Proteogenomic analysis of alternative splicing: the search for novel biomarkers for colorectal cancer

Gosia Komor
Collect clinical samples
- (Tumor) tissue
- Blood
- Stool

Perform molecular profiling
- DNA
- RNA
- Protein

Collect clinical information

Study tumor biology
- Preclinical models
Translational Gastrointestinal Oncology

Collect clinical samples
- (Tumor) tissue
- Blood
- Stool

Perform molecular profiling
- DNA
- RNA

Molecular profiling of colon tumors

Collect clinical information

Study tumor biology
- Preclinical models

Translation of molecular knowledge into clinical tests
Colorectal cancer (CRC)

Colorectal cancer is the 2\textsuperscript{nd} most common cancer type in the Netherlands
- Incidence rate of over 15 000 patients per year
- Most patients between 60-79 years old

Colorectal tumor progression

Figure adapted from Nature Reviews Cancer 9, 489–499 (2009)
Colorectal cancer has a high cure rate when diagnosed early.
Population wide screening for colorectal cancer implemented in the Netherlands

Fecal immunochemical test (FIT) → Colonoscopy

FIT performance *

- Specificity: ~95%
- Sensitivity CRC: ~79%
- Sensitivity precursor lesions (advanced adenomas): ~27%

Population wide screening for colorectal cancer implemented in the Netherlands

Fecal immunochemical test (FIT)  

FIT performance *
• Specificity: ~95%
• Sensitivity CRC: ~79%
• Sensitivity precursor lesions (advanced adenomas): ~27%

Colonoscopy

Clinical need for novel biomarkers

* Lee et al, Accuracy of Fecal Immunochemical Tests for Colorectal Cancer, Annals of Internal Medicine, 2014
AIM: Identify novel biomarkers for CRC screening

Tumor-specific molecular changes accompany tumor progression

DNA alterations, e.g.:
• Mutations (SNVs)
• Copy number aberrations
• Methylation

RNA alterations, e.g.:
• RNA Splicing
Tumor-specific molecular changes accompany tumor progression

DNA alterations, e.g.:
- Mutations (SNVs)
- Copy number aberrations
- Methylation

RNA alterations, e.g.:
- RNA Splicing

Figure adapted from Sveen et al. Oncogene 2016;35(19):2413-27
Tumor-specific molecular changes accompany tumor progression

DNA alterations, e.g.:
• Mutations (SNVs)
• Copy number aberrations
• Methylation

RNA alterations, e.g.:
• RNA Splicing

BCL2L1:
- Bcl-xL – anti-apoptotic
- Bcl-xS – pro-apoptotic

VEGFA:
- pro-angiogenic
- anti-angiogenic

AIM: Identify novel biomarkers for CRC screening

Figure adapted from Sveen et al. Oncogene 2016;35(19):2413-27
AIM: Identify novel biomarkers for CRC screening

Tumor-specific protein isoforms could complement or outperform hemoglobin in CRC screening.

pre-mRNA

alternatively spliced mRNA

protein isoforms

splicing

translation
AIM: Identify novel biomarkers for CRC screening

Tumor-specific protein isoforms could complement or outperform hemoglobin in CRC screening

Design an approach to identify tumor specific protein variants
With the use of available protein sequence databases, ~50% of mass spectra are still not identified.

Figure adapted from Duncan et al. Nat Biotechnol. 2010;28:659–664.
With the use of available protein sequence databases ~50% of mass spectra are still not identified.

Figure adapted from Duncan et al. Nat Biotechnol. 2010;28:659–664.
Experimental design
Down-modulation of splicing machinery to investigate differential splicing in a controlled setting

- CRC cell lines SW480
- siSRSF1
- siSF3B1
- siNonTargeting (siNT)
- RNA-seq
  - Illumina HiSeq
  - 2x125bp
- proteomics
  - LC-MS/MS
  - QExactive
Experimental design
Down-modulation of splicing machinery to investigate differential splicing in a controlled setting

CRC cell lines
SW480

siSRSF1
siSF3B1
siNonTargeting (siNT)

RNA-seq
Illumina HiSeq
2x125bp

proteomics
LC-MS/MS
QExactive

PacBio Iso-Seq
RSII, 4 fractions
0 - 50kb
Experimental design
Down-modulation of splicing machinery to investigate differential splicing in a controlled setting

