In [1]:

```
library('dryclean')
library('tidyverse')
library('GenomicRanges')
library('DNAcopy')
library('gUtils')
library('gridExtra')
library('DNAcopy')
library('skitools')
library('gTrack')
library('skidb')

setDTthreads(threads = 10)
getDTthreads()
```

```
Loading required package: data.table

── Attaching packages ─────────────────────── tidyverse 1.3.1 ──

✔ ggplot2 3.3.5     ✔ purrr   0.3.4
✔ tibble  3.1.4     ✔ dplyr   1.0.7
✔ tidyr   1.1.3     ✔ stringr 1.4.0
✔ readr   2.0.1     ✔ forcats 0.5.1

── Conflicts ────────────────────────── tidyverse_conflicts() ──
✖ dplyr::between()   masks data.table::between()
✖ dplyr::filter()    masks stats::filter()
✖ dplyr::first()     masks data.table::first()
✖ dplyr::lag()       masks stats::lag()
✖ dplyr::last()      masks data.table::last()
✖ purrr::transpose() masks data.table::transpose()

Loading required package: stats4

Loading required package: BiocGenerics

Loading required package: parallel


Attaching package: ‘BiocGenerics’


The following objects are masked from ‘package:parallel’:

    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
    clusterExport, clusterMap, parApply, parCapply, parLapply,
    parLapplyLB, parRapply, parSapply, parSapplyLB


The following objects are masked from ‘package:dplyr’:

    combine, intersect, setdiff, union


The following objects are masked from ‘package:stats’:

    IQR, mad, sd, var, xtabs


The following objects are masked from ‘package:base’:

    anyDuplicated, append, as.data.frame, basename, cbind, colnames,
    dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
    grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
    order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
    rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
    union, unique, unsplit, which, which.max, which.min
```

```
Loading required package: S4Vectors

Attaching package: ‘S4Vectors’

The following objects are masked from ‘package:dplyr’:

    first, rename

The following object is masked from ‘package:tidyr’:

    expand

The following objects are masked from ‘package:data.table’:

    first, second

The following object is masked from ‘package:base’:

    expand.grid

Loading required package: IRanges

Attaching package: ‘IRanges’

The following objects are masked from ‘package:dplyr’:

    collapse, desc, slice

The following object is masked from ‘package:purrr’:

    reduce

The following object is masked from ‘package:data.table’:

    shift

Loading required package: GenomeInfoDb

Attaching package: ‘gUtils’

The following object is masked from ‘package:ggplot2’:

    %+%

The following object is masked from ‘package:base’:

    %o%


Attaching package: ‘gridExtra’

The following object is masked from ‘package:BiocGenerics’:

    combine
```

```
The following object is masked from ‘package:dplyr’:

    combine


Loading required package: ComplexHeatmap

Loading required package: grid

========================================
ComplexHeatmap version 2.2.0
Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
Github page: https://github.com/jokergoo/ComplexHeatmap
Documentation: http://jokergoo.github.io/ComplexHeatmap-reference

If you use it in published research, please cite:
Gu, Z. Complex heatmaps reveal patterns and correlations in multidimensional
  genomic data. Bioinformatics 2016.
========================================


Loading required package: VariantAnnotation

Loading required package: SummarizedExperiment

Loading required package: Biobase

Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.


Loading required package: DelayedArray

Loading required package: matrixStats


Attaching package: ‘matrixStats’


The following objects are masked from ‘package:Biobase’:

    anyMissing, rowMedians


The following object is masked from ‘package:dplyr’:

    count


Loading required package: BiocParallel


Attaching package: ‘DelayedArray’


The following objects are masked from ‘package:matrixStats’:

    colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges


The following object is masked from ‘package:purrr’:

    simplify


The following objects are masked from ‘package:base’:

    aperm, apply, rowsum
```

```
Loading required package: Rsamtools
Loading required package: Biostrings
Loading required package: XVector

Attaching package: 'XVector'

The following object is masked from 'package:purrr':

    compact


Attaching package: 'Biostrings'

The following object is masked from 'package:base':

    strsplit


Attaching package: 'VariantAnnotation'

The following object is masked from 'package:stringr':

    fixed

The following object is masked from 'package:base':

    tabulate

Loading required package: htmlwidgets
Loading required package: devtools
Loading required package: usethis
Loading required package: plotly

Attaching package: 'plotly'

The following object is masked from 'package:VariantAnnotation':

    select

The following object is masked from 'package:XVector':

    slice

The following object is masked from 'package:ComplexHeatmap':

    add_heatmap

The following object is masked from 'package:IRanges':

    slice
```

```
The following object is masked from 'package:S4Vectors':

    rename

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout

Loading required package: reshape2

Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths

The following objects are masked from 'package:data.table':

    dcast, melt

Loading required package: igraph

Attaching package: 'igraph'

The following object is masked from 'package:plotly':

    groups

The following object is masked from 'package:Biostrings':

    union

The following objects are masked from 'package:DelayedArray':

    path, simplify

The following object is masked from 'package:GenomicRanges':

    union

The following object is masked from 'package:IRanges':

    union

The following object is masked from 'package:S4Vectors':

    union
```

```
The following objects are masked from 'package:BiocGenerics':

    normalize, path, union


The following objects are masked from 'package:dplyr':

    as_data_frame, groups, union


The following objects are masked from 'package:purrr':

    compose, simplify


The following object is masked from 'package:tidyr':

    crossing


The following object is masked from 'package:tibble':

    as_data_frame


The following objects are masked from 'package:stats':

    decompose, spectrum


The following object is masked from 'package:base':

    union


Warning message:
"replacing previous import 'GenomicRanges::shift' by 'data.table::shift' when loading 'gC
hain'"
Warning message:
"replacing previous import 'Matrix::%&%' by 'gUtils::%&%' when loading 'gChain'"
Warning message:
"replacing previous import 'GenomicRanges::union' by 'igraph::union' when loading 'skitoo
ls'"
Warning message:
"replacing previous import 'VariantAnnotation::select' by 'plotly::select' when loading '
skitools'"
Warning message:
"replacing previous import 'igraph::groups' by 'plotly::groups' when loading 'skitools'"
Warning message:
"replacing previous import 'ggplot2::last_plot' by 'plotly::last_plot' when loading 'skit
ools'"

Attaching package: 'skitools'


The following objects are masked from 'package:gUtils':

    gr.breaks, ra.duplicated, rebin, standardize_segs


The following object is masked from 'package:ggplot2':

    alpha


The following object is masked from 'package:stats':

    ccf


The following object is masked from 'package:utils':
```

```
    timestamp


Warning message:
"multiple methods tables found for 'seqinfo<-'"

Attaching package: 'gTrack'


The following object is masked from 'package:skitools':

    brewer.master


The following object is masked from 'package:SummarizedExperiment':

    seqinfo<-


The following object is masked from 'package:GenomicRanges':

    seqinfo<-


The following object is masked from 'package:GenomeInfoDb':

    seqinfo<-



Attaching package: 'skidb'


The following object is masked from 'package:skitools':

    read_hg
```

10

Load metadata files containing purity, ploidy of samples, as well as paths of tumor sample files (*tumor_cov* column), paired normal sample file (*normal_cov* column), T/N ratio (*cov* column) and dryclean output files (*dryclean* column).

In [2]:

```
metadata <- readRDS('/gpfs/commons/groups/imielinski_lab/projects/dryclean/MSK_IMPACT/db/pairs.rds')
metadata %>% glimpse()
```

```
Rows: 75
Columns: 13
$ pair         <chr> "P-0000584_1", "P-0003195_1", "P-0004835_1", "P-0009444_1…
$ tumor_sample <chr> "P-0000584-T03-IM6", "P-0003195-T02-IM6", "P-0004835-T02-…
$ Tumor_Type   <chr> "Breast Invasive Ductal Carcinoma", "Breast Invasive Duct…
$ purity       <dbl> 0.87, 0.73, 0.81, 0.23, 0.30, 0.29, 0.15, 0.17, 0.20, 0.2…
$ ploidy       <dbl> 2.1, 3.3, 3.1, 2.3, 4.3, 2.1, 2.2, 2.6, 4.8, 2.0, 2.0, 4.…
$ n_amps       <int> 2, 0, 8, 0, 8, 1, 0, 0, 1, 0, 0, 2, 6, 2, 0, 0, 0, 0, 0, …
$ n_homdels    <int> 0, 1, 0, 0, 0, 4, 0, 2, 0, 0, 4, 0, 0, 5, 0, 0, 0, 0, 3, …
$ frac_homdels <dbl> 0.00000, 0.00070, 0.00000, 0.00000, 0.00000, 0.00410, 0.0…
$ tumor_cov    <chr> "/gpfs/commons/groups/imielinski_lab/data/dryclean/MSK_IM…
$ normal_cov   <chr> "/gpfs/commons/groups/imielinski_lab/data/dryclean/MSK_IM…
$ idx          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 4, 1, …
$ cov          <chr> "/gpfs/commons/groups/imielinski_lab/data/dryclean/MSK_IM…
$ dryclean     <chr> "/gpfs/commons/groups/imielinski_lab/projects/dryclean/MS…
```

Comparing signal pre and post-dryclean

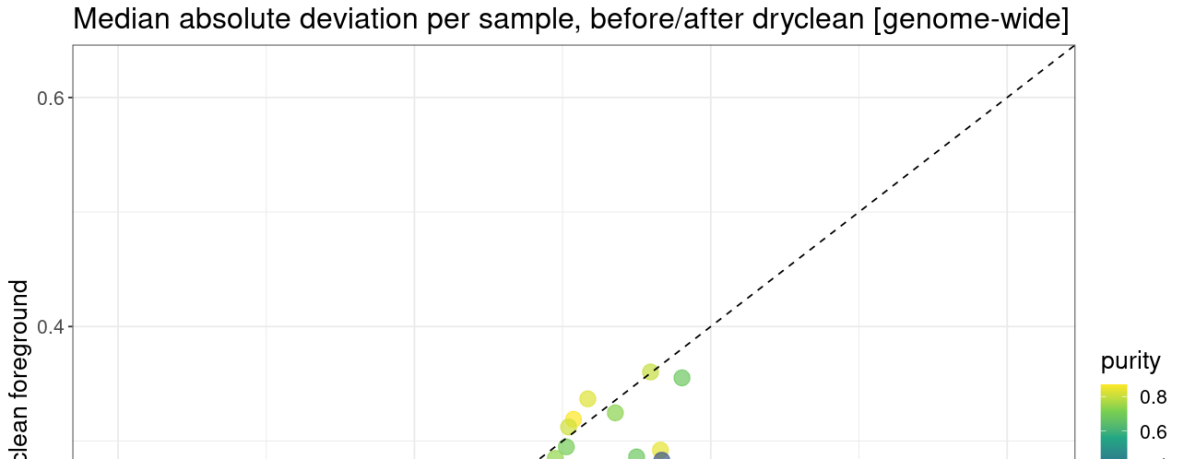
(1) Statistical dispersion of signal in pre-dryclean tumor/normal ratio vs. post-dryclean foreground

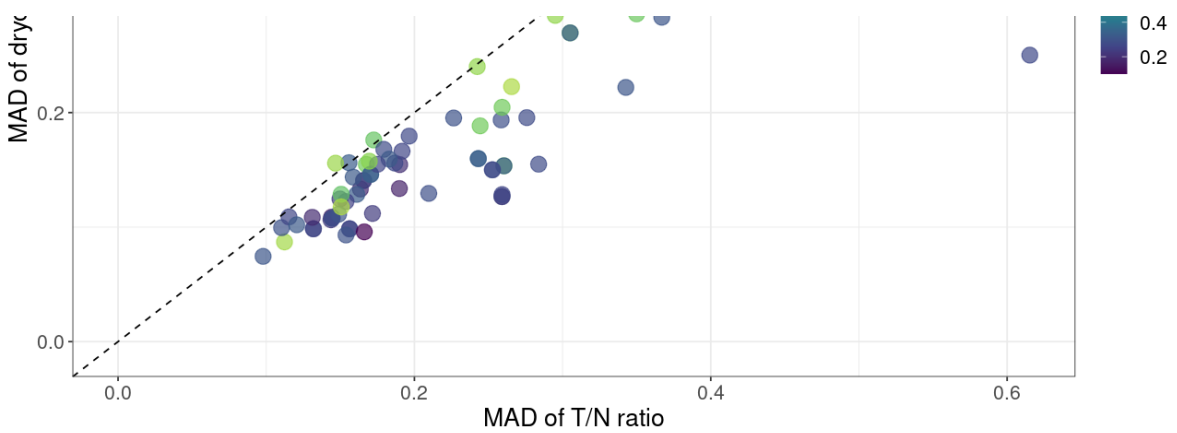
In [4]:

```
MAD_stats = NULL

for (idx in seq(1,dim(metadata)[1])){
    pre_sample <- readRDS(metadata[idx,]$cov) %>% as_tibble() %>%
        select(seqnames, start, end, ratio)
    post_sample <- readRDS(metadata[idx,]$dryclean) %>% as_tibble() %>%
        select(seqnames, start, end, foreground)

    data <- pre_sample %>% full_join(post_sample, by = c("seqnames", "start", "end")) %>
%
        gather(signal, value, -start, -end, -seqnames)
    total_stats <- data %>% group_by(signal) %>% summarise(MAD = mad(value)) %>% mutate(
seqnames='total')
    sample_stats <- data %>%
        group_by(signal, seqnames) %>%
        summarise(MAD = mad(value), .groups = 'drop') %>%
        full_join(total_stats, by=c('signal', 'seqnames', 'MAD')) %>%
        mutate(sample=metadata[idx,]$pair)

    if (is.null(MAD_stats)){
        MAD_stats <- sample_stats
    } else {
        MAD_stats <- MAD_stats %>% full_join(sample_stats, by=c('signal', 'seqnames', 'M
AD', 'sample'))
    }
}
```

In [18]:

