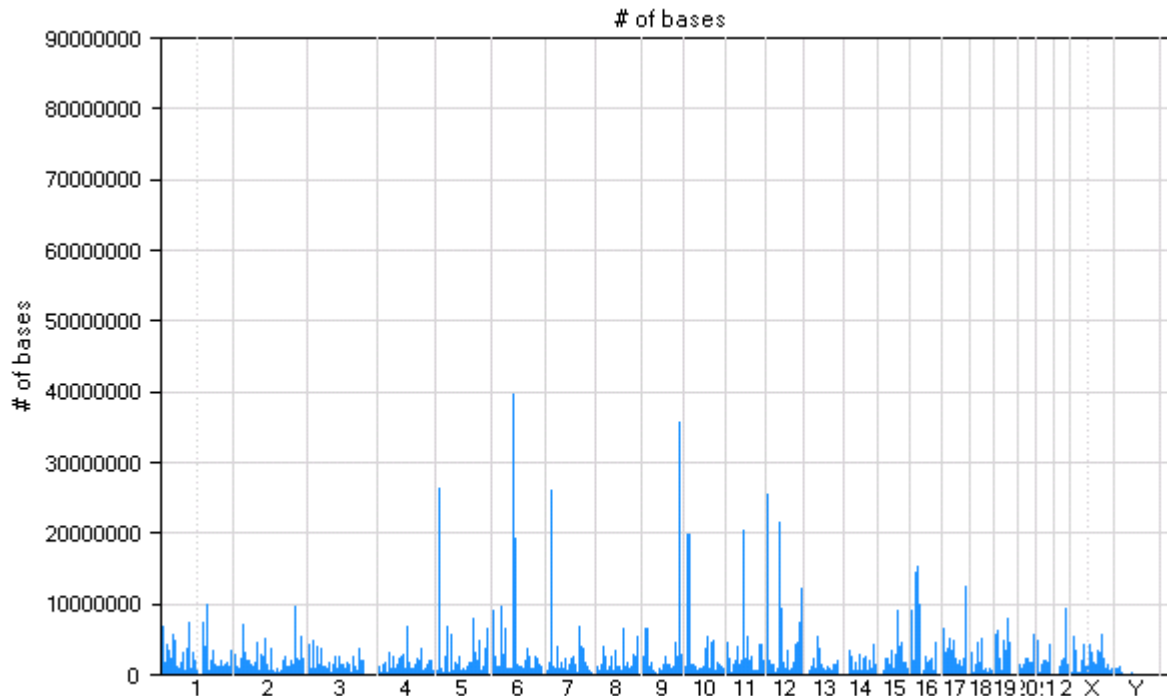
- Use quality score and base distribution to gauge quality of each set of reads

Alignment



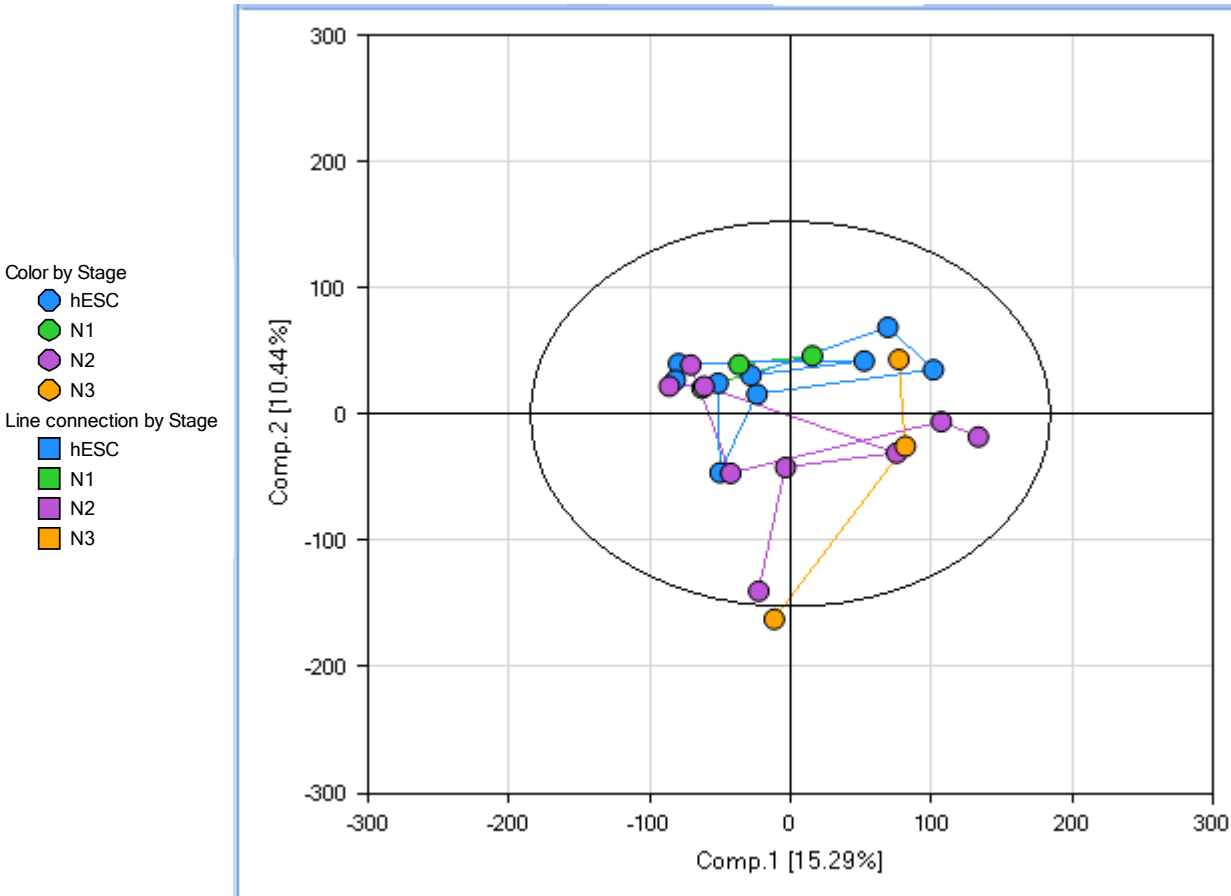
- Using OSA (Omicsoft Aligner Algorithm)
- Uses trimming of bases for quality to allow for a higher mapping percentage
- 30 hours for alignment and counting algorithms on a Windows Computer w/6 gb RAM and 4 cores

