Lightweight RNAseq analysis with BioConductor

A. Lesniewska[1,2] M.J. Okoniewski[2]

[1] Institute of Computer Science
Poznan University of Technology, Poland

[2] Functional Genomics Center
UNI ETH Zurich, Switzerland

Bioconductor Developer Meeting Europe - 17-18. 11. 2010

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

State of the technology

RNA-seq

- The coverage of `SOLID` starts to be enough to run whole transcriptomes RNAseq for higher species.
- 300-900M of reads per run
- Mapping is being constantly improved

State of the technology

RNA-seq

- The coverage of `SOLID` starts to be enough to run whole transcriptomes RNAseq for higher species.
- 300-900M of reads per run
- Mapping is being constantly improved

State of the technology

RNA-seq

- The coverage of `SOLID` starts to be enough to run whole transcriptomes RNAseq for higher species.
- 300-900M of reads per run
- Mapping is being constantly improved

Our assumptions

- We can use `database storage`
- Recent improvements in DB engines allow fast access: `indexing, partitioning`
- R as the analysis environment – good statistics, comparison to microarrays
- BioConductor library as the way of publishing the analytical API

Our assumptions

- We can use `database storage`
- Recent improvements in DB engines allow fast access: `indexing, partitioning`
- R as the analysis environment – good statistics, comparison to microarrays
- BioConductor library as the way of publishing the analytical API

Our assumptions

- We can use `database storage`
- Recent improvements in DB engines allow fast access: `indexing, partitioning`
- R as the analysis environment – good statistics, comparison to microarrays
- BioConductor library as the way of publishing the analytical API

Our assumptions

- We can use `database storage`
- Recent improvements in DB engines allow fast access: `indexing, partitioning`
- R as the analysis environment – good statistics, comparison to microarrays
- BioConductor library as the way of publishing the analytical API

Our assumptions

- We can use database storage
- Recent improvements in DB engines allow fast access: indexing, partitioning
- R as the analysis environment – good statistics, comparison to microarrays
- BioConductor library as the way of publishing the analytical API

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

Exonmap paradigms

- Database for accessing the annotations
- Gene or a group at a time – not everything
- Translation of `genes<->transcripts<->exons`
- Filtering of interesting genes and exons
- Splicing analyses and plots

It worked => Let's do the same for RNAseq…

Exonmap paradigms

- Database for accessing the annotations
- Gene or a group at a time – not everything
- Translation of `genes<->transcripts<->exons`
- Filtering of interesting genes and exons
- Splicing analyses and plots

It worked => Let's do the same for RNAseq...

Exonmap paradigms

- Database for accessing the annotations
- Gene or a group at a time – not everything
- Translation of `genes<->transcripts<->exons`
- Filtering of interesting genes and exons
- Splicing analyses and plots

It worked => Let's do the same for RNAseq. . .

Exonmap paradigms

- Database for accessing the annotations
- Gene or a group at a time – not everything
- Translation of `genes<->transcripts<->exons`
- Filtering of interesting genes and exons
- Splicing analyses and plots

It worked => Let's do the same for RNAseq. . .

Exonmap paradigms

- Database for accessing the annotations
- Gene or a group at a time – not everything
- Translation of `genes<->transcripts<->exons`
- Filtering of interesting genes and exons
- Splicing analyses and plots

It worked => Let's do the same for RNAseq. . .

Exonmap paradigms

- Database for accessing the annotations
- Gene or a group at a time – not everything
- Translation of `genes<->transcripts<->exons`
- Filtering of interesting genes and exons
- Splicing analyses and plots

It worked => Let's do the same for RNAseq…

Exonmap paradigms

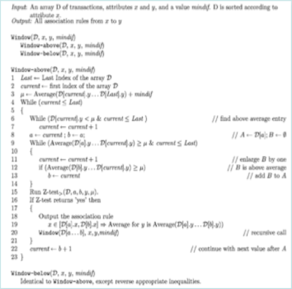
- Database for accessing the annotations
- Gene or a group at a time – not everything
- Translation of `genes<->transcripts<->exons`
- Filtering of interesting genes and exons
- Splicing analyses and plots

It worked => Let’s do the same for RNAseq. . .

