Supplementary tables, figures and information for

A comprehensive Comparisons of RNA-Seq based transcriptome analysis from reads to differential gene expression and cross comparison with microarrays

Intawat Nookaew [1,=], Marta Papini [1,=], Natapol Pornputtpong [1,], Gionata Scalcinati [1], Linn Fagerberg[2], Matthias Uhlén [2,3], Jens Nielsen [1,3,*]

[1] Novo Nordisk Foundation Center for Biosustainability, Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden.

[2] Novo Nordisk Foundation Center for Biosustainability, Department of Biotechnology, Royal Institute of Technology, Stockholm, Sweden

[3] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2970 Hørsholm, Denmark

[=] The authors equally contributed to this work

[*] Corresponding Author: nielsenj@chalmers.se

Table S1 Summary of RNA-seq data generated in this study

Sample	Number of pairend reads (milions)	Number of nucleotides (GB)
Batch r1	5.73	1.15
Batch r2	7.62	1.52
Batch r3	5.57	1.11
Chmostat r1	4.03	0.81
Chmostat r2	6.75	1.35
Chmostat r3	6.16	1.23
All	35.85	7.17
Average	5.97	1.19
sd	1.21	0.24

Table S2 Summary of genetic variations (SNV and INDEL) between *S. cerevisiae* CENPK 113-7D and reference strain S288c. The numbers show the genetic variation on specific region. The numbers in parenthesis represent the number of features containing at least one genetic variation in their ORF, upstream sequence, or probe.

Region	SNV	INDEL
In all chromosomes	28139	3520
1000 bp upstream	15494 (2469)	2704 (1175)
Open reading frame ORF	17139 (2269)	702 (335)
Microarray probe	2868 (2472)	151(119)

Table S3 Mapping statistic of reads using the three aligners based on high quality reads. Bold italic numbers are % value. rm dup = remove potention PCR duplicate, non-rm dup = not remove potential PCR duplicate.

Sample	Number of pairend reads (milions)	Number of high quality pairend reads (milions)	% mapping to reference genome								
			Gsnap			Stampy			TopHat		
			non-rm dup	rm dup	duplicate	non-rm dup	rm dup	duplicate	non-rm dup	rm dup	duplicate
Batch r1	5.73	5.64	***98.07***	***97.77***	***13.73***	***98.16***	***97.84***	***14.49***	***97.49***	***97.04***	***15.72***
Batch r2	7.62	7.51	***98.14***	***97.85***	***13.32***	***98.40***	***98.09***	***16.49***	***99.79***	***99.75***	***17.26***
Batch r3	5.57	5.48	***98.25***	***98.02***	***11.75***	***98.46***	***98.19***	***14.46***	***98.92***	***98.73***	***15.57***
Chmostat r1	4.03	3.97	***95.73***	***95.44***	***6.44***	***95.59***	***95.22***	***7.57***	***95.66***	***95.23***	***9.51***
Chmostat r2	6.75	6.65	***93.91***	***92.66***	***17.14***	***93.40***	***91.67***	***20.81***	***93.25***	***92.40***	***11.98***
Chmostat r3	6.16	6.06	***96.43***	***96.04***	***9.90***	***96.61***	***96.18***	***11.34***	***98.75***	***98.57***	***12.81***
All	35.85	35.30	**34.17**	**34.00**	**4.41**	**34.17**	**33.96**	**5.22**	**34.39**	**34.27**	**4.99**
average	5.97	5.88	***96.76***	***96.29***	***12.05***	***96.77***	***96.20***	***14.19***	***97.31***	***96.95***	***13.81***
sd	1.21	1.20	***1.74***	***2.08***	***3.65***	***2.01***	***2.52***	***4.50***	***2.45***	***2.73***	***2.88***

Table S4 Number of DGE (Q-values < 10e-5) from different methods (for microarray the number DGE is 1603)

Method	Gsnap	N.Gsnap	Stampy	N.Stampy	TopHat	N.TopHat	De novo
Cufdiff	2061	2172	1712	1741	1671	1726	1623
DESeq	2690	2731	2507	2503	2412	2432	2197
edgeR	3087	3161	2732	2742	2649	2673	2385
baySeq	1785	1807	1173	1198	1092	1133	1175
NOISeq*	2097	2070	1837	1784	1804	1754	1595

