LARGE SCALE EXPRESSION ANALYSIS

Evolution and Revolution

Current Topics in Genome Analysis 2014

Paul Meltzer

No Relevant Financial Relationships with Commercial Interests

WHOLE GENOME APPROACHES TO BIOLOGICAL QUESTIONS

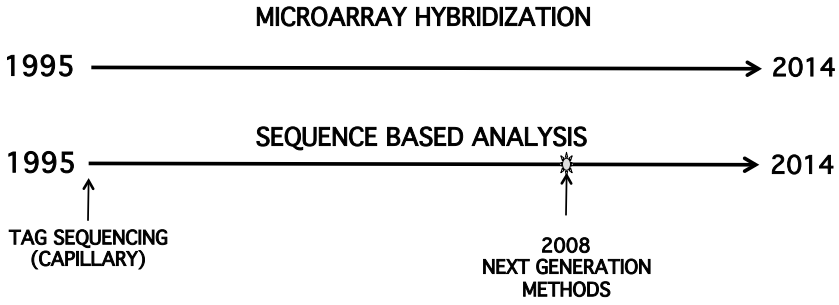
GENE EXPRESSION
GENE VARIATION
GENE FUNCTION

TRANSCRIPTOMICS IS IMPORTANT TO ALL OF THESE

SOME IMPORTANT ISSUES TO CONSIDER

- LARGE DYNAMIC RANGE
- LARGE NUMBER OF GENES
- SAMPLE NUMBER USUALLY MUCH SMALLER THAN GENE NUMBER
- NOT ALL TRANSCRIPTS ARE KNOWN
- ALL TECHNOLOGIES ARE IMPERFECT WITH VARIOUS LIMITATIONS AND IMPERFECTIONS
- ANALYTICAL TOOL DEVELOPMENT LAGS BEHIND DATA GENERATION TECH DEVELOPMENT

A BRIEF HISTORY OF LARGE SCALE TRANSCRIPTOME ANALYSIS



INCREASE IN FEATURE DENSITY

MICROARRAY TERMINOLOGY

- **Feature--an array element**
- **Probe--a feature corresponding to a defined sequence**
- **Target--a pool of nucleic acids of unknown sequence**

POSSIBLE ARRAY FEATURES

- **Synthetic Oligonucleotides**
- **PCR products from**
  - **Cloned DNAs**
  - **Genomic DNA**
- **Cloned DNA**

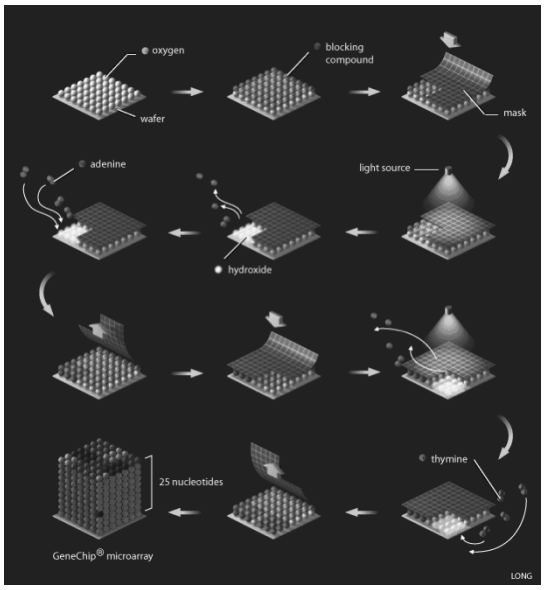
OLIGONUCLEOTIDE ARRAY DESIGN

- **Extremely flexible**
  - **3' bias**
  - **full length**
  - **exon specific**
  - **candidate transcripts**
  - **miRNAs**
- **Very high density possible**
- **Requires sequence data**

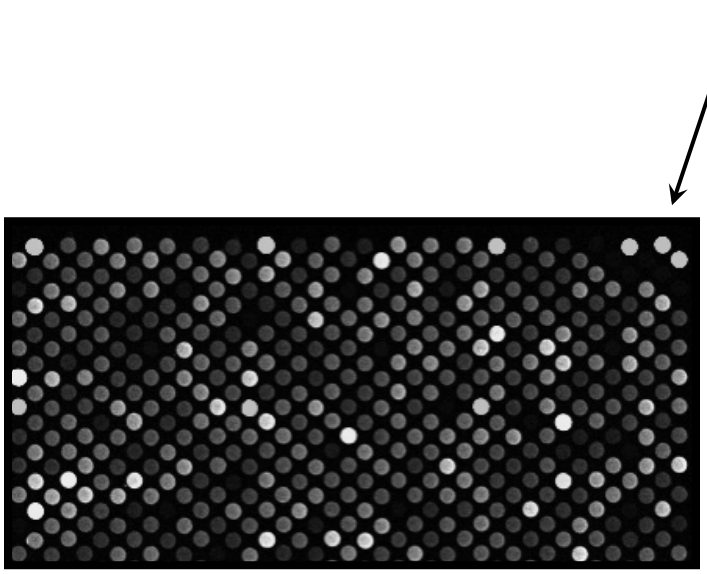
Microarray Manufacture

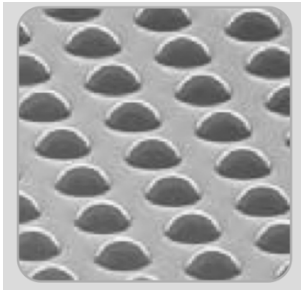
- **Printing**
- **Synthesis *in situ***
  - light directed
  - mechanically directed

LIGHT DIRECTED OLIGONUCLEOTIDE SYNTHESIS

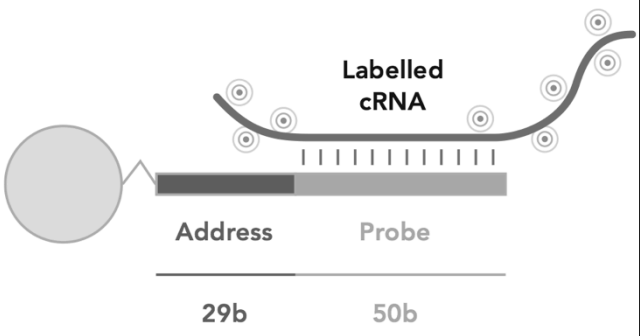


INK JET DIRECTED SYNTHESIS





RANDOMLY POSITIONED HIGH DENSITY
ARRAYS OF ADDRESSABLE OLIGONUCLEOTIDES
COUPLED TO BEADS



MICROARRAY READOUT

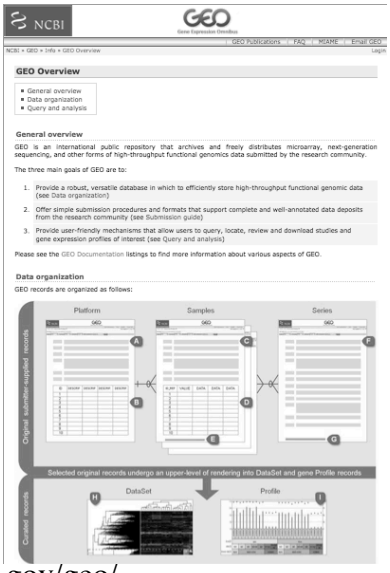
- **Determine quantity of target bound to each probe in a complex hybridization**
- **Must have high sensitivity, low background**
- **High spatial resolution essential**
- **Dual channel capability useful**
- **Fluorescent tags meet these demands**

Laboratory Essentials

- **Arrays**
- **Hybridization and Wash Equipment**
- **Scanner**
- **Software for processing array image**
- **Software for data analysis and display**
- **Bioinformatics collaborator**

Accessing Expression Data

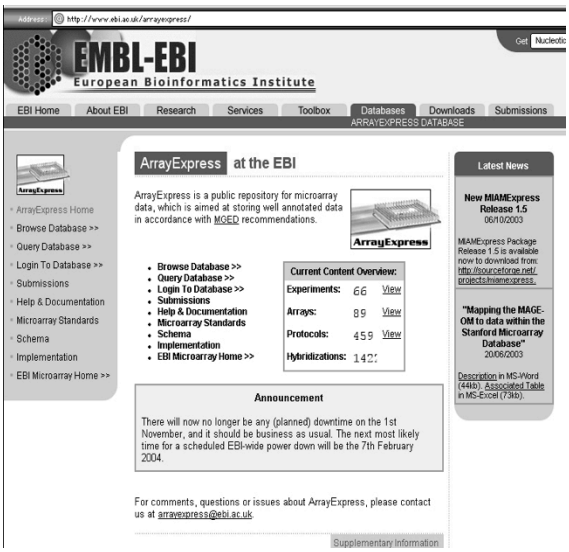
•Individual Lab and Journal Sites; public databases



Currently contains expression data on 1,131,582 samples.

