Supplementary Material for:

Precise breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers

Michal J Okoniewski1,2,*, Janine Meienberg3,4,*, Andrea Patrignani1, Alicja Szabelska1,5, Gabor Matyas3,4,#, Ralph Schlapbach1,#

1Functional Genomics Center Zurich, Zurich, Switzerland, 2Department of Neuroimmunology and Multiple Sclerosis Research, Neurology Clinic, University Hospital, Zurich, Switzerland, 3Center for Cardiovascular Genetics and Gene Diagnostics, Zurich, Switzerland, 4Zurich Center of Integrative Human Physiology, University of Zurich, Zurich, Switzerland, 5Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Poznan, Poland

*These authors contributed equally to this work
#These authors jointly directed this work

Keywords: sequencing, large deletions, amplicons, PacBio, Illumina, Sanger

Supplementary Figure S1. Both ends of the 26,887bp deletion on chromosome 15 in sample 44 displayed in the Integrative Genomics Viewer. The reads generated by the PacBio RS (upper panel) and the Illumina HiSeq 2000 (lower panel) were sorted by aligned position, base, and mapping quality, and compared with the results of Sanger sequencing (bottom). Sections of 22 reads are shown. Aligned reads are displayed as gray bars/arrow, letters indicate mismatched bases, purple vertical dashes insertions, and black horizontal lines deletions. Note that the total read counts (reads) and the percentage of reference bases (%) are given for the positions flanking the site where the coverage (gray bars) starts to lower. Uppercase letters represent the sequence in the region of the start point of the deletion; lowercase letters represent the sequence in the region of the deletion end point. Due to identical sequences at the site of breakpoints, the break and rejoining could have occurred at three positions, as indicated by open triangles. The dotted red line marks the most telomeric position of the possible breakpoints. For more details, see Supplementary Tables S1 and S2.
Supplementary Figure S2. Both ends of the 3,408,306bp deletion on chromosome 2 in sample 53B displayed in the Integrative Genomics Viewer. The reads generated by the PacBio RS (upper panel) and the Illumina HiSeq 2000 (lower panel) were sorted by aligned position, base, and mapping quality, and compared with the results of Sanger sequencing (bottom). Sections of 22 reads are shown. Aligned reads are displayed as gray bars/arrow, letters indicate mismatched bases, purple vertical dashes insertions, and black horizontal lines deletions. Note that the total read counts (reads) and the percentage of reference bases (%) are given for the positions flanking the site where the coverage (gray bars) starts to lower. Uppercase letters represent the sequence in the region of the start point of the deletion; lowercase letters represent the sequence in the region of the deletion end point. Due to identical sequences at the site of breakpoints, the break and rejoining could have occurred at three positions, as indicated by open triangles. The dotted red line marks the most telomeric position of the possible breakpoints. For more details, see Supplementary Tables S3 and S4.
DNA Sequencing

Methods

Genomics/Transcriptomics

Supplementary Figure S3. Counts of exact matches for different lengths of a GC-motif (red) at the site of deletion breakpoints in sample 44 with 5-, 10-, and 20-nucleotide flanking sequences for both PacBio and Illumina reads (denoted by different shades of green). Counts indicate that the true sequence includes 1 × GC (n=1) (see Supplementary Figure S1). Corresponding type II errors were calculated using the R script provided in the Supplementary Materials.

<table>
<thead>
<tr>
<th></th>
<th>n=0</th>
<th>n=1</th>
<th>n=2</th>
<th>n=3</th>
<th>n=4</th>
<th>type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flank 5</td>
<td>885</td>
<td>2270</td>
<td>11</td>
<td></td>
<td></td>
<td>&lt;2e-308</td>
</tr>
<tr>
<td><strong>Flank 10</strong></td>
<td>0</td>
<td>649</td>
<td>4</td>
<td></td>
<td></td>
<td>6e-61</td>
</tr>
<tr>
<td>Flank 20</td>
<td>0</td>
<td>74</td>
<td>1</td>
<td></td>
<td></td>
<td>4e-06</td>
</tr>
</tbody>
</table>

Supplementary Figure S4. Counts of exact matches for different lengths of a poly(C) motif (red) at the site of deletion breakpoints in sample 53B with 5-, 10-, and 20-nucleotide flanking sequences for both PacBio and Illumina reads (denoted by different shades of green). Counts indicate that the true sequence includes 2 × C (n=2) (see Supplementary Figure S1). Corresponding type II errors were calculated using the R script provided in the Supplementary Materials.