Coverage Statistics



- On average, 73% of reads mapped to genome (56-86%)
- For 36bp reads, this is good performance

Principal Component Analysis



Across over 49,733 ENSEMBL IDs, we can't see much separation between stages

Results → "Gene Expression Data"

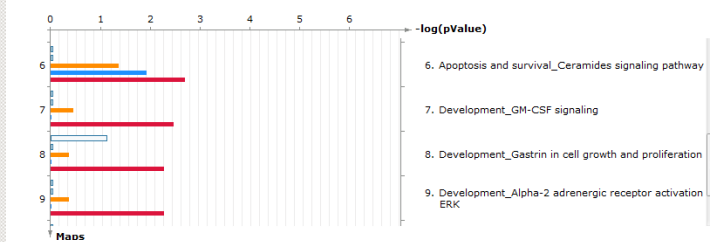
GeneID	N2 vs hESC.FoldC...	N2 vs hESC.RawP...	N3 vs hESC.FoldC...	N3 vs hESC.RawP...	GeneName	hESC	N1
ENSG000000...	4.51	0.0341	13.14	0.0123	RTN1	3.3580	3.3956
ENSG000000...	3.62	0.0448	8.53	0.0202	POU3F3	3.2320	3.0504
ENSG000000...	3.47	0.0295	8.50	0.0098	TSPAN7	3.7902	4.1305
ENSG000000...	2.96	0.0544	7.35	0.0153	CST3	6.8602	6.4405
ENSG000000...	2.68	0.0489	6.63	0.0100	DNER	3.6541	3.2311
ENSG000000...	2.60	0.0102	5.15	0.0026	NDRG2	3.6385	4.0710
ENSG000000...	2.21	0.0418	4.57	0.0080	TMEM59L	4.0603	4.7105
ENSG000000...	2.27	0.0635	4.50	0.0193	NDRG4	4.3861	4.2595
ENSG000000...	3.89	0.0036	4.07	0.0257	STMN2	4.0824	2.8156
ENSG000000...	2.04	0.0587	3.77	0.0156	CTNND2	4.9972	4.2521
ENSG000000...	2.37	0.0060	3.72	0.0035	SEPT3	4.8102	4.7584
ENSG000000...	3.29	0.0089	3.68	0.0366	CDR1	4.2653	3.0801
ENSG000000...	1.91	0.0692	3.47	0.0170	CPNE2	4.8549	4.6836
ENSG000000...	1.99	0.1059	3.46	0.0424	ATP1A2	5.8184	6.4045
ENSG000000...	2.05	0.0179	3.16	0.0082	PCDHGA12	5.2509	5.4914
ENSG000000...	1.56	0.1153	3.01	0.0095	C4orf48	4.9941	4.9591