http://www.ncbi.nlm.nih.gov/geo/

Accessing Expression Data



http://www.ebi.ac.uk/microarray-as/ae/

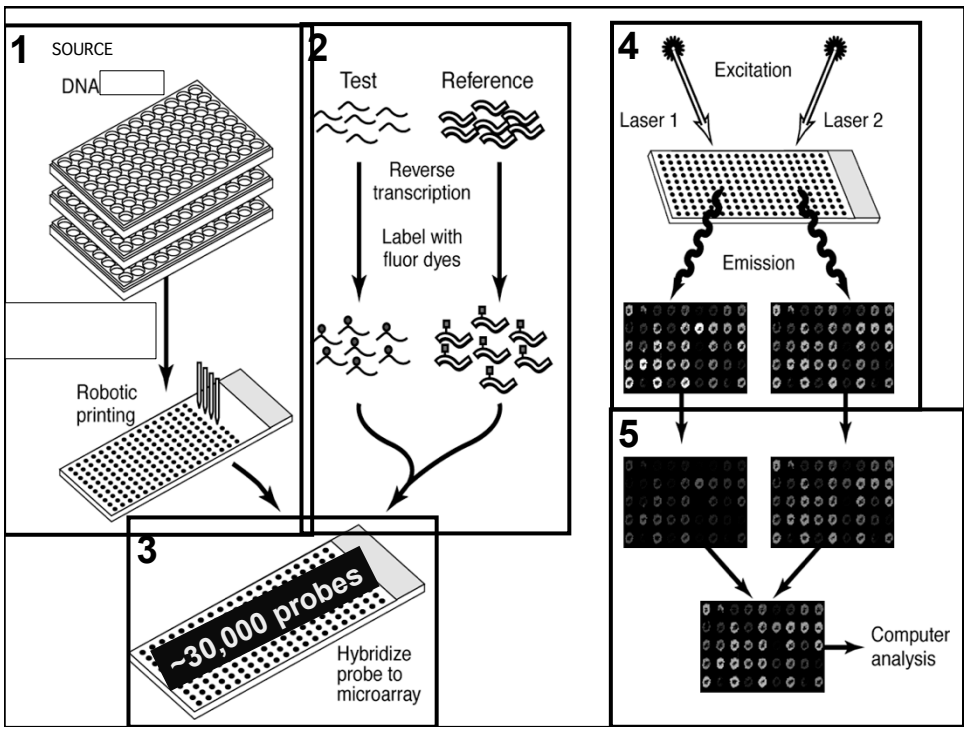
Publishing Expression Data

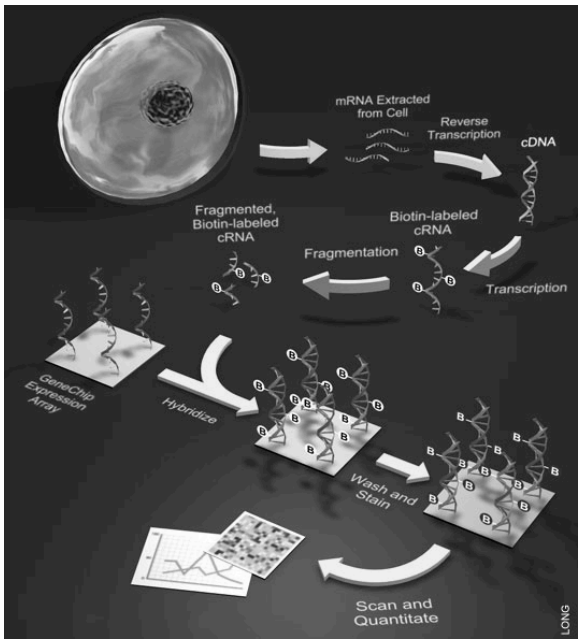
•MIAME standard

Minimum Information about a Microarray Experiment

- Format required by many journals
- Essential for database submissions

http://www.mged.org/Workgroups/MIAME/miame.html





ONE COLOR HYBRIDIZATION ON AN OLIGO ARRAY

Output of Microarray Analysis:

expression ratio
(2 color hybridization)

or

relative expression level
(1 color hybridization)

Both types of data can be analyzed with essentially the same tools.

APPLICATIONS OF EXPRESSION ARRAYS

- **Expression profiling of tissue specimens**
  - **Power arises from increasing sample number**

- **Direct comparisons (Induction, Knockdown etc.)**
  - **Biological system critical**

A RECURRING PROBLEM

Disease Genes

Transcription factors

Hormones/growth factors

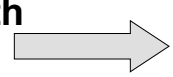
Drugs

Toxins

Infectious agents

Physical agents

siRNA's



?????

Downstream Genes

- **Direct targets**
- **Indirect targets**

EXPRESSION DATA ANALYSIS

•Large amount of data

Examples: 200 samples x 25000 probes= 5,000,000 data points

•Requires analysis and visualization tools

Overview of microarray bioinformatics:
Simon R, Curr Opin Biotechnol. 2008 Feb;19(1):26-9.

EXPRESSION DATA ANALYSIS

- **Check quality of individual experiments**
- **Preprocessing**
  - **Normalization**
  - **Remove genes which are not accurately measured**
  - **Remove genes which are similarly expressed in all samples**
- **Unsupervised Analysis**
- **Supervised Analysis**

Unsupervised Analysis

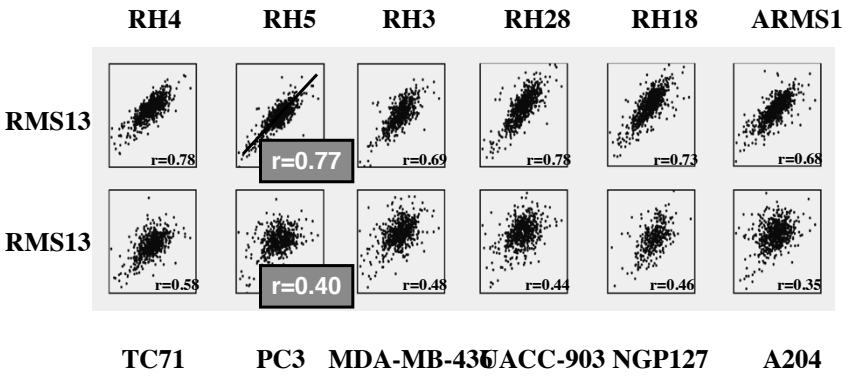
How do genes and samples cluster into groups?

Powerful method of data display.

Does not prove the validity of groups.

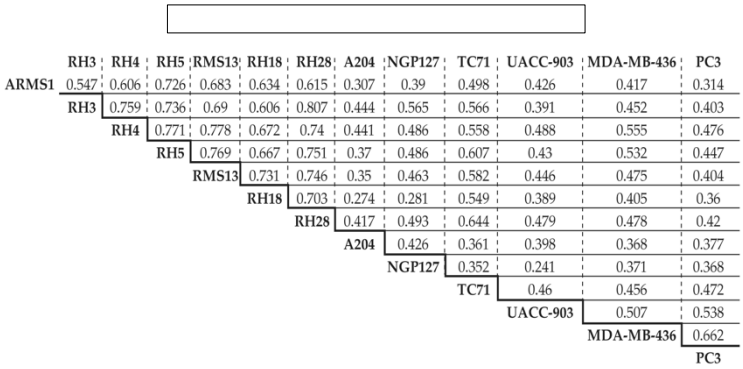
- **Clustered Samples Are Biologically Similar**
- **Clusters of Co-expressed genes**
  - **May be functionally related**
  - **May be enriched for pathways**

UNSUPERVISED CLUSTERING IS BASED ON A GLOBAL SIMILARITY METRIC



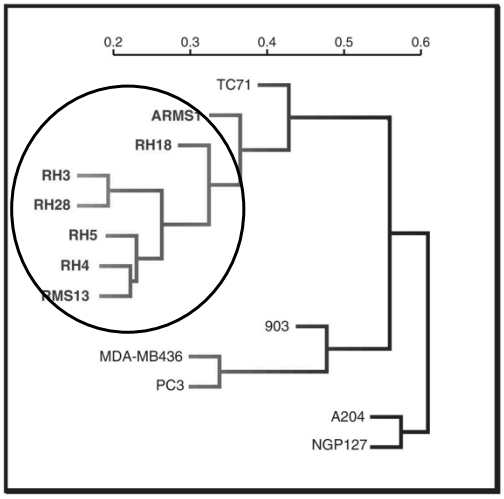
Khan et al. Can Res 58:5009

Matrix of Pearson Correlation Coefficients Distance Map

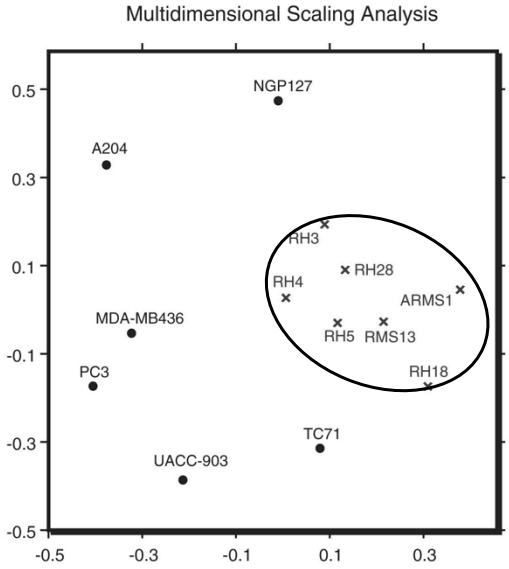


	RH3	RH4	RH5	RMS13	RH18	RH28	A204	NGP127	TC71	UACC-903	MDA-MB-436	PC3
ARMS1	0.547	0.606	0.726	0.683	0.634	0.615	0.307	0.39	0.498	0.426	0.417	0.314
RH3		0.759	0.736	0.69	0.606	0.807	0.444	0.565	0.566	0.391	0.452	0.403
RH4			0.771	0.778	0.672	0.74	0.441	0.486	0.558	0.488	0.555	0.476
RH5				0.769	0.667	0.751	0.37	0.486	0.607	0.43	0.532	0.447
RMS13					0.731	0.746	0.35	0.463	0.582	0.446	0.475	0.404
RH18						0.703	0.274	0.281	0.549	0.389	0.405	0.36
RH28							0.417	0.493	0.644	0.479	0.478	0.42
A204								0.426	0.361	0.398	0.368	0.377
NGP127									0.352	0.241	0.371	0.368
TC71										0.46	0.456	0.472
UACC-903											0.507	0.538
MDA-MB-436												0.662
PC3												