CRC cell lines
SW480

siSRSF1
siSF3B1
siNonTargeting (siNT)

RNA-seq
Illumina HiSeq
2x125bp

proteomics
LC-MS/MS
QExactive

PacBio Iso-Seq
RSII, 4 fractions
0 - 50kb

Proteogenomic pipeline
SPLICIFY – proteogenomic pipeline for differential splice variant identification

Short RNA-seq reads

Mass spectra
SPLICIFY – proteogenomic pipeline for differential splice variant identification

Short RNA-seq reads → Quality and adapter trimming (Trimmomatic) → Reads mapping (STAR) → Differential splicing analysis (rMATS) → Reference annotation → Differential splice variants on RNA level

Mass spectra
SPLICIFY – proteogenomic pipeline for differential splice variant identification

1. Short RNA-seq reads
2. Quality and adapter trimming (Trimmomatic)
3. Reads mapping (STAR)
4. Differential splicing analysis (rMATS)
5. Reference annotation

- Enriched protein database
- Human protein database
- Potential protein variants
- 3-frame translation
- Differential splice variants on RNA level
- Mass spectra
SPLICIFY – proteogenomic pipeline for differential splice variant identification

- Short RNA-seq reads
  - Quality and adapter trimming (Trimmomatic)
  - Reads mapping (STAR)
  - Differential splicing analysis (rMATS)
  - Reference annotation

- Human protein database
  - Potential protein variants
    - 3-frame translation
    - Enriched protein database

- Mass spectra
  - Identify MS/MS
  - Extract variant peptides
  - Differential protein isoforms

- Differential peptide expression (limma)
**SPLICIFY – proteogenomic pipeline for differential splice variant identification**

1. **Short RNA-seq reads**
   - Quality and adapter trimming (Trimmomatic)
   - Reads mapping (STAR)

2. **Differential splicing analysis (rMATS)**
   - Differential splice variants on RNA level
   - Potential protein variants
   - 3-frame translation
   - Extract variant peptides
   - Identify MS/MS

3. **Mass spectra**
   - Enriched protein database
   - Human protein database

4. **Differential protein isoforms**
   - Differential peptide expression (limma)
   - PacBio full-length transcripts
Differential splice variants identified on RNA level

Exon skipping:

\[
exclusion\_level = \frac{excl}{excl + incl}
\]

- exclusion spanning reads
- inclusion spanning reads
Differential splice variants identified on RNA level

**Exon skipping:**
- Exclusion (exclusion spanning reads)
- Inclusion (inclusion spanning reads)

**Alternatively splicing events**
- Skipped exon (SE)
- Mutually exclusive exons (MXE)
- Alternatively 5’ splice site (A5SS)
- Alternatively 3’ splice site (A3SS)
- Retained intron (RI)

\[
exclusion\_level = \frac{\text{excl}}{\text{excl} + \text{incl}}
\]

- excl - exclusion spanning reads
- incl - inclusion spanning reads
RT-qPCR validation of SPLICIFY results on RNA level

RNA-seq

OSBPL3 exon 9

- samples
  - siNT
  - siSF3B1

exclusion_level

- 0.20
- 0.30
- 0.40
- 0.50
RT-qPCR validation of SPLICIFY results on RNA level

RNA-seq

OSBPL3 exon 9

samples

<table>
<thead>
<tr>
<th>samples</th>
<th>siNT</th>
<th>siSF3B1</th>
</tr>
</thead>
<tbody>
<tr>
<td>exclusion_level</td>
<td>[Box plot]</td>
<td></td>
</tr>
</tbody>
</table>

RT-qPCR

Relative expression

<table>
<thead>
<tr>
<th></th>
<th>OSBPL3 inclusion</th>
<th>OSBPL3 exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>siNT</td>
<td>[Bar chart]</td>
<td>[Bar chart]</td>
</tr>
<tr>
<td>siSF3B1</td>
<td>[Bar chart]</td>
<td>[Bar chart]</td>
</tr>
</tbody>
</table>
RNA to protein translation: isoform-specific peptides

Exon skipping

Inclusion

Exclusion

- exon
RNA to protein translation: isoform-specific peptides

Exon skipping

inclusion

exclusion

- exon

XXXXXZZZZZ – exclusion specific split peptide
RNA to protein translation: isoform-specific peptides