<table>
<thead>
<tr>
<th></th>
<th>n=0</th>
<th>n=1</th>
<th>n=2</th>
<th>n=3</th>
<th>n=4</th>
<th>type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flank 5</td>
<td>116</td>
<td>52218</td>
<td>0</td>
<td></td>
<td></td>
<td>&lt;2e-308</td>
</tr>
<tr>
<td><strong>Flank 10</strong></td>
<td>0</td>
<td>44249</td>
<td>0</td>
<td></td>
<td></td>
<td>&lt;2e-308</td>
</tr>
<tr>
<td>Flank 20</td>
<td>0</td>
<td>32717</td>
<td>0</td>
<td></td>
<td></td>
<td>&lt;2e-308</td>
</tr>
</tbody>
</table>

TAAATTCTCCAACTCCATTA(GC)nTTTGTACCCAGGCTGGAGT

Supplementary Figure S4. Counts of exact matches for different lengths of a poly(C) motif (red) at the site of deletion breakpoints in sample 53B with 5-, 10-, and 20-nucleotide flanking sequences for both PacBio and Illumina reads (denoted by different shades of green). Counts indicate that the true sequence includes 2 × C (n=2) (see Supplementary Figure S1). Corresponding type II errors were calculated using the R script provided in the Supplementary Materials.
Supplementary Figure S5. Coverage plot from the PacBio software, SMRTportal. Plot is shown for sample 70 and is based upon a standard human genome reference.
**Supplementary Table S1. Read Depth and Percentage of Wild-type Allele in the Region Flanking the Breakpoint at the Start Site of the Deletion in Sample 44.**

<table>
<thead>
<tr>
<th>Location</th>
<th>Not Deleted</th>
<th>Wild-type Sequence</th>
<th>T</th>
<th>T</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>Deleted</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PacBio RS</td>
<td>7226 (99%)</td>
<td>6738 (99%)</td>
<td>6144 (99%)</td>
<td>5621 (99%)</td>
<td>5018 (100%)</td>
<td>70 (67%)</td>
<td>35 (40%)</td>
<td>43 (37%)</td>
<td>150 (96%)</td>
<td>39* (79%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiSeq 2000</td>
<td>9403 (100%)</td>
<td>7894 (100%)</td>
<td>6118 (100%)</td>
<td>4626 (100%)</td>
<td>3502 (100%)</td>
<td>1776 (4%)</td>
<td>845 (67%)</td>
<td>92 (20%)</td>
<td>Fifteen (53%)</td>
<td>6** (67%)</td>
</tr>
</tbody>
</table>

*No mapped reads 468 bases after the most telomeric breakpoint (read depth = 0).
**No mapped reads 14 bases after the most telomeric breakpoint (read depth = 0).

Bold letters indicate identical bases at the site of the breakpoint, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the three possible breakpoints. This is also the point where the read depth drops and the number of mismatches increases (see Supplementary Figure S1).

**Supplementary Table S2. Read Depth and Percentage of Wild-type Allele in the Region Flanking the Breakpoint at the End of the Deletion in Sample 44.**

<table>
<thead>
<tr>
<th>Location</th>
<th>Deleted</th>
<th>Wild-type Sequence</th>
<th>g</th>
<th>t</th>
<th>c</th>
<th>t</th>
<th>c</th>
<th>Not Deleted</th>
<th>g</th>
<th>c</th>
<th>t</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PacBio RS</td>
<td>77 (34%)</td>
<td>104 (72%)</td>
<td>71 (76%)</td>
<td>132 (72%)</td>
<td>185 (68%)</td>
<td>PacBio RS</td>
<td>77 (34%)</td>
<td>104 (72%)</td>
<td>71 (76%)</td>
<td>132 (72%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiSeq 2000</td>
<td>2200** (4%)</td>
<td>5594 (69%)</td>
<td>8918 (87%)</td>
<td>9609 (83%)</td>
<td>10347 (60%)</td>
<td>HiSeq 2000</td>
<td>2200** (4%)</td>
<td>5594 (69%)</td>
<td>8918 (87%)</td>
<td>9609 (83%)</td>
</tr>
</tbody>
</table>

*No mapped reads 368 bases before the most telomeric breakpoint (read depth = 0).
**No mapped reads 39 bases before the most telomeric breakpoint (read depth = 0).

Bold letters indicate identical bases at the site of the breakpoints, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the possible breakpoints. The most centromeric breakpoint, where the read depth drops and the number of mismatches increases, is indicated by a black bold line (see Supplementary Figure S1).

**Supplementary Table S3. Read Depth and Percentage of Wild-type Allele in the Region Flanking the Breakpoint at the Start Site of the Deletion in Sample 53B.**

<table>
<thead>
<tr>
<th>Location</th>
<th>Not Deleted</th>
<th>Wild-type Sequence</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>Deleted</th>
<th>C</th>
<th>T</th>
<th>T</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PacBio RS</td>
<td>4386 (99%)</td>
<td>4132 (99%)</td>
<td>3931 (99%)</td>
<td>3568 (99%)</td>
<td>2984 (99%)</td>
<td>PacBio RS</td>
<td>562 (98%)</td>
<td>493 (96%)</td>
<td>211 (60%)</td>
<td>190 (59%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiSeq 2000</td>
<td>20179 (100%)</td>
<td>18897 (100%)</td>
<td>17702 (100%)</td>
<td>14635 (100%)</td>
<td>14926 (100%)</td>
<td>HiSeq 2000</td>
<td>13551 (76%)</td>
<td>12489 (70%)</td>
<td>11602 (71%)</td>
<td>8538 (99%)</td>
</tr>
</tbody>
</table>

*No mapped reads 544 bases after the most telomeric breakpoint (read depth = 0).
**No mapped reads 13 bases after the most telomeric breakpoint (read depth = 0).

