RNA-Seq analysis in Galaxy

Advanced and alternative tools

Y. Hoogstrate[1] S. Hiltemann[1] H. Mei[2]

[1]Department of Bioinformatics & Department of Urology
ErasmusMC, Rotterdam

[2]Leiden University Medical Centre, Leiden

Galaxy community conference, 2014

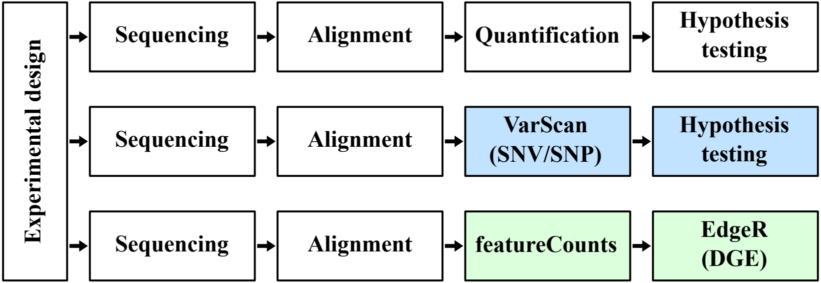
Overview

Introduction
- RNA-Seq

Single Nucleotide Variants (SNV)
- Introduction
- Tools
  - VarScan2
- Hands on

Differential Gene Expression (DGE) analysis
- Quantification
- Hypothesis testing
  - Sample pairing

RNA-seq analysis workflow(s)

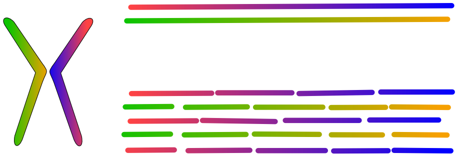


Single Nucleotide Variants

- 62,676,337 human SNPs in dbSNP (23-7-’13) [13]
- SNPs occur on average about every 100 to 300 bases (http://en.wikipedia.org/wiki/Human_genetic_variation)
- Contains information about heredity
- If expressed
  - Loss/change of protein
  - Loss of RNA 2D/3D structure
  - Affect alternative splicing

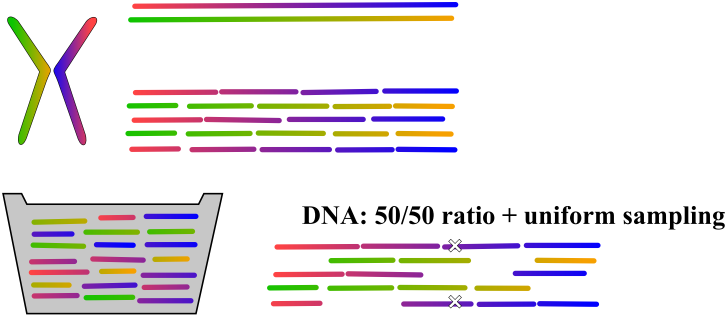
Major differences between SNVs in RNA-Seq & DNA-Seq

(allele specific) expression



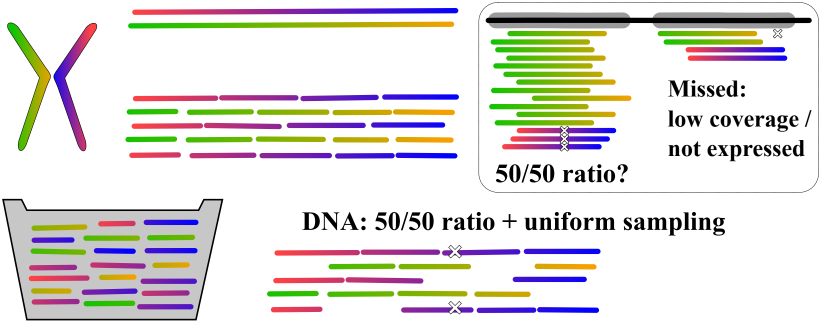
Major differences between SNVs in RNA-Seq & DNA-Seq

(allele specific) expression



Major differences between SNVs in RNA-Seq & DNA-Seq

(allele specific) expression

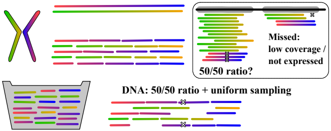


Missed:
low coverage /
not expressed

50/50 ratio?