Compare Experiments Workflow

#	Color	Experiment name	Species (Defined by genes origin)
1.	■	Log2 TPM.Tests_N3 vs hESC.FoldChange	Homo sapiens
2.	■	Log2 TPM.Tests_N2 vs hESC.FoldChange	Homo sapiens
3.	■	Log2 TPM.Tests_N1 vs hESC.FoldChange	Homo sapiens

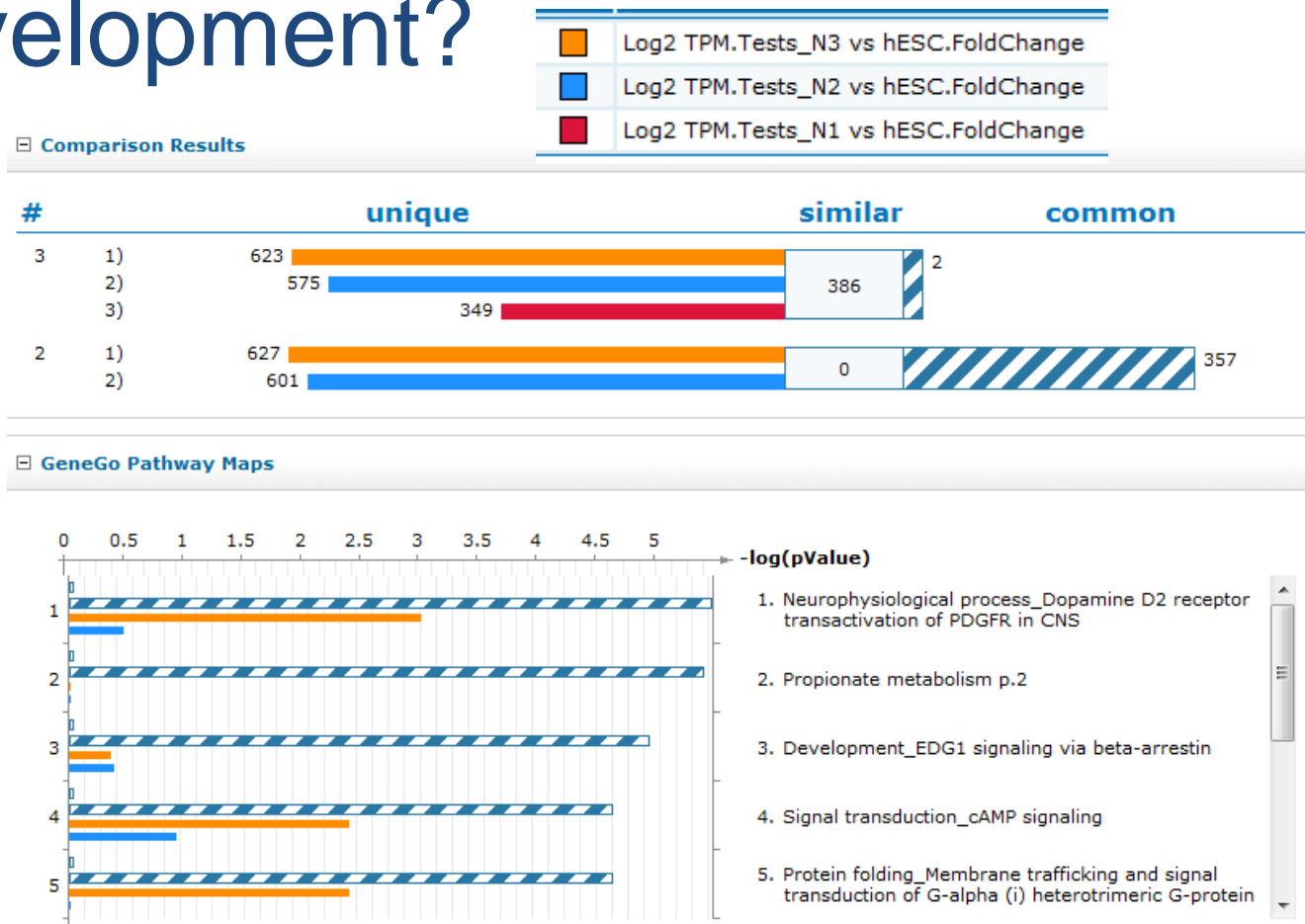
Define threshold and click "Apply"

GeneGo Pathway Maps



- After alignment, you are left with a type of "gene expression data". Can be analyzed in Array Studio and MetaCore