Hierarchical Clustering Dendrogram

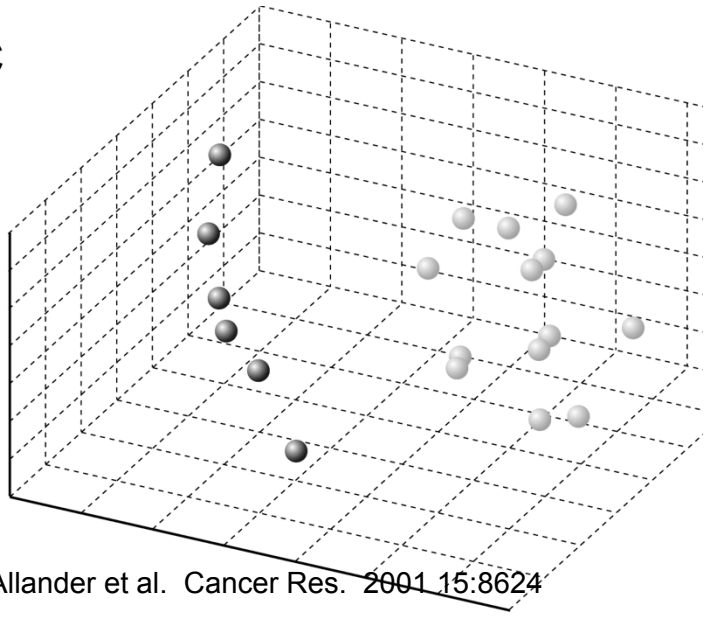


DATA DISPLAY BY MULTIDIMENSIONAL SCALING



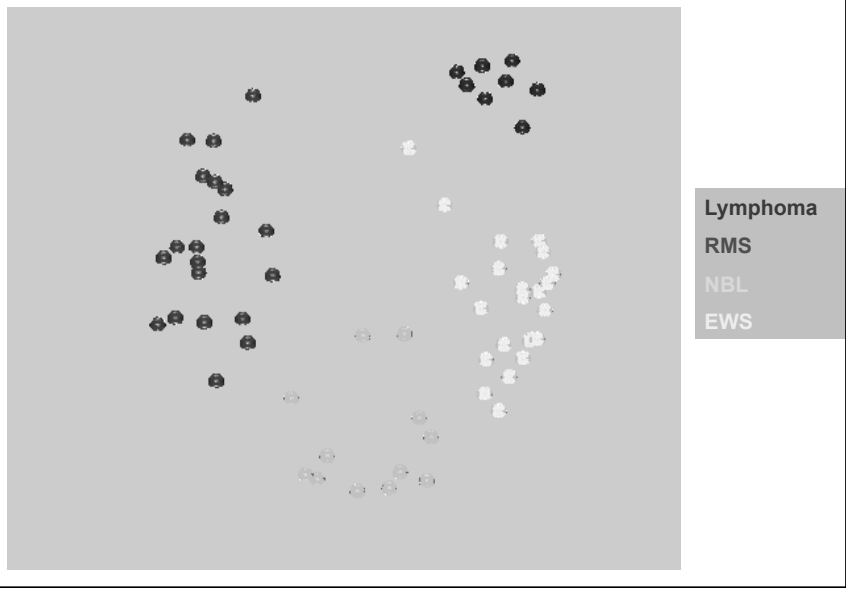
MDS PLOT

C



Allander et al. Cancer Res. 2001 15:8624

MULTIDIMENSIONAL SCALING OR PRINCIPLE COMPONENT ANALYSIS CAPTURE VARIATION IN SAMPLES AND ARE EXCELLENT VISUALIZATION TOOLS

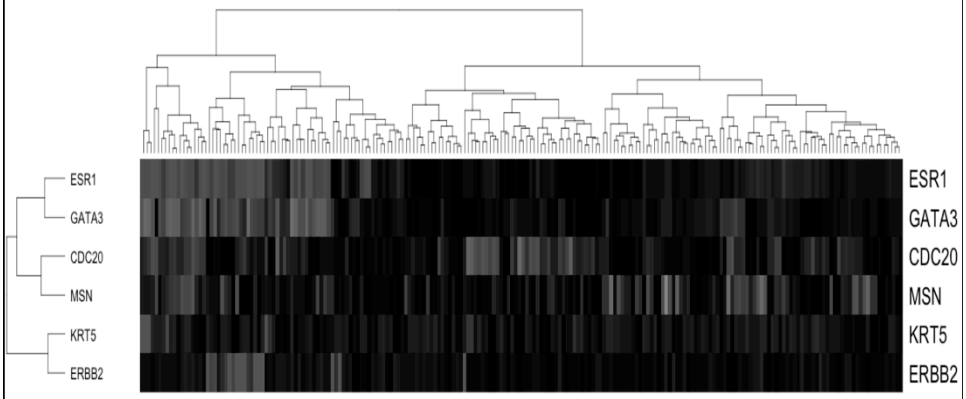


Khan et al
Nat Med
7:673

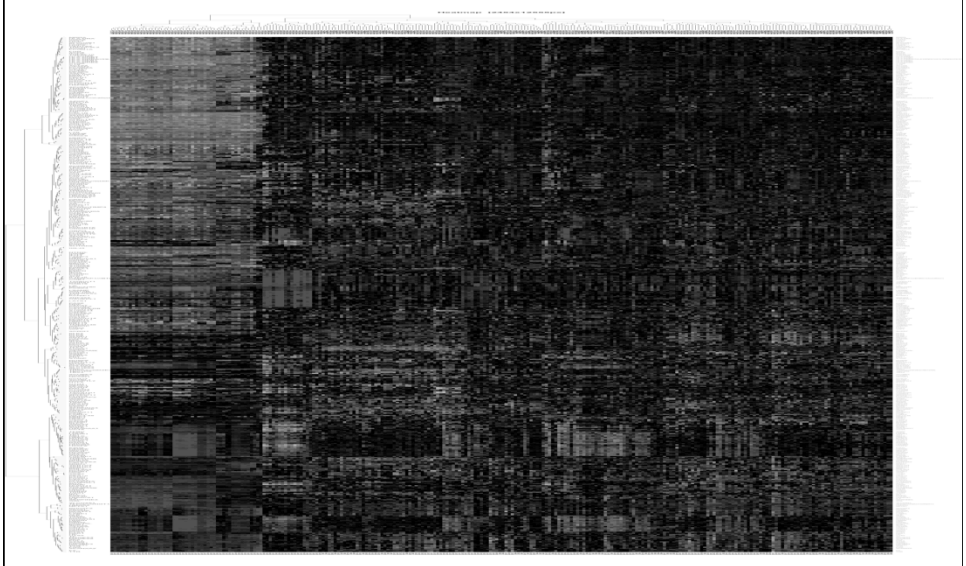
UNCLUSTERED DATA (6 GENES IN 200 SAMPLES)

CLUSTERING GENES ONLY

CLUSTERING GENES AND SAMPLES



CLUSTERING GENES AND SAMPLES



DATA FROM GEO

Supervised Analysis

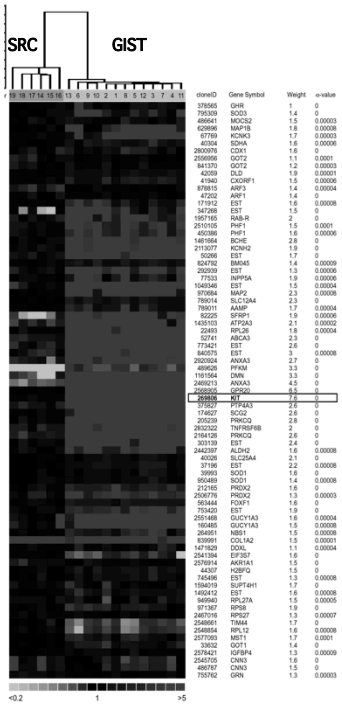
What genes distinguish samples in selected groups from each other?

- **Choice of groups can be based on any known property of the samples.**
- **Many possible underlying methods: t-test or F-statistic frequently used.**
- **Output includes ranked gene list.**
- **Leads to the development of classifiers which can be applied to unknown samples.**
- **Must address the problem of false discovery due to multiple comparisons and discrepancy between sample/gene numbers.**

SUPERVISED METHODS GENERATE RANKED GENE LISTS

TOP DISCRIMINATORS FOR GIST

Rank	Weight	Gene Description
1	7.55575	v-kit sarcoma oncogene
2	6.48306	G coupled receptor 20
3	4.60057	G coupled receptor 20
4	4.51681	annexin A3
5	3.33057	KIAA0353 protein
6	3.31734	phosphofructokinase
7	2.95095	DKFZP434N161 n
8	2.83435	protein kinase C, theta
9	2.79721	butyrylcholinesterase
10	2.72752	annexin A3

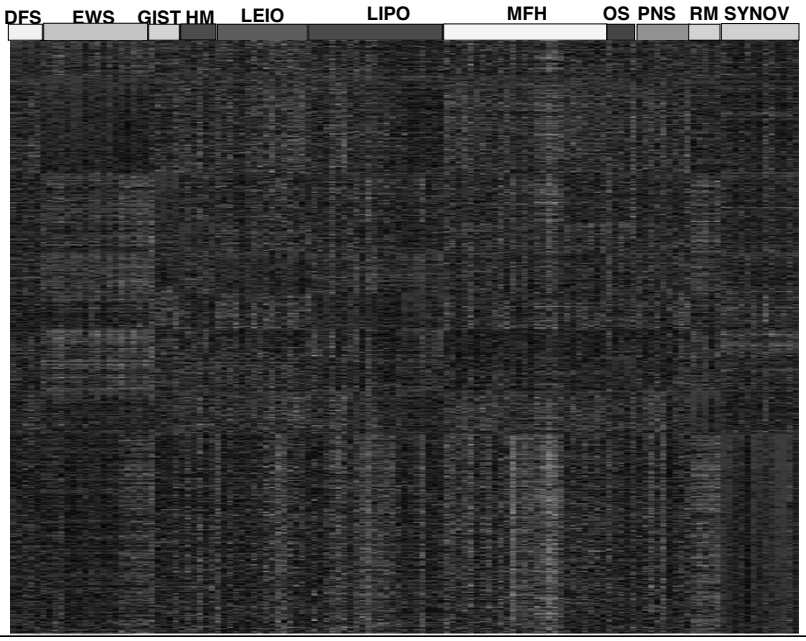


HIERARCHICAL CLUSTERING OF SAMPLES/GENES DISCRIMINATING GASTROINTESTINAL STROMAL TUMOR USING THE GENES SELECTED BY SUPERVISED ANALYSIS

KIT, THE MUTATIONAL AND THERAPEUTIC TARGET IS HIGHLY SIGNIFICANT.

