r-make: STAR

Version: STAR_2.1.4a_r178

Non-default Parameter:

outSAMunmapped Within
outFilterMultimapNmax 1
chimSegmentMin 20
chimScoreMin 1

Default Parameter:

outFilterScoreMin 0
outFilterScoreMinOverLread 0.66
outFilterMatchNmin 0
outFilterMatchNminOverLread 0.66

outFilterMultimapScoreRange 1
outFilterMismatchNmax 10
outFilterMismatchNoverLmax 0.3

outFilterIntronMotifs None

readMatesLengthsIn NotEqual

clip3pNbases 0
clip5pNbases 0

outSJfilterOverhangMin 30 12 12 12
outSJfilterCountUniqueMin 3 1 1 1
outSJfilterDistToOtherSJmin 10 0 5 10

scoreGap 0
scoreGapNoncan -8
scoreGapGCAG -4
scoreGapATAC -8
scoreGenomicLengthLog2scale -0.25
scoreDelOpen -2
scoreDelBase -2
scoreInsOpen -2
scoreInsBase -2
scoreStitchSJshift 1
seedSearchStartLmax 50
seedSearchStartLmaxOverLread 1.0
seedSearchLmax 0
seedMultimapNmax 10000
seedPerReadNmax 1000
seedPerWindowNmax 50
seedNoneLociPerWindow 10

alignIntronMin 21
alignIntronMax 0
alignMatesGapMax 0
alignSJoverhangMin 5
alignSplicedMateMapLmin 0
alignSplicedMateMapLminOverLmate 0.66
alignWindowsPerReadNmax 10000
alignTranscriptsPerWindowNmax 100
alignTranscriptsPerReadNmax 10000

chimScoreDropMax 20
chimScoreSeparation 10
chimScoreJunctionNonGTAG -1
chimJunctionOverhangMin 20

Rsamtools_1.8.6
GenomicRanges_1.8.13
rtracklayer_1.18.2

Subread

Read alignment and summarization were performed using programs included in Subread package (http://subread.sourceforge.net/) or Bioconductor R package Rsubread (http://bioconductor.org/packages/release/bioc/html/Rsubread.html).

(1) Build the index used by the Subread and Subjunc aligners:

For Illumina and Roche 454 data:

```
subread-buildindex -M 8000 -f 24 -o hg19_subread_index hg19_chr_sequences.fa
```

For SOLiD color space data:

```
subread-buildindex -c -M 8000 -f 24 -o hg19_subread_index hg19_chr_sequences.fa
```

These commands can be run on the command line of a Unix or Mac machine. Note that the index building can also be carried out in R using Bioconductor package Rsubread.

(2) Perform read alignments:

For expression analysis:

```
subread-align -i hg19_subread_index -u -H -n 10 -m 3 -p 1 -S fr -d 0 -D 600 -I 5
-P 6 -T 10 --DPGapOpen -1 --DPGapExt 0 --DPMismatch 0 --DPMatch 2 --trim5 0 --
trim3 0 -r reads1.txt -R reads2.txt -o subread_output.sam
```

For identification of exon-exon junctions:

```
subjunc -i hg19_subread_index -u -H -n 14 -m 1 -p 1 -S fr -d 0 -D 600 -I 5 -P 6 -
T 10 --trim5 0 --trim3 0-r reads1.txt -R reads2.txt -o subjunc_output.sam
```

The Phred-score offset used for BGI and COH sites is 64 ('-P 6'). For the other sequencing sites, the offset is 33 ('-P 3').

Note that read alignments also be carried out in R using Bioconductor package Rsubread.

(3) Read summarization

Read summarization was carried out in R using the featureCounts function included in Bioconductor package Rsubread. An inbuilt NCBI RefSeq annotation (build 37.2) was included in this package.

```
library(Rsubread)
fc <- featureCounts(samfiles, "hg19",annot.ext = NULL, isGTFAnnotationFile =
FALSE,
    GTF.featureType = "exon", GTF.attrType = "gene_id", useMetaFeatures = TRUE,
    allowMultiOverlap = FALSE, nthreads = 10, strandSpecific = 0,
    countMultiMappingReads = FALSE, minMQS = 0, isPairedEnd = TRUE,
    requireBothEndsMapped = FALSE, checkFragLength = TRUE, minFragLength = 0,
```

```
    maxFragLength = 2000, countChimericFragments = TRUE, chrAliases = NULL,
    reportReads = FALSE)
```

Note that the read summarization can also be carried out on the command line of a Unix or Mac machine using the featureCounts program included in the SourceForge Subread package. But users have to download the NCBI RefSeq gene annotation for human genome from NCBI ftp server and provide it to featureCounts program.

(4) Normalization

Normalization was performed in R using the voom function included in Bioconductor package limma.

```
y <-
voom(counts=fc$counts,design=design,lib.size=NULL,normalize.method="none",plot=FA
LSE,span=0.5)
```

SHRiMP2 + BitSeq

SHRiMP2 (v2.2.2)

ILM reads

```
FASTA=${PATH}/data/references/human_AceView.2010_ERCC_gtf.fasta

INDEX=${PATH}/data/references/human_AceView.2010_ERCC_gtf-ls

$SHRIMP_FOLDER/utils/project-db.py --shrimp-mode ls $FASTA


$SHRIMP_FOLDER/gmapper-ls -L ${INDEX} -Q -E -N 16 -o 60 --max-alignments 120 -h
90% -p opp-in --ignore-qvs --insert-size-dist ${mean}","${sd} -1 ${R1} -2 ${R2}
>${OUT} 2>${LOG}

samtools view -bS ${OUT} > ${OUT_BAM}
```

or

LIF reads

```
FASTA=${PATH}/data/references/human_AceView.2010_ERCC_gtf.fasta

INDEX=${PATH}/data/references/human_AceView.2010_ERCC_gtf-cs;

$SHRIMP_FOLDER/utils/project-db.py --shrimp-mode cs $FASTA


$SHRIMP_FOLDER/gmapper-cs -L ${INDEX} -Q -E -N 16 -o 60 --max-alignments 120 -h
90% -p opp-in --ignore-qvs --insert-size-dist ${mean}","${sd} -1 ${R1} -2 ${R2}
>${OUT} 2>${LOG}

samtools view -bS ${OUT} > ${OUT_BAM}
```

NOTE: fragment lengths mean and SD were calculated from Bowtie alignments to the transcript sequences