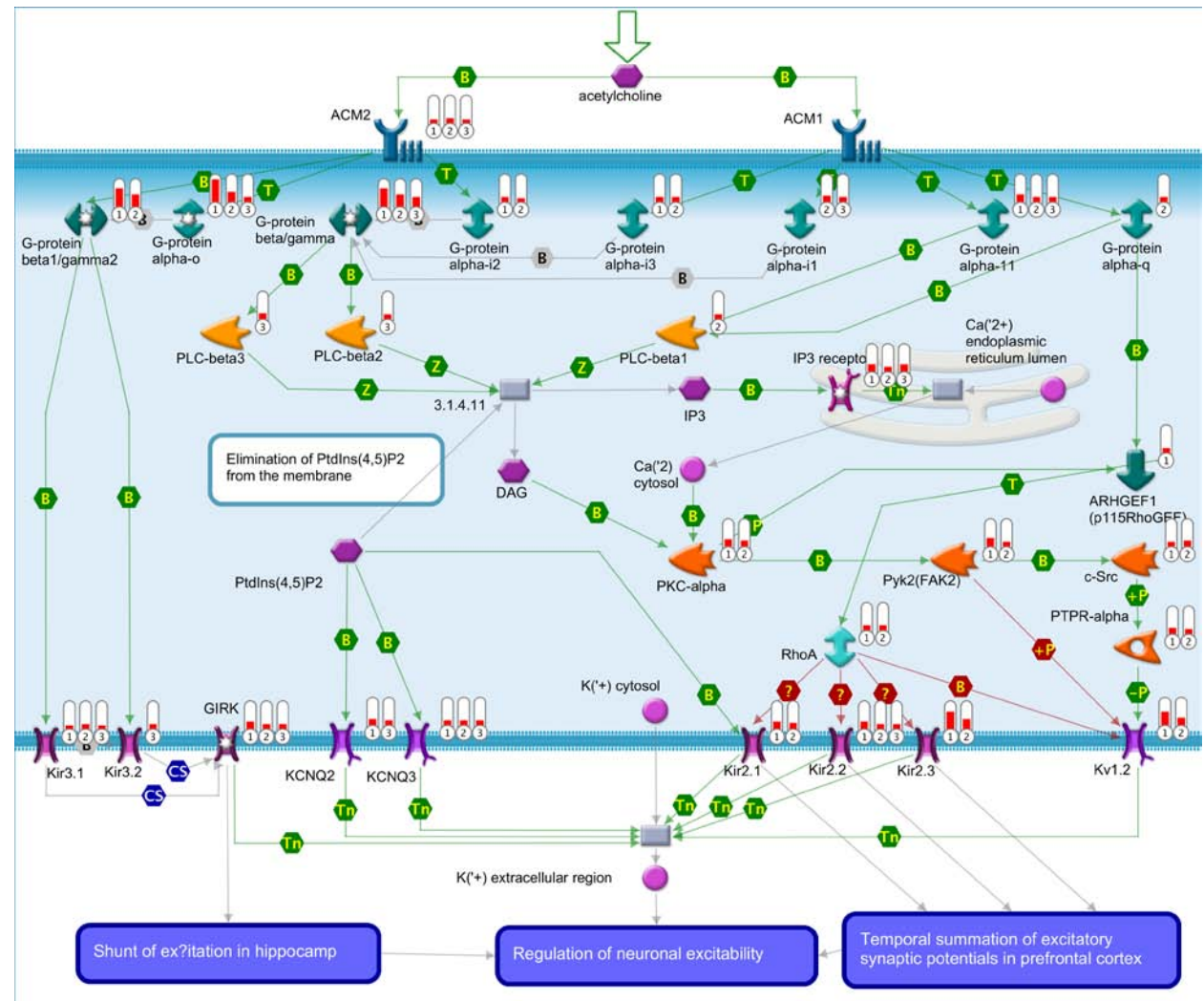
How early can we detect neuronal development?