Allander et al. Cancer Res. 2001 15:8624

CHARACTERISTIC PATTERNS OF GENE EXPRESSION IN DIFFERENT SARCOMAS



BAIRD ET AL. CANCER RES 2005

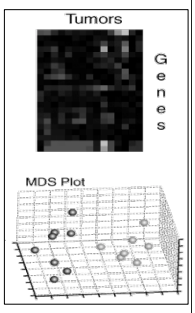
GENOMICS FROM BENCH TO BEDSIDE

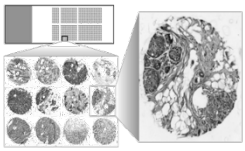
WHOLE GENOME

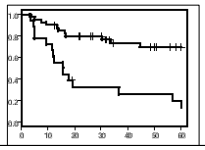
GENE SELECTION

GENE VALIDATION

ASSAY DEVELOPMENT







SIGNAL STRENGTH VARIES IN TISSUE PROFILING EXPERIMENTS

THE MOST INTERESTING QUESTIONS TEND TO BE ASSOCIATED WITH WEAKER SIGNAL.

A CONTINUUM OF POSSIBLE OUTCOMES FROM MICROARRAY RESEARCH

• SOME FEATURES WILL SEPARATE SAMPLES EASILY INTO CLASSES, AND MIGHT BE REDUCED TO SINGLE GENE TESTS, IMPLEMENTED IN A CONVENTIONAL FASHION.

• OTHERS WILL BE MORE DIFFICULT, AND REQUIRE MULTIPLE GENE MEASUREMENTS.

• MANY CLINICALLY RELEVANT FEATURES APPEAR TO FALL WITHIN THIS DIFFICULT GROUP.

A CONTINUUM OF POSSIBLE OUTCOMES FROM MICROARRAY RESEARCH

• SOME GENES WILL SHOW DIFFERENCES BETWEEN GROUPS OF SAMPLES BY CHANCE ALONE.

• THERE MAY BE NO ONE GENE WHICH SEPARATES GROUPS RELIABLY.

• FIND THE MOST INFORMATIVE GENES AND USE THEM IN COMBINATION .

RISK OF OVERFITTING IN CLINICAL STUDIES WITH SMALL SAMPLE SETS

NEED INDEPENDENT VALIDATION SETS.

J Natl Cancer Inst. 2007 Jan 17;99(2):147-57.
Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.
Dupuy A, Simon RM.

BACKGROUND: Both the validity and the reproducibility of microarray-based clinical research have been challenged. There is a need for critical review of the statistical analysis and reporting in published microarray studies that focus on cancer-related clinical outcomes. METHODS: Studies published through 2004 in which microarray-based gene expression profiles were analyzed for their relation to a clinical cancer outcome were identified through a Medline search followed by hand screening of abstracts and full text articles. Studies that were eligible for our analysis addressed one or more outcomes that were either an event occurring during follow-up, such as death or relapse, or a therapeutic response. We recorded descriptive characteristics for all the selected studies. A critical review of outcome-related statistical analyses was undertaken for the articles published in 2004. RESULTS: Ninety studies were identified, and their descriptive characteristics are presented. Sixty-eight (76%) were published in journals of impact factor greater than 6. A detailed account of the 42 studies (47%) published in 2004 is reported. Twenty-one (50%) of them contained at least one of the following three basic flaws: 1) in outcome-related gene finding, an unstated, unclear, or inadequate control for multiple testing; 2) in class discovery, a spurious claim of correlation between clusters and clinical outcome, made after clustering samples using a selection of outcome-related differentially expressed genes; or 3) in supervised prediction, a biased estimation of the prediction accuracy through an incorrect cross-validation procedure. CONCLUSIONS: The most common and serious mistakes and misunderstandings recorded in published studies are described and illustrated. Based on this analysis, a proposal of guidelines for statistical analysis and reporting for clinical microarray studies, presented as a checklist of "Do's and Don'ts," is provided.

MICROARRAY STUDIES GENERATE ORGANIZED LIST OF GENES

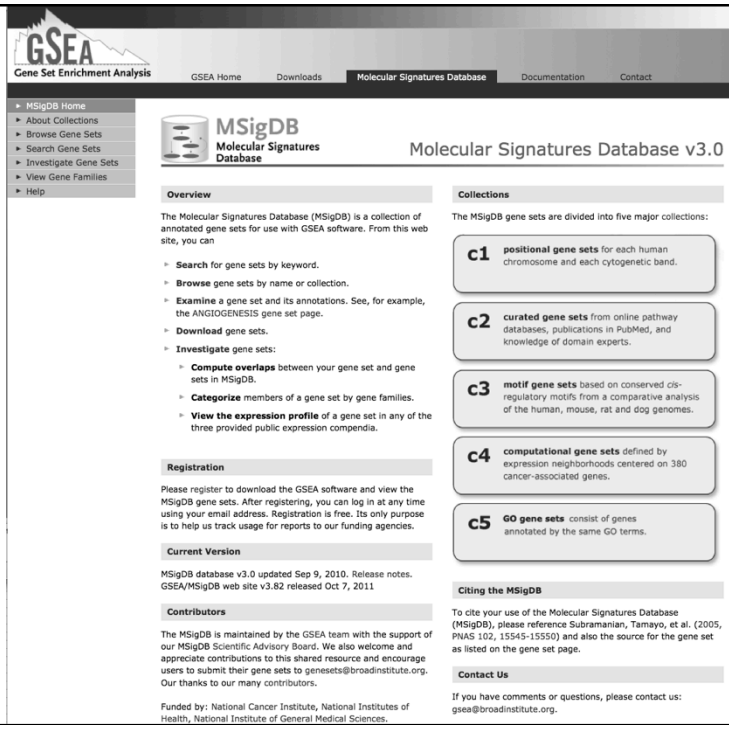
- **Often cryptic and hard to interpret.**
- **Hypothesis generating, but this is often rather subjective.**
- **Seldom provide strong evidence for a specific mechanism.**
- **Expression data is intrinsically limited.**

GETTING BEYOND GENE LISTS

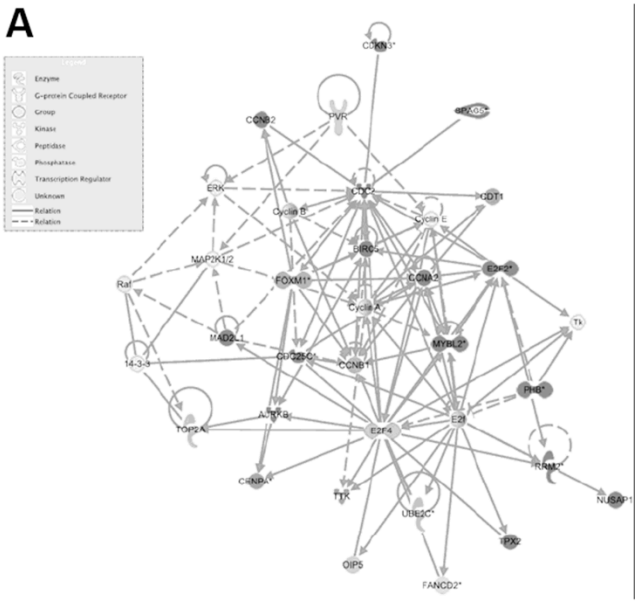
- **Optimal use of gene annotations.**
  - **Gene Ontology**
    (http://david.abcc.ncifcrf.gov/)
- **Optimizing use of public data.**

.

- **GENE SIGNATURE BASED METHODS (Gene Set Enrichment Analysis).**



Pathway Analysis



Balleine et al. 14:8244 Clin Can Res 2008

WHAT TO LOOK FOR IN CLINICAL CORRELATIVE STUDIES USING MICROARRAYS

- WELL DEFINED QUESTION AND PATIENT SAMPLE.
- HIGH QUALITY ARRAY MEASUREMENTS (HARD TO ASSESS WITHOUT REFERENCE TO PRIMARY DATA---SHOULD BE MADE PUBLIC).
- APPROPRIATE AND RIGOROUS STATISTICAL ANALYSIS OF ARRAY DATA.
- FORMAL CLASSIFIER THAT CAN BE APPLIED TO NEW SAMPLES.
- VALIDATION SAMPLE SET.

WHAT TO LOOK FOR IN CLINICAL
CORRELATIVE STUDIES
USING MICROARRAYS

• GOAL SHOULD BE TO SEEK AND VALIDATE CLINICALLY RELEVANT SIGNATURES WITHIN DEFINED PATIENT GROUPS FOR WHICH NO CURRENT FEATURES ADEQUATELY ANSWER THE CLINICAL QUESTION POSED.