DNA: 50/50 ratio + uniform sampling

Single Nucleotide Polymorphisms in RNA-Seq



- Major difference(s) between DNA-Seq:
  - Detected SNPs are expressed
    - Biological context
    - SNPs RNA-Seq only within exons and ncRNAs
    - Allele specific expression profiles
- Detection:
  - Expression affects coverage; in DNA-seq coverage should be uniform

Single Nucleotide Polymorphisms in RNA-Seq

Detection tools

- Alignment
  - TopHat [15, 5]
  - STAR [3]
  - ... many many more
- SNV calling
  - VarScan2 [6]
  - samtools [7]
  - exactSNP *(part of subread [9] package)*
  - GATK [17]

SNV detection in RNA using VarScan2

Requires samtools [7] for intermediate mpileup files

- Samtools: BAM file → mpileup file
- VarScan2
  - Compare alignment to reference genome
    - mpileup file + ref. fasta file → VCF file
    - Galaxy: *VarScan*
    - Galaxy: *VarScan ... optimized for direct BAM/SAM input*
  - Compare alignment to other alignment
    - 2×mpileup file → VCF file
    - Galaxy: *VarScan*
- *"statistical significance ... is computed by Fisher's exact test of the read counts supporting each allele (reference and variant) compared to the expected distribution based on sequencing error alone" [6]*

Single Nucleotide Polymorphisms in RNA-Seq

Using: Samtools, VarScan

reference	A	C	T	G	A	
read1	a	c	c	g	c	Alignment
read2	a	c	t	g	a	
read3	a	c	c	g	a	
read4	a	c	c	g	a	
read5	a	c	t	a	a	
read6		c	c	g	a	
read7			c	g	a	
read quality (q)	0.99	0.99	0.85	0.8	0.99	
aligned	5	6	7	7	7	Expected (based on quality)
q*aligned	4.95	5.94	5.95	5.6	6.93	
(1-q)*aligned	0.05	0.06	1.05	1.4	0.07	
exp match (abs)	5	6	6	6	7	
exp mismatch (abs)	0	0	1	1	0	
obs match	5	6	2	6	6	Observed
obs mismatch	0	0	5	1	1	
P(obs|exp) fisher exact	1.000	1.000	0.049	0.538	0.500	Hypothesis testing
P < 0.05	REF	REF	SNP	REF	REF	

Single Nucleotide Polymorphisms in RNA-Seq

Covered examples during hands-on

- SNP: Artificial alignment (hg19)
- InDel: MCF7 alignment (hg18) [10, 1]

Differential gene expression

- Triggered by
  - Stimuli (signal molecules)
  - Genetics (mutation)
- Genes interact in a network, effect spreads out
  - Genetic redundancy / biological robustness (often multiple changes necessary to cause a disease)

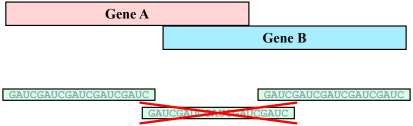
Differential gene expression analysis tools

- Alignment
  - TopHat [15, 5]
  - STAR [3]
  - ... many many more
- Measuring expression (quantification)
  - HTSeq-count [2]
  - Cufflinks [16]
  - featureCounts [8]
- Group-wise comparison (hypothesis testing)
  - EdgeR [12]
  - DESeq2 [11]
  - Cuffdiff [14]

Measure expression levels in RNA-Seq data

1. Align read to reference genome
2. Measuring expression = counting aligned reads
   - Count in annotated exons
   - Positive integers (read counts of 3.1415 or −42 are impossible)
   - Quantitative (read count has an absolute meaning)

- Observation (read count) must be statistically independent
  - No multi-map reads
  - Skip overlapping gene annotations



Measure expression levels in RNA-Seq data

In Galaxy

- featureCounts [8]
  - Pro's
    - Fast
    - Flexible
    - Free (GPL)
    - Accepts both BAM and SAM files
    - Only requires name-sorted files when mate-pairs are counted together (name-sorting is slow)
  - Con's
    - Built-in name-sorting supports no threading – rather do this with samtools [7]