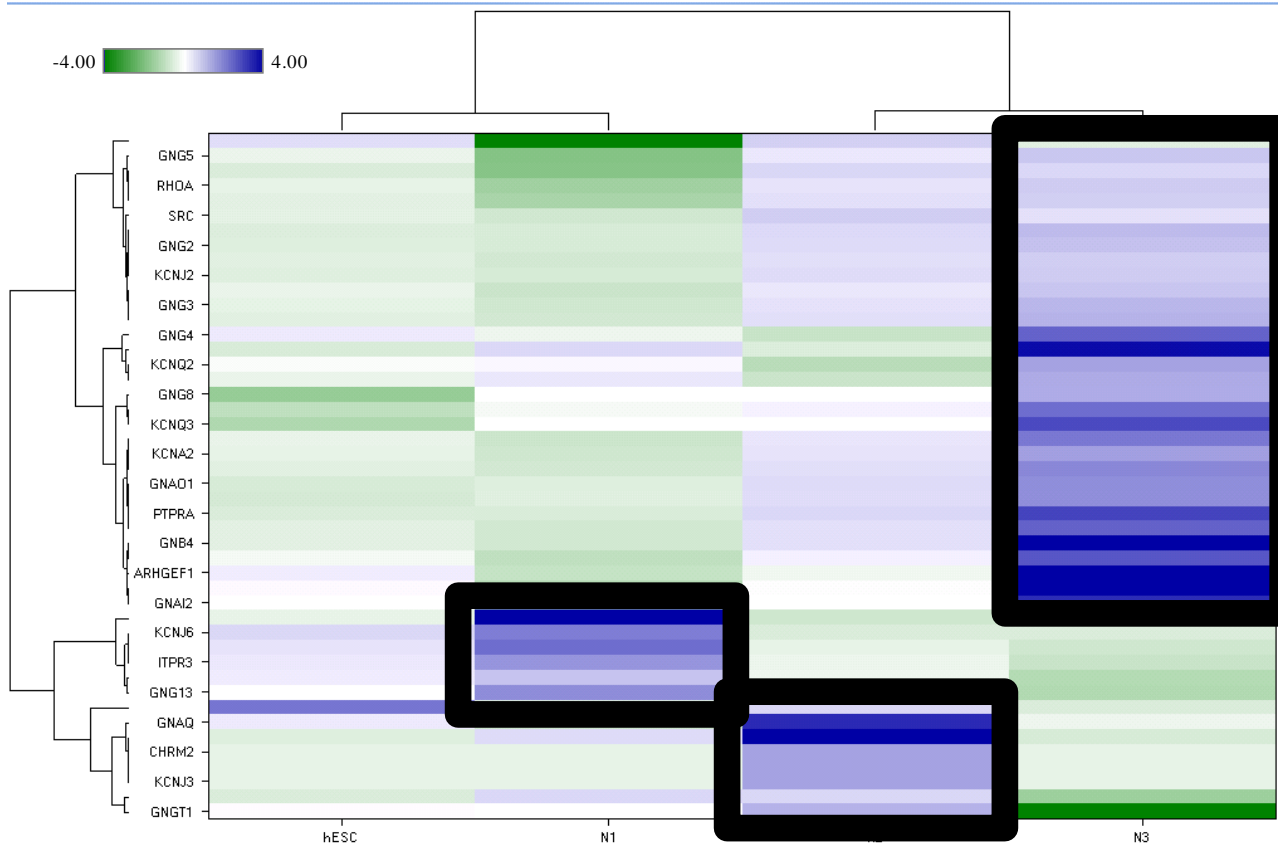
Up-regulation of neuronal development pathways can be detected early as N2 stage, but not N1.

Neurophysiological process: ACM1 and ACM2 in neuronal membrane polarization

- Clear evidence of increase in pathway at N2 and N3 stages
- Possibly evidence at N1 stage (not statistically significant)