EXPRESSION PROFILING IN THE CLINIC?

PROBLEMS:

- **SPECIALIZED TECHNOLOGY**
- **RNA IS UNSTABLE**
- **FROZEN TISSUE NOT PART OF USUAL OR SAMPLE FLOW**

EXPRESSION PROFILING IN THE CLINIC?

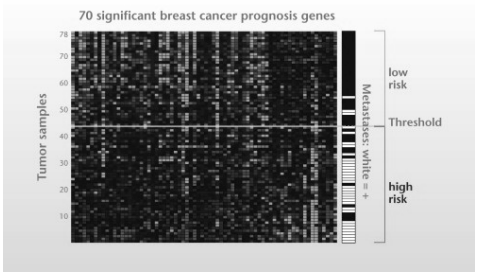
OPTIONS:

- **REFERENCE LABORATORIES**
- **RNA PRESERVATIVES**
- **USE OF PARAFFIN EMBEDDED MATERIALS.**

•USE ARRAYS FOR DISCOVERY TO EXTRACT SIGNATURES WHICH CAN BE ASSAYED WITH ALTERNATIVE TECHNOLOGIES.

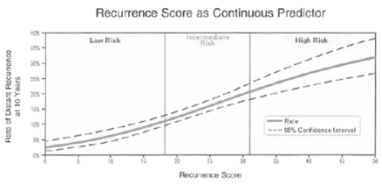
FDA APPROVED TESTS FOR BREAST CANCER BASED ON EXPRESSION STUDIES

70 GENE MICROARRAY SIGNATURE



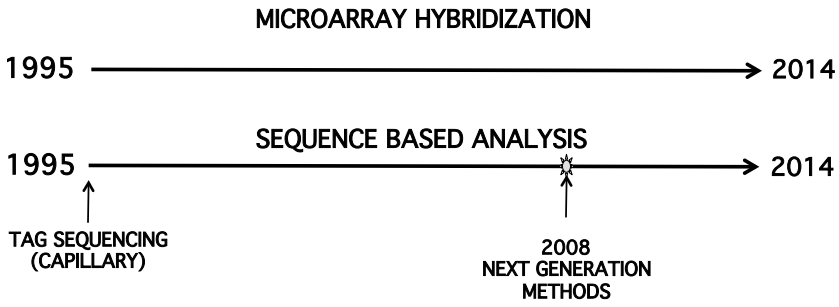
Van de Vijver et al
NEJM 347:1999 .

Muiltgene RT-PCR SIgnature

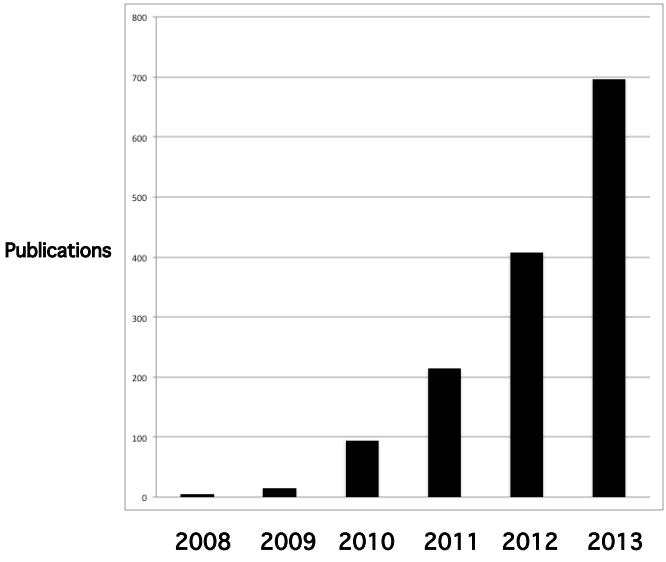


Paik et al NEJM 351:2817

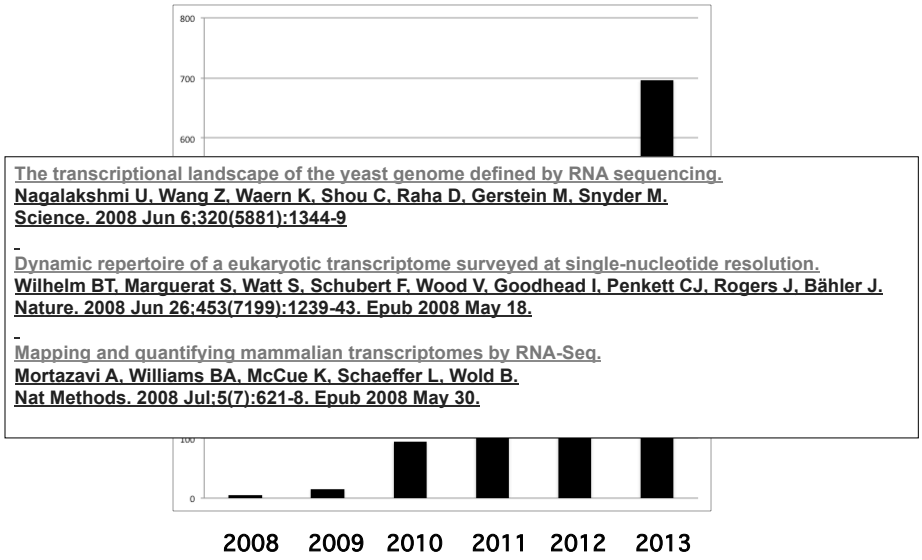
A BRIEF HISTORY OF LARGE SCALE TRANSCRIPTOME ANALYSIS



PubMed Citations for RNA-Seq



PubMed Citations for RNA-Seq



The transcriptional landscape of the yeast genome defined by RNA sequencing.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M.
Science. 2008 Jun 6;320(5881):1344-9

Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.
Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J.
Nature. 2008 Jun 26;453(7199):1239-43. Epub 2008 May 18.

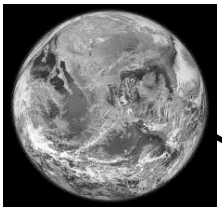
Mapping and quantifying mammalian transcriptomes by RNA-Seq.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.
Nat Methods. 2008 Jul;5(7):621-8. Epub 2008 May 30.

ARRAYS VS. NEXT GENERATION SEQUENCING

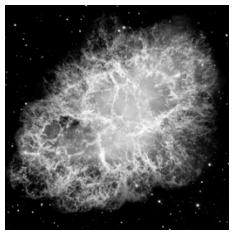
• ARRAY TECHNOLOGIES MEASURE THE RELATIVE ABUNDANCE OF NUCLEIC ACIDS OF DEFINED SEQUENCE IN A COMPLEX MIXTURE.

- **SEQUENCING CAN ACCOMPLISH THE SAME THING.**

When Worlds Collide



Microarray Expression Profiling



- MANY NEW OPPORTUNITIES
- MANY NEW CHALLENGES

ARRAYS VS. NEXT GENERATION SEQUENCING

MICROARRAYS		SEQUENCING
• READILY AVAILABLE MATURE TECHNOLOGY • RELATIVELY INEXPENSIVE • EFFECTIVE WITH VERY COMPLEX SAMPLES • HUNDREDS OF SAMPLES PRACTICAL • CAN TARGET SUBSET OF GENOME	PROS	• WHOLE GENOME DATA • RELATIVELY UNIFORM ANALYTICAL PIPELINE • FREE OF HYBRIDIZATION ARTIFACTS • LARGE DYNAMIC RANGE • POSSIBILITY OF ONE PLATFORM FOR ALL APPLICATIONS
MICROARRAYS	CONS	SEQUENCING
• REQUIRE PLATFORM AND APPLICATION SPECIFIC DATA PROCESSING • PRONE TO PLATFORM SPECIFIC ARTIFACTS • LIMITED DYNAMIC RANGE • SOME PROBES PERFORM POORLY • MANY SOURCES OF NOISE • WHOLE GENOME STUDIES MAY REQUIRE MANY ARRAYS, INCREASING SAMPLE REQUIREMENTS AND COMPLICATING ANALYSIS		• EVOLVING TECHNOLOGY • TECHNOLOGY SPECIFIC ARTIFACTS • RESOURCE INTENSIVE • COMPUTATIONALLY INTENSIVE • NO STANDARD ANALYSIS YET • LOWER SAMPLE THROUGHPUT

MEASURING GENE EXPRESSION BY RNA SEQUENCING

ADVANTAGES

- RNA SEQUENCE VARIATIONS DETECTED AT SINGLE NUCLEOTIDE RESOLUTION
  - -ALLELE SPECIFIC EXPRESSION
  - -MUTATIONS
  - -RNA EDITING
- RNA STRUCTURE: SPLICING, START SITES, TERMINATION SITES; REARRANGEMENTS
- DETECTED SIGNALS ARE RELATIVELY UNAMBIGUOUS; POTENTIAL TO OUTPERFORM MICROARRAY
- TRANSCRIPT DISCOVERY