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

Lindell&Aumann window algorithm

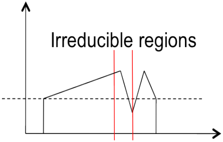


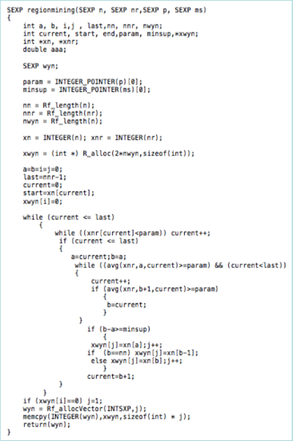
Input: An array D of transactions, attributes x and y, and a value *mindif*. D is sorted according to attribute x.
Output: All association rules from x to y

```
Window($\mathcal{D}$, $x$, $y$, mindif)
    Window-above($\mathcal{D}$, $x$, $y$, mindif)
    Window-below($\mathcal{D}$, $x$, $y$, mindif)

Window-above($\mathcal{D}$, $x$, $y$, mindif)
Last ← Last Index of the array $\mathcal{D}$
current ← first Index of the array $\mathcal{D}$
$\mu \leftarrow$ Average($\mathcal{D}[current].y \ldots \mathcal{D}[Last].y$) + mindif
While ($current \leq Last$)
{
    While ($\mathcal{D}[current].y < \mu$ & $current \leq Last$ )      // find above average entry
        current ← current + 1
    $a \leftarrow current$ ; $b \leftarrow a$;                            // $A \leftarrow \mathcal{D}[a]$; $B \leftarrow \emptyset$
    While (Average($\mathcal{D}[a].y \ldots \mathcal{D}[current].y$) $\geq \mu$ & $current \leq Last$)
    {
        current ← current + 1                                   // enlarge $B$ by one
        if (Average($\mathcal{D}[b].y \ldots \mathcal{D}[current].y$) $\geq \mu$)  // $B$ is above average
            $b \leftarrow current$                                      // add $B$ to $A$
    }
    Run Z-test$_>$($\mathcal{D}, a, b, y, \mu$).
    If Z-test returns 'yes' then
    {
        Output the association rule
            $x \in [\mathcal{D}[a].x, \mathcal{D}[b].x] \Rightarrow$ Average for $y$ is Average($\mathcal{D}[a].y \ldots \mathcal{D}[b].y$))
        Window($\mathcal{D}[a \ldots b]$, $x$, $y$,mindif)                // recursive call
    }
    current ← $b + 1$                                           // continue with next value after $A$
}

Window-below($\mathcal{D}$, $x$, $y$, mindif)
    Identical to Window-above, except reverse appropriate inequalities.
```





```
SEXP regionmining(SEXP n, SEXP nr,SEXP p, SEXP ms)
{
    int a, b, i,j , last,nn, nnr, nwyn;
    int current, start, end,param, minsup,*xwyn;
    int *xn, *xnr;
    double aaa;

    SEXP wyn;

    param = INTEGER_POINTER(p)[0];
    minsup = INTEGER_POINTER(ms)[0];

    nn = Rf_length(n);
    nnr = Rf_length(nr);
    nwyn = Rf_length(n);

    xn = INTEGER(n); xnr = INTEGER(nr);

    xwyn = (int *) R_alloc(2*nwyn,sizeof(int));

    a=b=i=j=0;
    last=nnr-1;
    current=0;
    start=xn[current];
    xwyn[i]=0;

    while (current <= last)
        {
            while ((xnr[current]<param)) current++;
             if (current <= last)
             {
                a=current;b=a;
                 while ((avg(xnr,a,current)>=param) && (current<last))
                 {
                     current++;
                     if (avg(xnr,b+1,current)>=param)
                        {
                         b=current;
                        }
                  }
                    if (b-a>=minsup)
                        {
                     xwyn[j]=xn[a];j++;
                    if  (b==nn) xwyn[j]=xn[b-1];
                     else xwyn[j]=xn[b];j++;
                        }
                    current=b+1;
             }
        }
    if (xwyn[i]==0) j=1;
    wyn = Rf_allocVector(INTSXP,j);
    memcpy(INTEGER(wyn),xwyn,sizeof(int) * j);
    return(wyn);
}
```

Figure: algorithm & implementation

Lindell&Aumann window algorithm

- Linear complexity
- Finds irreducible regions
- Applicable directly to coverage on genome data
- Follows biological intuitions
- Biological interpretation of consistent “exonic” region