Bold letters indicate identical bases at the site of the breakpoints, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the possible breakpoints. This is also the point where the read depth drops and the number of mismatches increases (see Supplementary Figure S2).

**Supplementary Table S4. Read Depth and Percentage of Wild-type Allele in the Region Flanking the Breakpoint at the End of the Deletion in Sample 53B.**

<table>
<thead>
<tr>
<th>Location</th>
<th>Deleted</th>
<th>Wild-type Sequence</th>
<th>t</th>
<th>t</th>
<th>t</th>
<th>t</th>
<th>t</th>
<th>Not Deleted</th>
<th>c</th>
<th>c</th>
<th>t</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PacBio RS</td>
<td>223* (81%)</td>
<td>230 (79%)</td>
<td>240 (61%)</td>
<td>331 (91%)</td>
<td>459 (74%)</td>
<td>PacBio RS</td>
<td>4421 (99%)</td>
<td>5192 (100%)</td>
<td>5678 (100%)</td>
<td>6066 (100%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HiSeq 2000</td>
<td>7818** (73%)</td>
<td>9888 (100%)</td>
<td>10675 (57%)</td>
<td>13335 (89%)</td>
<td>14923 (54%)</td>
<td>HiSeq 2000</td>
<td>16393 (100%)</td>
<td>17832 (100%)</td>
<td>19429 (100%)</td>
<td>20646 (100%)</td>
</tr>
</tbody>
</table>

*No mapped reads 411 bases after the most telomeric breakpoint (read depth = 0).
**No mapped reads 11 bases after the most telomeric breakpoint (read depth = 0).

Bold letters indicate identical bases at the site of the breakpoints, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the possible breakpoints. The most centromeric breakpoint, where the read depth drops and the number of mismatches increases, is indicated by a black bold line (see Supplementary Figure S2).
**AWK Script (Example for sample 70)**

```awk
BEGIN {c8=0; c7=0; c6=0; c5=0; c4=0; c3=0; c2=0; c1=0; c0=0; sekw="C"; l=0}
{ if ($0 ~ />/)   
  { if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c8=c8+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c7=c7+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c6=c6+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c5=c5+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c4=c4+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c3=c3+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c2=c2+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c1=c1+1;  
    if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c0=c0+1;  
    l=l+length(sekw)  
    sekw = ""; } }
else sekw=sekw$0; 
} END { print (c0, " ",c1, " ",c2, " ",c3, " ",c4, " ",c5, " ",c6, " ",c7, " ",c8, " ",l) }
else sekw=sekw$0;
}
END { print (c0, " ",c1, " ",c2, " ",c3, " ",c4, " ",c5, " ",c6, " ",c7, " ",c8, " ,l)}
```

**R Script**

```r
dane<-read.csv('polyCmotifs.csv', header=T) #loading data with counts of the exact matches with 5,10,20-nucleotide flanking region

a<0.01 # significance level of the test
z<-qnorm(1-a) # z statistics needed for the calculations of the type II error
n<-rowSums(dane) # total number of the number of
l1<-40
l1<-c(10,20,40) #length of flanking region
p<0.12 # probability of mismatch
N<83830 # total number of reads
psum1<dnbinom(l1,1,p) #estimated probability that there is no error in the sequence, negative binomial distribution is assumed
n.reads1=N*psum1 # expected number of exact matches with chosen flanking region

#calculation of the type II error (probability that the null hypothesis was wrongly accepted)
beta<NULL
x<seq(0,8,by=1) #considered number of deletions
for (i in 1:3) {
  lambda<as.numeric(colnames(dane)[which(dane[i,]==max(dane[i,]))]) # choice of the motif with maximun counts of the exact matches
  x<[-x][lambda] #obtaining possible alternative hypotheses by exculding the lambda from considered cases
  #calculations of errors has to be divided in 2 cases, when alternative is smaller or higher than null hypothesis
  y<x[lambda]
  b<1-pnorm((lambda-z*sqrt(lambda/n[i])-y)/sqrt(y/n[i])) # type II error for alternative hypotheses < lambda
  y<x[lambda]
  b<c(b,pnorm((lambda+z*sqrt(lambda/n[i])-y)/sqrt(y/n[i])))# type II error for alternative hypotheses > lambda
  beta<beta+sum(b) # type II error
}
print(beta)
```