* use cutoff by probability value > 0.875

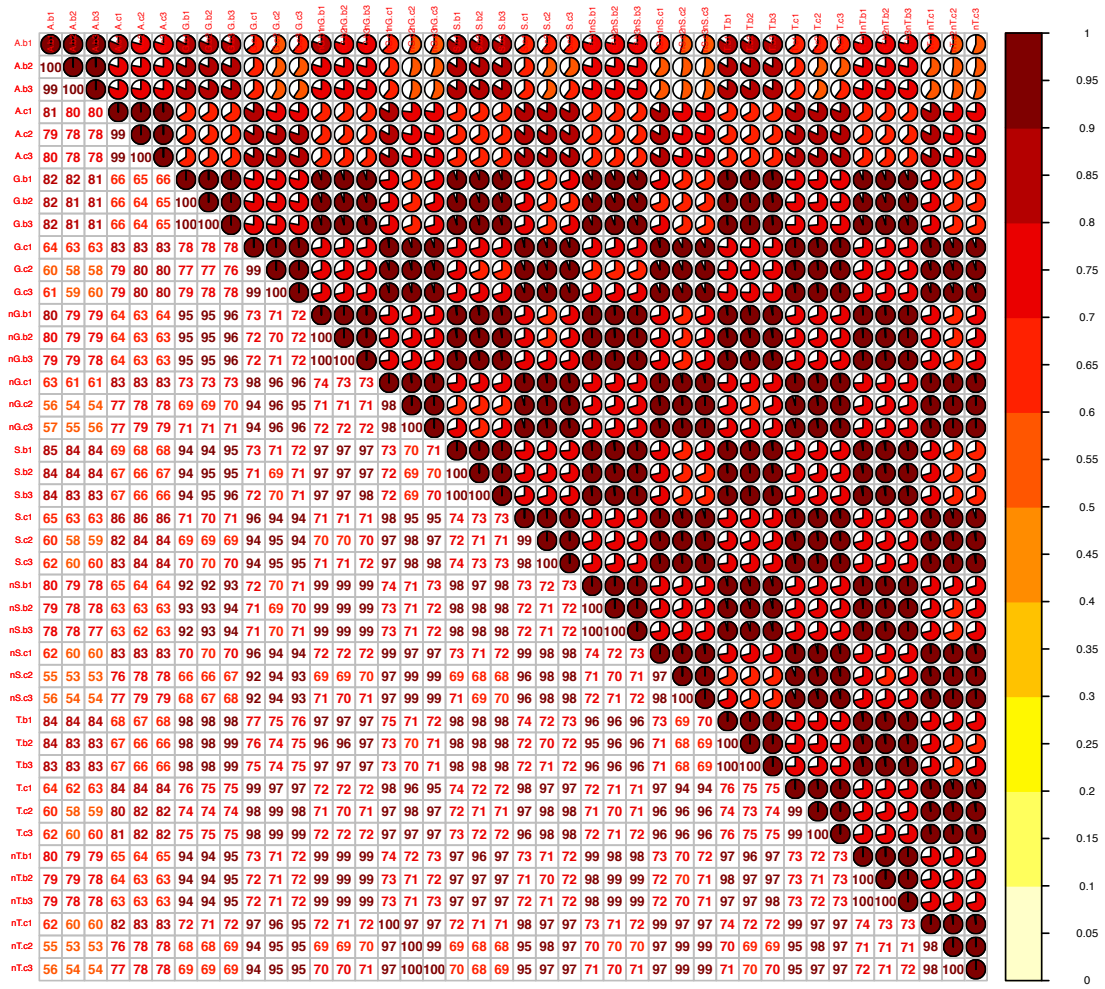

Figure S1 Sample-wise correlation analysis. A = array, G= Gsnap, S = Stampy, T = TopHat. The non-removed potential PCR duplicates calculations indicate by "n."