Lindell&Aumann window algorithm

- Linear complexity
- Finds irreducible regions
- Applicable directly to coverage on genome data
- Follows biological intuitions
- Biological interpretation of consistent “exonic” region

Lindell&Aumann window algorithm

- Linear complexity
- Finds irreducible regions
- Applicable directly to coverage on genome data
- Follows biological intuitions
- Biological interpretation of consistent “exonic” region

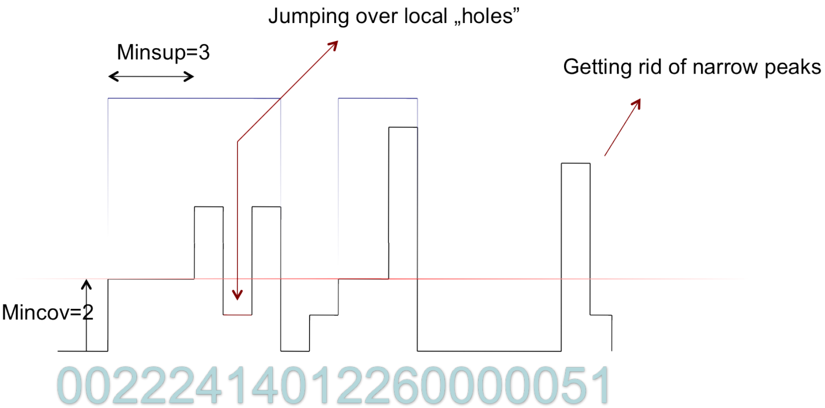
Lindell&Aumann window algorithm

- Linear complexity
- Finds irreducible regions
- Applicable directly to coverage on genome data
- Follows biological intuitions
- Biological interpretation of consistent “exonic” region

Lindell&Aumann window algorithm

- Linear complexity
- Finds irreducible regions
- Applicable directly to coverage on genome data
- Follows biological intuitions
- Biological interpretation of consistent “exonic” region

Irreducible region



Jumping over local „holes”

Minsup=3

Getting rid of narrow peaks

Mincov=2

0022224140122600000051

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

How it works?

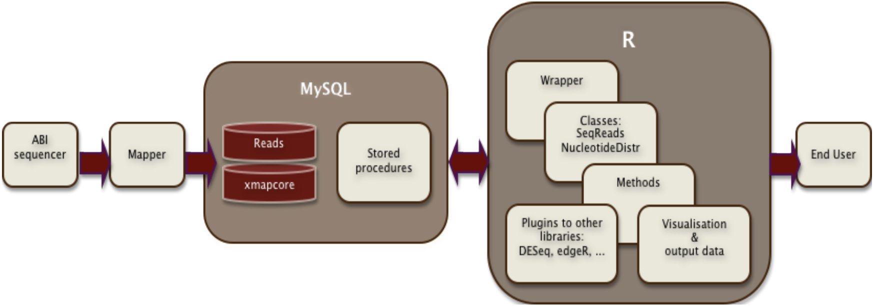

ABI sequencer
Mapper
MySQL
Reads
xmapcore
Stored procedures
R
Wrapper
Classes: SeqReads NucleotideDistr
Methods
Plugins to other libraries: DESeq, edgeR, ...
Visualisation & output data
End User

Figure: The flow of RNA seq data processing in the xmapcore database and the rnaSeqMap library.

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

Data cleaning and preparation

- Libraries prepared and sequenced
- Raw data files transferred
- Colorspace reads mapped
- Samtools
- AWK script to get the simple, but biiiiig tables
- Import into MySQL

Data cleaning and preparation

- Libraries prepared and sequenced
- Raw data files transferred
- Colorspace reads mapped
- Samtools
- AWK script to get the simple, but biiiiig tables
- Import into MySQL

Data cleaning and preparation

- Libraries prepared and sequenced
- Raw data files transferred
- Colorspace reads mapped
- Samtools
- AWK script to get the simple, but biiiiig tables
- Import into MySQL

Data cleaning and preparation

- Libraries prepared and sequenced
- Raw data files transferred
- Colorspace reads mapped
- Samtools
- AWK script to get the simple, but biiiiig tables
- Import into MySQL

Data cleaning and preparation

- Libraries prepared and sequenced
- Raw data files transferred
- Colorspace reads mapped
- Samtools
- AWK script to get the simple, but biiiiig tables
- Import into MySQL