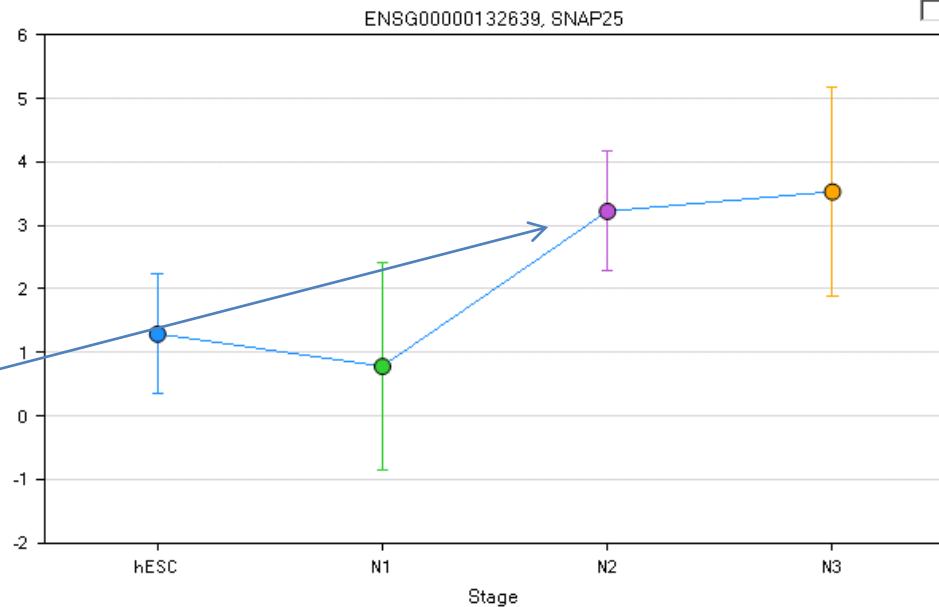
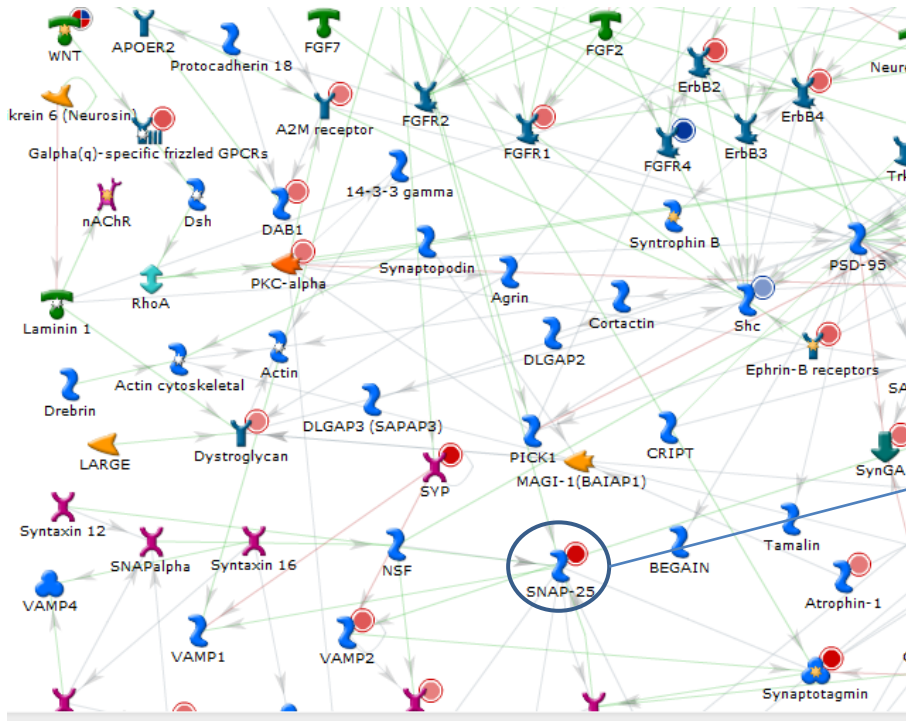


Neurophysiological process: ACM1 and ACM2 in neuronal membrane polarization Heatmap



Specific genes in the pathway are activated throughout the development process

Development Neurogenesis: Synaptogenesis



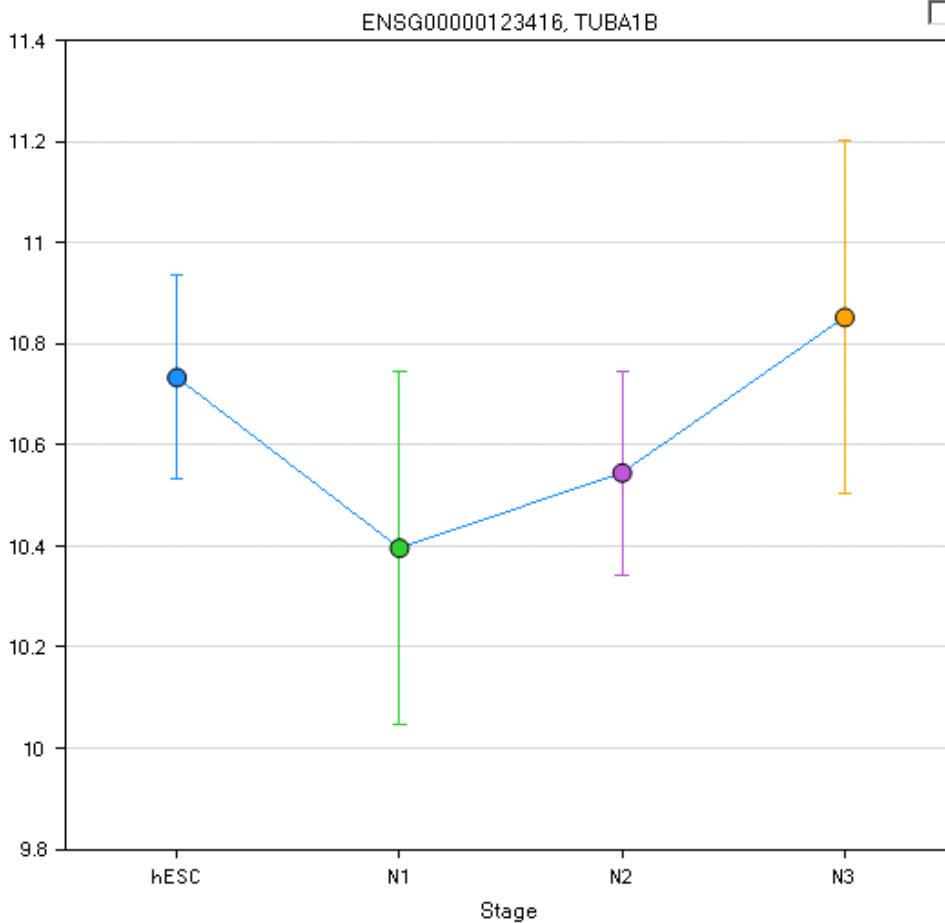
Multiple markers in the network are increased at N2 and N3 stages of development. Markers cannot be detected at N1 stage