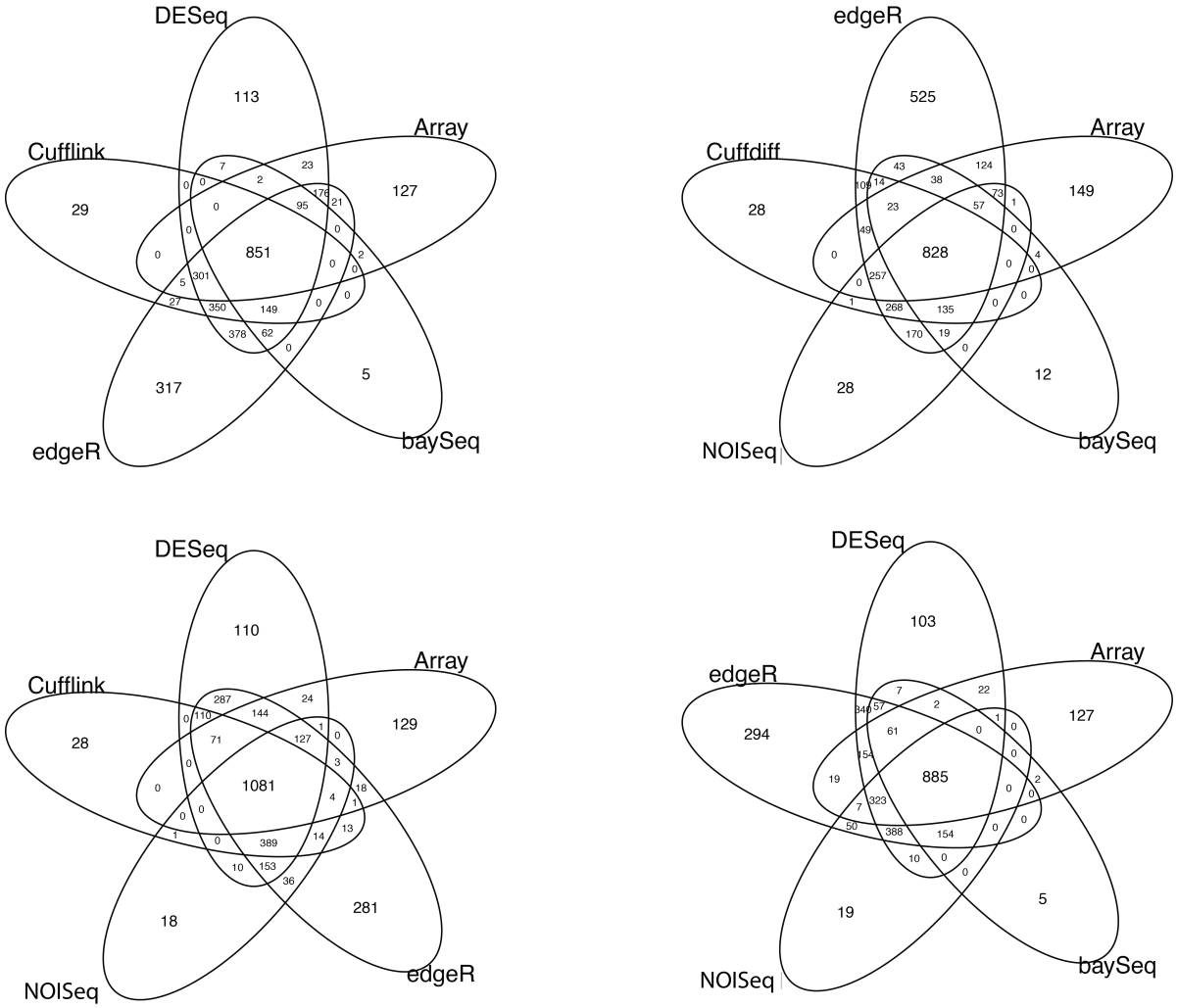

Figure S2 Venn diagram comparison of DGE derived from different statistical analysis with the DGE results from microarray in similar way as Figure 3B