Data cleaning and preparation

- Libraries prepared and sequenced
- Raw data files transferred
- Colorspace reads mapped
- Samtools
- AWK script to get the simple, but biiiiig tables
- Import into MySQL

Database back-end

- MySQL >= 5.1
- Xmapcore database (denormalized Ensembl)
- Seq_reads table – with experiment number and genome coordinates of each read
- Indexed
- Partitioned into chromosome
- Average genome range query: 30s laptop, 5s fgcz-s-024

Database back-end

- MySQL >= 5.1
- Xmapcore database (denormalized Ensembl)
- Seq_reads table – with experiment number and genome coordinates of each read
- Indexed
- Partitioned into chromosome
- Average genome range query: 30s laptop, 5s fgcz-s-024

Database back-end

- MySQL >= 5.1
- Xmapcore database (denormalized Ensembl)
- Seq_reads table – with experiment number and genome coordinates of each read
- Indexed
- Partitioned into chromosome
- Average genome range query: 30s laptop, 5s fgcz-s-024

Database back-end

- MySQL >= 5.1
- Xmapcore database (denormalized Ensembl)
- Seq_reads table – with experiment number and genome coordinates of each read
- Indexed
- Partitioned into chromosome
- Average genome range query: 30s laptop, 5s fgcz-s-024

Database back-end

- MySQL >= 5.1
- Xmapcore database (denormalized Ensembl)
- Seq_reads table – with experiment number and genome coordinates of each read
- Indexed
- Partitioned into chromosome
- Average genome range query: 30s laptop, 5s fgcz-s-024

Database back-end

- MySQL >= 5.1
- Xmapcore database (denormalized Ensembl)
- Seq_reads table – with experiment number and genome coordinates of each read
- Indexed
- Partitioned into chromosome
- Average genome range query: 30s laptop, 5s fgcz-s-024

Database back-end

Databases:

- xmapcore
- or basic (3 tables gene,trenscript,exon) in xmapcore-like format
- maybe easily produced from non-Ensembl annotation for rare-species

Database back-end

Databases:

- xmapcore
- or basic (3 tables gene,trenscript,exon) in xmapcore-like format
- maybe easily produced from non-Ensembl annotation for rare-species

Database back-end

Databases:

- xmapcore
- or basic (3 tables gene,trenscript,exon) in xmapcore-like format
- maybe easily produced from non-Ensembl annotation for rare-species

Stored procedures

- Region reads in given sample
- `Gene <-> Transcript <-> Exon <-> reads`
- Genes on a chromosome
- Intergenic regions on a chromosome

Stored procedures

- Region reads in given sample
- `Gene <-> Transcript <-> Exon <-> reads`
- Genes on a chromosome
- Intergenic regions on a chromosome

Stored procedures

- Region reads in given sample
- `Gene <-> Transcript <-> Exon <-> reads`
- Genes on a chromosome
- Intergenic regions on a chromosome

Stored procedures

- Region reads in given sample
- `Gene <-> Transcript <-> Exon <-> reads`
- Genes on a chromosome
- Intergenic regions on a chromosome

Classes in R

- SeqReads – a collection of reads for samples in a given genomic region
- NucleotideDistribution (S3 class) – nucleotide by nucleotide distribution of measured feature
  - Coverage of reads
  - Fold change
  - Splicing Index
  - Significant regions

Classes in R

- SeqReads – a collection of reads for samples in a given genomic region
- NucleotideDistribution (S3 class) – nucleotide by nucleotide distribution of measured feature
    - Coverage of reads
    - Fold change
    - Splicing Index
    - Significant regions

Classes in R

- SeqReads – a collection of reads for samples in a given genomic region
- NucleotideDistribution (S3 class) – nucleotide by nucleotide distribution of measured feature
    - Coverage of reads
    - Fold change
    - Splicing Index
    - Significant regions

Classes in R

- SeqReads – a collection of reads for samples in a given genomic region
- NucleotideDistribution (S3 class) – nucleotide by nucleotide distribution of measured feature
  - Coverage of reads
  - Fold change
  - Splicing Index
  - Significant regions

Classes in R

- SeqReads – a collection of reads for samples in a given genomic region
- NucleotideDistribution (S3 class) – nucleotide by nucleotide distribution of measured feature
  - Coverage of reads
  - Fold change
  - Splicing Index
  - Significant regions

Interesting genes

- Good coverage
- Good coverage of exons
- Interesting splicing index
- Interesting new regions – novel exons

More algorithms to establish within the framework!!