BitSeq (v0.4.2)

```
# FASTA file with transcript sequences

FASTA=${PATH}"/data/references/human_AceView.2010_ERCC_gtf.fasta";

# file containing information about transcript lengths - created by BitSeq

REF=${PATHD}"/data/references/AceView.tr";

# default BitSeq parameters
```

```
PAR=${BitSeq}"/parameters.txt";


samtools merge -f -n $OUT_BAM $PATH/*.bam;

$BitSeq/parseAlignment -f BAM $OUT_BAM -o $OUT_PROB --trSeqFile $FASTA --verbose
-P 3

$BitSeq/estimateExpression --verbose $OUT_PROB -o $OUT --outType RPKM -p $PAR -t
$REF -P 16

$BitSeq/getVariance --verbose -o ${OUT_MEAN} ${OUT}".rpkm"
```

TopHat 2 + CuffDiff 2

TopHat (v2.0.3)

```
# index build on the genome sequences by Bowtie (letter or colour space
accordingly)

INDEX=${PATH}/data/references/hg19_genome_ERCC;

tophat --bowtie1 --report-secondary-alignments --bowtie-n -n 2 -r ${mean} --mate-
std-dev ${sd} -p 15 -o ${OUT} ${INDEX} ${R1} ${R2} &> ${LOG}
```

NOTE: fragment lengths mean and SD were calculated from Bowtie alignments to the transcript sequences

CuffDiff (v2.0.0)

```
cuffdiff -q -o CuffDiff_TH2_STD/ -p 30 -u -c 1 --no-update-check
../human_AceView.2010_ERCC_exon.gtf
BAM/SEQC_ILM_BGI_A_1_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_A_2_TopHat_2.0_std.bam,B
AM/SEQC_ILM_BGI_A_3_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_A_4_TopHat_2.0_std.bam
BAM/SEQC_ILM_BGI_B_1_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_B_2_TopHat_2.0_std.bam,B
AM/SEQC_ILM_BGI_B_3_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_B_4_TopHat_2.0_std.bam
BAM/SEQC_ILM_BGI_C_1_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_C_2_TopHat_2.0_std.bam,B
AM/SEQC_ILM_BGI_C_3_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_C_4_TopHat_2.0_std.bam
BAM/SEQC_ILM_BGI_D_1_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_D_2_TopHat_2.0_std.bam,B
AM/SEQC_ILM_BGI_D_3_TopHat_2.0_std.bam,BAM/SEQC_ILM_BGI_D_4_TopHat_2.0_std.bam
```

NOTE: We were using the CuffDiff for expression estimates to ensure cross replicates and cross-sample library size normalization

TopHat 2 'G' + CuffDiff 2

TopHat (v2.0.3)

```
# index build on the genome sequences by Bowtie (letter or colour space
accordingly)

INDEX=${PATH}/data/references/hg19_genome_ERCC;

# location of transcriptome index build by TopHat based on GTF file with -G
option (option set and GTF file provided only during 'first' execution)

INDEX2=${PATH}/data/references/TopHat/human_AceView.2010_ERCC_TopHat

tophat --bowtie1 --transcriptome-index=${INDEX2} --report-secondary-alignments --
bowtie-n -n 2 -r ${mean} --mate-std-dev ${sd} -p 15 -o ${OUT} ${INDEX} ${R1}
${R2} &> ${LOG}
```

NOTE: fragment lengths mean and SD were calculated from Bowtie alignments to the transcript sequences

CuffDiff (v2.0.0)

```
cuffdiff -q -o CuffDiff_TH2_G/ -p 30 -u -c 1 --no-update-check
../human_AceView.2010_ERCC_exon.gtf
BAM/SEQC_ILM_BGI_A_1_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_A_2_TopHat_2.0_g.bam,BAM/S
EQC_ILM_BGI_A_3_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_A_4_TopHat_2.0_g.bam
BAM/SEQC_ILM_BGI_B_1_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_B_2_TopHat_2.0_g.bam,BAM/S
EQC_ILM_BGI_B_3_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_B_4_TopHat_2.0_g.bam
BAM/SEQC_ILM_BGI_C_1_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_C_2_TopHat_2.0_g.bam,BAM/S
EQC_ILM_BGI_C_3_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_C_4_TopHat_2.0_g.bam
BAM/SEQC_ILM_BGI_D_1_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_D_2_TopHat_2.0_g.bam,BAM/S
EQC_ILM_BGI_D_3_TopHat_2.0_g.bam,BAM/SEQC_ILM_BGI_D_4_TopHat_2.0_g.bam
```

NOTE: We were using the CuffDiff for expression estimates to ensure cross replicates and cross-sample library size normalization

Methods developed and used for Magic analysis of the SEQC main and neuroblastoma RNA-seq projects

The Magic RNA-seq analysis pipeline was developed at NCBI for the double purpose of analyzing RNA-seq data in clinical cohorts, in the hope of identifying features associated to disease conditions, and to annotate the genes for the public AceView genes database. The SEQC projects, with built in truths and controls, and multiple teams involved in parallel was ideal to optimize the concepts and thoroughly test the code. Limitations due to the technical noise and biases inherent to the technique being understood, Magic is now able to provide specific expression measures for transcripts and genes and their differential, together with a measured false discovery rate. The Magic code is available on the AceView NCBI website www.aceview.org in the Downloads, Software tab, Magic at ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic . The package, written in C, is completely original, self-standing and open source. It includes its own aligner and all the necessary software to perform the integrative data analysis. Once experiments are described in the associated acedb/AceView object-oriented meta-database, the pipeline runs automatically, using parameters optimized for that experiment and protocol. It generates data tables for gene and transcripts expression, new exon, intron and transcripts discovery, SNVs and rearrangements as well as multiple quality control reports.

Here we describe the application of the Magic pipeline to the analysis of the ~10 terabases of SEQC RNA-seq data.

The Magic pipeline

The Magic pipeline, detailed in Figure 1, includes quality control, alignment, gene annotation, quantification, normalization, selection of differentially expressed features and discovery of new genes and transcripts. The 2013 SEQC main study focused on quantification of the expression of genes, transcripts and exon junction discovery. In addition, the Neuroblastoma study identified differential genes and transcripts between phenotypic classes (by disease stages and MYCN amplification), and characterized new transcripts and new genes.