Detection of Alternative Transcription

- Unique exons can be used to detect alternative splicing
 - Not many unique exons for a comprehensive gene model like Ensembl
- Detection of exon-exon junctions is done during the alignment stage
 - Junctions can potentially be used to look for alternative splicing
 - Novel genes
 - Novel isoforms

TUBA1B Expression

Not
differentially
expressed
between stages

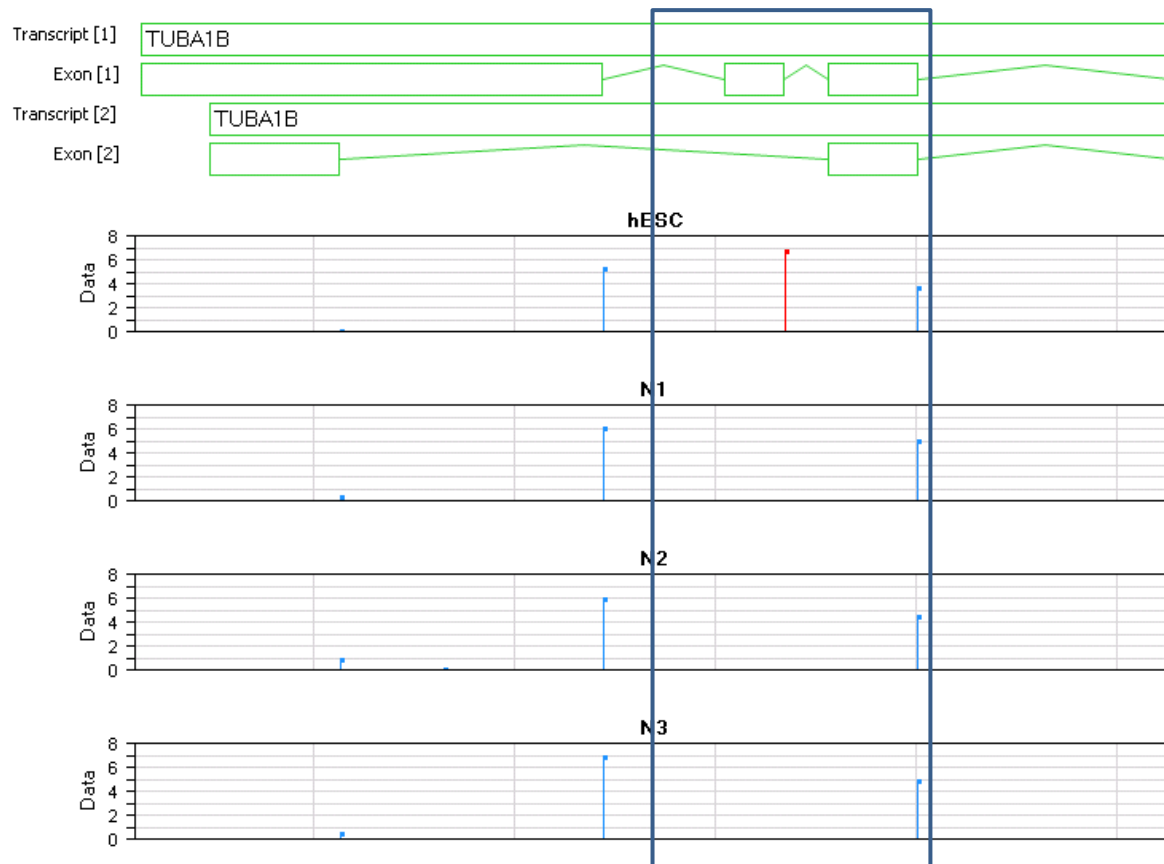


Protein Details

TUBA1B HUMAN

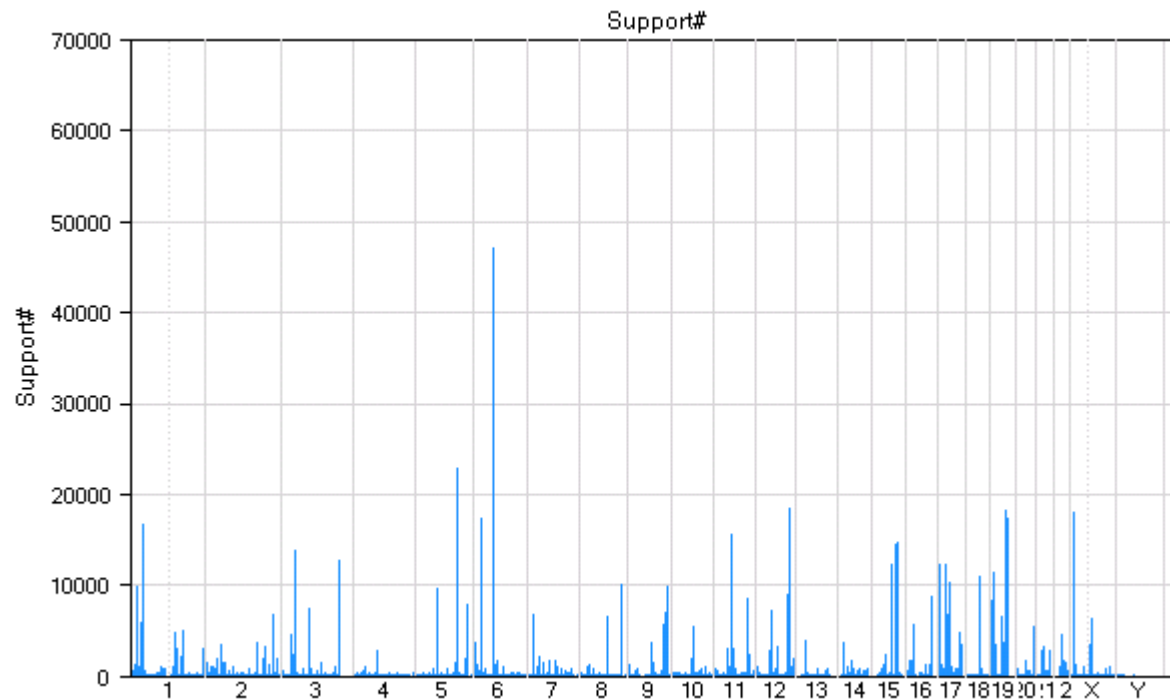
Name	TUBA1B_HUMAN / Tubulin alpha-1B chain
Synonyms	Alpha-tubulin ubiquitous, TUBA1B_HUMAN, Tubulin alpha-1B chain, Tubulin alpha-ubiquitous chain, Tubulin K-alpha-1
Description	Tubulin is the major constituent of microtubules. It binds two moles of GTP, one at an exchangeable site on the beta chain and one at a non-exchangeable site on the alpha-chain.
Molecular Weight	50152
Relations (Parents)	Tubulin alpha (HUMAN) (Protein group) , Tubulin (in microtubules) (HUMAN) (Group of complexes)
Localization	microtubule