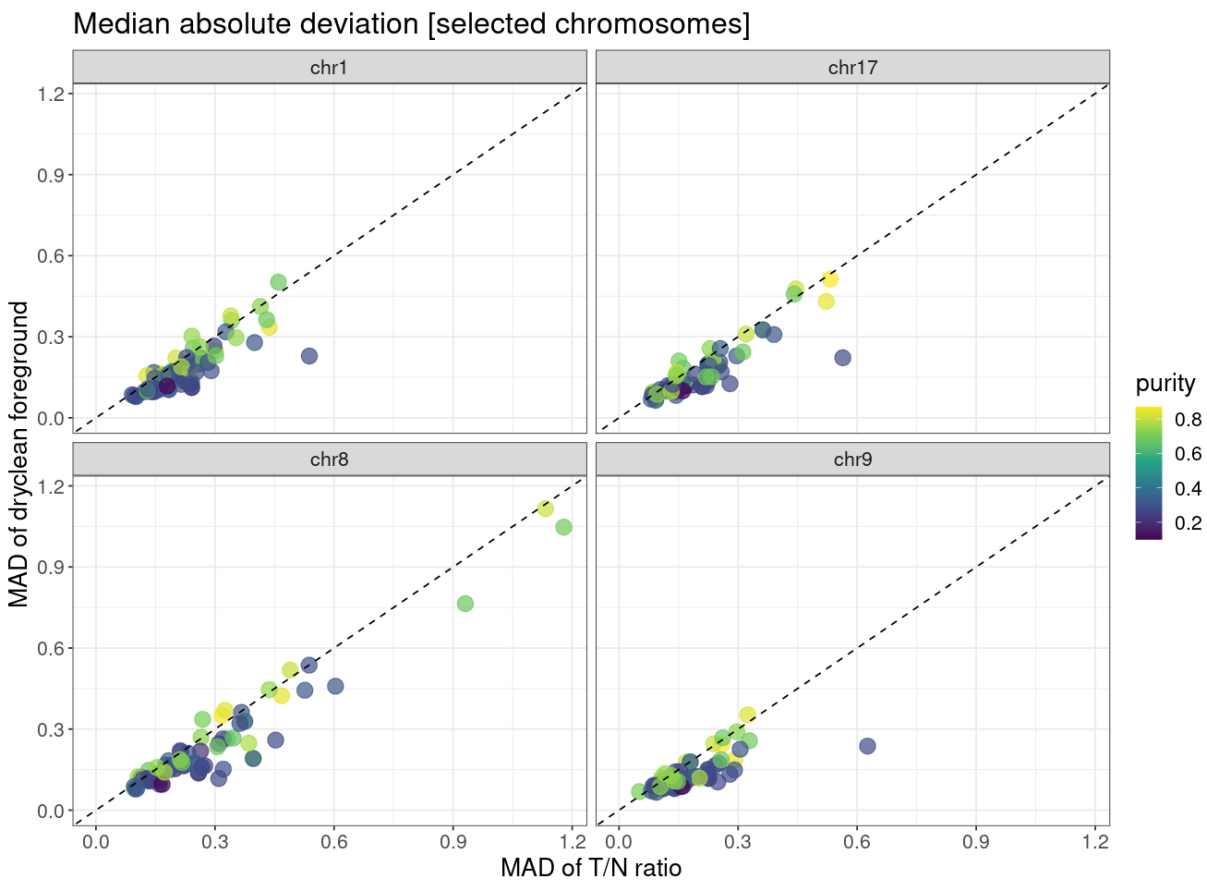
```
options(repr.plot.width=10.7, repr.plot.height=8)

mad_plot_data <- metadata %>%
    select(pair, purity, ploidy) %>%
    right_join(MAD_stats, by=c('pair'='sample')) %>%
    spread(signal, MAD) %>%
    filter(seqnames=="total")

mad_plot_data %>%
    ggplot() +
        geom_point(aes(x=ratio, y=foreground, color=purity, size=1), alpha=0.7) +
        geom_abline(intercept=0, slope=1, linetype='dashed') +
        #ggrepel::geom_text_repel(aes(x=ratio, y=foreground, label=purity)) +
        xlim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        ylim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        theme_bw() + theme(legend.position="right", text = element_text(size=16)) +
        guides(size="none") +
        labs(x="MAD of T/N ratio", y="MAD of dryclean foreground",
             title="Median absolute deviation per sample, before/after dryclean [genome-w
ide]") +
        scale_color_viridis_c()
```





In [17]:

```
options(repr.plot.width=11, repr.plot.height=8)

mad_plot_data <- metadata %>%
    select(pair, purity, ploidy) %>%
    right_join(MAD_stats, by=c('pair'='sample')) %>%
    spread(signal, MAD) %>%
    filter(seqnames %in% c('1', '17', '8', '9'))

mad_plot_data %>%
    ggplot() +
        geom_point(aes(x=ratio, y=foreground, color=purity, size=1), alpha=0.7) +
        geom_abline(intercept=0, slope=1, linetype='dashed') +
        #ggrepel::geom_text_repel(aes(x=ratio, y=foreground, label=purity)) +
        xlim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        ylim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        theme_bw() + theme(legend.position="right", text = element_text(size=16)) +
        guides(size="none") +
        labs(x="MAD of T/N ratio", y="MAD of dryclean foreground",
             #title="Median absolute deviation per sample, before/after dryclean [selecte
d chromosomes]"
             title="Median absolute deviation [selected chromosomes]"
            ) +
        scale_color_viridis_c() +
        facet_wrap(paste0("chr", seqnames)~.)
```



(2) Signal distribution in pre-dryclean samples [possible oversampling]

In [21]:

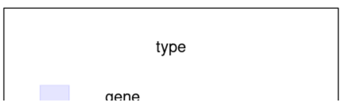
```
Sys.setenv(SKI_DB_ROOT="~/DB/")
Sys.setenv(SKI_SOFTWARE_ROOT="~/Software/")

ge = read_gencode()
exons = ge %Q% (type == 'exon')
genes = ge %Q% (type == 'gene')

gt.ge = track.gencode() #hg19
```

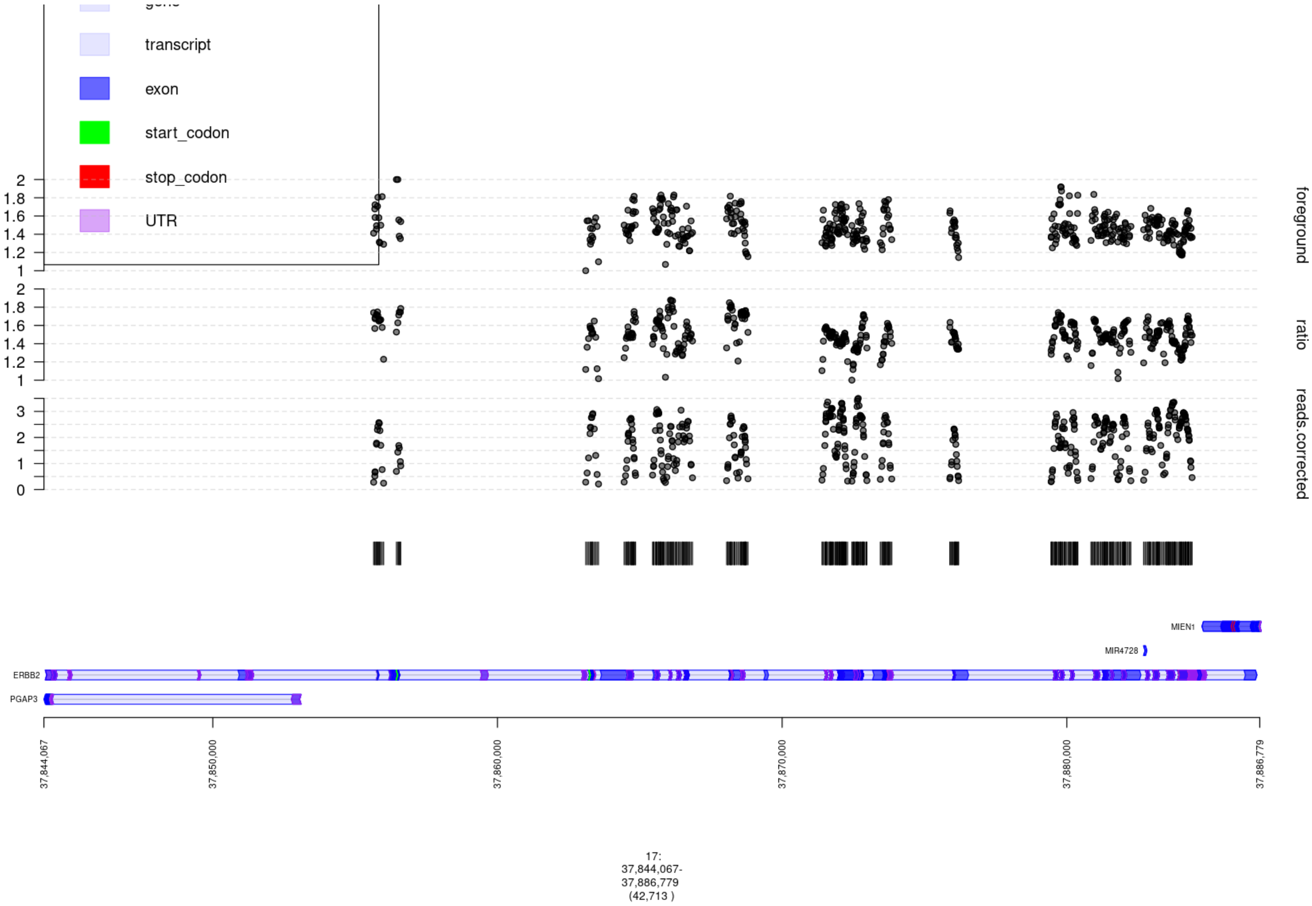
```
Warning message in track.gencode():
“Downloading GENCODE track for genome build hg19 from http://mskilab.com/gTrack/hg19//gen
code.composite.collapsed.rds - for quicker loading, download this file locally and point
GENCODE_DIR environment variable to the enclosing directory.  To generate and use a new c
ached gTrack object from a GENCODE .gtf or .gff3 set GENCODE_DIR env variable to an exist
ing local directory and run track.gencode(url_or_path_to_gencode_gtf) once, and then use
track.gencode() subsequently to access that cached gTrack object.”
Warning message in track.gencode():
“Pulling gencode annotations from http://mskilab.com/gTrack/hg19//gencode.composite.colla
psed.rds”
```

In [25]:

```
options(repr.plot.width=15, repr.plot.height=13)

idx <- 1

tcov = metadata[idx, tumor_cov] %>% readRDS()
ncov = metadata[idx, normal_cov] %>% readRDS()
cov = metadata[idx, cov] %>% readRDS()
dc = metadata[idx, dryclean] %>% readRDS()

gt.ncov = gTrack(ncov, 'reads.corrected', circle=TRUE, lwd.border=0.8)
gt.tcov = gTrack(tcov, 'reads.corrected', circle=TRUE, lwd.border=0.8)
gt.cov = gTrack(cov, 'ratio', circle=TRUE, lwd.border=0.8)
gt.dcf = gTrack(dc, 'foreground', circle=TRUE, lwd.border=0.8)
gt.dcb = gTrack(dc, 'background', circle=TRUE, lwd.border=0.8)

win = (genes %Q% (gene_name == 'ERBB2') + 1e2) %&% exons %Q% (1)

gtr = gTrack(reduce(cov))

plot(c(gt.ge, gtr, gt.tcov, gt.cov, gt.dcf), win)
```

```
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
```

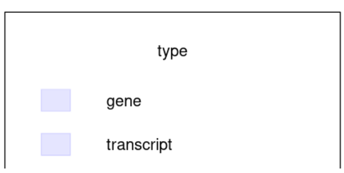


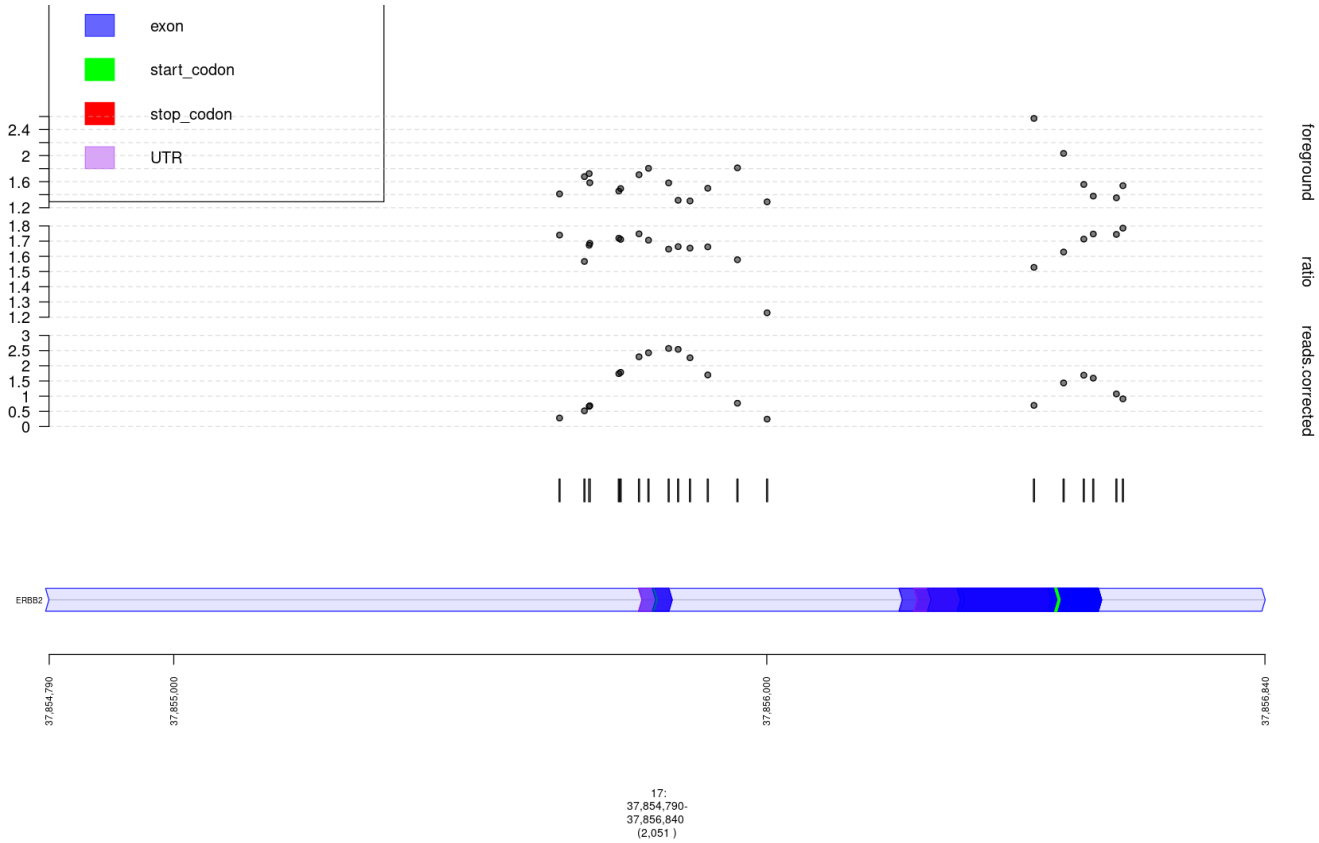


In [29]:

```
options(repr.plot.width=15, repr.plot.height=13)
win = exons %&% reduce(cov) %&% ((genes %Q% (gene_name == 'ERBB2') + 1e3)) %&% exons %Q%
(1) + 1e3

plot(c(gt.ge, gtr, gt.tcov, gt.cov, gt.dcf), win)
```

```
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(gr, windows):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(gr, windows):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(gr, windows):
"findOverlaps applied to ranges with non-identical seqlengths"
```





The “bell-shape curved” signal for exon regions in raw data. Neighbouring probes could be collapsed (while averaging the signal).