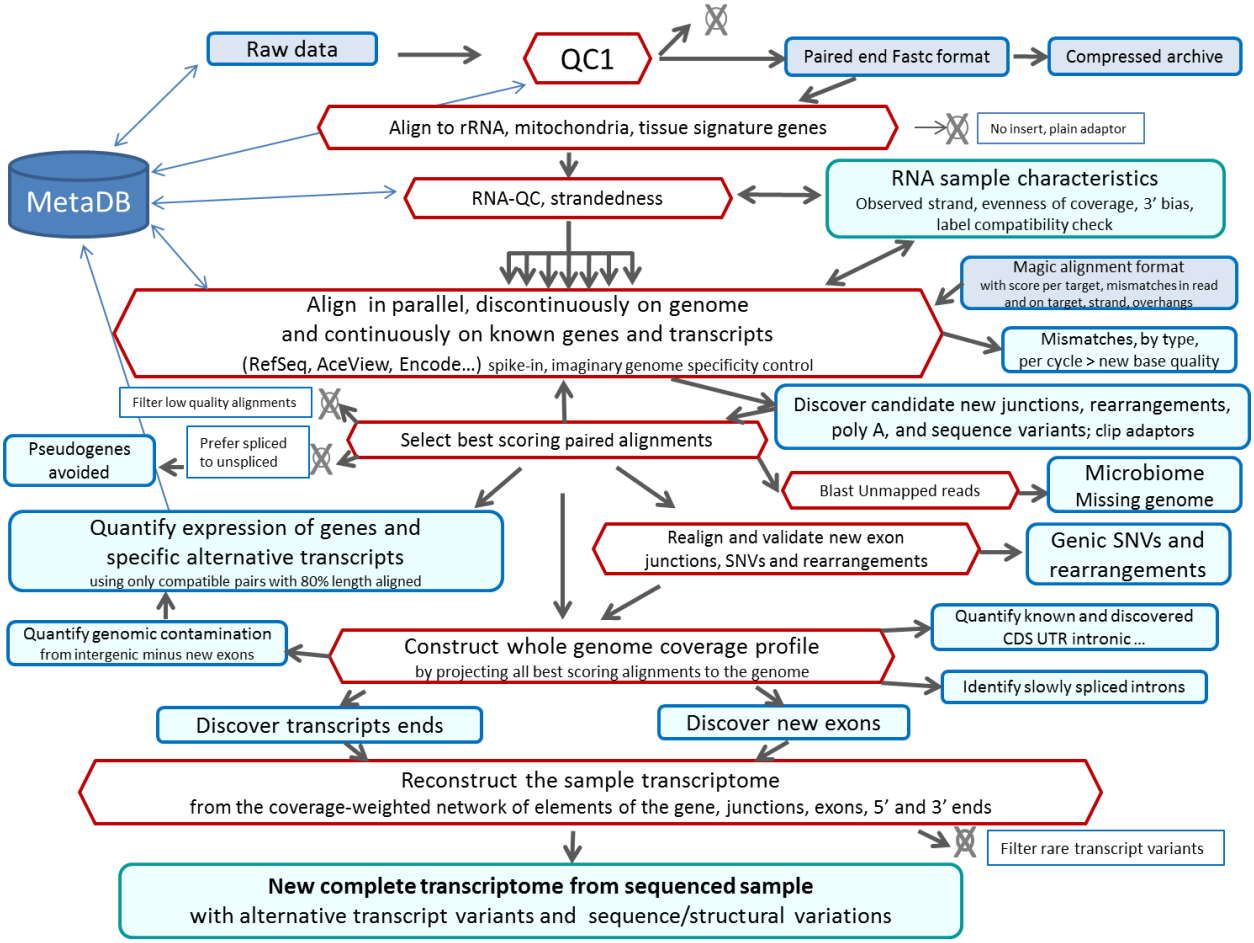


Figure 1: The complete Magic pipeline dataflow for RNA-seq analysis.

Read alignments and target selection

Read pairs are sorted alphabetically and aligned as single objects, using the Fastc modification of the fasta format, representing pairs on a single line as ATGNC><CCTATTT. The mapping strategy is to hash a dense set of seeds of length 16 bases and scan the targets using a fast seed-extension automaton auto-adaptable to the mismatch profile of the platform. Alignments may include substitutions, short insertions or deletions. RNA is aligned discontinuously on the genome, allowing for discovery of exon junctions, structural rearrangements or, simply, polyA addition sites and sequence adaptors, which are recognized in the unaligned overhanging ends of reads. The eventual adaptor sequences, and poly A tails are trimmed dynamically during the alignment, verifying that the trimmed part does not match the target genome or gene, and eventually discovering the adaptor sequence.

Magic maps the reads in parallel on multiple targets, and then selects the best paired alignments across all targets. The target sequences used are available at ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic/TARGET.human.2013_06_15.tar.gz , see the fasta files in TARGET/Targets/hs.*.fasta.gz. For SEQC studies done in 2013, we used

- the reference genome (GRCh 37, without patches) and three gene annotations:
- AceView genes (264,387 transcripts, from 55,836 genes, 2011 subset of the 2010 version): AceView 2010 genes integrate and summarize all human cDNA sequences in GenBank, dbEST and RefSeq NM/NR transcript models and all Entrez gene models from 2010. Entrez gene names are used when available; otherwise genes are named using an AceView identifier, stable across releases. cDNA sequences in GenBank define about twice as many genes and close to twice as many exon junctions than RefSeq only (370,000 rather than 198,000). AceView 2010 annotates 200,000 spliced transcript variants, that is, 7 times more alternatively spliced variants than RefSeq.
- RefSeq v104 (40,894 transcripts from 24,536 genes, January 2013)
- Encode v15/Ensembl 37.70 (cdna.all and ncRNA, minus rRNAs: 194,703 transcripts from 56,071 genes, released January 2013)].

- We also used as target the mitochondrial genome (NC_012920.1),
- the ribosomal RNA 5S and 45S precursor (NR_046235 and NR_023363),
- small non coding RNA genes (including 8611 genes from Ensembl 37.70: 1923 snRNA, 1529 snoRNA, 3110 miRNA and 2049 misc-RNA, complemented by 625 tRNAs from Lowe, UCSC tRNA track, February 2012),
- then the 92 ERCC RNA spike-in sequences, and
- finally the imaginary genome as a mapping specificity control. The imaginary whole genome is constructed by complementing the bases (exchange A:T and G:C), but not reversing the order: it has exactly the same statistics as the genome but is a completely alien decoy: hits to this target are false positives and characteristics of their alignments are used to tune the alignment quality filters.