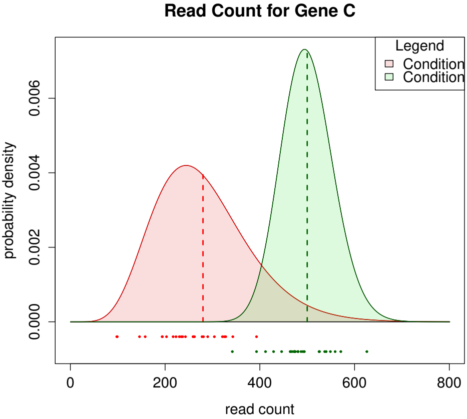
Differential gene expression analysis

edgeR

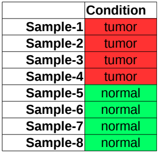
- edgeR [12]
  - Differential gene expression analysis
  - Free R Package (GPL2)
  - Galaxy wrapper does normalizations for you
    - Use raw reads, do NOT use FPKM/RPKM!
- "Limma" for count data
  - Not Gaussian (normal) distributed like e.g. micro-array data – but negative binomial

Differential gene expression analysis

Read counts: negative binomial distributed

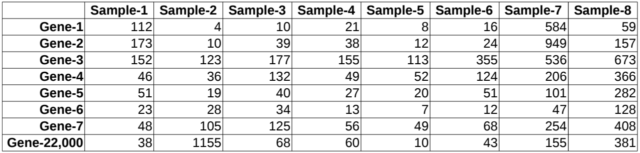


Differential gene expression analysis



	Condition
Sample-1	tumor
Sample-2	tumor
Sample-3	tumor
Sample-4	tumor
Sample-5	normal
Sample-6	normal
Sample-7	normal
Sample-8	normal

Design matrix



	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8
Gene-1	112	4	10	21	8	16	584	59
Gene-2	173	10	39	38	12	24	949	157
Gene-3	152	123	177	155	113	355	536	673
Gene-4	46	36	132	49	52	124	206	366
Gene-5	51	19	40	27	20	51	101	282
Gene-6	23	28	34	13	7	12	47	128
Gene-7	48	105	125	56	49	68	254	408
Gene-22,000	38	1155	68	60	10	43	155	381

Expression matrix

Differential gene expression analysis

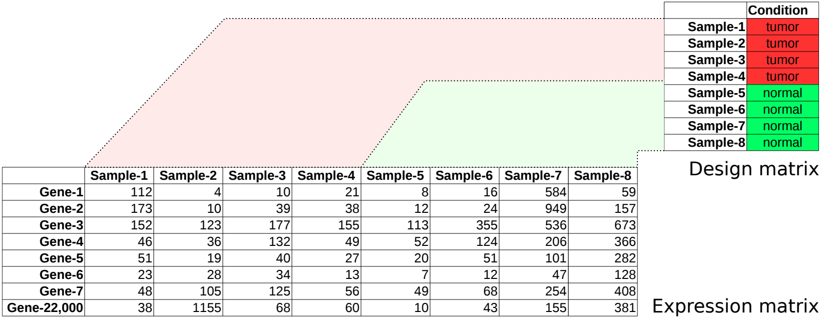
	Condition
Sample-1	tumor
Sample-2	tumor
Sample-3	tumor
Sample-4	tumor
Sample-5	normal
Sample-6	normal
Sample-7	normal
Sample-8	normal

Design matrix

	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8
Gene-1	112	4	10	21	8	16	584	59
Gene-2	173	10	39	38	12	24	949	157
Gene-3	152	123	177	155	113	355	536	673
Gene-4	46	36	132	49	52	124	206	366
Gene-5	51	19	40	27	20	51	101	282
Gene-6	23	28	34	13	7	12	47	128
Gene-7	48	105	125	56	49	68	254	408
Gene-22,000	38	1155	68	60	10	43	155	381

Expression matrix

Differential gene expression analysis



	Condition
Sample-1	tumor
Sample-2	tumor
Sample-3	tumor
Sample-4	tumor
Sample-5	normal
Sample-6	normal
Sample-7	normal
Sample-8	normal