MEASURING GENE EXPRESSION BY RNA SEQUENCING

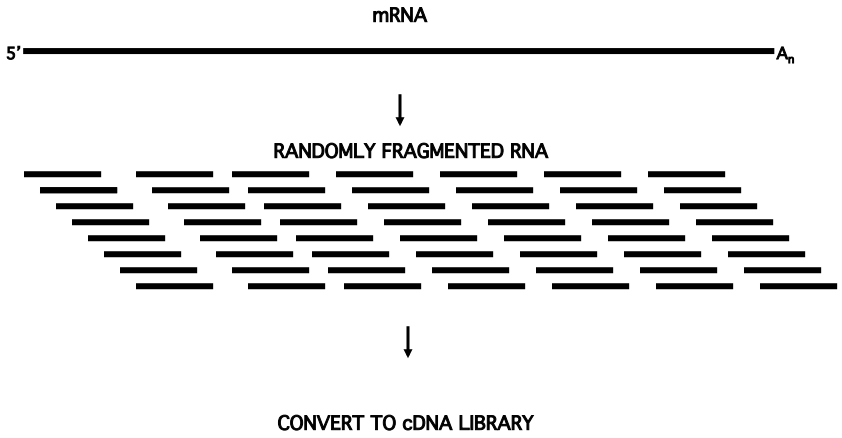
- FULL LENGTH mRNA----RNA-Seq
- TAG SEQUENCING (SAGE-LIKE)
- PolyA vs. Total (ribosomal depleted)
- Strand specific vs. non-strand specific
- miRNA
- lincRNA

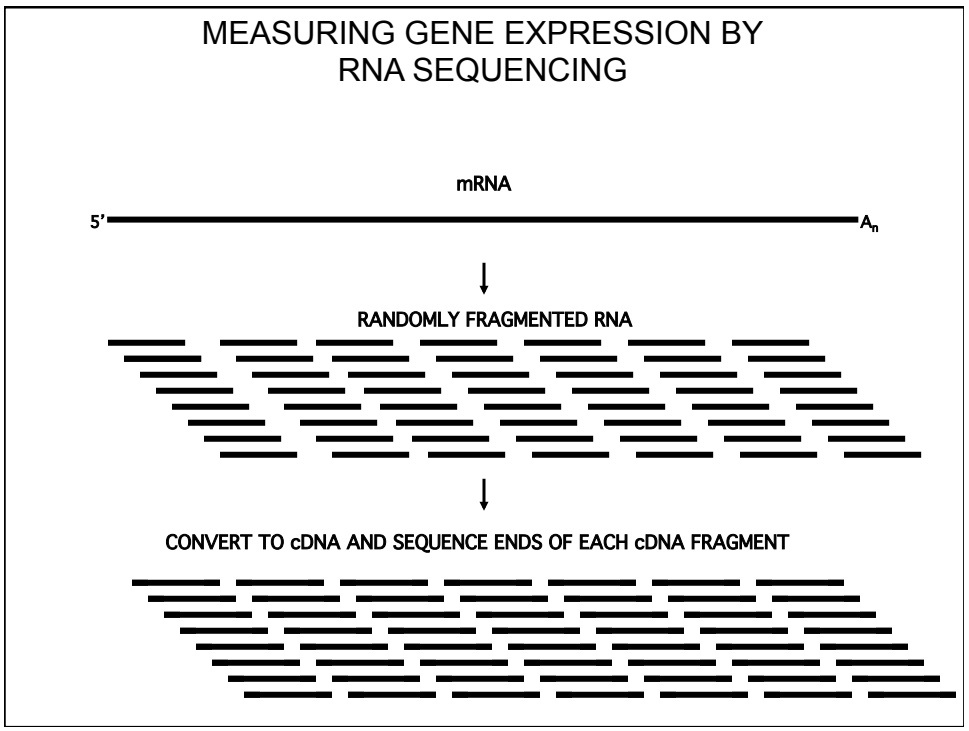
MEASURING GENE EXPRESSION BY RNA SEQUENCING: PROS AND CONS

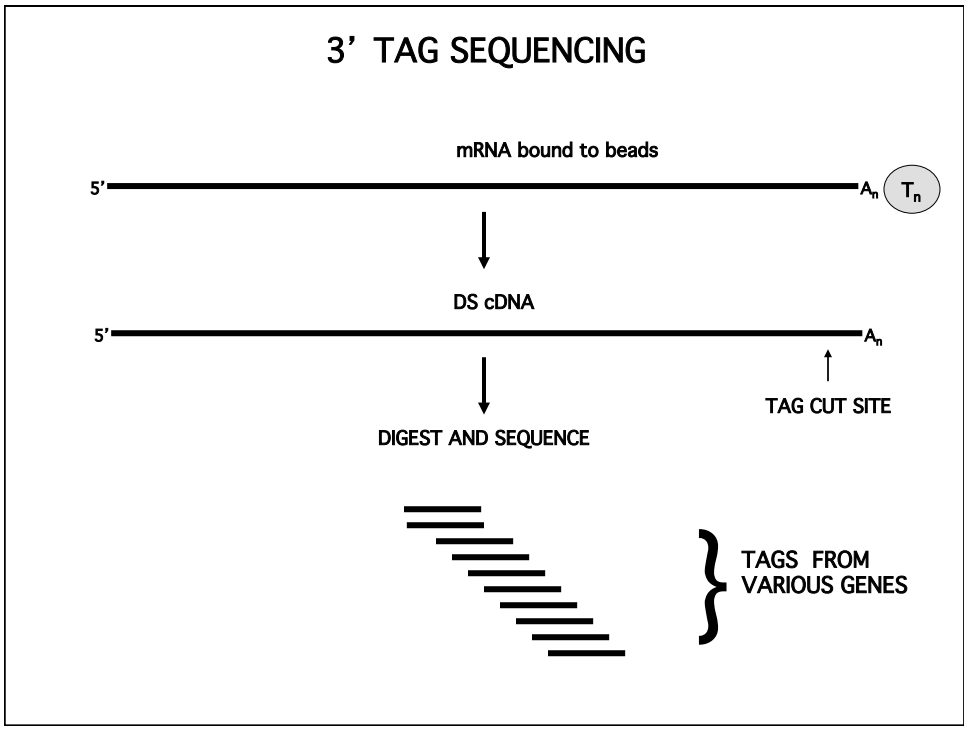
LIMITATIONS

- LOWER LIMIT OF DETECTION IS CONSTRAINED BY THE mRNA ABUNDANCE DISTRIBUTION AND THE NUMBER OF ALIGNED READS PER SAMPLE.
- LARGE SAMPLE NUMBERS DIFFICULT TO PROCESS WITHOUT AUTOMATION.
- SOFTWARE IS STILL EVOLVING: REQUIRES SOPHISTICATED BIOINFORMATICS COLLABORATION.
- COMPUTATIONAL HARDWARE REQUIREMENTS ARE SUBSTANTIAL
- LIBRARY PREP METHODS EVOLVING.
- DATA COMPARISON PROBLEMATIC IF METHODS DIFFER.

MEASURING GENE EXPRESSION BY RNA SEQUENCING




MEASURING GENE EXPRESSION BY
RNA SEQUENCING
mRNA
5'
An
RANDOMLY FRAGMENTED RNA
CONVERT TO cDNA AND SEQUENCE ENDS OF EACH cDNA FRAGMENT


3' TAG SEQUENCING
mRNA bound to beads
5'
An
Tn
DS cDNA
5'
An
TAG CUT SITE
DIGEST AND SEQUENCE
TAGS FROM
VARIOUS GENES

3' TAG SEQUENCING

- SEQUENCES ALIGNED AND COUNTED
- LIBRARIES OF TAGS FROM MANY SAMPLES CAN BE IDENTIFIED BY ADDING A "BARCODE" AND POOLED BEFORE SEQUENCING
- POTENTIAL TO ANALYZE LARGE NUMBERS OF SAMPLES IN PARALLEL

RNASeq COMPUTATIONAL WORKFLOW

RAW READS

⇩

ALIGN

RNASeq COMPUTATIONAL WORKFLOW

RAW READS

⇩

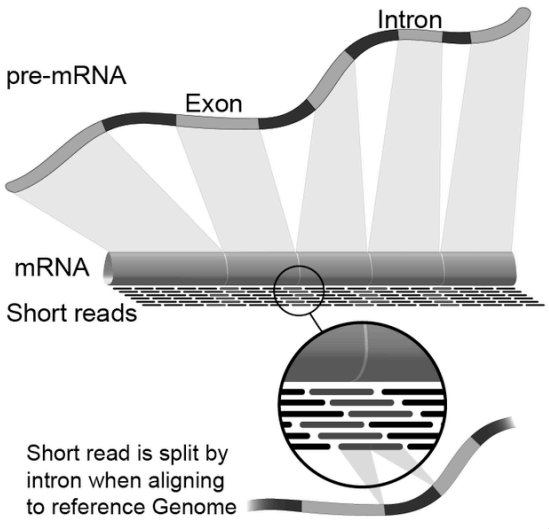
ALIGN

TO WHAT???

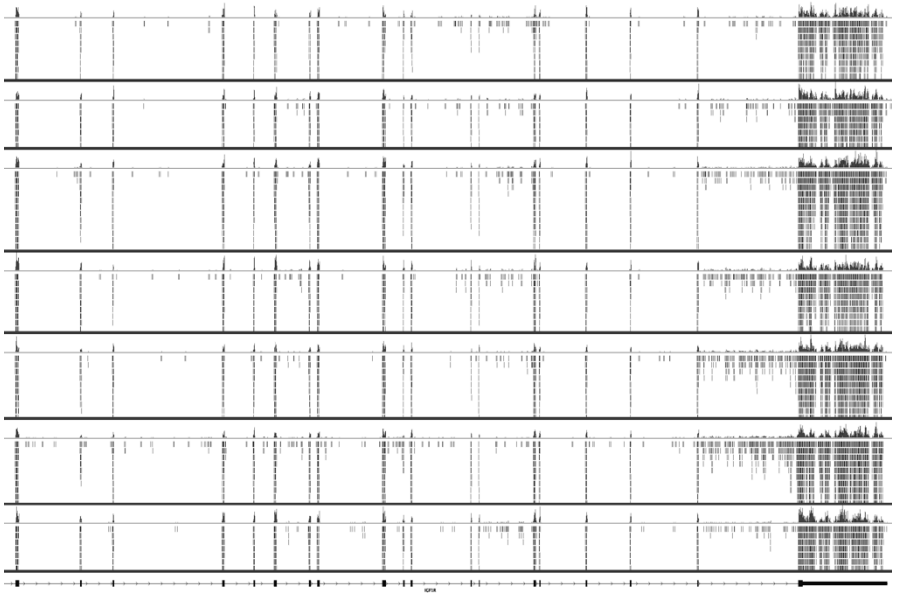
- **GENOME?**
- **TRANSCRIPTOME?**
- **WHICH TRANSCRIPTOME? REFSEQ? EMBL? CUSTOM?**

WHAT ALIGNER????