Once fragments have been aligned to all targets, a universal scoring system allows Magic to select the best alignments for each read pair. For RNA-Seq, each alignment is scored by adding one point per base aligned and removing eight points per mismatch (base substitution, insertion or deletion). The length of the aligned region is then recursively clipped to optimize this score. As a result, no mismatch is reported closer than 8 bases to the edge of the aligned segment, allowing reliable recognition of substitutions or indels and precise cooperative discovery of exon-intron boundaries. For whole genome or exome sequencing, the penalty for mismatches is reduced to 4, since we only expect rare discontinuities in the alignment. For each read pair, alignments on all targets, genes or genome, are scored, and only the best scores, often on multiple targets, are kept, giving preference to spliced over unspliced targets in order to disfavor pseudogenes. Global quality filters are then applied, more tolerant to mismatches in full length unique alignments. In the quasi-unique mode (used in the NB MAV and Main 2012 files), read pairs mapping over their entire length and equally well to several genes or genomic loci, up to 9 sites, are assigned in Bayesian proportion seeded on the number of reads uniquely mapping to these loci. For example, in the NB cohort of 498 individuals sequenced at BGI as Illumina HiSeq (60 million reads PE100 bp), 95.41% of the reads mapped uniquely, 1.43% mapped equally well to 2 to 9 loci, an additional 2.02% mapped to a single genomic site but to two antisense transcripts and could not be attributed to a single gene, because the libraries did not preserve the RNA strand. Only 0.79% of the reads were unmapped and unjustified, the remaining 0.36% being rejected for quality reasons.

Magic exports a richer and more explicit alignment format than Bam/Sam, so that the downstream tracking of mismatches, SNPs, discovered splice sites and structural rearrangements is simplified. Analysis is fast: typically, 1 gigabase of RNA sequences are aligned on all targets using about 6 hours of CPU, 8 GB of RAM and 2 gigabytes of disk space.

In practice there are no parameters to specify as Magic is auto-adaptive. The reads are described in the object oriented AceView/AceDB run database, essentially locating the original fastq file and specifying the platform (ILM, Roche, SOLiD), the protocol (Total or polyA, stranded or not) and the eventual existence of paired reads. Groups of reads are manually created in the database, for instance in SEQC main, by sample, platform, site or library, and in NB by phenotype. Actions to be performed such as sample comparisons, DEG calls or covariance analysis, titration analysis and so on are specified for each group. The pipeline is then ran very simply as "MAGIC ALIGN GENE_EXPRESSION INTRON_DISCOVERY SNP ",.

A typical outcome is illustrated by the NB project: of the 30.7 billion reads (paired), 98.9% were aligned at high quality to genes or the genome, with on average 98 nt aligned per 100 nt read and 99.703% identity (1.16 transitions, 1.64 transversions, and 0.17 insertions/deletions per kb aligned, but the majority of those mismatches are lane dependent sequencing noise: in regions covered at least 10 fold, where SNPs supported by at least 4 variant pairs and 20% of the pairs bearing the SNV can be called, only 14% of the mismatches contribute to SNVs). In contrast, only 0.00126% of the reads mapped to the imaginary genome and only partly, aligning over an average of 37 nt.

Coverage plots, introns and exons discovery

Best alignments to the various targets are then summarized and merged by projection on the genome, yielding a coverage density plot with sharp exon boundaries and no leaking into introns. The genome is then partitioned into coding regions (1.84%, of which successively 1.09 RefSeq + 0.50 AceView + 0.02 Ensembl + 0.23 new), UTR/non-coding exonic regions (6.86%, of which 1.10 RefSeq + 3.22 AceView + 0.39 Ensembl + 2.15 new), intronic (50.13%, of which 32.85 RefSeq + 12.72 AceView + 2.27 Ensembl + 2.30 new) and intergenic regions (41.16%), using hierarchically the annotations from RefSeq, AceView, Encode and the Magic new transcriptome (derived from RNA-seq), and giving priority to coding over non coding over introns over intergene.

Candidate new exons are considered when coverage exceeds a threshold well above the background noise (50 fold coverage in the NB experiment, where intergenic is covered on average 10.2 fold). At 50 fold coverage, 22% of the genome is expressed, 5% in previously annotated exons (RefSeq, AceView or Ensembl) and 12% in introns. New exons and new exon junctions get validated only

if they naturally integrate into full transcripts with commensurable support over all elements. For example, a new exon junction supported only 300 times in a gene with local coverage 100,000 is rejected. Indeed, we observed that a low level (0.5% to 1%) of artefacts, rearrangements and deletions, affects the RNA/cDNA molecules structure during RNA-seq library preparation. This estimate is down from 3% in standard cloned cDNA libraries, yet because of the size of RNA-seq experiments, these disguised false new variants flourish and the absolute number of supporting reads is less relevant than the local relative support: in very highly expressed genes, one observes a myriad of deceptive artefactual 'alternative variants'. Similarly, a new exon junction mapped in an area apparently untranscribed and with no independent exonic support is likely a (mapping/gene family/pseudogene/structural) artefact, because real transcription is processive and transcripts ends must be found. As a result of these sanity filters, only a fraction around 20% of the newly discovered elements will contribute to a reliable RNA-seq based gene annotation. For example, NB new exons finally cover 68.5 Mb of genome (including 7 Mb of new protein coding region and 61.5 of UTR or non-coding) and absorb 4.38% of the 2.7 Tb of RNA sequences aligned (outside mitochondrial and ribosomal RNA genes). 71 Mb of new intronic sequences located outside of previously annotated gene boundaries – in previously 'intergenic' regions - are defined in the process and add up to the 1.48 Gb of previously annotated intronic sequences (in non-stranded RNA-seq).

The 'background noise' used in the expression index computation is measured in each run as the average coverage of the genome outside the known and new premessenger annotations. We consider it mainly reflects genomic contamination, because in the SEQC main study, the intergenic coverage in the very same sample sequenced in different libraries, platforms and sites fluctuates from library to library (Figure 2).