(3) Collapsing regions - parameter selection

In [30]:

```
width_data = NULL

for (i in c(seq(10,40,10), seq(50,500,50))){
    probes = (reduce(cov + i)-i)
    probes_tbl <- width(probes) %>%
        as_tibble() %>%
        mutate(pad=i)
    if (is.null(width_data)){
        width_data <- probes_tbl
    } else {
        width_data <- width_data %>% full_join(probes_tbl, by=c('value', 'pad'))
    }
}
```

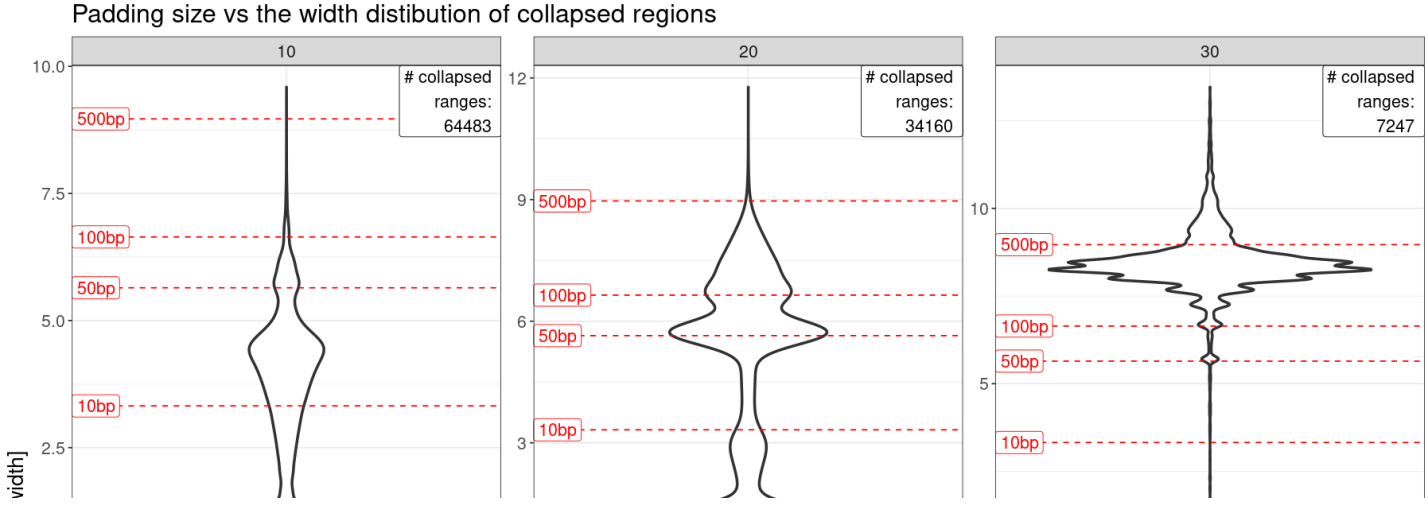
```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 24 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 28 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
```

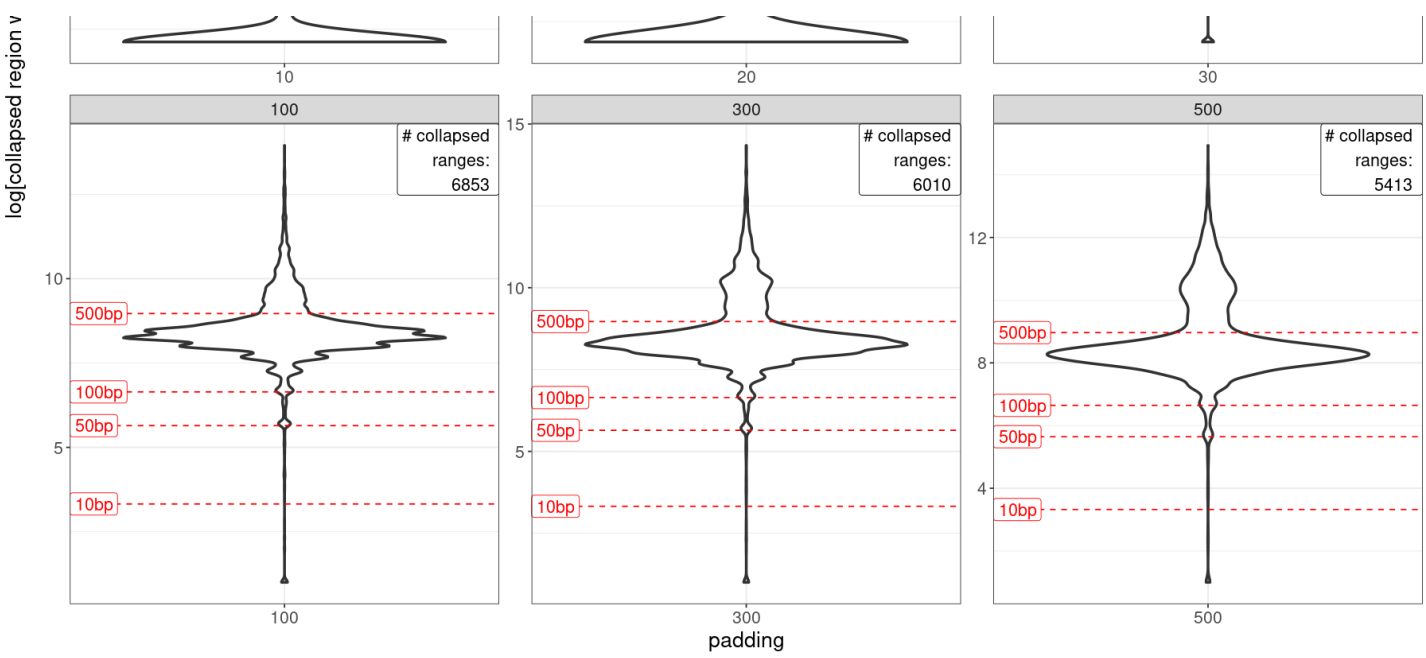
```
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 32 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 35 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 39 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 143 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 182 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 211 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 229 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 239 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
```

```
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 245 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 255 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 258 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

In [33]:

```
options(repr.plot.width=15, repr.plot.height=12)
annot_width = tibble(width=c(log2(10), log2(50), log2(100), log2(500)),
                     label=c("10bp", "50bp", "100bp", "500bp"))
num_ranges_tbl <- width_data %>%
    filter(pad %in% c(10,20,30,100, 300, 500)) %>%
    mutate(pad=as.factor(as.character(pad))) %>%
    group_by(pad) %>%
    summarise(num_ranges=length(value))

width_data %>%
    filter(pad %in% c(10,20,30,100, 300, 500)) %>%
    mutate(log_width=log2(value), pad=as.character(pad)) %>%
    mutate(pad = factor(pad, levels = c('10', '20', '30', '100', '300', '500'))) %>%
    ggplot() +
        geom_violin(aes(x=pad, y=log_width, group=pad), size=1) +
        geom_hline(yintercept = annot_width$width, linetype="dashed", color="red") +
        geom_label(data=annot_width, aes(x=-Inf, y=width, label=label, size=10), color="
red", hjust="inward") +
        geom_label(data=num_ranges_tbl, aes(x=Inf, y=Inf, label=paste0("# collapsed \nra
nges: \n", num_ranges," "),
                                            size=10), hjust="inward", vjust="inward") +
        theme_bw() +
        facet_wrap(pad~., scales="free") +
        theme(text = element_text(size=16), legend.position="none") +
        labs(x="padding", y="log[collapsed region width]",
            title="Padding size vs the width distibution of collapsed regions"
            )
```





In [34]:

```
num_ranges_data = tibble(pad=0,
                         num_ranges=reduce(cov) %>% length(),
                         num_ranges_over50=(reduce(cov) %Q% (width>50) %>% length()))
for (i in c(seq(10,40,10), seq(50,500,50))){
    num_ranges_data <- add_row(num_ranges_data, pad=i,
                               num_ranges=((reduce(cov + i)-i) %>% length()),
                               num_ranges_over50=((reduce(cov + i)-i) %Q% (width>50) %>%
length()))
}
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 24 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 24 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 28 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 28 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 32 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
```

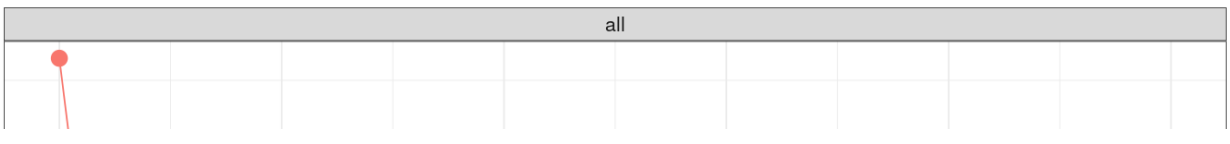
```
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 32 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 35 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 35 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 39 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 39 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 143 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 143 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
```

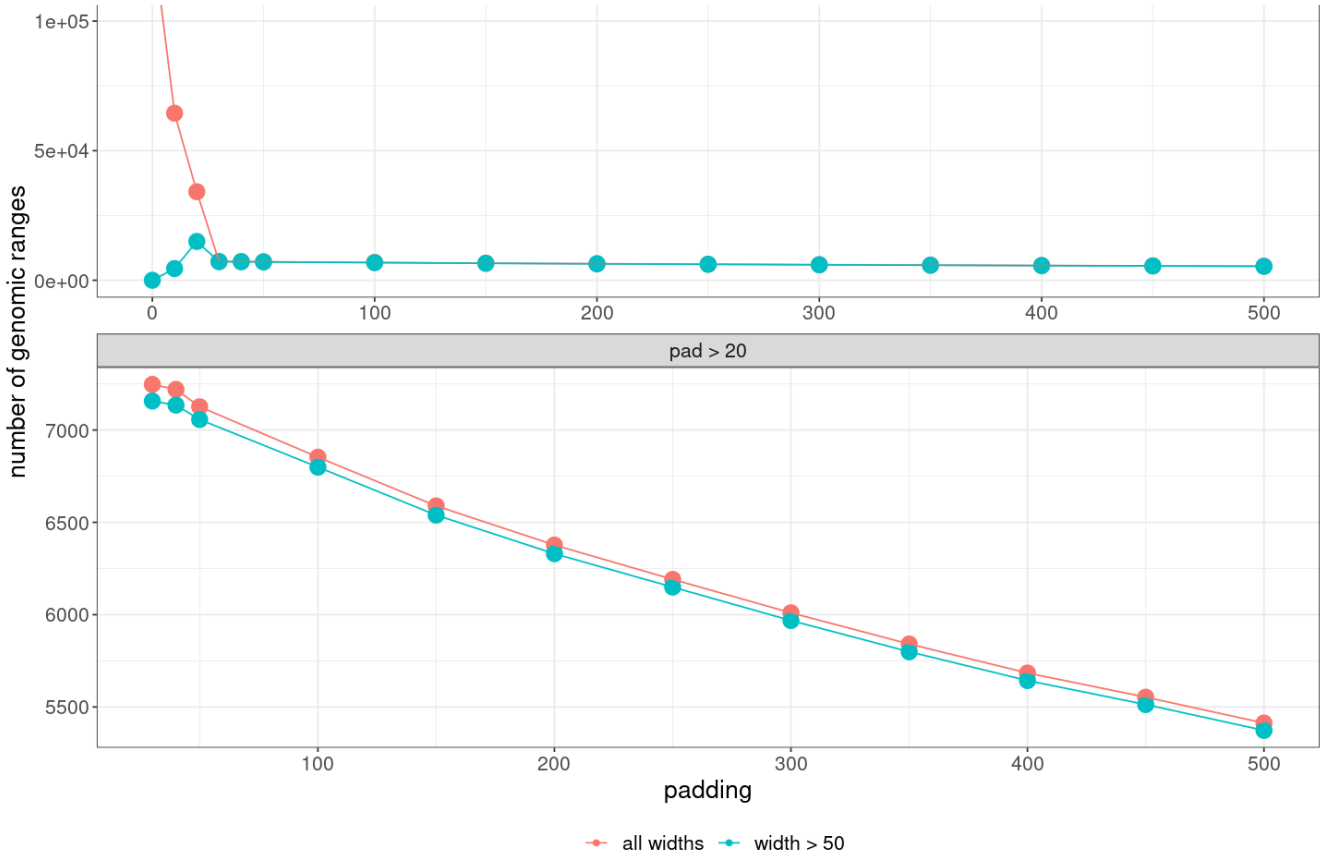
```
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 182 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 182 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 211 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 211 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 229 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 229 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 239 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 239 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 245 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
```

```
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 245 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 255 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 255 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 258 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 258 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

In [35]:

```
options(repr.plot.width=12, repr.plot.height=9)
num_ranges_plot_data <- num_ranges_data %>%
    filter(pad > 20) %>%
    mutate(plot_facet = "pad > 20") %>%
    full_join(num_ranges_data %>% mutate(plot_facet="all"),
              by = c("pad", "num_ranges", "plot_facet", "num_ranges_over50")) %>%
    gather(type, val, -pad, -plot_facet) %>%
    mutate(type = as.factor(type))

num_ranges_plot_data$type <- recode_factor(num_ranges_plot_data$type, num_ranges='all wid
ths',
                                           num_ranges_over50='width > 50')

num_ranges_plot_data %>%
    ggplot() +
        geom_point(aes(x=pad, y=val, size=1, group=type, color=type)) +
        geom_line(aes(x=pad, y=val, group=type, color=type)) +
        theme_bw() +
        theme(text = element_text(size=16), legend.position="bottom") +
        facet_wrap(plot_facet~., ncol=1, scales="free") +
        labs(y = "number of genomic ranges", x="padding", color="") +
        guides(size="none")
```