Interesting genes

- Good coverage
- Good coverage of exons
- Interesting splicing index
- Interesting new regions – novel exons

More algorithms to establish within the framework!!

Interesting genes

- Good coverage
- Good coverage of exons
- Interesting splicing index
- Interesting new regions – novel exons

More algorithms to establish within the framework!!

Interesting genes

- Good coverage
- Good coverage of exons
- Interesting splicing index
- Interesting new regions – novel exons

More algorithms to establish within the framework!!

Interesting genes

- Good coverage
- Good coverage of exons
- Interesting splicing index
- Interesting new regions – novel exons

More algorithms to establish within the framework!!

Interesting intergenic regions

- Irreducible regions with good coverage
- We treat them as novel genes and run gene-style analysis
- Looking for exons

More algorithms to establish within the framework!!

Interesting intergenic regions

- Irreducible regions with good coverage
- We treat them as novel genes and run gene-style analysis
- Looking for exons

More algorithms to establish within the framework!!

Interesting intergenic regions

- Irreducible regions with good coverage
- We treat them as novel genes and run gene-style analysis
- Looking for exons

More algorithms to establish within the framework!!

Interesting intergenic regions

- Irreducible regions with good coverage
- We treat them as novel genes and run gene-style analysis
- Looking for exons

More algorithms to establish within the framework!!

Output

- Iranges objects – for interesting regions
- DESeq object – gene/exon level expression - for the significance analysis with DESeq
- Lists of interesting features

Output

- Iranges objects – for interesting regions
- DESeq object – gene/exon level expression - for the significance analysis with DESeq
- Lists of interesting features

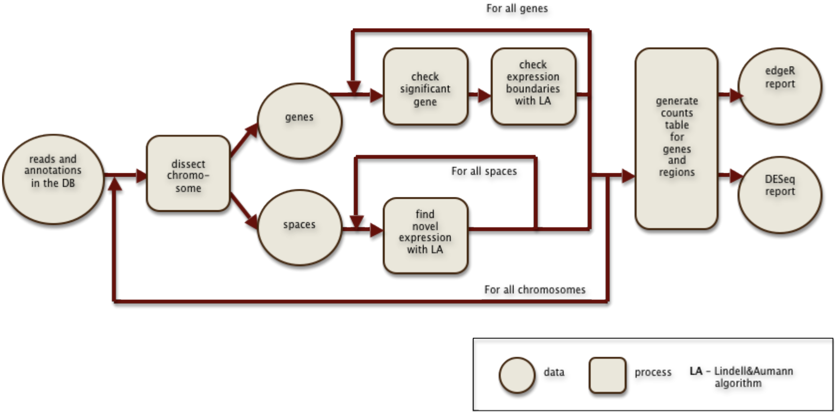
Output

- Iranges objects – for interesting regions
- DESeq object – gene/exon level expression - for the significance analysis with DESeq
- Lists of interesting features

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

An example of rnaSeqMap analysis pipeline


For all genes
reads and annotations in the DB
dissect chromo-some
genes
spaces
check significant gene
check expression boundaries with LA
For all spaces
find novel expression with LA
For all chromosomes
generate counts table for genes and regions
edgeR report
DESeq report
data
process
LA - Lindell&Aumann algorithm

Analysis pipelines

- Get all the genes from a chromosome
  - Check for interesting features
  - Check possible gene extensions – expression closely around the gene
- Get all the intergenic regions on chromosome
  - Find novel expressed regions
  - Describe the regions

Analysis pipelines

- Get all the genes from a chromosome
  - Check for interesting features
  - Check possible gene extensions – expression closely around the gene
- Get all the intergenic regions on chromosome
  - Find novel expressed regions
  - Describe the regions

Analysis pipelines

- Get all the genes from a chromosome
  - Check for interesting features
  - Check possible gene extensions – expression closely around the gene
- Get all the intergenic regions on chromosome
  - Find novel expressed regions
  - Describe the regions

Analysis pipelines

- Get all the genes from a chromosome
  - Check for interesting features
  - Check possible gene extensions – expression closely around the gene
- Get all the intergenic regions on chromosome
  - Find novel expressed regions
  - Describe the regions