Design matrix

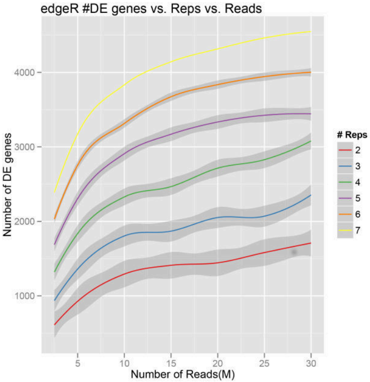
	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8
Gene-1	112	4	10	21	8	16	584	59
Gene-2	173	10	39	38	12	24	949	157
Gene-3	152	123	177	155	113	355	536	673
Gene-4	46	36	132	49	52	124	206	366
Gene-5	51	19	40	27	20	51	101	282
Gene-6	23	28	34	13	7	12	47	128
Gene-7	48	105	125	56	49	68	254	408
Gene-22,000	38	1155	68	60	10	43	155	381

Expression matrix

Contrast = tumor ↔ normal

Differential gene expression analysis

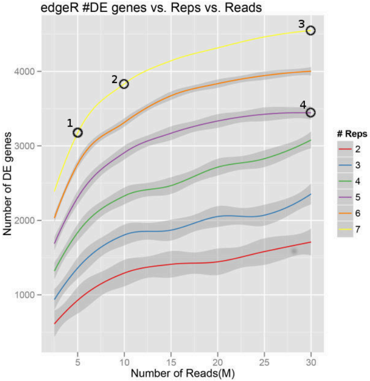
MCF7 cell line



[10]

Differential gene expression analysis

MCF7 cell line



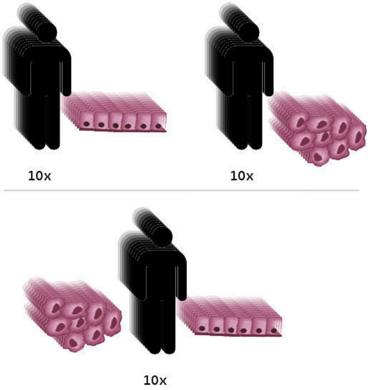
[10]

Differential gene expression analysis: sample pairing

- Sample pairing (do not confuse with PE-reads!) / batch effects
- Goal: correction for patient / batch specific expression profiles
- Examples:
    - 10×Tumour & Normal (both of same patient)
    - 3 populations: African, American & Asian
    - 2 batches: 1 at Monday, 1 at Friday

Differential gene expression analysis

Prostate cancer and normal prostate



[4]

References I

[1] Aleksandra Adomas, Sara Grimm, Christine Malone, Motoki Takaku, Jennifer Sims, and Paul Wade. Breast tumor specific mutation in gata3 affects physiological mechanisms regulating transcription factor turnover. *BMC Cancer*, 14(1):278, 2014.

[2] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq: A python framework to work with high-throughput sequencing data. *bioRxiv*, 2014.

[3] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[4] Kalpana Kannan, Liguo Wang, Jianghua Wang, Michael M. Ittmann, Wei Li, and Laising Yen. Recurrent chimeric rnas enriched in human prostate cancer identified by deep sequencing. *Proceedings of the National Academy of Sciences*, 108(22):9172–9177, 2011.

[5] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013.

[6] Daniel C. Koboldt, Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 2009.

[7] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

References II

[8] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 2013.

[9] Yang Liao, Gordon K. Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108, 2013.

[10] Yuwen Liu, Jie Zhou, and Kevin P. White. Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

[11] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *bioRxiv*, 2014.

[12] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[13] Elizabeth M. Smigielski, Karl Sirotkin, Minghong Ward, and Stephen T. Sherry. dbsnp: a database of single nucleotide polymorphisms. *Nucleic Acids Research*, 28(1):352–355, 2000.

[14] Cole Trapnell, David G. Hendrickson, Martin Sauvageau, Loyal Goff, John L. Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nat Biotech*, 31(1):46–53, Jan 2013.

[15] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[16] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5):511–515, May 2010.

References III

[17] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*, chapter 11:11.10, pages 11.10.1–11.10.33. John Wiley and Sons, Inc., 2002.

More links

- http://www.bioinformatics.babraham.ac.uk/training/RNA-Seq_analysis_course.pptx
- http://galaxy.ctmm-trait.nl/
- http://toolshed.dtls.nl/