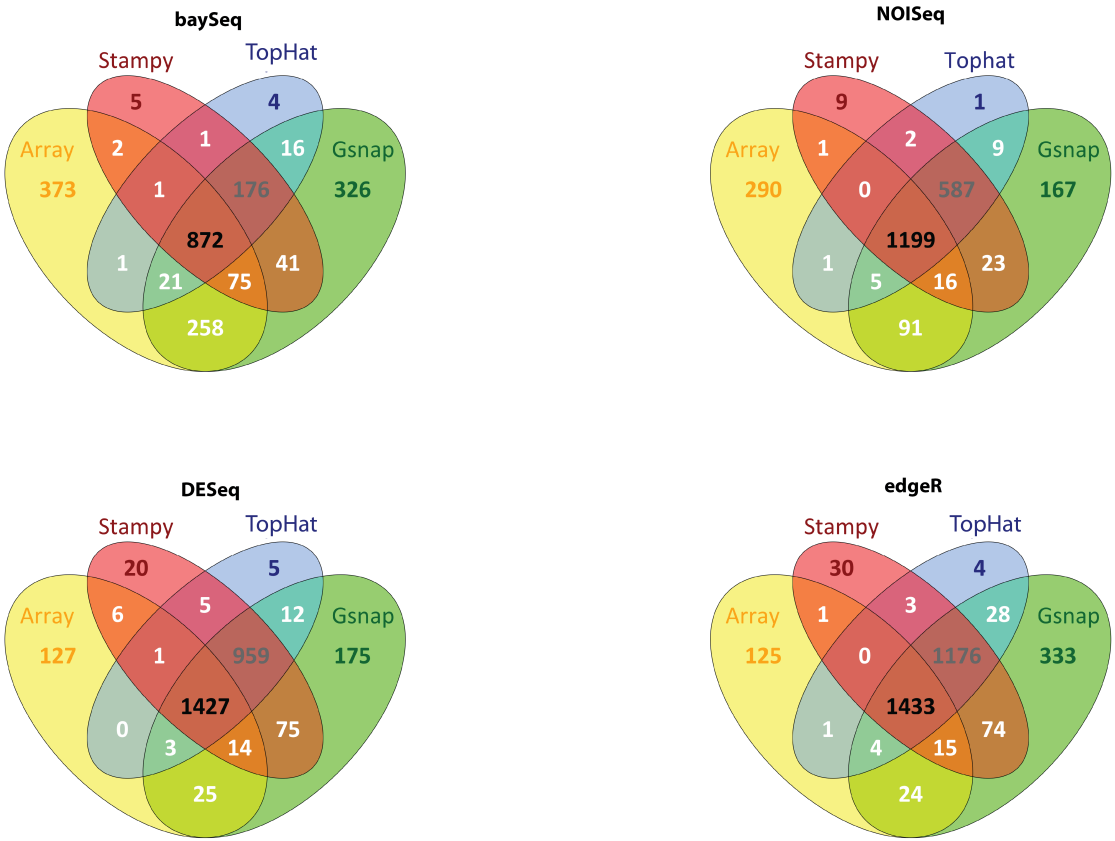

Figure S3 Comparison of DGE using different statistical methods with DGE from microarray in similar way as Figure 3C