Analysis pipelines

- Get all the genes from a chromosome
  - Check for interesting features
  - Check possible gene extensions – expression closely around the gene
- Get all the intergenic regions on chromosome
  - Find novel expressed regions
  - Describe the regions

Analysis pipelines

- Get all the genes from a chromosome
  - Check for interesting features
  - Check possible gene extensions – expression closely around the gene
- Get all the intergenic regions on chromosome
  - Find novel expressed regions
  - Describe the regions

Analysis pipelines - code

```
test.gene<-function(g,exps,nsums,mi,ms)
{
 rs <- newSeqReadsFromGene(g)
 rs <- addExperimentsToReadset(rs,exps)
 nd.cov <- getCoverageFromRS(rs,exps)
 nd.cov <- normalizeBySum(nd.cov, nsums)
 nd.reg <- findRegionsAsND(nd.cov,as.int(mi),ms=ms)
 ir.reg <- findRegionsAsIR(nd.cov,as.int(mi),ms=ms)
 cat("region search algorithm...\n")
 out <- g
 out <- c(out, apply(distribs(nd.cov),2,max))
 out <- c(out, apply(distribs(nd.cov),2,mean))
 out <- c(out, apply(distribs(nd.reg),2,max))
}
```

Analysis pipelines - code

```
test.space<-function(exps,ch,st,en,str,nsums,mi,ms)
{
g.ch <- rnaSeqMap:::.chromosome.number(ch)
rs <- newSeqReads(g.ch,st,en,str)
rs <- addExperimentsToReadset(rs,exps)
nd.cov <- getCoverageFromRS(rs,exps
nd.cov <- normalizeBySum(nd.cov, nsums)
nd.reg <- findRegionsAsND(nd.cov,as.int(mi),ms=ms)
out <- c(ch,st, en, str)
out <- c(out, apply(distribs(nd.cov),2,max))
out <- c(out, apply(distribs(nd.cov),2,mean))
out <- c(out, apply(distribs(nd.reg),2,max))
}
```

Analysis pipelines - code

```
my.genes<-geneInChromosome(22, 200000, 204000,1)
my.spaces<-spaceInChromosome(22, 200000, 204000,1)
    interesting.genes <- NULL
for (i in 1:length(my.genes))
{     cat ("Running gene ", i , "----------\n")
    interesting.genes <- rbind(interesting.genes,
      test.gene(my.genes[i], 1:6, nsums))}
   interesting.spaces <- NULL
for (i in 1:(dim(my.spaces))[1])
{      cat ("Running space ", i , "----------\n")
    interesting.spaces <- rbind(interesting.spaces,
       test.space(1:2, 22,my.spaces[i,1],
       my.spaces[i,2],my.spaces[i,3] ))}
```

Advantages of rnaSeqMap

- Complex analysis of huge data on a small machine - awk, MySQL, R do not have big requirements
- Flexible and fine-grained approach to transcriptomics
  - Not a single nucleotide can hide, if it is expressed
  - Flexible boundaries of expression regions – we rely on Ensembl, but do not have to trust it blindly

Advantages of rnaSeqMap

- Complex analysis of huge data on a small machine - awk, MySQL, R do not have big requirements
- Flexible and fine-grained approach to transcriptomics
  - Not a single nucleotide can hide, if it is expressed
  - Flexible boundaries of expression regions – we rely on Ensembl, but do not have to trust it blindly

Advantages of rnaSeqMap

- Complex analysis of huge data on a small machine - awk, MySQL, R do not have big requirements
- Flexible and fine-grained approach to transcriptomics
  - Not a single nucleotide can hide, if it is expressed
  - Flexible boundaries of expression regions – we rely on Ensembl, but do not have to trust it blindly

Challenges

- Size and allocation of RAM memory to run big regions – we have to run one chromosome at a time
- Speed of queries for reads data – not bad now
- Speed of analysis – optimized by rewriting in C
- Installation is not simple – but still simpler than many other systems

Challenges

- Size and allocation of RAM memory to run big regions – we have to run one chromosome at a time
- Speed of queries for reads data – not bad now
- Speed of analysis – optimized by rewriting in C
- Installation is not simple – but still simpler than many other systems

Challenges

- Size and allocation of RAM memory to run big regions – we have to run one chromosome at a time
- Speed of queries for reads data – not bad now
- Speed of analysis – optimized by rewriting in C
- Installation is not simple – but still simpler than many other systems