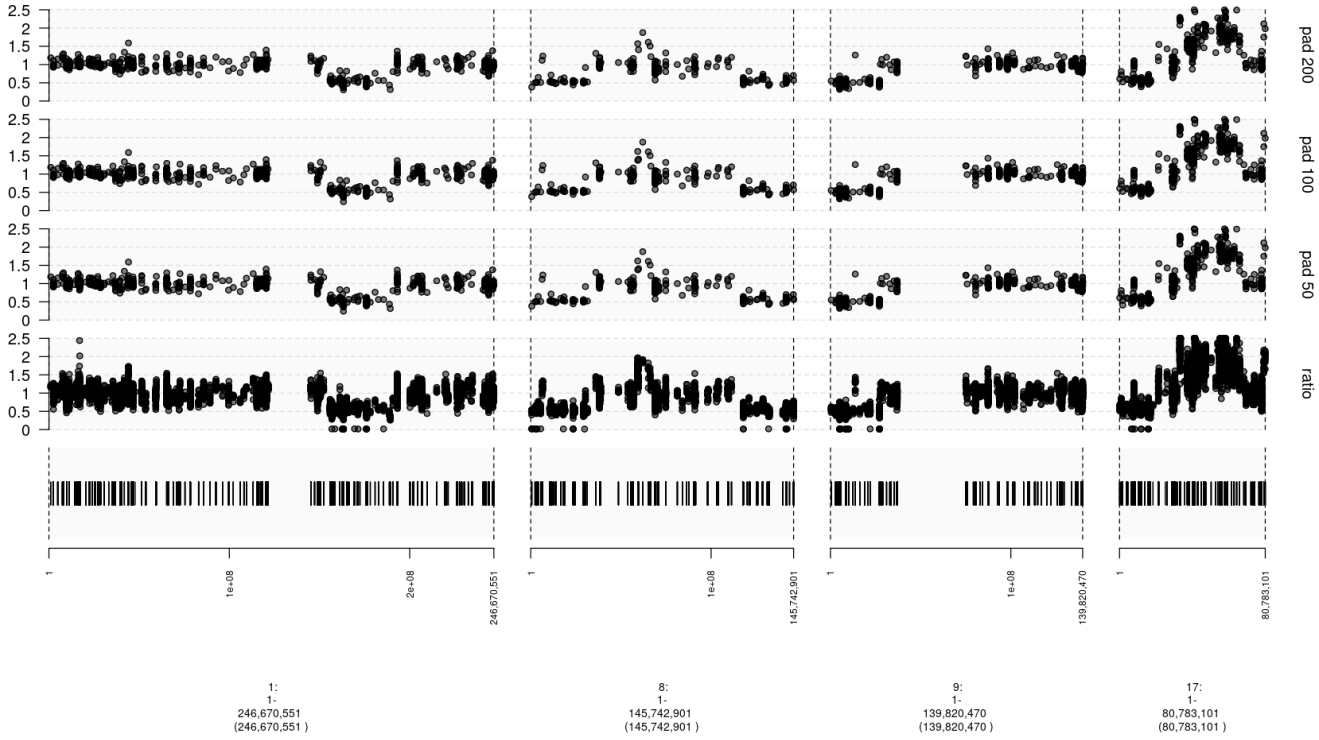


In [36]:

```
options(repr.plot.width=15, repr.plot.height=13)

pad = 50
probes50 = (reduce(cov + pad)-pad) %Q% (order(width)) %Q% (width>50)
pad = 100
probes100 = (reduce(cov + pad)-pad) %Q% (order(width)) %Q% (width>50)
pad = 200
probes200 = (reduce(cov + pad)-pad) %Q% (order(width)) %Q% (width>50)

gt.probes = gTrack(probes, height = 10)

covp = probes %$% cov[,'ratio']
tcovp = probes %$% tcov[,'reads.corrected']
ncovp = probes %$% ncov[,'reads.corrected']

gt.ncovp = gTrack(ncovp, 'reads.corrected', circle=TRUE, lwd.border=0.8)
gt.tcovp = gTrack(tcovp, 'reads.corrected', circle=TRUE, lwd.border=0.8)
gt.covp = gTrack(covp, 'ratio', circle=TRUE, lwd.border=0.8)

plot(c(gtr,
       gt.cov,
       gTrack(probes50 %$% cov[,'ratio'], 'ratio', circle=TRUE, lwd.border=0.8, name="pa
d 50"),
       gTrack(probes100 %$% cov[,'ratio'], 'ratio', circle=TRUE, lwd.border=0.8, name="p
ad 100"),
       gTrack(probes200 %$% cov[,'ratio'], 'ratio', circle=TRUE, lwd.border=0.8, name="p
ad 200")),
     c(1, 8, 9, 17))
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 39 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
```

```
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 182 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```



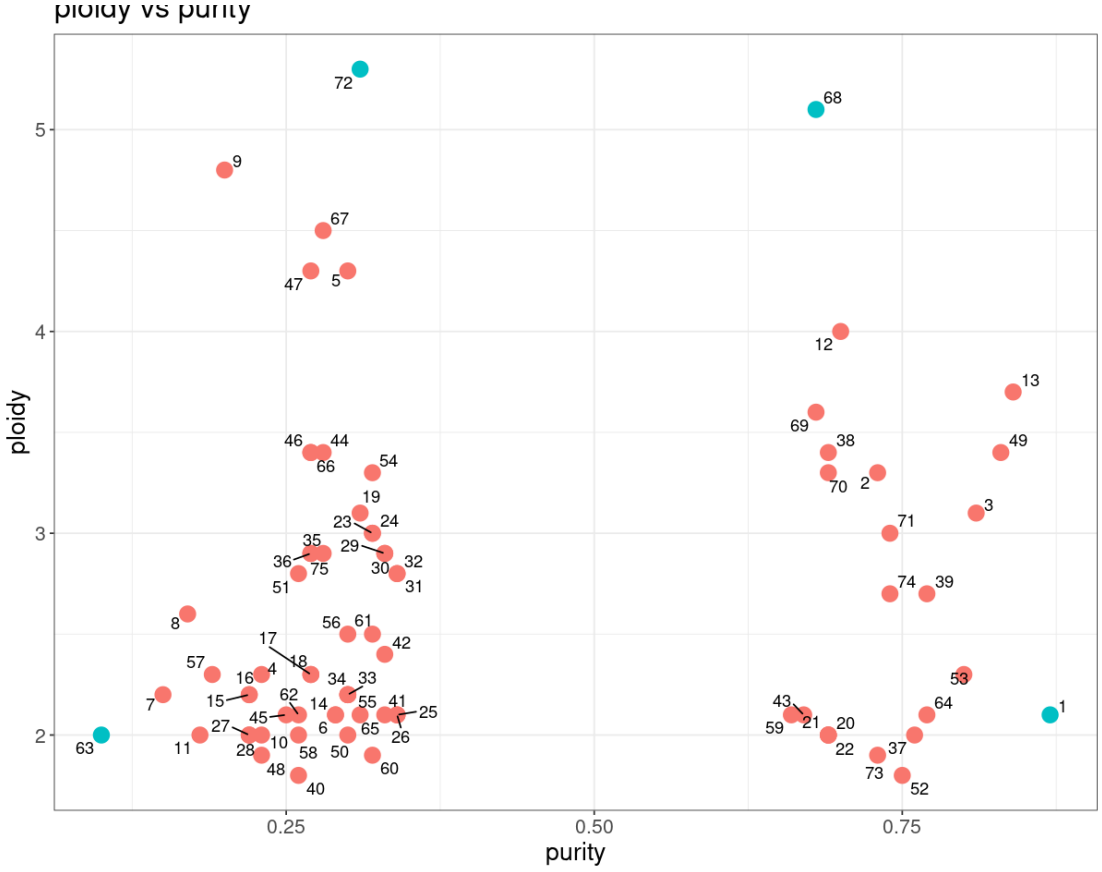
Inspecting change in signal on samples with different purity/ploidy.

In [37]:

```
options(repr.plot.width=10, repr.plot.height=8)
metadata %>%
    mutate(idx=1:dim(metadata)[1], selected=idx %in% c(1, 68, 72, 63)) %>%
    ggplot() +
    geom_point(aes(x=purity, y=ploidy, size=1, color=selected)) +
    ggrepel::geom_text_repel(aes(x=purity, y=ploidy, label=idx)) +
    theme_bw() +
    theme(text=element_text(size=16), legend.position="none") +
    labs(title="ploidy vs purity")
```

```
Warning message:
“ggrepel: 6 unlabeled data points (too many overlaps). Consider increasing max.overlaps”
```

ploidy vs purity

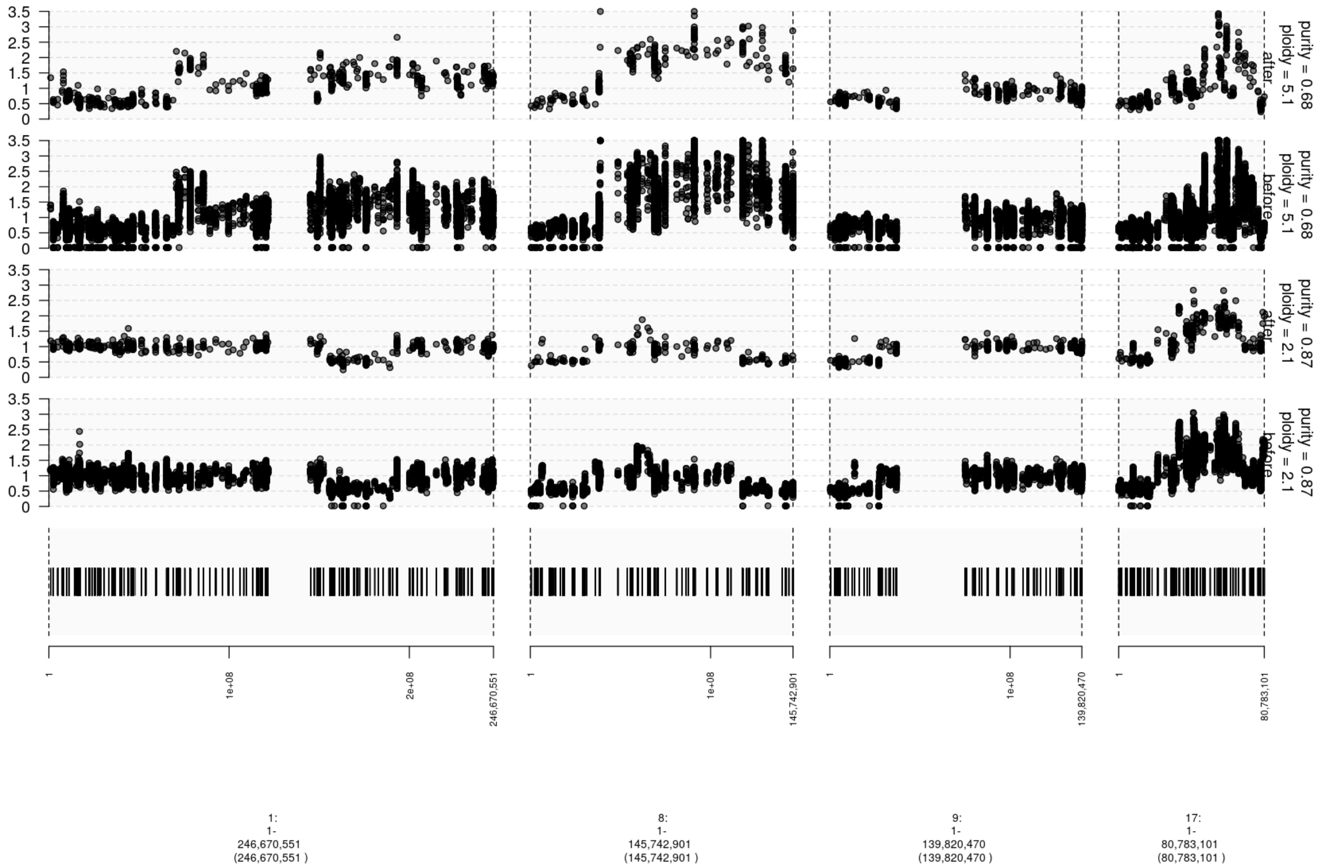


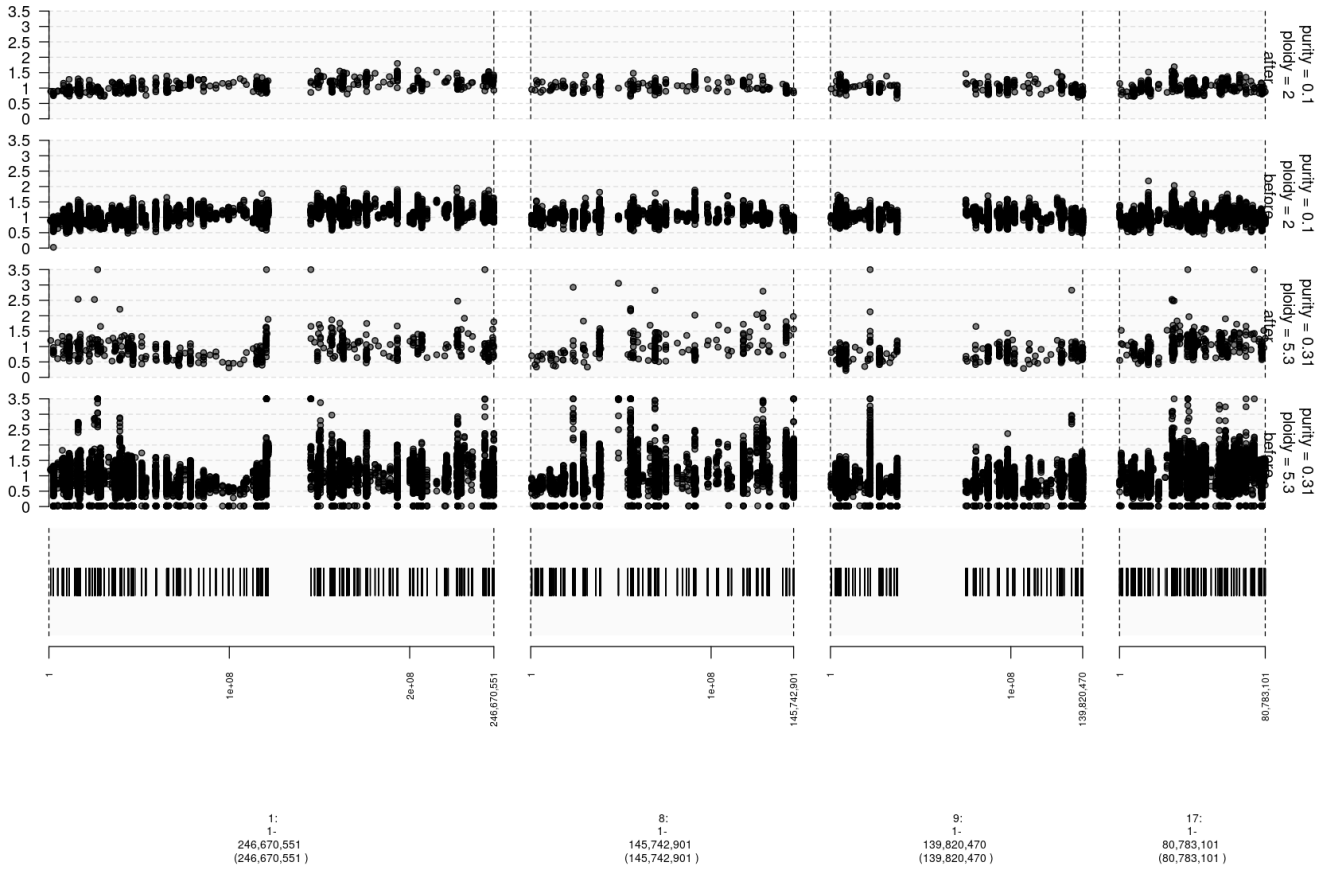
In [38]:

```
options(repr.plot.width=15, repr.plot.height=15)
get_tracks <- function(idx, pad=100, w=50){
    sample = metadata[idx, ]
    tn_cov = sample$cov %>% readRDS()
    probes = (reduce(tn_cov + pad)-pad) %Q% (order(width)) %Q% (width>w)
    return(
        list(collapsed=gTrack(probes %$% tn_cov[,'ratio'], 'ratio', circle=TRUE,
                              lwd.border=0.8, name=paste0("purity = ", sample$purity,
                                                          "\nploidy = ", sample$ploidy,
                                                          "\n after"), y0 = 0, y1 = 3.5
),
             ratio=gTrack(tn_cov, 'ratio', circle=TRUE,
                          lwd.border=0.8, name=paste0("purity = ", sample$purity,
                                                          "\nploidy = ", sample$ploidy,
                                                          "\n before"), y0 = 0, y1 = 3.
5))
        )
}


tracks1 <- get_tracks(idx=1)
tracks2 <- get_tracks(idx=68)
tracks3 <- get_tracks(idx=72)
tracks4 <- get_tracks(idx=63)

plot(c(gtr,
       tracks1$ratio, tracks1$collapsed,
       tracks2$ratio, tracks2$collapsed),
     c(1, 8, 9, 17))

plot(c(gtr,
       tracks3$ratio, tracks3$collapsed,
       tracks4$ratio, tracks4$collapsed),
     c(1, 8, 9, 17))
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
```