Exon skipping

XXX

ZZZZZ

XXX

YYYY

inclusion

exclusion

XXX

YYYY

XX

exon

XX

exclusion specific split peptide

XX

inclusion specific split peptide
RNA to protein translation: isoform-specific peptides

**Exon skipping**

- **Exclusion specific split peptide:** $XXXXXZZZZZ$
- **Inclusion specific peptide on target:** $YYYY$

**Exon inclusion**

- **Exclusion specific split peptide:** $XXXXXXYYYY$
- **Inclusion specific split peptide:** $YYVV$
RNA to protein translation: isoform-specific peptides

Exon skipping

- Exclusion specific split peptide: XXXXXXXZZZZZ
- Inclusion specific split peptide: XXXXYYYY YYVV
- Inclusion specific peptide on target: YYVV

Retained intron

- Exclusion specific split peptide: XXXXAAAAA AABB
- Inclusion specific spanning peptide: XXXXYYYY YYVV
- Inclusion specific peptide on target: AABB
## Differential isoform identified on protein level

<table>
<thead>
<tr>
<th>Experiment</th>
<th>On target</th>
<th>Spanning peptide</th>
<th>Split peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>siSF3B1 vs siNT</td>
<td>3278</td>
<td>9</td>
<td>1794</td>
</tr>
<tr>
<td>siSRSF1 vs siNT</td>
<td>217</td>
<td>3</td>
<td>154</td>
</tr>
</tbody>
</table>
**Differential isoform identified on protein level**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>On target</th>
<th>Spanning peptide</th>
<th>Split peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>siSF3B1 vs siNT</td>
<td>3278</td>
<td>9</td>
<td>1794</td>
</tr>
<tr>
<td>siSRSF1 vs siNT</td>
<td>217</td>
<td>3</td>
<td>154</td>
</tr>
</tbody>
</table>

**RNA → translation → Protein**

**Graphs:**
- **siSF3B1** and **siSRSF1**
  - Event type (RNA level)
  - Event type (protein level)
Quantitative differences on RNA and protein level

RefSeq Genes

RNA isoforms
Quantitative differences on RNA and protein level

RefSeq Genes

RNA isoforms

RNA-seq

OSBPL3 exon 9

samples

<table>
<thead>
<tr>
<th></th>
<th>siINT</th>
<th>siSF3B1</th>
</tr>
</thead>
<tbody>
<tr>
<td>exclusion_level</td>
<td>0.20</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Quantitative differences on RNA and protein level

RefSeq Genes

RNA isoforms

Isoform-specific peptides

RNA-seq

OSBPL3 exon 9

samples

<table>
<thead>
<tr>
<th>samples</th>
<th>sINT</th>
<th>siSF3B1</th>
</tr>
</thead>
<tbody>
<tr>
<td>exclusion_level</td>
<td>0.20</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Netherland Cancer Institute
Quantitative differences on RNA and protein level

RefSeq Genes

RNA isoforms

Isoform-specific peptides

RNA-seq

OSBPL3 exon 9

samples

OSBPL3 isoform specific peptide intensities

LC-MS/MS

exclusion

inclusion

samples

siNT vs siSF3B1

peptide

1

2

3

Intensity

7.0

7.5

8.0

group

siNT vs siSF3B1
PacBio full length transcripts used as annotation to identify novel events
siSF3B1 vs siNT

SPLICIFY with reference annotation

SPLICIFY with PacBio full-length transcripts

Full length transcript used as annotation to quantify Illumina reads
- Comparison to the standard SPLICIFY with reference annotation
- Both approaches include Illumina reads for the differential analysis
PacBio Iso-Seq provides a number of novel alternatively spliced events
differential splicing analysis of siSF3B1 vs siNT1

Alternatively splicing events

- Skipped exon (SE)
- Mutually exclusive exons (MXE)
- Alternatively 5’ splice site (A5SS)
- Alternatively 3’ splice site (A3SS)
- Retained intron (RI)
PacBio Iso-Seq provides a number of novel alternatively spliced events
differential splicing analysis of siSF3B1 vs siNT1