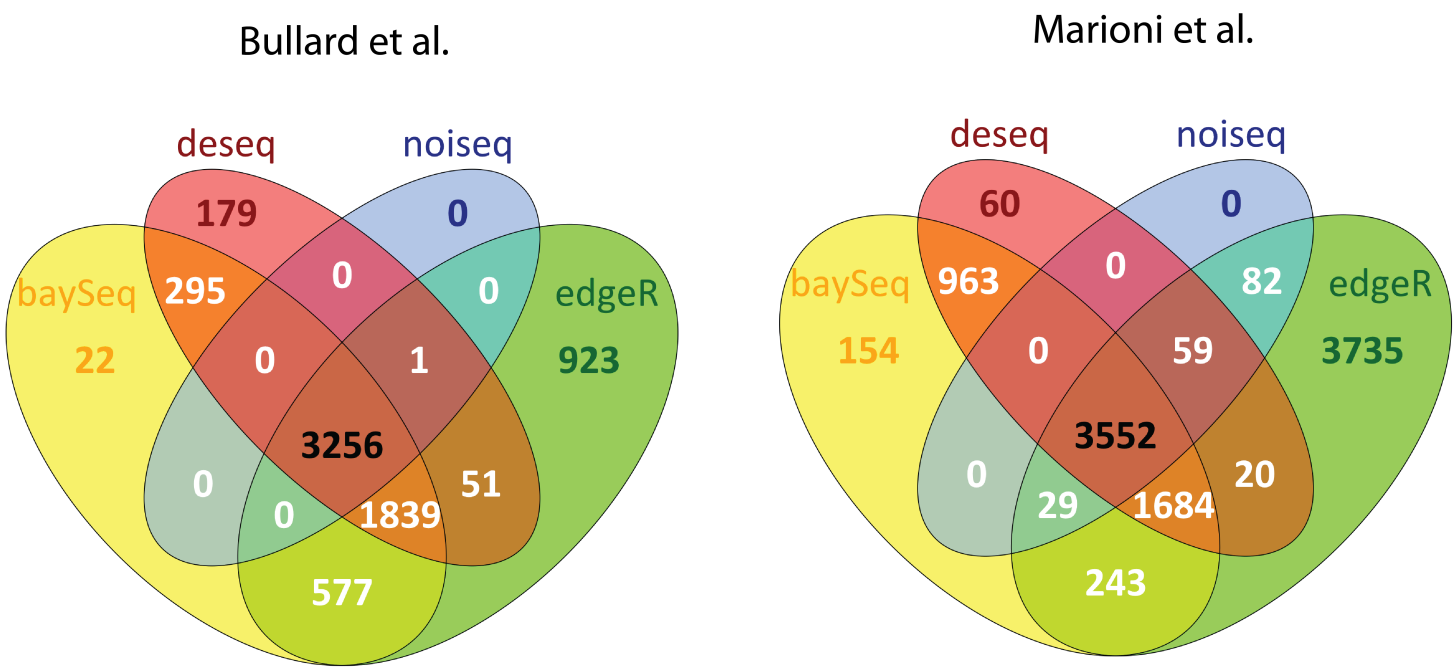

Figure S4 Comparisons of DGE identification using edgeR, baySeq, DESeq and NOISeq methods. The read count table downloaded from 2 published mammalian data sets of Bullard et al (BMC Bioinformatics. 2010 Feb 18;11:94) and Marioni et al. (Genome Res. 2008 Sep;18(9):1509-17.). The Venn's diagram were made with the cut-off Q<1e-5 and Pr > 0.875. Cuffdiff method was not included in the comparison because alignment results (SAM/BAM), which are required as input for Cuffdiff, are not provided with the articles.

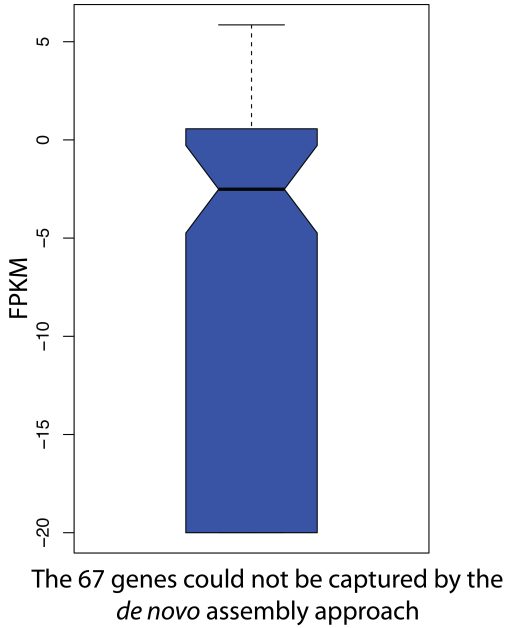

Figure S5 Boxplot shows distribution of gene expression values (FPKM from reference mapping analysis from both batch and chemostat cultures) of the 67 genes could not be captured by the *de novo* assembly approach as full-length transcripts.