```
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```



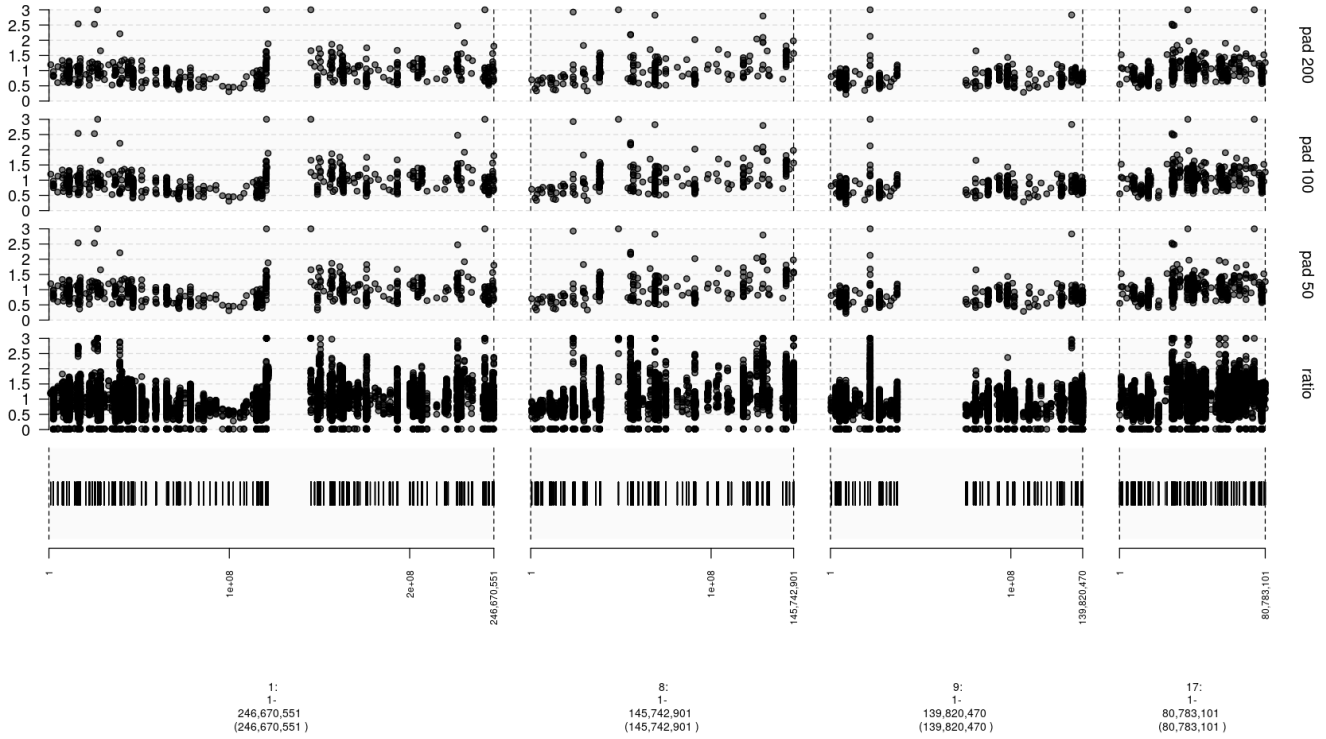


In [39]:

```
options(repr.plot.width=15, repr.plot.height=13)

idx <- 72
sample = metadata[idx, ]
tn_cov = sample$cov %>% readRDS()
probes50 = (reduce(tn_cov + 50)-50) %Q% (order(width)) %Q% (width>50)
probes100 = (reduce(tn_cov + 100)-100) %Q% (order(width)) %Q% (width>50)
probes200 = (reduce(tn_cov + 200)-200) %Q% (order(width)) %Q% (width>50)

gt.ratio = gTrack(tn_cov, 'ratio', circle=TRUE, lwd.border=0.8, name='ratio', y0 = 0, y1
= 3)
gt.p50 = gTrack(probes50 %$% tn_cov[,'ratio'], 'ratio', circle=TRUE,
                            lwd.border=0.8, name='pad 50', y0 = 0, y1 = 3)
gt.p100 = gTrack(probes100 %$% tn_cov[,'ratio'], 'ratio', circle=TRUE,
                            lwd.border=0.8, name='pad 100', y0 = 0, y1 = 3)
gt.p200 = gTrack(probes200 %$% tn_cov[,'ratio'], 'ratio', circle=TRUE,
                            lwd.border=0.8, name='pad 200', y0 = 0, y1 = 3)

plot(c(gtr, gt.ratio, gt.p50, gt.p100, gt.p200), c(1, 8, 9, 17))
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 39 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 182 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```



Selected parameters: collapsing probes within a 100bp from each other and filtering out regions smaller than 50bp. That gives 6799 regions.

(4) Collapsing regions

In [44]:

```
# collapse normal samples (for PON)
normal_table <- readRDS('/gpfs/commons/groups/imielinski_lab/data/dryclean/MSK_IMPACT/nor
mal_1k/normal_table.rds')
normal_out_dir <- '/gpfs/commons/home/ksienkiewicz/dryclean/BRCA_normal_pad100_width50'

pad = 100
w = 50

updated_paths = c()
nreg = c()
for (idx in 1:dim(normal_table)[1]){
    ncov <- normal_table[idx, normal_cov] %>% readRDS

    reduced_ncov <- (reduce(ncov + pad, with.revmap = TRUE) - pad) %Q% (order(width)) %Q
% (width>w)
    mcols(reduced_ncov)$reads.corrected <- aggregate(ncov, mcols(reduced_ncov)$revmap,
                                                     reads.corrected = mean(reads.corre
cted, na.rm=TRUE))$reads.corrected

    nreg <- c(nreg, length(reduced_ncov))
    fpath = file.path(normal_out_dir, paste0(normal_table[idx, sample], '.rds'))
    updated_paths = c(updated_paths, fpath)
    saveRDS(reduced_ncov, file = fpath)
}

normal_table_updated <- normal_table
normal_table_updated$normal_cov <- updated_paths
saveRDS(normal_table_updated, file.path(normal_out_dir, 'normal_table.rds'))
```

In [46]:

```
# collapse BRCA tumor samples, normal samples and ratios
tumor_out_dir <- '/gpfs/commons/home/ksienkiewicz/dryclean/BRCA_pad100_width50'

pad = 100
w = 50

stats <- NULL
updated_paths_tumor = c()
updated_paths_normal = c()
updated_paths_tn = c()
nreg_tumor = c()
nreg_tn = c()
for (idx in 1:dim(metadata)[1]){
    print(idx)
    sample <- metadata[idx, pair]

    ncov <- metadata[idx, normal_cov] %>% readRDS
    reduced_ncov <- (reduce(ncov + pad, with.revmap = TRUE) - pad) %Q% (order(width)) %Q
% (width>w)
    mcols(reduced_ncov)$reads.corrected <- aggregate(ncov, mcols(reduced_ncov)$revmap,
                                                     reads.corrected = mean(reads.corre
cted, na.rm=TRUE))$reads.corrected
    normal_path = file.path(tumor_out_dir, 'normal_samples', paste0(sample, '_normal.rds
'))
    saveRDS(reduced_ncov, file = normal_path)

    tcov <- metadata[idx, tumor_cov] %>% readRDS
    reduced_tcov <- (reduce(tcov + pad, with.revmap = TRUE) - pad) %Q% (order(width)) %Q
% (width>w)
    mcols(reduced_tcov)$reads.corrected <- aggregate(tcov, mcols(reduced_tcov)$revmap,
                                                     reads.corrected = mean(reads.corre
cted, na.rm=TRUE))$reads.corrected
    tumor_path = file.path(tumor_out_dir, 'tumor_samples', paste0(sample, '_tumor.rds'))
    saveRDS(reduced_tcov, file = tumor_path)

    tncov <- metadata[idx, cov] %>% readRDS
    reduced_tncov <- (reduce(tncov + pad, with.revmap = TRUE) - pad) %Q% (order(width))
```

```
%Q% (width>w)
    mcols(reduced_tncov)$ratio <- aggregate(tncov, mcols(reduced_tncov)$revmap, ratio =
mean(ratio, na.rm=TRUE))$ratio
    mcols(reduced_tncov)$log.ratio <- aggregate(tncov, mcols(reduced_tncov)$revmap,
                                                log.ratio = mean(log.ratio, na.rm=TRUE)
)$log.ratio
    mcols(reduced_tncov)$reads.corrected.tum <- aggregate(tncov, mcols(reduced_tncov)$rev
map,
                                                          reads.corrected.tum = mean(re
ads.corrected.tum, na.rm=TRUE))$reads.corrected.tum
    mcols(reduced_tncov)$reads.corrected.norm <- aggregate(tncov, mcols(reduced_tncov)$r
evmap,
                                                           reads.corrected.norm = mean(
reads.corrected.norm, na.rm=TRUE))$reads.corrected.norm
    tn_path = file.path(tumor_out_dir, 'TN_samples', paste0(sample, '_TN.rds'))
    saveRDS(reduced_tncov, file = tn_path)

    sample_tbl <- tibble(pair=sample, normal_cov=normal_path, tumor_cov=tumor_path, cov=
tn_path,
                         nreg_normal=length(reduced_ncov), nreg_tumor=length(reduced_tcov
),
                          nreg_tn=length(reduced_tncov))
    if(is.null(stats)){
        stats <- sample_tbl
    } else {
        stats <- stats %>% full_join(sample_tbl, by=c('pair', 'normal_cov', 'tumor_cov',
'cov',
                                                      'nreg_normal', 'nreg_tumor', 'nreg
_tn'))
    }
}

metadata_updated <- metadata
metadata_updated$normal_cov <- stats$normal_cov
metadata_updated$tumor_cov <- stats$tumor_cov
metadata_updated$cov <- stats$cov
```

```
[1] 1
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 2
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 3
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 4
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
```

```
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 5
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 6
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 7
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 8
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 9
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 10
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 11
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 12
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 13
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 14
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 15
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 16
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 17
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 18
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 19
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 20
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 21
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 22
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 23
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 24
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
```

```
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 25
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 26
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 27
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 28
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 29
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 30
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 31
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
```

```
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 32
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 33
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 34
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 35
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 36
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 37
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 38
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
```

```
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 39
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 40
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 41
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 42
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 43
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 44
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 45
```

- -

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 46
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 47
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 48
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 49
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 50
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 51
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

ranges. See ?`trim,GenomicRanges-method` for more information."

```
[1] 52
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 53
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 54
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 55
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 56
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 57
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 58
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
```

```
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 59
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 60
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 61
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 62
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 63
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 64
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 65
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
```

```
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 66
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 67
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 68
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 69
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 70
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 71
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

```
[1] 72
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
```

```
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 73
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 74
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

```
[1] 75
```

```
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
"GRanges object contains 90 out-of-bound ranges located on sequences 1,
  2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
  22, and X. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information."
```

In [50]:

```
stats %>% select(matches("nreg")) %>% unique()
```

A tibble: 1 × 3

nreg_normal	nreg_tumor	nreg_tn
<int>	<int>	<int>
6799	6799	6799

(5) Creating PON for collapsed regions

In [ ]:

```
normal_out_dir <- '/gpfs/commons/home/ksienkiewicz/dryclean/BRCA_normal_pad100_width50'

detergent = prepare_detergent(normal.table.path = file.path(normal_out_dir, 'normal_table
.rds'),
             path.to.save = normal_out_dir, num.cores = 10, use.all = TRUE, balance = FA
LSE)
saveRDS(detergent, file.path(normal_out_dir, 'pon.rds'))
```

In [52]:

```
dryclean_paths <- c()
for (idx in 1:dim(metadata_updated)[1]){
    sample <- metadata_updated[idx, pair]
    tumor.sample <- readRDS(metadata_updated[idx, tumor_cov])
```

```
    drycleaned_path = file.path(tumor_out_dir, 'dryclean', paste0(sample, '_drycleaned.t
umor_cov.rds'))
    cov_out = start_wash_cycle(cov = tumor.sample, mc.cores = 10,
                               detergent.pon.path = file.path(normal_out_dir, 'pon.rds')
, germline.filter = FALSE)
    saveRDS(cov_out, file=drycleaned_path)
    dryclean_paths <- c(dryclean_paths, drycleaned_path)
}
metadata_updated$dryclean <- dryclean_paths
```

(6) Collapsing results

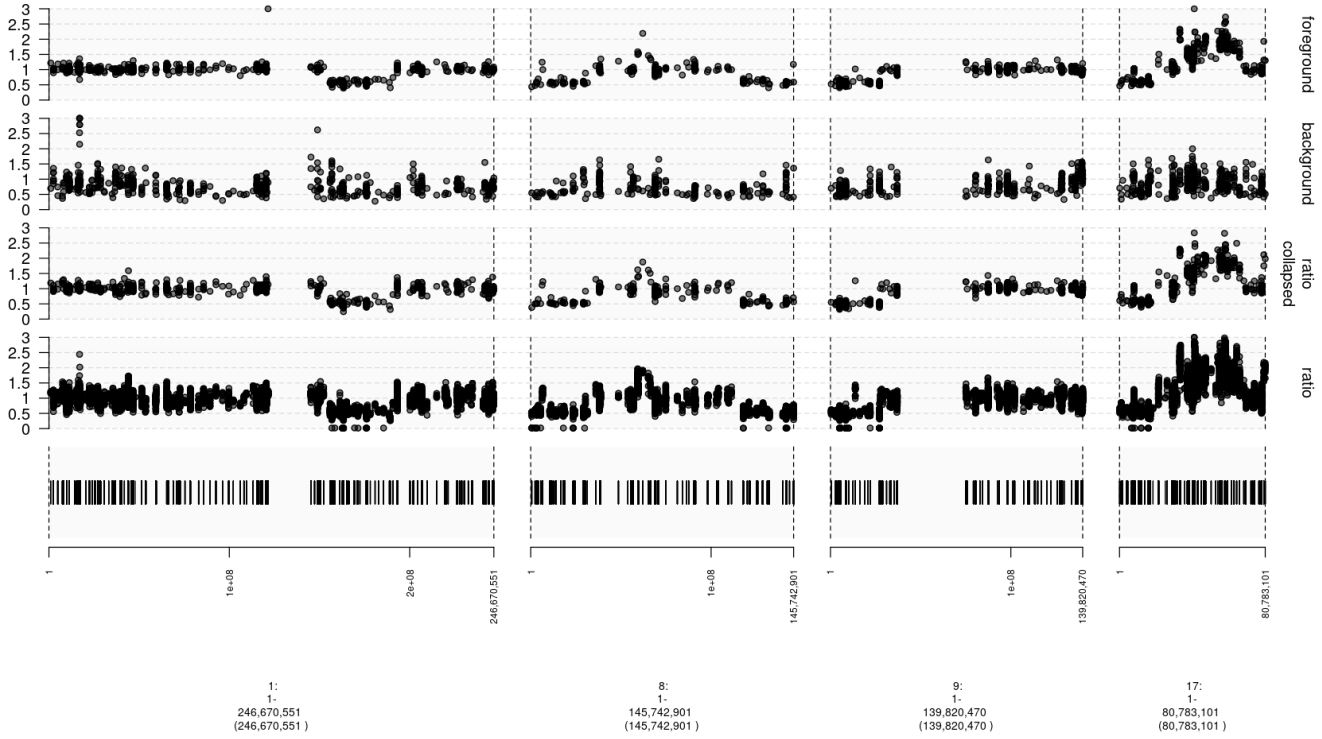
High-purity sample

In [55]:

```
options(repr.plot.width=15, repr.plot.height=13)

idx <- 1
cov_before = metadata[idx, cov] %>% readRDS()
cov_after = metadata_updated[idx, cov] %>% readRDS()
dcln = metadata_updated[idx, dryclean] %>% readRDS()

gt.ratio.before = gTrack(cov_before, 'ratio', circle=TRUE, lwd.border=0.8, name='ratio',
y0 = 0, y1 = 3)
gt.ratio.after = gTrack(cov_after, 'ratio', circle=TRUE, lwd.border=0.8, name='ratio\nco
llapsed', y0 = 0, y1 = 3)
gt.fg = gTrack(dcln, 'foreground', circle=TRUE, lwd.border=0.8, name='foreground', y0 =
0, y1 = 3)
gt.bg = gTrack(dcln, 'background', circle=TRUE, lwd.border=0.8, name='background', y0 =
0, y1 = 3)

plot(c(gtr, gt.ratio.before, gt.ratio.after, gt.bg, gt.fg), c(1, 8, 9, 17))
```

```
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 4 out-of-bound ranges located on sequences 1,
  8, 9, and 17. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 4 out-of-bound ranges located on sequences 1,
  8, 9, and 17. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 4 out-of-bound ranges located on sequences 1,
  8, 9, and 17. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
```

flags of the underlying sequences). You can use trim() to trim these
ranges. See ?`trim,GenomicRanges-method` for more information."



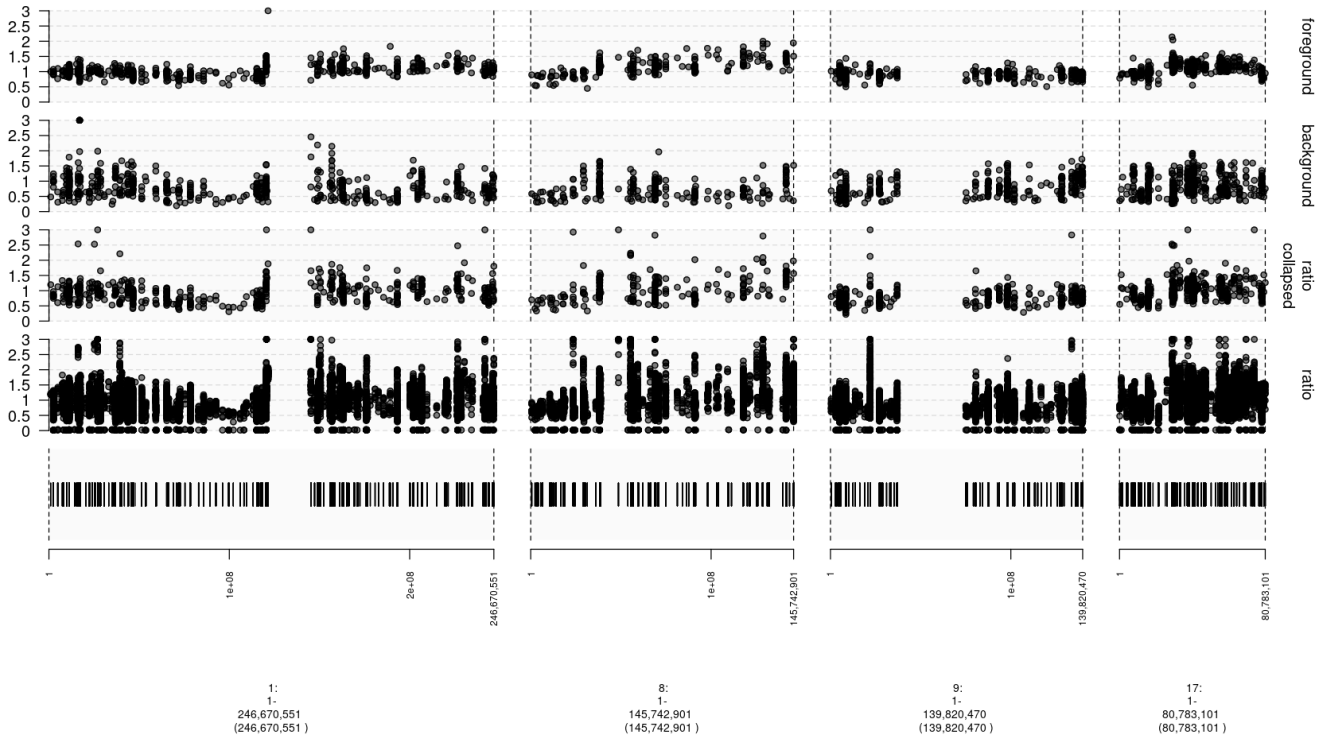
Low-purity sample

In [56]:

```
options(repr.plot.width=15, repr.plot.height=13)

idx <- 72
cov_before = metadata[idx, cov] %>% readRDS()
cov_after = metadata_updated[idx, cov] %>% readRDS()
dcln = metadata_updated[idx, dryclean] %>% readRDS()

gt.ratio.before = gTrack(cov_before, 'ratio', circle=TRUE, lwd.border=0.8, name='ratio',
y0 = 0, y1 = 3)
gt.ratio.after = gTrack(cov_after, 'ratio', circle=TRUE, lwd.border=0.8, name='ratio\nco
llapsed', y0 = 0, y1 = 3)
gt.fg = gTrack(dcln, 'foreground', circle=TRUE, lwd.border=0.8, name='foreground', y0 =
0, y1 = 3)
gt.bg = gTrack(dcln, 'background', circle=TRUE, lwd.border=0.8, name='background', y0 =
0, y1 = 3)

plot(c(gtr, gt.ratio.before, gt.ratio.after, gt.bg, gt.fg), c(1, 8, 9, 17))
```

Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(query, subject, ...):
"findOverlaps applied to ranges with non-identical seqlengths"
Warning message in gr.findoverlaps(gr, windows):
"findOverlaps applied to ranges with non-identical seqlengths"

```
findOverlaps applied to ranges with non-identical seqlengths”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 4 out-of-bound ranges located on sequences 1,
  8, 9, and 17. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 4 out-of-bound ranges located on sequences 1,
  8, 9, and 17. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
Warning message in gr.findoverlaps(query, subject, ...):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in gr.findoverlaps(gr, windows):
“findOverlaps applied to ranges with non-identical seqlengths”
Warning message in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE):
“GRanges object contains 4 out-of-bound ranges located on sequences 1,
  8, 9, and 17. Note that ranges located on a sequence whose length is
  unknown (NA) or on a circular sequence are not considered out-of-bound
  (use seqlengths() and isCircular() to get the lengths and circularity
  flags of the underlying sequences). You can use trim() to trim these
  ranges. See ?`trim,GenomicRanges-method` for more information.”
```

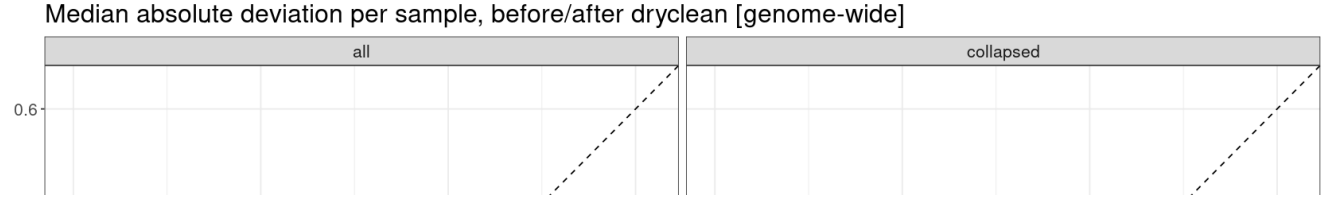


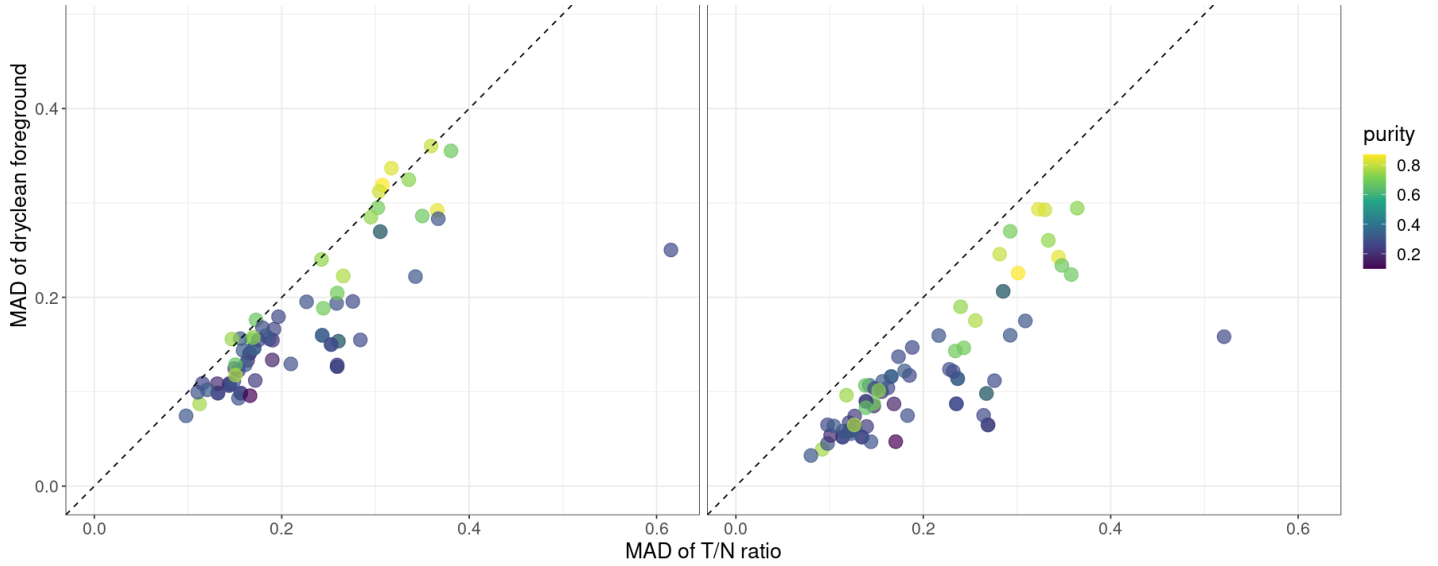
In [57]:

```
MAD_stats_collapsed = NULL

for (idx in seq(1,dim(metadata_updated)[1])){
    pre_sample <- readRDS(metadata_updated[idx,]$cov) %>% as_tibble() %>%
        select(seqnames, start, end, ratio)
    post_sample <- readRDS(metadata_updated[idx,]$dryclean) %>% as_tibble() %>%
        select(seqnames, start, end, foreground)

    data <- pre_sample %>% full_join(post_sample, by = c("seqnames", "start", "end")) %>
%
        gather(signal, value, -start, -end, -seqnames)
    total_stats <- data %>% group_by(signal) %>% summarise(MAD = mad(value)) %>% mutate(
seqnames='total')
    sample_stats <- data %>%
        group_by(signal, seqnames) %>%
        summarise(MAD = mad(value), .groups = 'drop') %>%
        full_join(total_stats, by=c('signal', 'seqnames', 'MAD')) %>%
        mutate(sample=metadata[idx,]$pair)

    if (is.null(MAD_stats_collapsed)){
        MAD_stats_collapsed <- sample_stats
    } else {
        MAD_stats_collapsed <- MAD_stats_collapsed %>% full_join(sample_stats, by=c('sig
nal', 'seqnames', 'MAD', 'sample'))
    }
}

MAD_all <- MAD_stats_collapsed %>%
    spread(signal, MAD) %>%
    mutate(foreground_collapsed = foreground, ratio_collapsed=ratio) %>%
    select(-foreground, -ratio) %>%
    full_join(MAD_stats %>% spread(signal, MAD), by=c('seqnames', 'sample'))
```

In [60]:

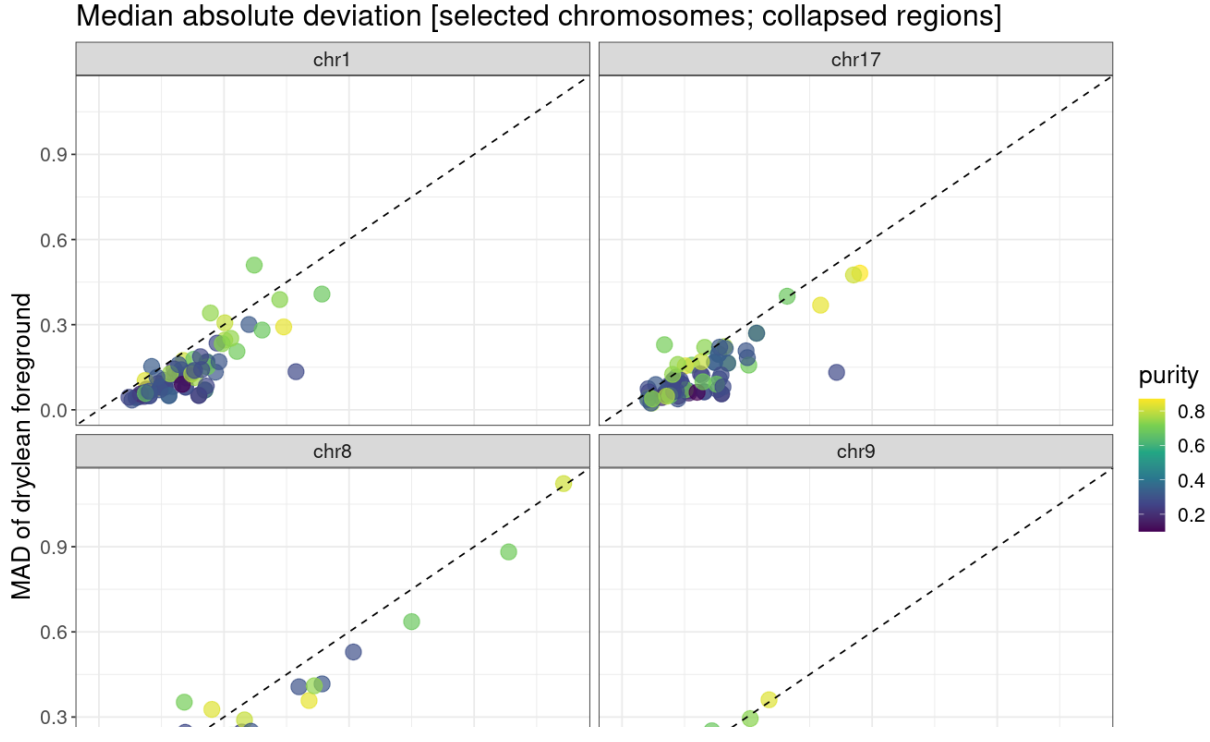
```
options(repr.plot.width=15, repr.plot.height=8)

mad_plot_data <- metadata_updated %>%
    select(pair, purity, ploidy) %>%
    right_join(MAD_all, by=c('pair'='sample')) %>%
    filter(seqnames=="total") %>%
    gather(signal, MAD, -pair, -purity, -ploidy, -seqnames) %>%
    mutate(type=ifelse(grepl('collapsed', signal), 'collapsed', 'all')) %>%
    separate(signal, c('signal'), sep="_", extra="drop") %>%
    spread(signal, MAD)


mad_plot_data %>%
    ggplot() +
        geom_point(aes(x=ratio, y=foreground, color=purity, size=1), alpha=0.7) +
        geom_abline(intercept=0, slope=1, linetype='dashed') +
        #ggrepel::geom_text_repel(aes(x=ratio, y=foreground, label=purity)) +
        xlim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        ylim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        theme_bw() + theme(legend.position="right", text = element_text(size=16)) +
        guides(size="none") +
        labs(x="MAD of T/N ratio", y="MAD of dryclean foreground",
             title="Median absolute deviation per sample, before/after dryclean [genome-w
ide]") +
        scale_color_viridis_c() +
        facet_wrap(type~.)
```

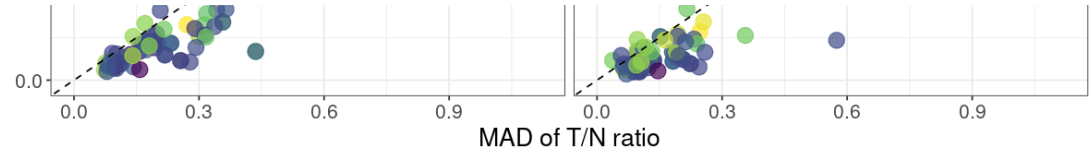




In [61]:

```
options(repr.plot.width=11, repr.plot.height=8)

mad_plot_data <- metadata_updated %>%
    select(pair, purity, ploidy) %>%
    right_join(MAD_all, by=c('pair'='sample')) %>%
    gather(signal, MAD, -pair, -purity, -ploidy, -seqnames) %>%
    mutate(type=ifelse(grepl('collapsed', signal), 'collapsed', 'all')) %>%
    separate(signal, c('signal'), sep="_", extra="drop") %>%
    spread(signal, MAD) %>%
    filter(seqnames %in% c('1', '17', '8', '9')) %>%
    filter(type == "collapsed")

mad_plot_data %>%
    ggplot() +
        geom_point(aes(x=ratio, y=foreground, color=purity, size=1), alpha=0.7) +
        geom_abline(intercept=0, slope=1, linetype='dashed') +
        #ggrepel::geom_text_repel(aes(x=ratio, y=foreground, label=purity)) +
        xlim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        ylim(0, max(c(mad_plot_data$foreground, mad_plot_data$ratio))) +
        theme_bw() + theme(legend.position="right", text = element_text(size=16)) +
        guides(size="none") +
        labs(x="MAD of T/N ratio", y="MAD of dryclean foreground",
             title="Median absolute deviation [selected chromosomes; collapsed regions]"
            ) +
        scale_color_viridis_c() +
        facet_wrap(paste0("chr", seqnames)~., nrow=2)
```





(7) Comparison of copy-number states (adjusting for purity and ploidy)

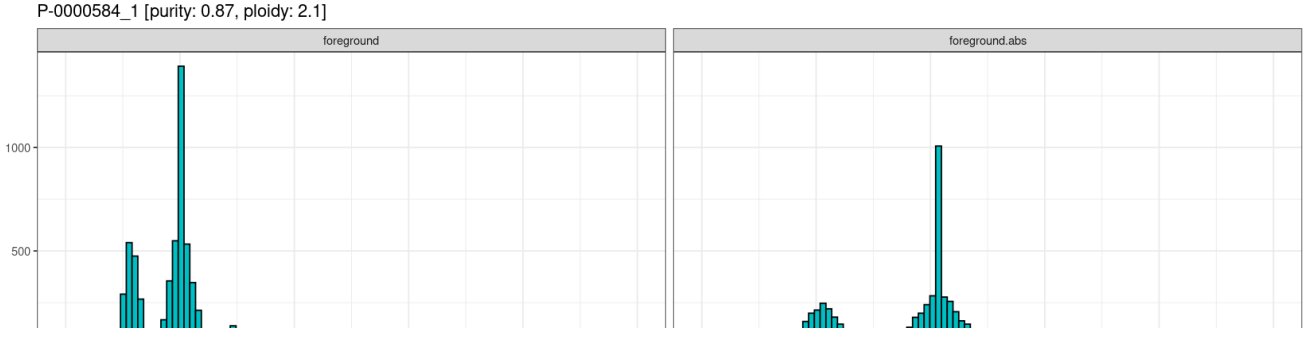
In [62]:

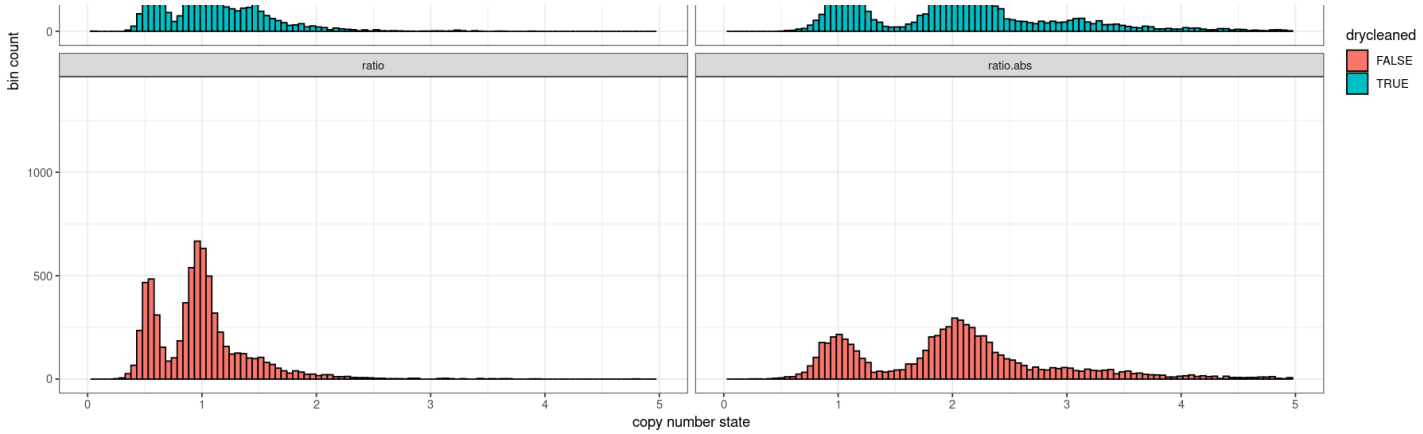
```
get_abs_collapsed <- function(pair_name){
    idx <- which(metadata_updated$pair == pair_name)
    stopifnot(length(idx) == 1)

    pre_sample <- readRDS(metadata_updated[idx,]$cov)
    pre_sample <- pre_sample  %>%
        as_tibble() %>%
        mutate(ratio.abs=rel2abs(gr=pre_sample, purity=metadata_updated[idx,]$purity,
                                 ploidy=metadata_updated[idx,]$ploidy, field='ratio'))
    post_sample <- readRDS(metadata_updated[idx,]$dryclean)
    post_sample <- post_sample  %>%
        as_tibble() %>%
        mutate(foreground.abs=rel2abs(gr=post_sample, purity=metadata_updated[idx,]$puri
ty,
                                 ploidy=metadata_updated[idx,]$ploidy, field='foreground
'))

    data <- pre_sample %>%
        full_join(post_sample, by = c("seqnames", "start", "end", "width", "strand")) %>
%
        select("seqnames", "start", "end", matches("ratio"), matches("ground"))
    return(data)

}
```

In [64]:

```
options(repr.plot.width=15, repr.plot.height=8)

idx <- 1
sample <- get_abs_collapsed(pair_name=metadata_updated[idx,]$pair)
sample %>%
    select(seqnames, start, ratio, ratio.abs, foreground, foreground.abs) %>%
    gather(signal, val, -seqnames, -start) %>%
    mutate(drycleaned=str_detect(signal, 'foreground')) %>%
    ggplot() +
        geom_histogram(aes(x=val, fill=drycleaned), bins=100, color="black") +
        facet_wrap(signal ~.) +
        theme_bw() +
        xlim(0,5) +
        labs(title=paste0(metadata_updated[idx,]$pair," [purity: ",
                          metadata_updated[idx,]$purity,", ploidy: ", metadata_updated[i
dx,]$ploidy,"]"),
             x="copy number state", y="bin count")
```

```
Warning message:
“Removed 259 rows containing non-finite values (stat_bin).”
Warning message:
“Removed 8 rows containing missing values (geom_bar).”
```

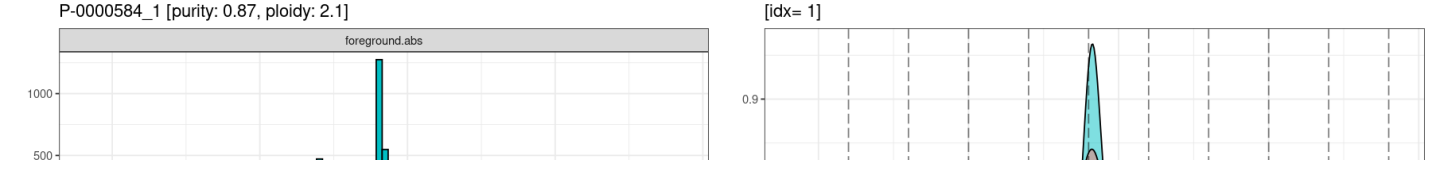


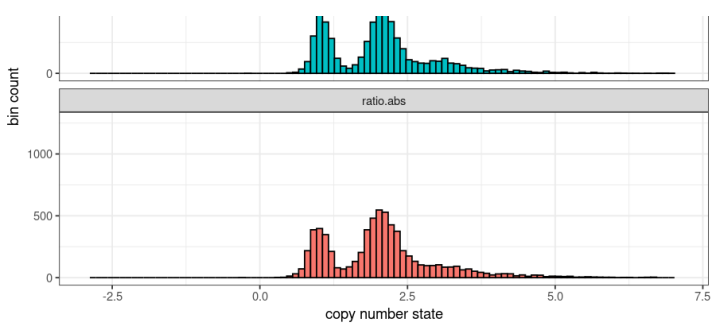


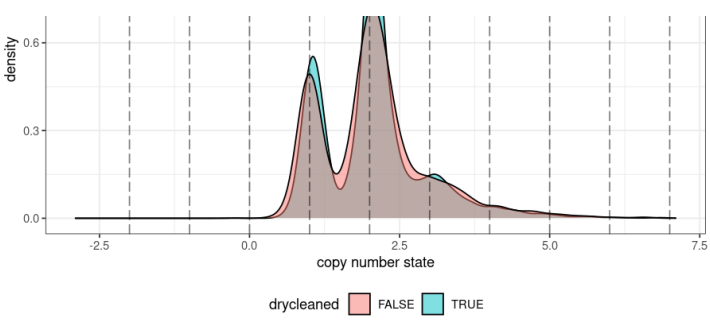
In [65]:

```
options(repr.plot.width=15, repr.plot.height=5)
for (idx in 1:10){
    sample <- get_abs_collapsed(pair_name=metadata_updated[idx,]$pair)

    p1 <- sample %>%
        select(seqnames, start, ratio.abs, foreground.abs) %>%
        gather(signal, val, -seqnames, -start) %>%
        mutate(drycleaned=str_detect(signal, 'foreground')) %>%
        ggplot() +
            geom_histogram(aes(x=val, fill=drycleaned), bins=100, color="black") +
            facet_wrap(signal ~., ncol=1) +
            theme_bw() +
            xlim(metadata_updated[idx,]$ploidy-5, metadata_updated[idx,]$ploidy+5) +
            labs(title=paste0(metadata_updated[idx,]$pair," [purity: ",
                              metadata_updated[idx,]$purity,", ploidy: ", metadata_updat
ed[idx,]$ploidy,"]"),
                 x="copy number state", y="bin count") +
            theme(legend.position="null")
    p2 <- sample %>%
    select(seqnames, start, ratio.abs, foreground.abs) %>%
    gather(signal, val, -seqnames, -start) %>%
    mutate(drycleaned=str_detect(signal, 'foreground')) %>%
    ggplot() +
        geom_density(aes(x=val, fill=drycleaned, group=signal), color="black", alpha=0.5
) +
        geom_vline(xintercept = seq(as.integer(metadata_updated[idx,]$ploidy-5),
                                    as.integer(metadata_updated[idx,]$ploidy+5)),
                   alpha=0.5, linetype = "longdash") +
        #facet_wrap(signal ~.) +
        theme_bw() +
        xlim(metadata_updated[idx,]$ploidy-5, metadata_updated[idx,]$ploidy+5) +
        labs(title=paste0("[idx= ",idx,"]"),
             x="copy number state") +
            theme(legend.position="bottom")
    grid.arrange(p1, p2, ncol=2)
}
```

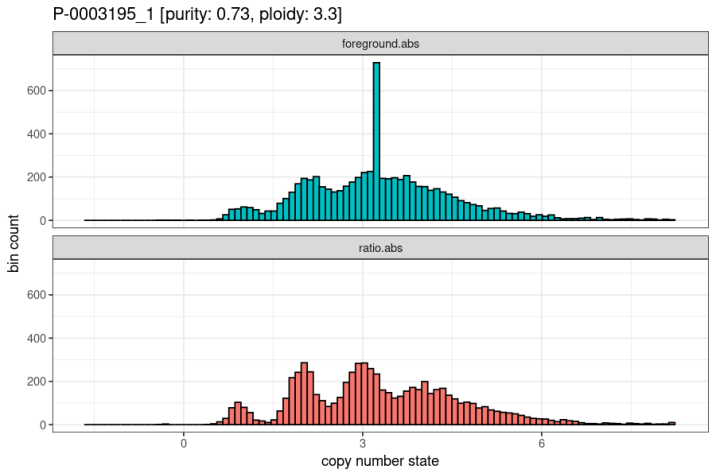
```
Warning message:
“Removed 86 rows containing non-finite values (stat_bin).”
Warning message:
“Removed 4 rows containing missing values (geom_bar).”
Warning message:
“Removed 86 rows containing non-finite values (stat_density).”
Warning message:
“Removed 110 rows containing non-finite values (stat_bin).”
Warning message:
“Removed 4 rows containing missing values (geom_bar).”
Warning message:
“Removed 110 rows containing non-finite values (stat_density).”
```

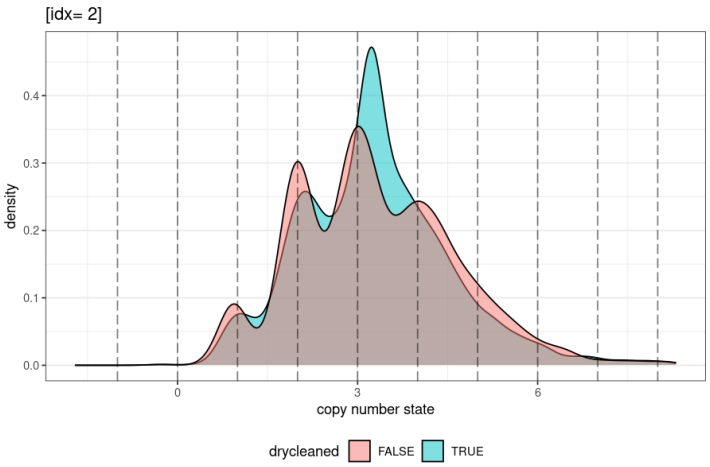




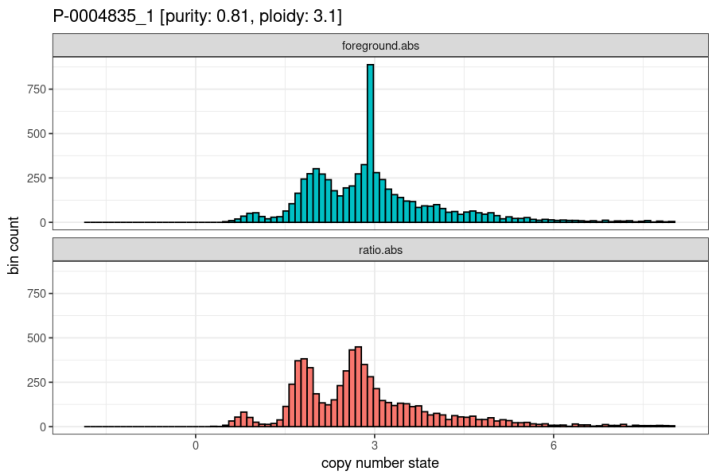


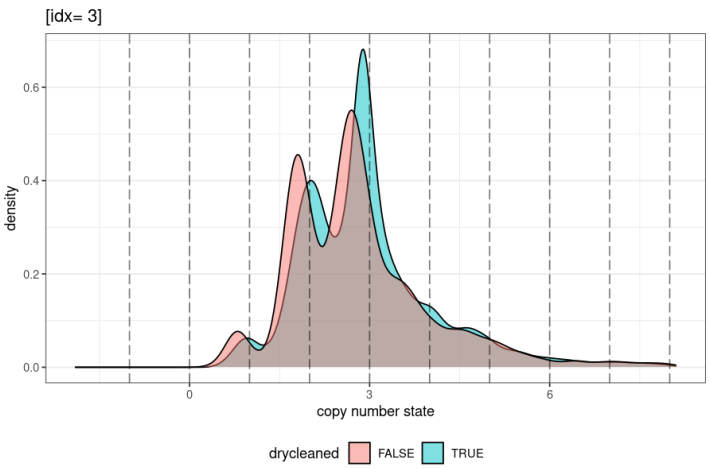

```
Warning message:
“Removed 334 rows containing non-finite values (stat_bin).”
Warning message:
“Removed 4 rows containing missing values (geom_bar).”
Warning message:
“Removed 334 rows containing non-finite values (stat_density).”
```



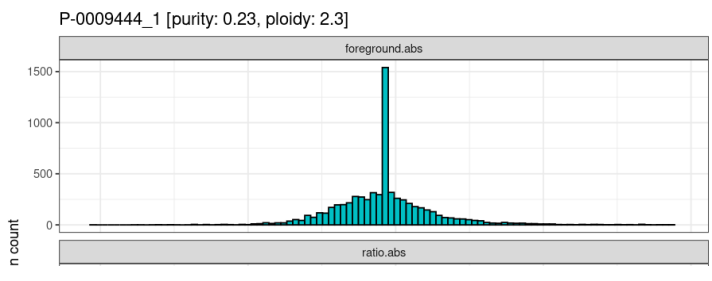


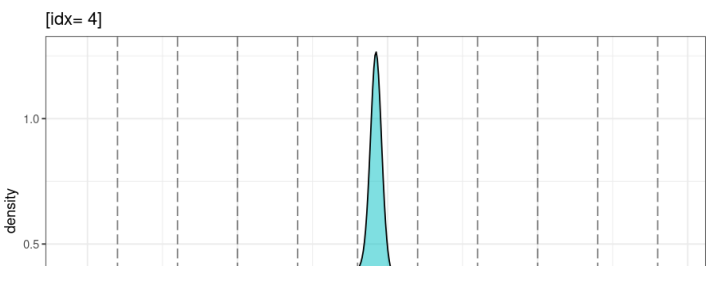