RNA-seq Alignment



RNA-seq IGF1R



RNASeq COMPUTATIONAL WORKFLOW

RAW READS

⇩

ALIGN TO TRANSCRIPTOME

⇩

ALIGNED READS

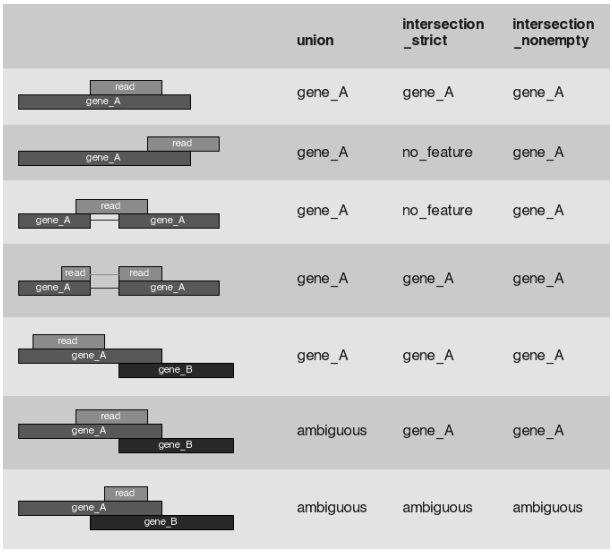
⇩

NORMALIZED READ COUNT

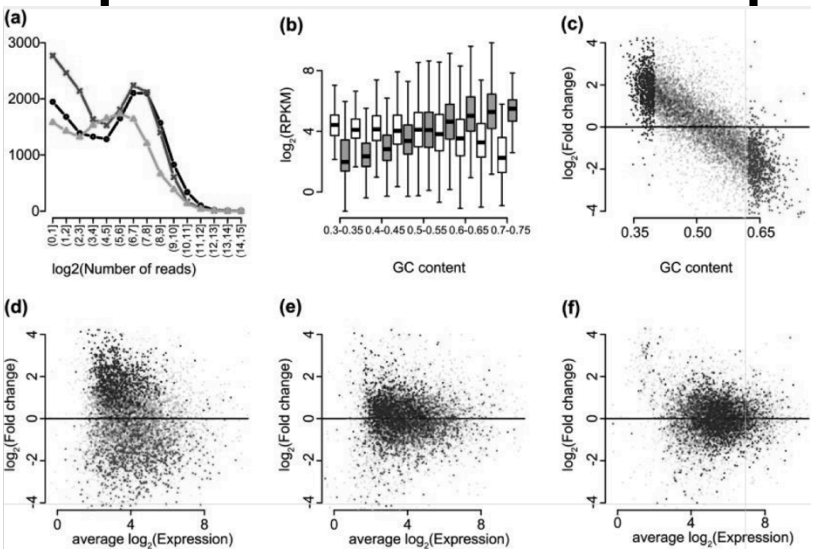
Models for RNA-seq

- Count-based models
- Multi-reads (isoform resolution)
- Paired-end reads (include length resolution step)
- Positional bias along transcript length
- Sequence bias

Read Counting



Sequence Bias in RNA-Seq



Hansen et al. Biostatistics 13:204-216 (2012)

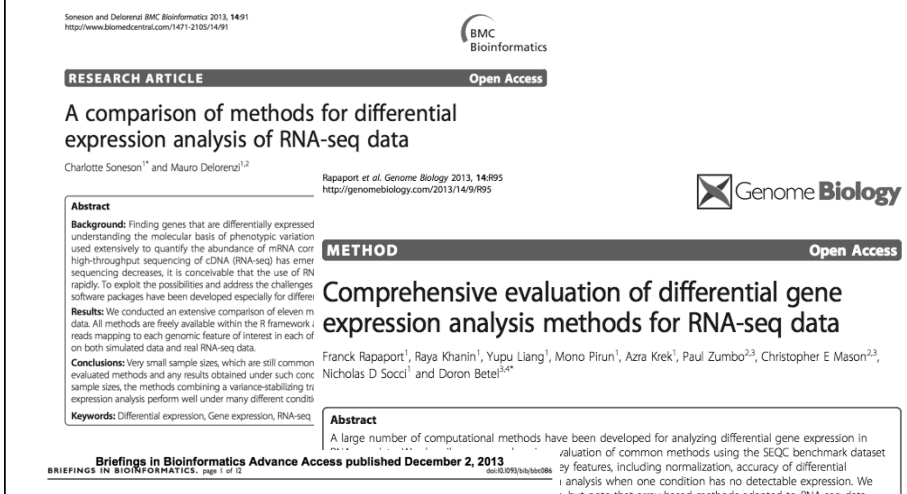
Result of Quantification

CountTable.txt

	A	B	C	D	E	F	G
1	Ensembl.Gene	HGNC.symbol	length	adipose	adrenal	brain	breast
2	ENSG00000000	TSPAN6	5150	6.5102076	4.4399499	4.47541785	6.19290729
3	ENSG00000000	TNMD	1881	4.77434788	3.03741631	-2.30427166	6.20206438
4	ENSG00000000	DPM1	7899	4.67436267	5.14359648	5.0350925	5.03635496
5	ENSG00000000	SCYL3	12929	2.61653185	3.44062654	3.04429556	3.19008304
6	ENSG00000000	C1orf112	21973	2.23079247	2.2412008	1.59208827	2.33934858
7	ENSG00000000	FGR	15989	5.13218607	4.64955569	1.82875376	3.22232919
8	ENSG00000000	CFH	17278	7.31413745	7.75995914	5.29604675	7.69713191
9	ENSG00000000	FUCA2	4576	5.99712898	5.74841489	5.01716452	6.08121442
10	ENSG00000000	GCLC	17290	5.41823808	5.95937652	6.49950722	6.56422871
11	ENSG00000000	NFYA	5471	4.59377563	4.23783092	4.78376968	4.50642521
12	ENSG00000000	C1orf201	25360	2.59820197	2.71831835	4.21443513	3.15709964
13	ENSG00000000	NIPAL3	15545	4.40544708	4.85233419	7.83920432	5.08638068
14	ENSG00000000	LAS1L	15066	5.11482973	5.55638125	5.0521603	4.73990806
15	ENSG00000000	ENPP4	9173	5.97080384	5.20063279	7.33245181	6.00612757
16	ENSG00000000	SEMA3F	15469	6.47286323	4.83802467	2.34795406	3.82755053
17	ENSG00000000	CFTR	18020	-1.01862313	-8.20277874	0.06918673	-4.77202424
18	ENSG00000000	ANKIB1	11241	5.44434018	5.60447554	6.78072994	6.04619478
19	ENSG00000000	CYP51A1	9653	1.01631067	-0.13131638	1.28723368	0.27236988
20	ENSG00000000	KRIT1	39656	4.59068842	5.25652479	5.09598679	4.93862567
21	ENSG00000000	RAD52	28855	2.36874662	3.83785346	2.80235076	3.13103594
22	ENSG00000000	MYH16	11845	-6.69104847	-3.34479775	-3.44177519	-3.98352835
23	ENSG00000000	BAD	5112	2.99253103	3.126457	2.6329015	3.44052741
24	ENSG00000000	LAP3	8517	6.71111978	7.70371262	6.75735727	6.77254509
25	ENSG00000000	CD99	6647	8.30711588	8.62102385	5.96837933	8.5496861

Differential Gene Expression

- A LARGE NUMBER OF VARIABLES INTRINSIC TO RNA-Seq ACCOMPANY THE DATA.
- THESE POSE A NEW SET OF COMPUTATIONAL PROBLEMS WHICH DIFFER SUBSTANTIALLY FROM THOSE ENCOUNTERED IN THE ANALYSIS OF MICROARRAY DATA.



Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

BMC Bioinformatics

RESEARCH ARTICLE Open Access

A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson[1*] and Mauro Delorenzi[1,2]

Abstract

Background: Finding genes that are differentially expressed
understanding the molecular basis of phenotypic variation
used extensively to quantify the abundance of mRNA corr
high-throughput sequencing of cDNA (RNA-seq) has emer
sequencing decreases, it is conceivable that the use of RN
rapidly. To exploit the possibilities and address the challenges
software packages have been developed especially for differe

Results: We conducted an extensive comparison of eleven m
data. All methods are freely available within the R framework
reads mapping to each genomic feature of interest in each of
on both simulated data and real RNA-seq data.

Conclusions: Very small sample sizes, which are still common
evaluated methods and any results obtained under such conc
sample sizes, the methods combining a variance-stabilizing tra
expression analysis perform well under many different conditi

Keywords: Differential expression, Gene expression, RNA-seq

Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

Genome Biology

METHOD Open Access

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Abstract

A large number of computational methods have been developed for analyzing differential gene expression in
valuation of common methods using the SEQC benchmark dataset
ey features, including normalization, accuracy of differential
analysis when one condition has no detectable expression. We
, but note that array-based methods adapted to RNA-seq data
RNA-seq. Our results demonstrate that increasing the number of
on power over increased sequencing depth.