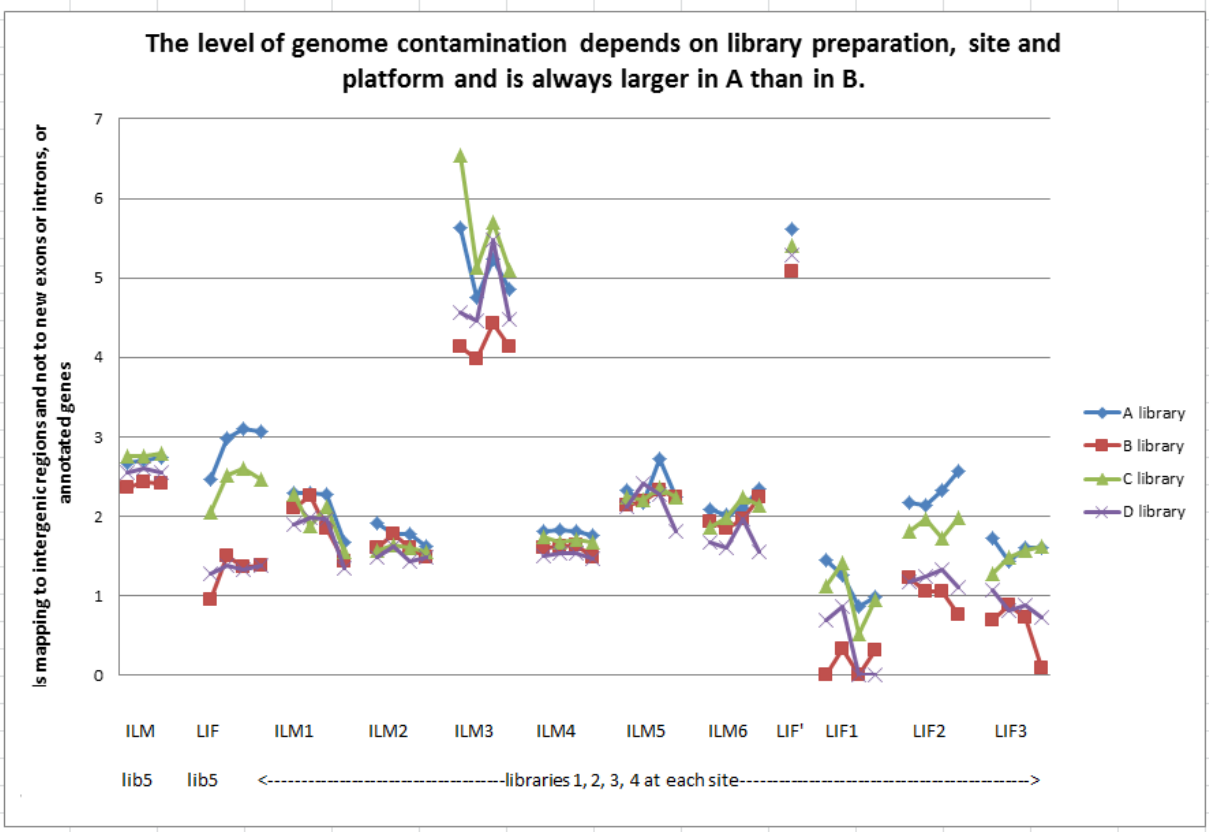
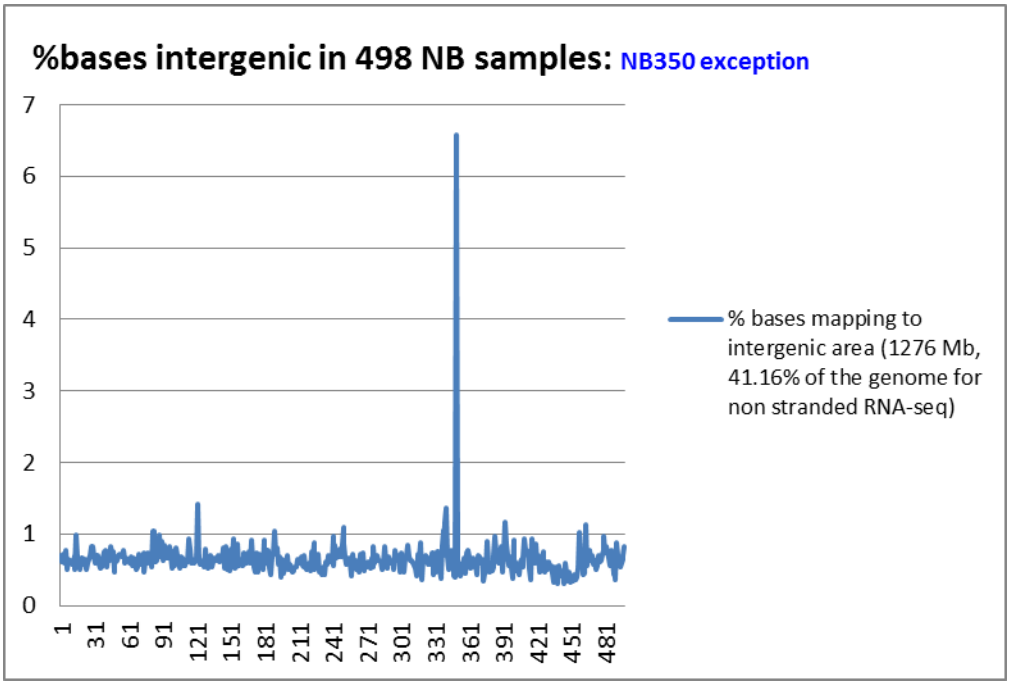
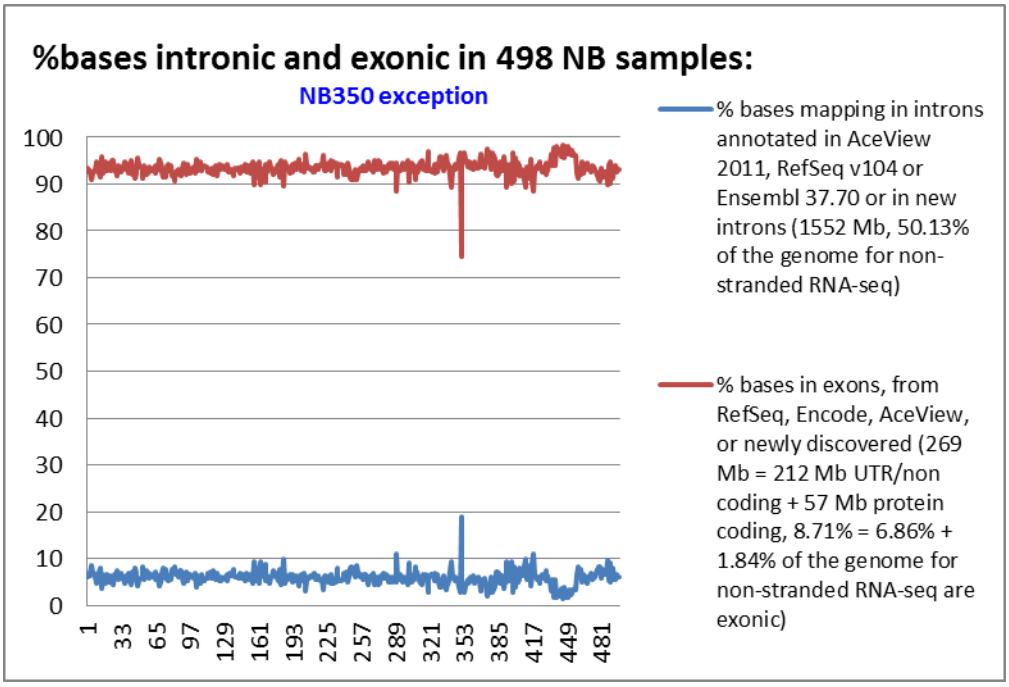