Challenges

- Size and allocation of RAM memory to run big regions – we have to run one chromosome at a time
- Speed of queries for reads data – not bad now
- Speed of analysis – optimized by rewriting in C
- Installation is not simple – but still simpler than many other systems

Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

Numeric results

- In total of 38546 genes and pseudogenes, there are:
  - 6863 genes with expression regions >10 for all 6 patients
  - 24172 genes with expression >10 at least for one patient
  - 14375 genes with no irreducible regions >10 in any patient
  - 9912 genes with at least 100 reads mapped in total in 6 samples
  - 5822 genes with no reads at all

Similar to detection on microarrays, however coverage is still too low to detect splicing in most cases. . .

Numeric results

- In total of 38546 genes and pseudogenes, there are:
  - 6863 genes with expression regions >10 for all 6 patients
  - 24172 genes with expression >10 at least for one patient
  - 14375 genes with no irreducible regions >10 in any patient
  - 9912 genes with at least 100 reads mapped in total in 6 samples
  - 5822 genes with no reads at all

Similar to detection on microarrays, however coverage is still too low to detect splicing in most cases. . .

Numeric results

- In total of 38546 genes and pseudogenes, there are:
  - 6863 genes with expression regions >10 for all 6 patients
  - 24172 genes with expression >10 at least for one patient
  - 14375 genes with no irreducible regions >10 in any patient
  - 9912 genes with at least 100 reads mapped in total in 6 samples
  - 5822 genes with no reads at all

Similar to detection on microarrays, however coverage is still too low to detect splicing in most cases. . .

Numeric results

- In total of 38546 genes and pseudogenes, there are:
  - 6863 genes with expression regions >10 for all 6 patients
  - 24172 genes with expression >10 at least for one patient
  - 14375 genes with no irreducible regions >10 in any patient
  - 9912 genes with at least 100 reads mapped in total in 6 samples
  - 5822 genes with no reads at all

Similar to detection on microarrays, however coverage is still too low to detect splicing in most cases. . .

Numeric results

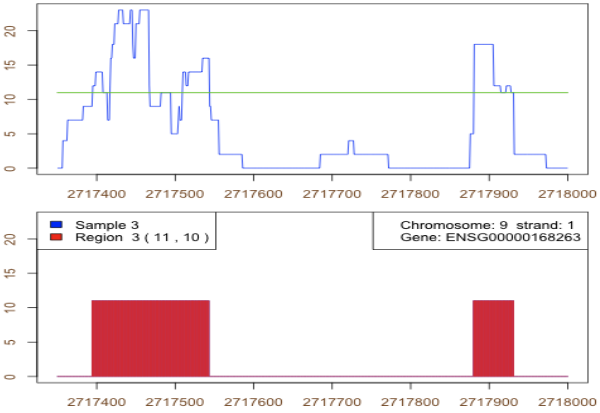
- In total of 38546 genes and pseudogenes, there are:
  - 6863 genes with expression regions >10 for all 6 patients
  - 24172 genes with expression >10 at least for one patient
  - 14375 genes with no irreducible regions >10 in any patient
  - 9912 genes with at least 100 reads mapped in total in 6 samples
  - 5822 genes with no reads at all

Similar to detection on microarrays, however coverage is still too low to detect splicing in most cases. . .

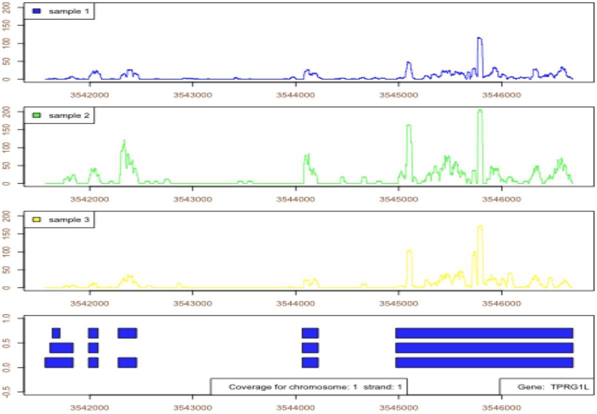
Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

