Structural variant detection in colorectal cancer

E van den Broek¹, JC Haan¹, MH Jansen¹, B Carvalho¹, MA van de Wiel², ID Nagtegaal³, CJA Punt⁴, B Ylstra¹, S Abeln⁵, GA Meijer¹, RJA Fijneman¹

¹ Dept. of Pathology, VU University Medical Center, Amsterdam, The Netherlands
² Dept. of Epidemiology & Biostatistics, VU University Medical Center, Amsterdam, The Netherlands
³ Dept. of Pathology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands
⁴ Dept. of Medical Oncology, Academic Medical Center, University of Amsterdam, The Netherlands
⁵ Dept. of Computer Science, VU University, Amsterdam, The Netherlands
Colorectal cancer (CRC)

• Colorectal cancer is a major health concern worldwide

• Second cause of cancer related death
  – The incidence worldwide is 1,200,000
  – The incidence in the US is 144,000

• Mortality rates

<table>
<thead>
<tr>
<th>Stage</th>
<th>Mortality Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>&lt; 10 %</td>
</tr>
<tr>
<td>Stage 2</td>
<td>25 - 30 %</td>
</tr>
<tr>
<td>Stage 3</td>
<td>45 – 50 %</td>
</tr>
<tr>
<td>Stage 4</td>
<td>&gt; 90 %</td>
</tr>
</tbody>
</table>
CRC research

Clinical needs for biomarker discovery

Clinical needs:

1. Screening  
   Diagnostic biomarkers

2. Predict recurrence  
   Prognostic biomarkers

3. Personalized therapy  
   Predictive biomarkers

Key molecular features:

- normal colon
- adenoma
- progressive adenoma
- localized CRC
- CRC liver metastasis

CRC stages:

- Stage I+II
- Stage III
- Stage IV

 CRC research
Chromosomal Instability
a hallmark of CRC

SKY: numerical & structural aberrations

M Hermsen et al., Oncogene 2005
CAIRO & CAIRO2 studies

Phase III clinical trials
In total 1575 patients were included

CApecitabine, IRinotecan, Oxaliplatin
in advanced colorectal cancer


DNA from 356 patients: primary tumor and matched normal
  – Representative group
  – Isolated from FFPE
Comparative Genomic Hybridization (CGH)
Agilent, 180k array CGH

Array CGH: 356 CAIRO & CAIRO2 samples

1. Segmentation
2. Calling
3. Copy numbers

Numerical aberrations
Segmentation - array CGH
Profile of one tumor with 180k probes

Segmentation was performed using Circular Binary Segmentation algorithm (DNAcopy. Olshen et al. 2004)
Calling - array CGH
Profile of one tumor with 180k probes

Calling was performed using CGHcall
(CGHcall. vd Wiel et al. 2007)
Structural Variants (SV) in cancer

Hematological disorders
- Philadelphia chromosome
  - t(9;22)
  - Fusion gene: BCR-ABL
  - Drug: Imatinib / Gleevec

Epithelial cancers
- TMPRSS2-ERG in prostate cancers
- VTI1A-TCF7L2 is confirmed in 3% of 97 CRCs
  - Bass et al., Nature Genetics 2011
AIM

TO IDENTIFY RECURRENT SOMATIC STRUCTURAL GENOMIC VARIANTS THAT CAUSE CRC
Breakpoint (BP) detection
Based on array CGH

Array CGH: 356 CAIRO & CAIRO2 samples

1. Segmentation
2. Calling
3. Copy numbers
   Numerical aberrations

1. Segmentation
2. Breakpoint detection
3. Candidate genes
   Structural variants
BP detection in array CGH
Profile of one tumor with 180k probes

Breakpoints are defined by the start position of the first probe of each segment

Breakpoint annotation per gene
Results based on array CGH

**BP detection**

- Total number of genes with BPs: 5,737 genes
- 482 candidate genes were identified with recurrent BP (FDR < 0.1)

![Bar chart showing candidate genes (top 15) with MACROD2 highlighted.](image-url)
Overall survival: MACROD2

Recurrent BP (1) versus no-BP (0)

Log rank $P = 0.08$
Results based on array CGH

*BP detection*

- Total number of genes with BPs: 5,737 genes
- 482 candidate genes were identified with recurrent BP (FDR < 0.1)

**Limitations breakpoint determination using array CGH:**
- Location BP is estimation (average probe distance is ~17 kb)
- DNA structure is unknown
- Balanced events will be missed
482 candidate genes were identified with recurrent BP (FDR < 0.1)

Validation array CGH BPs

NGS data from TCGA

- 482 candidate genes were identified with recurrent BP (FDR < 0.1)

Candidate validation is required

CRC samples (COAD & READ)

Whole Genome DNA Seq from paired tumor-normal samples
482 candidate genes were identified with recurrent BP (FDR < 0.1)

Validation array CGH BPs

NGS data from TCGA

- Candidate validation is required
- Structural Variant (SV) detection

Candidate driven

Negative Control Genes (no BP)
Computational methods

Focus on candidate genes

Based on paired-end NGS data
- Read-pair approach
  - Discordance: location / bridge length / orientation reads

Discordant pairs (DP) types
- Translocation: > different chromosomes
- Insertion: > bridge length
- Deletion: > bridge length
- Inversion: > orientation
- Eversion: > orientation
- Single mapped: could indicate a breakpoint
Computational methods

Focus on candidate genes

Based on paired-end NGS data
1. Read-pair approach

Combined with:
2. Read-depth
3. Define breakpoint location
4. Determine tumor specific events
Translocation

IGV

MACROD2
• Discordant pairs
• Breakpoints

Fusion partner
Based on DP groups:

Approximately 5 fold higher number of translocation-DP groups for candidate genes compared to control genes.
Translocation-DP groups per candidate gene in TCGA samples

Putative translocations

Candidate genes (top 20)

Frequency of translocation-DP groups (au)

Candidate genes

MACROD2
Correlation per candidate gene

- Frequency of samples with BP based on array CGH
- Frequency of translocation-DP groups in TCGA data
Conclusions

• 482 candidate genes with recurrent breakpoints were identified in a large cohort of 356 CRC samples, based on array CGH analysis.

• The Cancer Genome Atlas provided an essential CRC reference dataset (COAD, READ) to validate Structural Variants in candidate genes with recurrent breakpoints.

• Identification of BPs based on array CGH is correlated with SV detection in TCGA CRC NGS data.

• Further studies will be performed to investigate clinical and functional significance of validated candidate genes.