Figure 2: Coverage of intergenic region in the SEQC main 4 samples depends on site, protocol and library preparation

In the neuroblastoma study, one of the RNA samples, NB350, was likely partly degraded, leading to enrichment of intergenic (genomic DNA contamination) relative to exonic and intronic (Figure 3):





Quantification of RNA-seq expression data

To ensure specificity of the mapping, only 'compatible' pairs, where both reads are aligned over 80% of their length in the same gene or as a gene extension, in close vicinity on the genome, and in compatible orientation, are used to compute the expression indexes presented in the RNA-seq expression tables. Actually, 95.82% of the fragments in the NB project map as compatible pairs. Reads mapping equally well to several alternative transcripts of the same gene are counted only once in the gene. Regarding the transcripts, the critically desirable feature is to measure the distinctive part of each alternative transcript. This is achieved by a

simple procedure we call hierarchic mapping. Each read mapping equally well to multiple transcripts (say in exons common to variants .a, .c and .f) is attributed to the variant with the first alphabetic name (variant .a). AceView transcripts are named by decreasing order of protein coding potential or of length. As a result, if transcripts .a and .b share 2 kb of exons, but differ only by a splice junction, only the reads mapping specifically to this structural difference, characteristic of transcript .b, but absent from .a, will score in .b. Similarly, only the part of transcript .v that is not present in any of the variants .a .b .c .d …. u will score in .v. Differential use of a particular promotor, termination signal or splice pattern will be cleanly reported in the first variant possessing this feature. Contrary to the Cufflink procedure, this method is numerically stable because it does not involve a matrix inversion. It is valid even if additional alternative forms, not yet annotated, exist in vivo. The three NB or SEQC Main transcript tables dated December 2013 (RefSeq, AceView and Ensembl) use this hierarchic mapping strategy, which proved in our hands to yield more powerful predictors than genes, non-hierarchical transcripts or exon junctions.

The Magic normalized RNA-Seq index of expression is directly comparable to a normalized microarray logarithmic luminosity. The logarithm is introduced because gene expression is mostly log-normal. As usual, the index is normalized to be independent of the size of the experiment and of the length of the gene. But several corrections are quantified run by run and introduced to compensate for undesirable batch effects: the insert length of the library, which introduces a steric constraint on the coverage of short genes; the 3' bias, which limits the coverage of long genes; the level of genomic contamination, estimated from the density of intergenic reads, excluding new exons discovered in the experiment; and the eventual presence of extremely highly expressed genes. Correcting for these measurable effects, which are library dependent and occasionally strong, stabilizes the Magic index across samples. Finally, the sampling fluctuations, which affect the counts in low expressed genes, are absorbed by the log(sqrt) function, also known as inverse hyperbolic sine. All these corrections affect the normalization factor z, defined below, while maintaining the read counts n on the original scale, to correctly control the sampling fluctuations. Neglecting these refinements and corrections, the Magic index would be equal to $\log_2$(1000 FPKM), so FPKM=1 would correspond to an index of 9.96.

The index is computed a

$$\text{Index} = \log_2\left(n + \sqrt{(5+b)^2 + n^2}\right) - 1 + \log_2(z) \quad , \text{ with } z = 10^{12} r/N'l'$$

- n is the number of reads aligned in the gene. A read partially aligned counts as a fraction.
- b is the background noise over the length of the gene, mainly corresponding to genomic contamination.
- r is the average aligned length of the reads, in bases.
- $N'/10^{12}$ is the corrected number of terabases aligned to the known transcriptome, after excluding the ribosomal and mitochondrial genes and the genes gathering more than 2% of the total number of bases aligned (31 instances in this experiment, 16 EEF1A1, 7 DDX1, 4 ALB, 2 COL1A1, 1 COL1A2 and 1 TRIB2). This correction compensates an apparent variation of the index of all other genes in the presence of highly-variable extremely-expressed genes, such as albumin in liver or hemoglobins in blood.
- l' is the length of the gene effectively sequenceable in the particular experiment. First, to evaluate the 3' bias, the coverage of very long ubiquitous genes (such as EEF2 (3.5kb), FLNA (8.5 kb), MALAT1 (7kb), AHNAK (18 kb) or NEAT (22 kb)) is measured. In polyA selected experiments, RNA degradation usually leads to incomplete coverage of these long genes. The covered length is often around 3 kb, sometimes much longer, but may also be below 1 kb. This length is used in each run as an upper bound l_{max}. Second, the average length λ of the insert is measured from the aligned read pairs, and subtracted from the length l of each gene to compensate for the steric constraint. Indeed $l'=(l-\lambda)$ is the number of possible positions of the insert in the gene. This correction matters for short genes, under-represented in libraries with long insert size. Finally, l' is restricted to the interval $[\lambda, l_{max}]$.
- If b is small and n is large, the index is approximately $\log_2(z.n)$ representing the $\log_2$ of the coverage of the gene per terabase aligned

For genes that are not 'significantly expressed', i.e. with less than 4 reads above the experimental noise, the index is flagged in the tables as NA/value, indicating a tradeoff between low precision due to high sampling error and low accuracy due to the proximity of the zero counts plateau. For zero counts, the index is lower for larger genes and deeper experiments, expressing higher confidence that the gene is not expressed.

When the same biological sample is sequenced in multiple runs (e.g. the 15 NB samples with duplicate runs), reads mapped to the same gene are cumulated before computing the index. In phenotypic groups consisting of different individuals, the index of the group is computed as the average index of all participants. Thus in additive groups, we compute the log of the average of the number of reads, while in phenotypic groups we compute the average of the log.

The MAV tables provided for the AceView Magic predictions, dated 20120828, 20121127 and 20121210, used a less stringent version of the expression index, allowing quasi unique alignments, some partial, non-hierarchic mapping to transcripts and a different quantification of the zero counts: the index corresponding to read counts below 4 were linearly interpolated between 0 for 0 reads and the computed index value for 4 reads. The newer more refined expression index files, using uniquely mapped pairs and hierarchic quantification of the transcripts, are available for AceView, RefSeq and Encode genes as supplementary tables dated _December2013.

All tables are sorted by descending maximal index of expression across the 498 samples. The first columns contain metadata connecting AceView gene to RefSeq, GeneID, Agilent expression array probes and genomic coordinates on the human reference build 37. In the December 2013 NB files, the relation between the expression index and the average gene coverage is given by the plot in Figure 4.