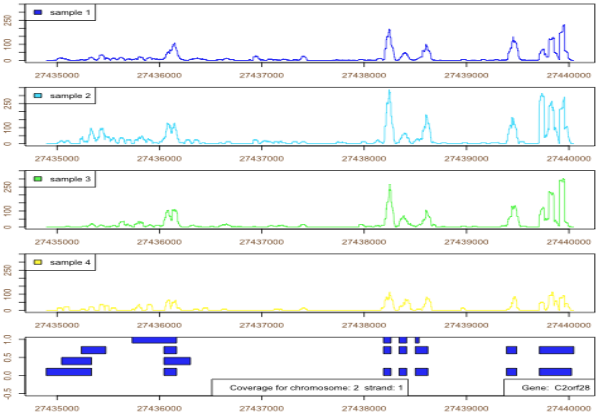
Irreducible regions of coverage



Examplary plot



Examplary plot



Outline

1. Motivation
   - State of the technology
   - Exonmap paradigms
   - Data Mining
2. Contribution
   - Schema of the library
   - Processing
   - Analysis pipelines
3. Summary and future developments
   - Numeric results
   - Examplary plots
   - Splicing index

Splicing indeks

- Similar to original in Gardina et al.
- Normalized to +/- 1
- Calculated on each nucleotide

Splicing indeks

- Similar to original in Gardina et al.
- Normalized to +/- 1
- Calculated on each nucleotide

Splicing indeks

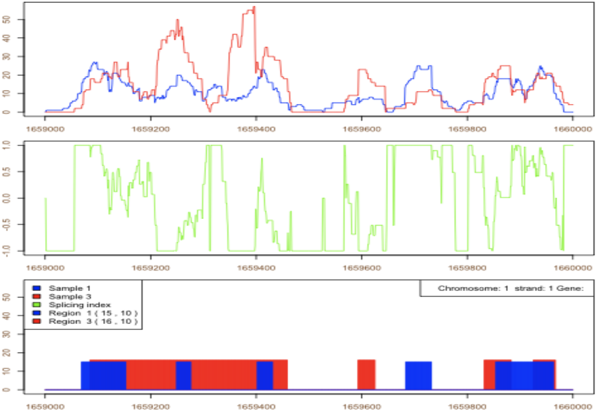
- Similar to original in Gardina et al.
- Normalized to +/- 1
- Calculated on each nucleotide

Splicing index

$$SI(n) = \begin{cases} 0, \mathit{if} & (E_{1n} = 0 \wedge E_{2n} = 0) \\ 1, \mathit{if} & (E_{1n} = 0 \wedge E_{2n} = 0) \quad \vee \left(\frac{E_{1n}}{G_{1n}} \cdot \frac{E_{2n}}{G_{2n}} > 2\right) \\ -1, \mathit{if} & (E_{1n} = 0 \wedge E_{2n} = 0) \quad \vee \left(\frac{E_{1n}}{G_{1n}} \cdot \frac{E_{2n}}{G_{2n}} < 0.5\right) \\ log_2\left(\frac{E_{1n}}{G_{1n}} \cdot \frac{E_{2n}}{G_{2n}}\right) & \text{in all other cases} \end{cases}$$

Where E_{1n} and E_{2n} are the coverage values for a given nucleotide, while G_{1n} and G_{2n} are the counts of reads in the region or gene.

Splicing index



Future developments

- exon/isoform discovery
- paired end reads
- new splicing index forms
- parallel execution with snow, multicore,. . .
- . . . etc

Summary

- The library rnaSeqMap in Bioconductor 2.7
- ...
- Have fun!!!

For Further Reading I

- Aumann Y, Lindell Y:
  J. Intell. Inf. Syst. 2003, **20**(3):255–283.
- Gardina et al.:
  BMC Genomics 2006, 7:325.
- Yates T, Okoniewski MJ, Miller CJ
  Nucleic Acids Research 2008, **36(suppl 1)**:D780–D786.
- Okoniewski M, Yates T, Dibben S, Miller C
  Genome Biology 2007, **8(5)**:R79.

Acknowledgements

- Functional Genomic Center Zurich:
  Marzanna Künzli-Gontarczyk
  Sirisha Aluri
  Weihong Qi
  Hubert Rehrauer
  Tanguy Le Carrour
  Remy Bruggmann
  Hansruedi Baetschmann
- Kinderspital Zürich:
  Beat Scheaffer
  Marco Wachtel
- Institute of Molecular Systems Biology, ETH:
  Lucia Bautista Borrego
- PICR Manchester:
  Tim Yates