Alternatively splicing events

- Skipped exon (SE)
- Mutually exclusive exons (MXE)
- Alternatively 5’ splice site (A5SS)
- Alternatively 3’ splice site (A3SS)
- Retained intron (RI)

Skipped exon

Mutually exclusive exons
PacBio Iso-Seq provides a number of novel alternatively spliced events
differential splicing analysis of siSF3B1 vs siNT1

Alternatively splicing events
- Skipped exon (SE)
- Mutually exclusive exons (MXE)
- Alternatively 5’ splice site (A5SS)
- Alternatively 3’ splice site (A3SS)
- Retained intron (RI)

![Graph showing PacBio full-length transcripts and reference annotation](image)

### Alternative 5’ splice site
- PacBio transcripts: 124
- Reference annotation: 7
- Overlap: 29

### Alternative 3’ splice site
- PacBio transcripts: 207
- Reference annotation: 26
- Overlap: 71

### Retained intron
- PacBio transcripts: 399
- Reference annotation: 10
- Overlap: 63

Netherlands Cancer Institute
Antoni van Leeuwenhoek
PacBio Iso-Seq provides a number of novel alternatively spliced events
differential splicing analysis of siSF3B1 vs siNT1

Alternatively splicing events

- Skipped exon (SE)
- Mutually exclusive exons (MXE)
- Alternatively 5’ splice site (A5SS)
- Alternatively 3’ splice site (A3SS)
- Retained intron (RI)

![Diagram showing alternatively splicing events]

Alternative 5’ splice site

<table>
<thead>
<tr>
<th>PacBio transcripts</th>
<th>Reference annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>124</td>
<td>7</td>
</tr>
</tbody>
</table>

Alternative 3’ splice site

<table>
<thead>
<tr>
<th>PacBio transcripts</th>
<th>Reference annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>207</td>
<td>26</td>
</tr>
</tbody>
</table>

Retained intron

<table>
<thead>
<tr>
<th>PacBio transcripts</th>
<th>Reference annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>399</td>
<td>10</td>
</tr>
</tbody>
</table>

![Graph showing PacBio full-length transcripts vs Reference Annotation]

- Number of events
- Event type: SE, MXE, A5SS, A3SS, RI
Novel isoforms identified with PacBio Iso-seq are expressed on protein level

### Isoform-specific peptides

<table>
<thead>
<tr>
<th></th>
<th>On target</th>
<th>Spanning peptides</th>
<th>Split peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>PacBio transcripts</td>
<td>3090</td>
<td>23</td>
<td>2350</td>
</tr>
<tr>
<td>Reference Annotation</td>
<td>2518</td>
<td>6</td>
<td>1964</td>
</tr>
</tbody>
</table>
Novel isoforms identified with PacBio Iso-seq are expressed on protein level

Retained intron

Read coverage in siSF3B1

Read coverage in siNT

RefSeq Genes

PacBio transcripts

RNA-seq isoforms

Isoform-specific peptides
Novel isoforms identified with PacBio Iso-seq are expressed on protein level

Alternative 3’ splice site

Read coverage in siSF3B1

Read coverage in siNT

RefSeq Genes

PacBio transcripts

RNA-seq isoforms

Isoform-specific peptides
Conclusions

• Established SPLICIFY
  • Proteogenomic pipeline combining RNA-seq and LC-MS/MS data for differential splice variant identification
  
• Confirmation of the splice variants on RNA level
  • RT-qPCR
  • by PacBio full-length transcripts
  
• [https://github.com/NKI-TGO/SPLICIFY](https://github.com/NKI-TGO/SPLICIFY) 
  • will be available soon

• Novel splicing events identified with PacBio Iso-seq
  • confirmation on the protein level
Future plans

Organoids
- Healthy colon tissue
- adenomas
- CRCs

Human tissues

mRNA isolation
Illumina RNA-seq

protein isolation
LC-MS/MS

SPLICIFY proteogenomic pipeline
Future plans

Organoids
- Healthy colon tissue
- adenomas
- CRCs

Human tissues
- mRNA isolation
  - Illumina RNA-seq
- protein isolation
  - LC-MS/MS

SPLICIFY proteogenomic pipeline

Stool samples

FIT samples

Antibody-based assay for the best candidates
Acknowledgements

This research was financially supported by a grant from the Dutch Cancer Society
Grant number: NKI 2013-6025