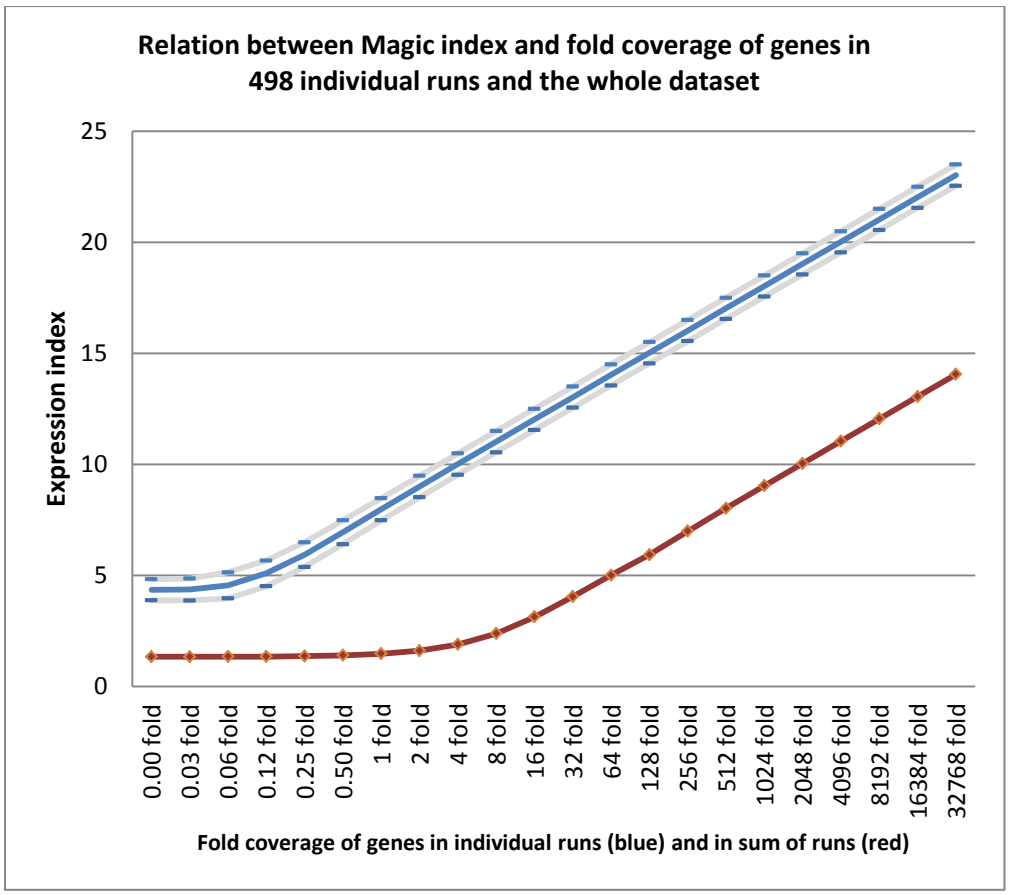


Figure 4: Relation between the expression index and the average gene coverage in individual samples (blue/gray) and in the sum of all samples (red). The variations per sample reflect the fact that some experiments (low grey/blue) have twice as many reads as others (high grey/blue).

Selection of differentially expressed genes or transcripts in RNA-seq or arrays

Selection of differentially expressed genes, or transcript variants, or exons or junctions or microarray probes between conditions A and B is delicate, and needs to circumvent the difficulties associated to sampling errors affecting the low expressors.

1. The method needs to robustly recognize differential genes, including those expressed in only one condition, to be resilient to outliers, and to be insensitive to the precise normalization of the low expressed genes.
2. When sampling a natural population, the heterogeneity across individuals may be multimodal, due to functional redundancy, as there are multiple ways to achieve the same phenotype. This fact makes conventional statistics based on generalized Gaussian assumptions inapplicable.
3. Finally an estimation of significance and false discovery rate is needed.

Examining the distribution of indexes across two groups to be compared, the general idea is to delineate for each gene or transcript the eventual existence of two distinct zones of expression, the A and B zones, where either A or B widely predominates. For the most discriminative DEGs, the A and B zones are fully separated: all A samples fall in index zone A, all B samples in index zone B. But in general, the two conditions overlap, leaving a gray zone in the center.

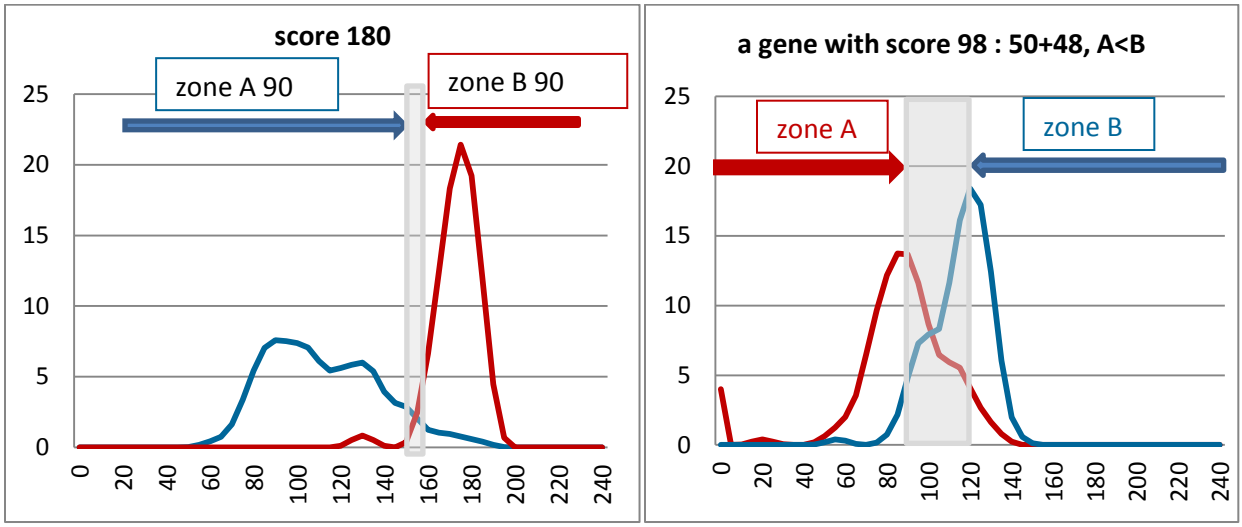


Figure 5: Selection of differentially expressed genes between two conditions, A red and B blue: the normalized distribution of indexes of expression for each condition is plotted versus the index of expression (x10). A cumulated count, starting from the extremities of the distribution, identifies the zones of dominance of the red, or the blue, and the choice of the purity ratio (here set at 4) allows to define the less informative (gray) index zone, where none of the conditions predominates. This method is insensitive to the exact measure of low expressors as it cumulates with equal voting power all individuals to the left of the gray zone The left panel shows a quasi-perfect DEG with a spread out distribution, and score 90+90: most of the index range is discriminative.