Briefings in Bioinformatics Advance Access published December 2, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 12 doi:10.1093/bib/bbt086

Comparison of software packages for detecting differential expression in RNA-seq studies

Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

Submitted: 20th August 2013; Received (in revised form): 9th October 2013

PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

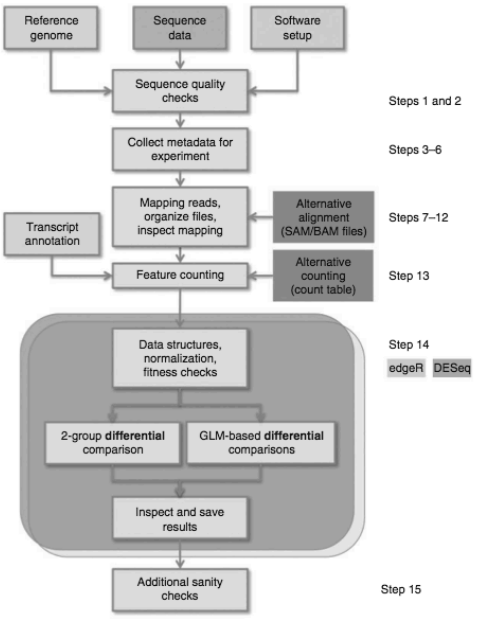
Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

[1]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Department of Statistics, University of Oxford, Oxford, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, Victoria, Australia. [5]Department of Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. [6]Functional Genomics Center UNI ETH, Zurich, Switzerland. [7]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia. [8]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [9]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. Correspondence should be addressed to M.D.R. (mark.robinson@imls.uzh.ch) or W.H. (whuber@embl.de).

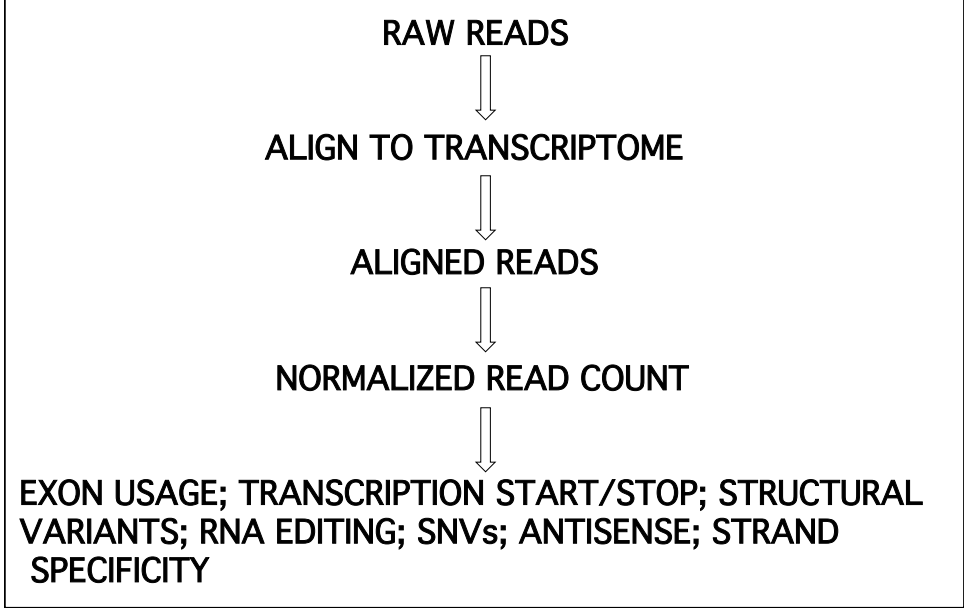
Published online 22 August 2013; doi:10.1038/nprot.2013.099

RNA sequencing (RNA-seq) has been rapidly adopted for the profiling of transcriptomes in many areas of biology, including studies into gene regulation, development and disease. Of particular interest is the discovery of differentially expressed genes across different conditions (e.g., tissues, perturbations) while optionally adjusting for other systematic factors that affect the data-collection process. There are a number of subtle yet crucial aspects of these analyses, such as read counting, appropriate treatment of biological variability, quality control checks and appropriate setup of statistical modeling. Several variations have been presented in the literature, and there is a need for guidance on current best practices. This protocol presents a state-of-the-art computational and statistical RNA-seq differential expression analysis workflow largely based on the free open-source R language and Bioconductor software and, in particular, on two widely used tools, DESeq and edgeR. Hands-on time for typical small experiments (e.g., 4–10 samples) can be <1 h, with computation time <1 d using a standard desktop PC.

RNA-seq workflow as proposed by Anders et al. in Nature Protocols

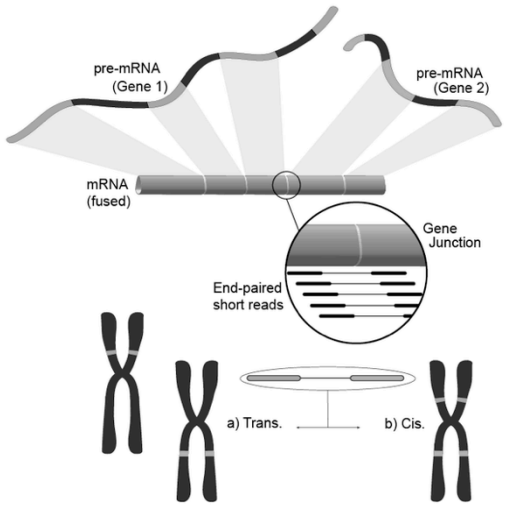


RNASeq COMPUTATIONAL WORKFLOW



Fusion Gene Detection

Fusion gene schematic



Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 340620, 6 pages
http://dx.doi.org/10.1155/2013/340620

Hindawi

Research Article

State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity

Matteo Carrara,[1] Marco Beccuti,[2] Fulvio Lazzarato,[3] Federica Cavallo,[1] Francesca Cordero,[2] Susanna Donatelli,[2] and Raffaele A. Calogero[1]

[1] *Department of Molecular Biotechnology and Health Sciences, University of Torino, Via Nizza 52, 10126 Torino, Italy*
[2] *Department of Computer Science, University of Torino, C.So Svizzera 185, 10149 Torino, Italy*
[3] *Unit of Cancer Epidemiology, Department of Biomedical Sciences and Human Oncology, University of Torino, 10126 Torino, Italy*

Correspondence should be addressed to Raffaele A. Calogero; raffaele.calogero@unito.it

Received 4 October 2012; Revised 11 January 2013; Accepted 15 January 2013

Fusion Detection

TABLE 1: Filtering steps embedded in the algorithms.

Filters	Fusion finders							
	FF	THF	MS	FM	FH	DF	BF	CS
Pair distance	X					X	X	X
Anchor length		X			X			X
Read-through	X	X		X	X		X	
Junction-spanning				X	X		X	
PCR artifact				X	X		X	
Homology	X	X					X	
Quality			X	X				

FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH: FusionHunter; DF: deFuse; BF: Bellerophontes; CS: ChimeraScan.

False Positive Fusion Detection

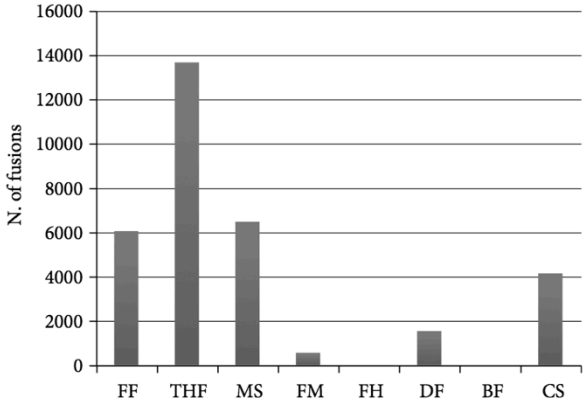


FIGURE 4: *False positive fusion detected using a synthetic dataset without chimeras.* FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH: FusionHunter; DF: defuse; BF: Bellerophontes; CS: ChimeraScan.

WHEN IS RNA-Seq PREFERRED?

IT DEPENDS ON THE EXPERIMENTAL GOALS.

CURRENTLY

- RNA-Seq IS THE PREFERRED METHOD FOR ASSESSING TRANSCRIPT STRUCTURE AND SEQUENCE VARIATION AT GENOME SCALE.

- ROLE OF RNA-Seq FOR ROUTINE COUNT BASED EXPRESSION ANALYSIS IS LESS CLEAR AT THIS TIME.

- AS SEQUENCE THROUGHPUT INCREASES, COSTS DECLINE, AND AS STANDARDIZED ANALYTICAL PIPELINES FOR SPECIFIC EXPERIMENTAL GOALS ARE DEVELOPED, RNA-Seq WILL BECOME INCREASINGLY ATTRACTIVE FOR GENERAL USE.

USEFUL WEB SITES

MGEGD The Microarray Gene Expression Data Society:

http//www.mged.org/

NCBI Gene Expression Omnibus:

http://ncbi.nih.gov/geo/

NCBI Sequence Read Archive (SRA):

http://www.ncbi.nlm.nih.gov/sra

EBI Microarray informatics:

http://www.ebi.ac.uk/microarray/index.html

Stanford Microarray Database:

http://smd.stanford.edu/

UCSF DeRisi lab:

http://derisilab.ucsf.edu/data/microarray/index.html

Broad Institute:

Gene Set Enrichment Analysis (GSEA)
http://www.broadinstitute.org/gsea/

Connectivity Map:

http://www.broadinstitute.org/cmap/

Bioconductor

http://bioconductor.org

Biostars

http://biostars.org

RNA-seq blog

http://www.rna-seqblog.com/