Organ/Tissue Expression (RNA)	Adrenal Glands; Bone Marrow; Brain; Colon; Fetal brain; Fetal kidney; Fetal liver; Fetal thymus; Heart; Intestine, Small; Kidney; Liver; Lung; Lymphocytes; Mammary Glands, Human; Muscle, Skeletal; Ovary; Palatine Tonsil; Pancreas; Placenta; Prostate; Retina; Salivary Glands; Skin; Spinal Cord; Spleen; Testis; Thymus Gland; Thyroid Gland; Trachea; Uterus

TUBA1B Alternative Transcription



- Using detection of exon-exon junctions, we can detect that there is a switch in transcription between the two splice variants
- Transcript[1] expressed in hESC stage, but not N1, N2, or N3 stages

Unannotated exon junctions



Much to be learned

- Novel genes
- Novel isoforms

Summary

- It is possible to take NGS data from the public domain, align that data to the genome, and then look for gene expression patterns
 - Differential expression
 - Alternative transcription
- Detection of neuronal pathways does not occur easily until the N2 and N3 stages in differentiation
- Alternative transcription can be detected by looking at exon-exon junctions

Other Questions

- Pathways involved in keeping embryonic stem cells from differentiating
- Novel genes/isoforms
- Gene Fusion
- Mutations linked to gene expression changes

Questions on Methods

For questions on methods for analysis or more information, contact matt.newman@omicsoft.com