In practice, we construct the smoothed histogram of the index of expression of each gene for conditions A and B. Each individual contributes an elemental Gaussian of area 1 and sigma 0.5 centered on the measured value. The width is not supposed to catch the sampling error, but rather to represent some biological incertitude in the measure. The 2 histograms are normalized to a global area of 100. We then select a heuristic 'purity ratio' R which controls the width of the gray zone (usually between 1 and 3; for example, with R=3, 75% individuals in zone A are A). To locate the gray zone boundaries, we start from the left of the distribution, and count the numbers a and b of A and B individuals up to a moving frontier, and stop when (a-b) is maximal, (b-a) is then also maximal on the right. We then extend the gray zone around this frontier until on the A side, a>Rb and on the B side, b>Ra.

The differential scores are measured on each side of the gray zone: as the percentage of A samples minus B samples in the A zone, and as the percentage of B-A in the B zone. The sum of the A and B scores is the differential score of the gene or feature. It varies from 0 to 200, a perfect score of 100+100 indicates a DEG for which the 2 distributions do not overlap. A score of 120 = 30 + 90 would indicate that about 30% of A and 90% of B samples are classifying.

We then add two conditions: the difference of average index between the A and B distributions is chosen (at least 1 in the per stage comparisons, i.e. the fold change is at least 2). Also, the low limit of the high zone must be at least 3 fold above the average ‘significantly expressed’ index in the group (index around 7.5 to 10, depending on the sequencing depth of the run).

This simple method has three important advantages: It does not depend on the precise quantification of the low expressors; in fact the method mostly depends on the qualitative ordering of the samples by their index of expression. It is robust against multimodal distributions. It is insensitive to outliers that can be biological exceptions or mislabeled samples. In fact, in every large scale experiment from the SEQC or other projects, where comparison to microarrays, to biological replicates, or just to the annotated sex was possible, 1 to 5% of the samples were found to be mislabeled. In the course of the present study, 3 samples were re-assigned.

Estimating the False Discovery Rate of differentially expressed features

Depending on the endpoint, the number of genes, transcripts or SNPs measured as differential between the 2 conditions, A and B, varies greatly from just a few in the High-Risk Die-Survive case, to many thousands in the Favorable-Unfavorable case. But because the number of potentially tested elements is extremely large, fifty thousand gene, three hundred thousand transcripts, millions of SNPs, it is necessary to evaluate the number of false positives expected by chance when testing a multiple hypothesis.

There are two methods: theory and simulation. In the theoretical approach, a mathematical model of the experiment is developed and the number of elements exceeding by chance a given threshold is computed ab initio. This is delicate because the premises of the model may not apply: for example, expression of neighboring genes can be correlated, contradicting the implicit hypothesis of independence of the measured elements. In the simulated "Monte-Carlo" approach, the number of false positive is estimated by randomly shuffling the classes and counting the apparent differential features using the exact same protocol as in the real experiment.

Given 2 phenotypic groups A and B of size m and n, we randomly partition N times the A+B cohort into 2 random strata α_I and β_I, also of size m and n, both containing the same proportion m/n of A and B elements and we compare the histogram h(s) of the number of elements with differential score s between A and B to the noise histogram k(s) using the average of the N=10 random strata, rounding each average value to the closest superior integer. The false discovery rate is then evaluated as the ratio of the noise to experiment observations at or above threshold t

$$FDR(t) = H(t) / K(t), \quad \text{with } H(t) = \sum_{s=t}^{200} h(s); \; K(t) = \sum_{s=t}^{200} k(s)$$

For the easy endpoints, we observe thousands of highly reliable differential elements with FDR below 1/1000. For the difficult end points, we only observe a few elements while accepting a FDR around 20%.

The underlying logic in this method is that although the measure of the index of expression of each gene may be noisy, these variations will affect in the same way the A/B experiment and the random strata, so that at low score, most differential elements probably represent random sampling fluctuations. The resampling stratification allows to choose a threshold auto-adapted to the measures in each experiment. Interestingly, the method is applicable even if the classes A and B are of different size (m > n) and even if we do not equalize the random strata relative to external characteristics (propensity sampling). In the worst case, the number of differential elements in the random strata is increased, leading to an overestimation of the false discovery rate.

As an illustration, in the case of MYCN amplified (MNA) versus not amplified differentially expressed transcripts, the system selects a threshold above the noise evaluated by 10 iterations (red curve), yielding a FDR below 10^{-3}.

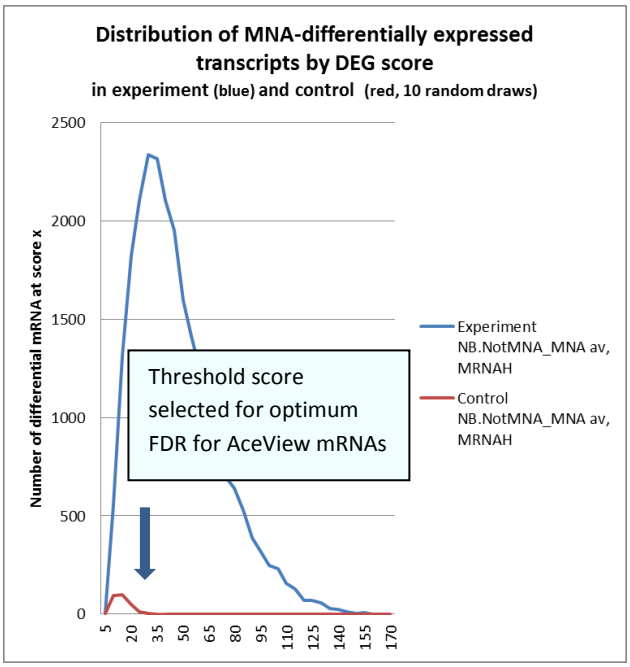


Figure 6: Selection of threshold minimizing FDR while keeping a maximum of differential features, on the example of transcripts expressed less in MNA than in non-MNA. The distribution of transcripts that might be called differential is shown for the experiment (bluish curves) and the controls (10 random draws; red curve low and to the left).