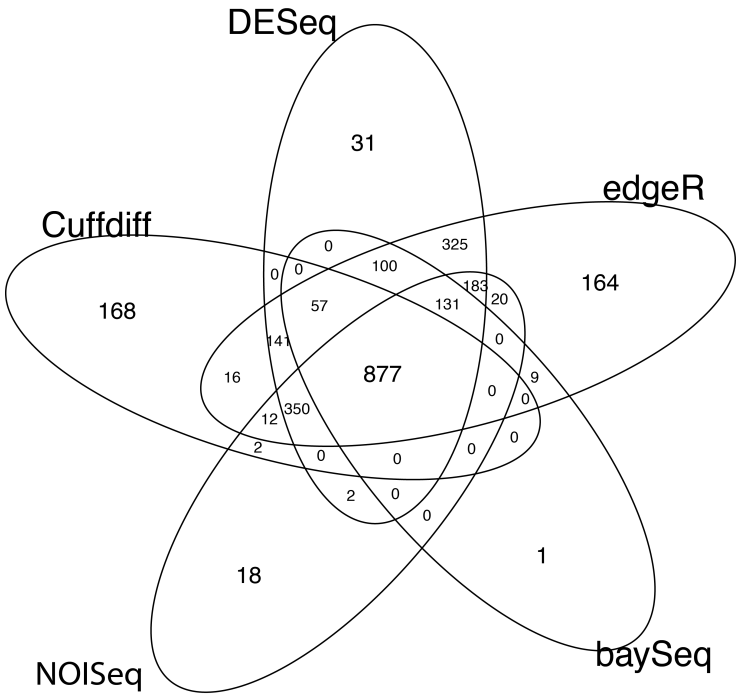

Figure S6 Venn's diagram of the comparison of differential gene expression based on RNA-seq data (result from de novo assembly approach) through five different statistical methods: Cuffdiff, DESeq, NOISeq, edgeR and baySeq.

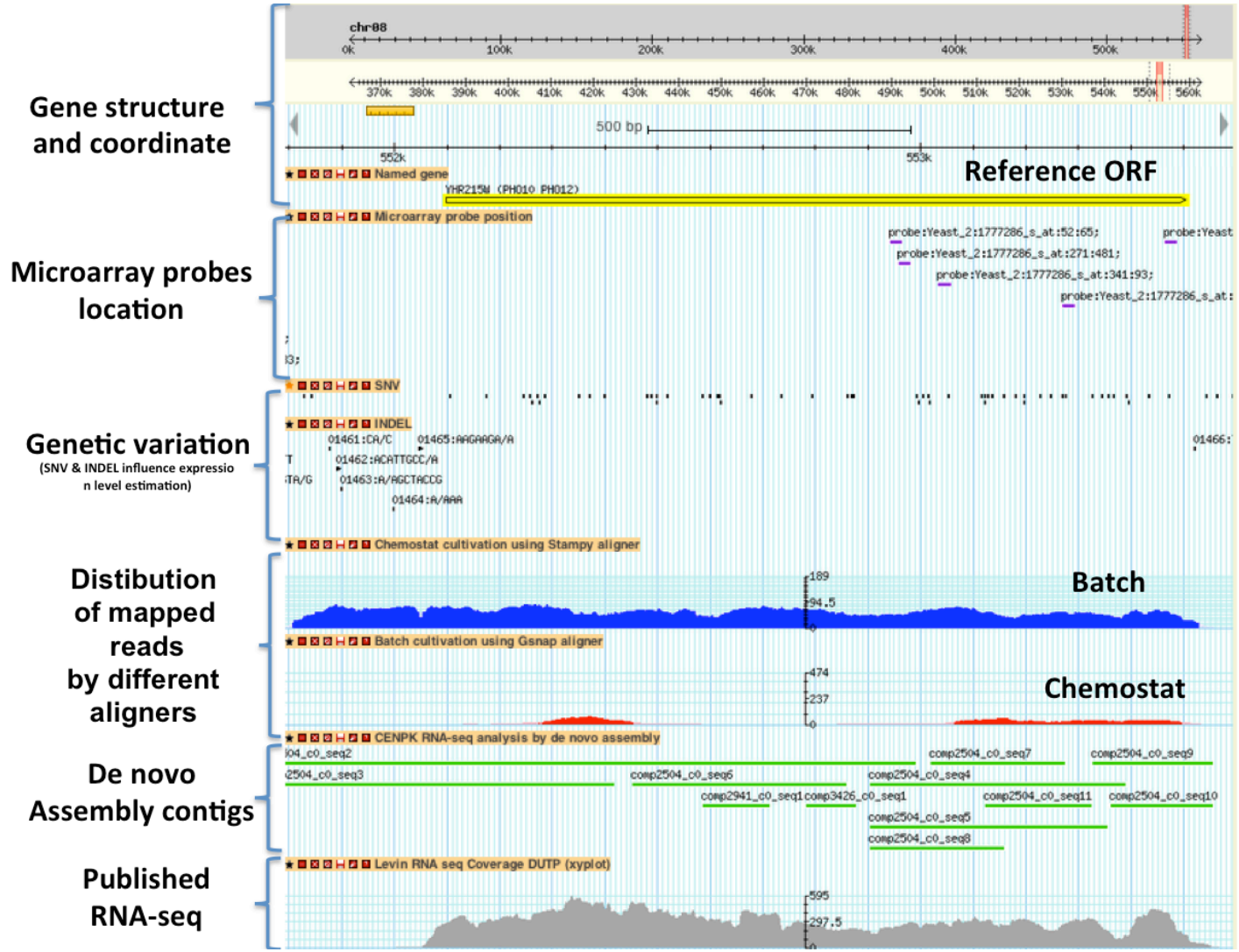

Figure S7 Screen shot of the Yseq browser (www.sysbio.se/Yseq)

Supplementary information

Alignment parameters

Gsnap
gsnap -d **databsename** -D **genomedatabase_location** -A **sam fastq.1 fastq.2** > **outfile**

Stampy
stampy.py -f **sam --solexa** -o **outfile** -g **genomename** -h **genomehash** -M **fastq.1 fastq.2**

TopHat
tophat -o **outfile indexfile fastq.1 fastq.2**

Calculate FPKM
cufflink -o **outfile** –G **gfffile** -b **DNAfastafiles bamfile**

Generate read count table by HTseq
htseq-count -m **intersection-strict** -s **no** -t **gene** -i **ID samfile gff_file** > **outfile**

Remove PCR duplicate
java -jar MarkDuplicates.jar I=**input** O=**output** REMOVE_DUPLICATES=**true**

De novo assembly
Trinity.pl --seqType fq --jaccard_clip --kmer_method inchworm --CPU 8 --output **outfile** --left **infile_1** --right **infile_2**

Statistical analysis parameters

X = count table
baySeq
```
basyseq.test = function(x){
repl = c(rep(1,3),rep(2,3))
```

```
groups <- list(NDE = c(1,1,1,1,1,1), DE = c(1,1,1,2,2,2))
CD <- new("countData", data =as.matrix(x), replicates = repl, groups = groups)
CD@libsizes <- getLibsizes(CD)
CDP.NBML <- getPriors.NB(CD, samplesize = 1e6, estimation = "QL", cl = cl)
CDPost.NBML <- getLikelihoods.NB(CDP.NBML, pET = 'BIC', cl = cl)
return(CDPost.NBML)
}
```

DESeq

```
deseq.test = function(x){
conds = c("b","b","b","c","c","c")
cds<-newCountDataSet(x,conds)
cds<-estimateSizeFactors(cds)
sizeFactors(cds)
cds <- estimateDispersions (cds)
res <- nbinomTest (cds, "b", "c")
return(res)
}
```

edgeR

```
edgeR.test = function(x){
group = c("b","b","b","c","c","c")
dat = DGEList(counts=x,group=group)
design <- model.matrix(~factor(group))
dat <- estimateGLMTrendedDisp(dat,design)
dat <- estimateGLMTagwiseDisp(dat,design)
fit <- glmFit(dat,design)
stat.Glm <- glmLRT(dat,fit)
return(stat.Glm)
}
```

NOISeq

```
noiseq.test = function(x) {
mydata <- vector("list", length = 2)
mynames <- rownames(x)
mydata[[1]] <- as.matrix(x[,1:3])
mydata[[2]] <- as.matrix(x[,4:6])
rownames(mydata[[1]]) <- rownames(mydata[[2]]) <- mynames
myresults <- noiseq(mydata[[1]], mydata[[2]], repl = "bio", k = 0.5, norm = "rpkm", long = mylength)
}
```

Cufdiff

cuffdiff -o **outfile gfffile** -b **DNAfastafiles b1.bam,b2.bam,b3.bam c1.bam,c2.bam, c3.bam**

Software version

FASTX ver. 0.0.13
SolexaQA ver. 1.10
ATLAS2 ver. 1.0
Bowtie ver 0.12.7
BEDTools ver. 2.14.3
SAMTool ver 0.1.18
Gsnap ver. 2012-01-11
GMAP ver. 2012-01-11
Stampy ver. 1.0.13
TopHat ver. 1.4.1
Cufflink/Cuffdiff ver 1.3.0
HTseq ver. 0.5.3p3
Trinity ver. R2012-01-25
baySeq ver. 1.8.3
DESeq ver. 1.6.1
edgeR ver. 2.4.6
NOISeq ver. 2011/04/9