```
Warning message:
“Removed 57 rows containing non-finite values (stat_bin).”
Warning message:
“Removed 4 rows containing missing values (geom_bar).”
Warning message:
“Removed 57 rows containing non-finite values (stat_density).”
```

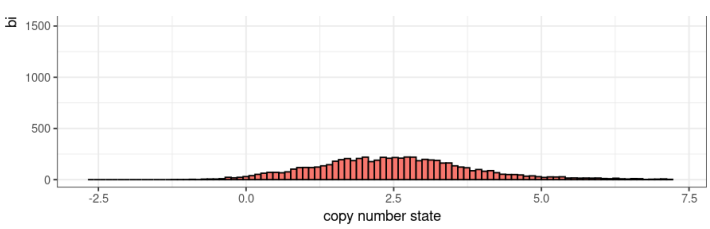


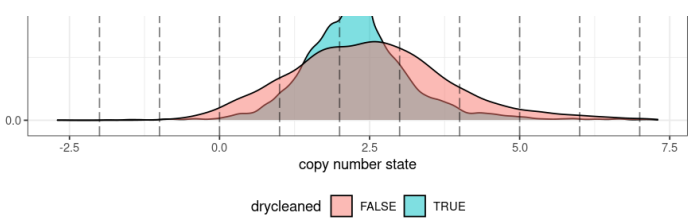



```
Warning message:
“Removed 568 rows containing non-finite values (stat_bin).”
Warning message:
“Removed 4 rows containing missing values (geom_bar).”
Warning message:
“Removed 568 rows containing non-finite values (stat_density).”
```

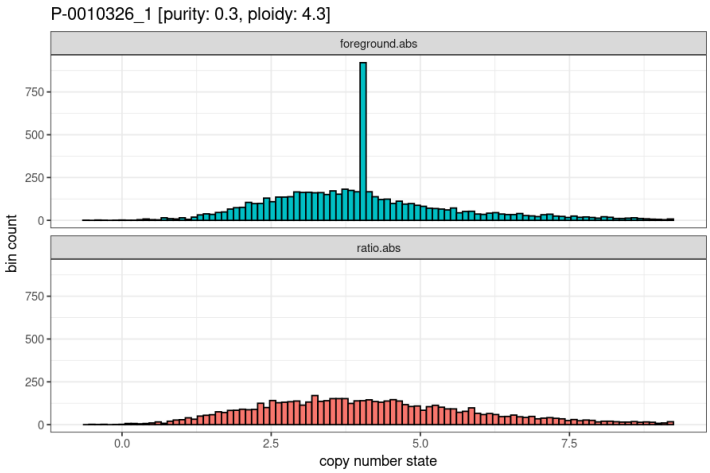


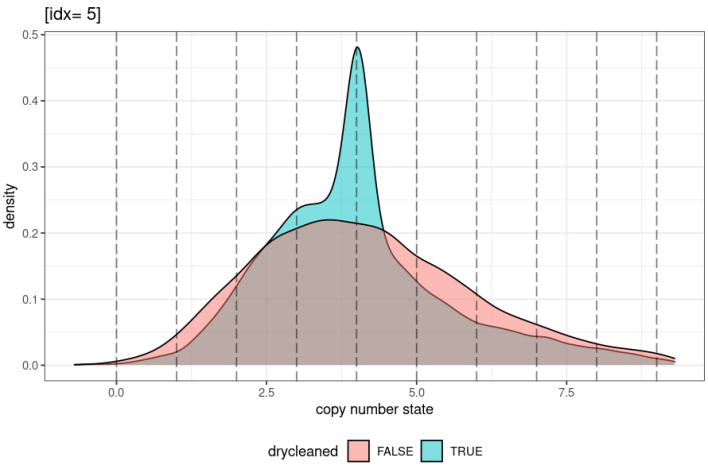




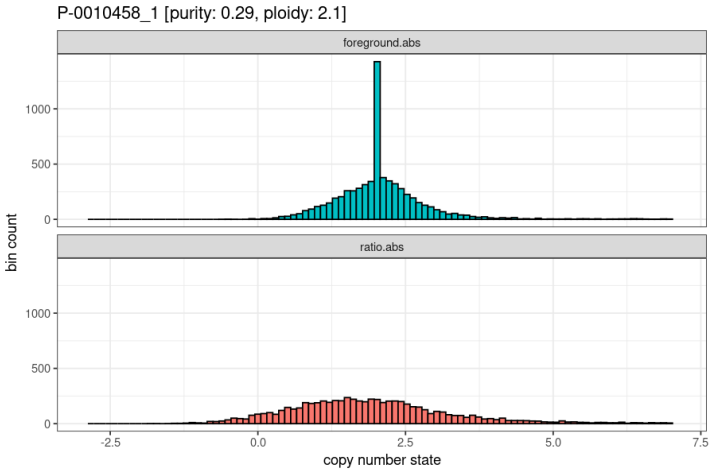


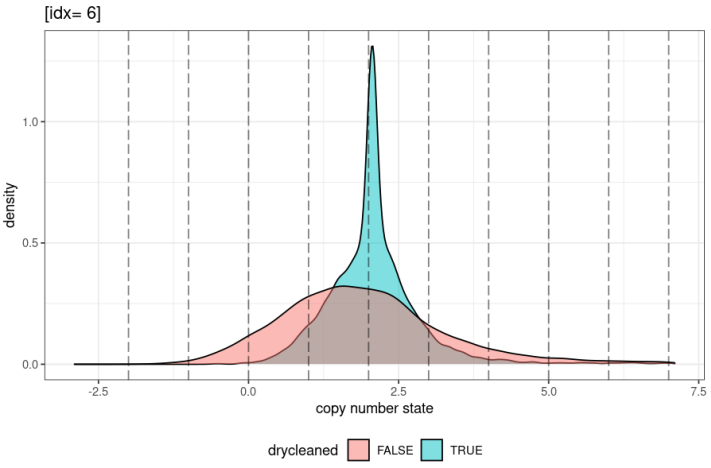

```
Warning message:
"Removed 219 rows containing non-finite values (stat_bin)."
Warning message:
"Removed 4 rows containing missing values (geom_bar)."
Warning message:
"Removed 219 rows containing non-finite values (stat_density)."
```



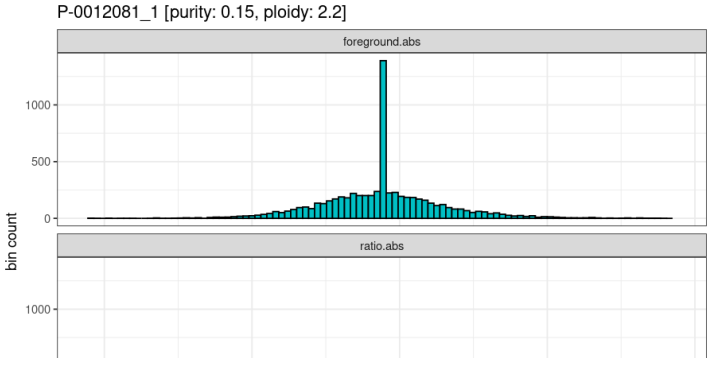


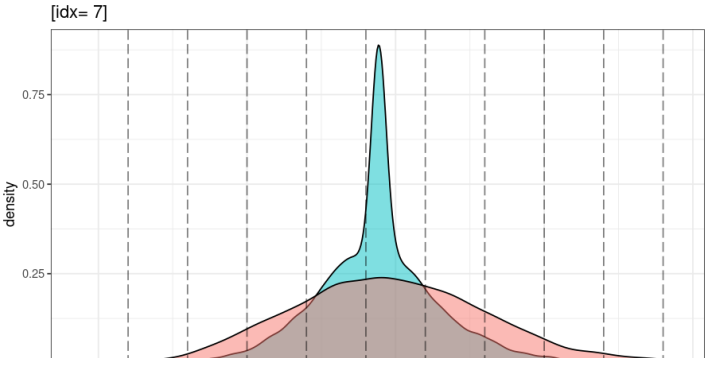

```
Warning message:
"Removed 100 rows containing non-finite values (stat_bin)."
Warning message:
"Removed 4 rows containing missing values (geom_bar)."
Warning message:
"Removed 100 rows containing non-finite values (stat_density)."
```

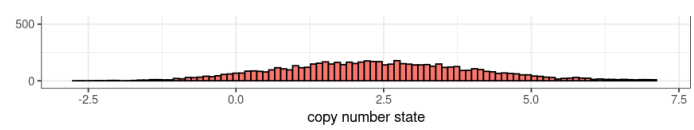




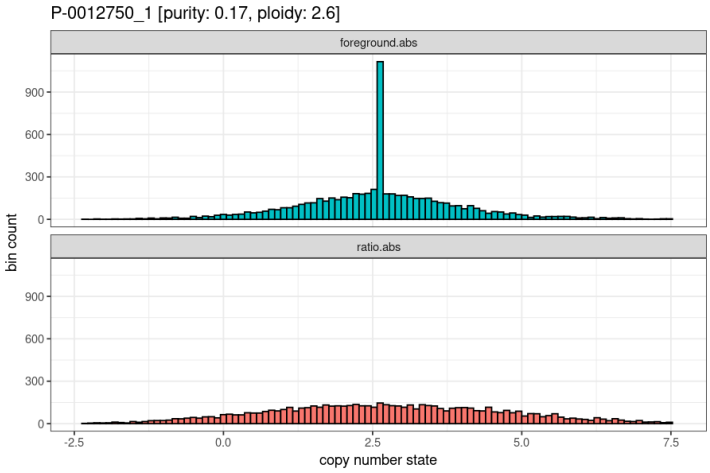

```
Warning message:
"Removed 200 rows containing non-finite values (stat_bin)."
Warning message:
"Removed 4 rows containing missing values (geom_bar)."
Warning message:
"Removed 200 rows containing non-finite values (stat_density)."
```

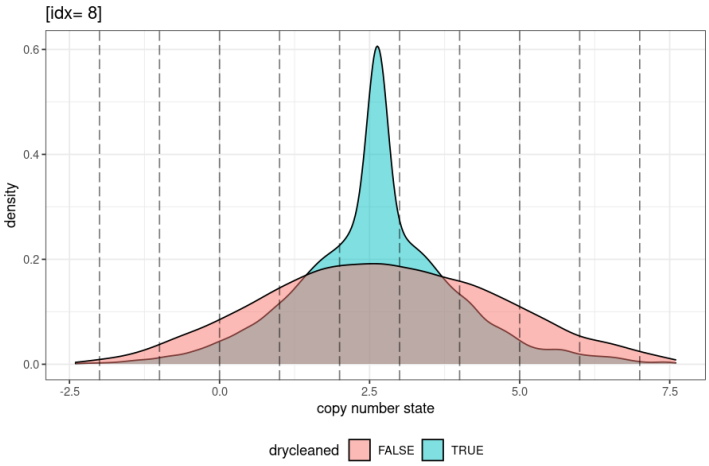




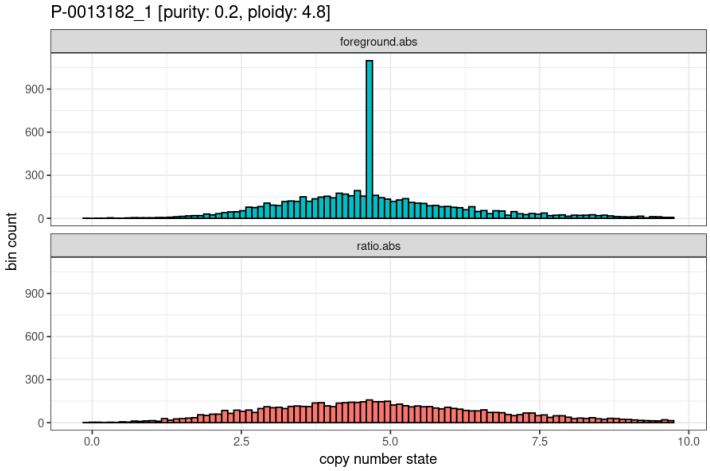


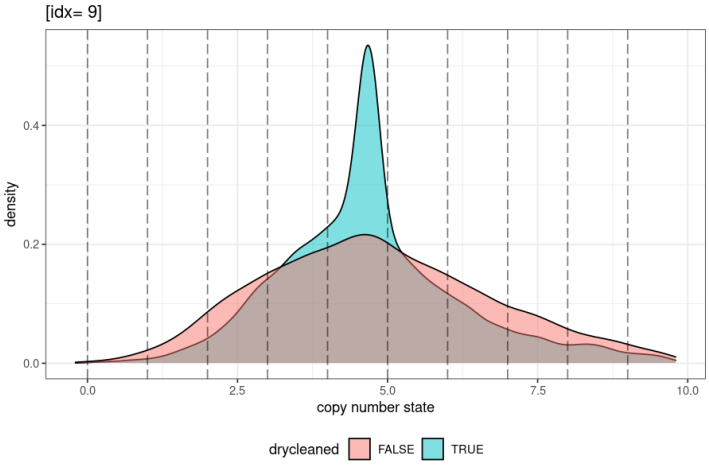

```
Warning message:
"Removed 403 rows containing non-finite values (stat_bin)."
Warning message:
"Removed 4 rows containing missing values (geom_bar)."
Warning message:
"Removed 403 rows containing non-finite values (stat_density)."
```

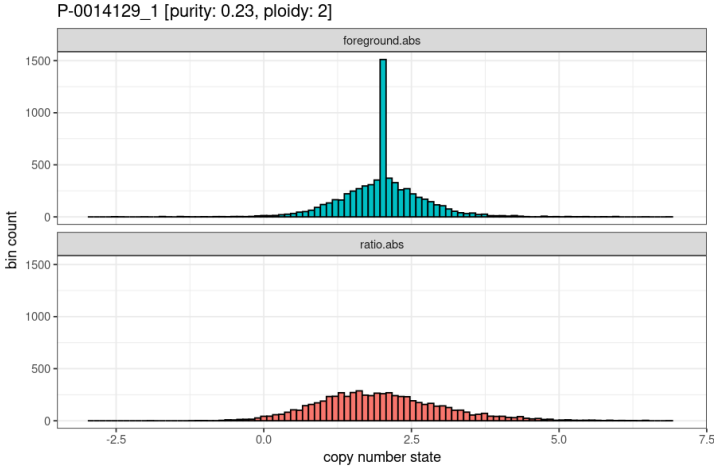


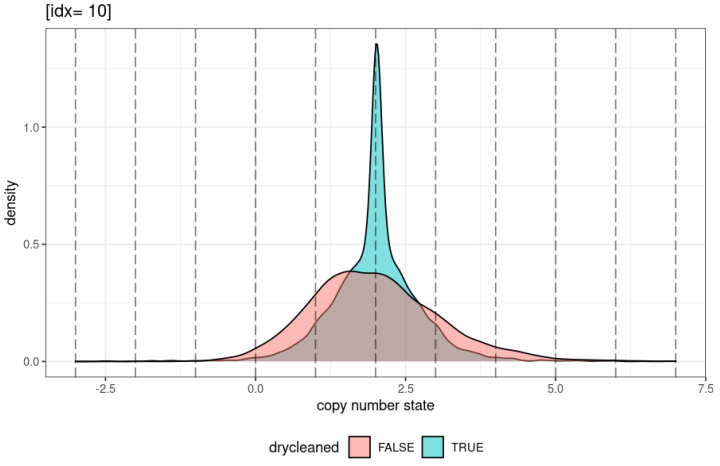



```
Warning message:
"Removed 72 rows containing non-finite values (stat_bin)."
Warning message:
"Removed 4 rows containing missing values (geom_bar)."
Warning message:
"Removed 72 rows containing non-finite values (stat_density)."
```









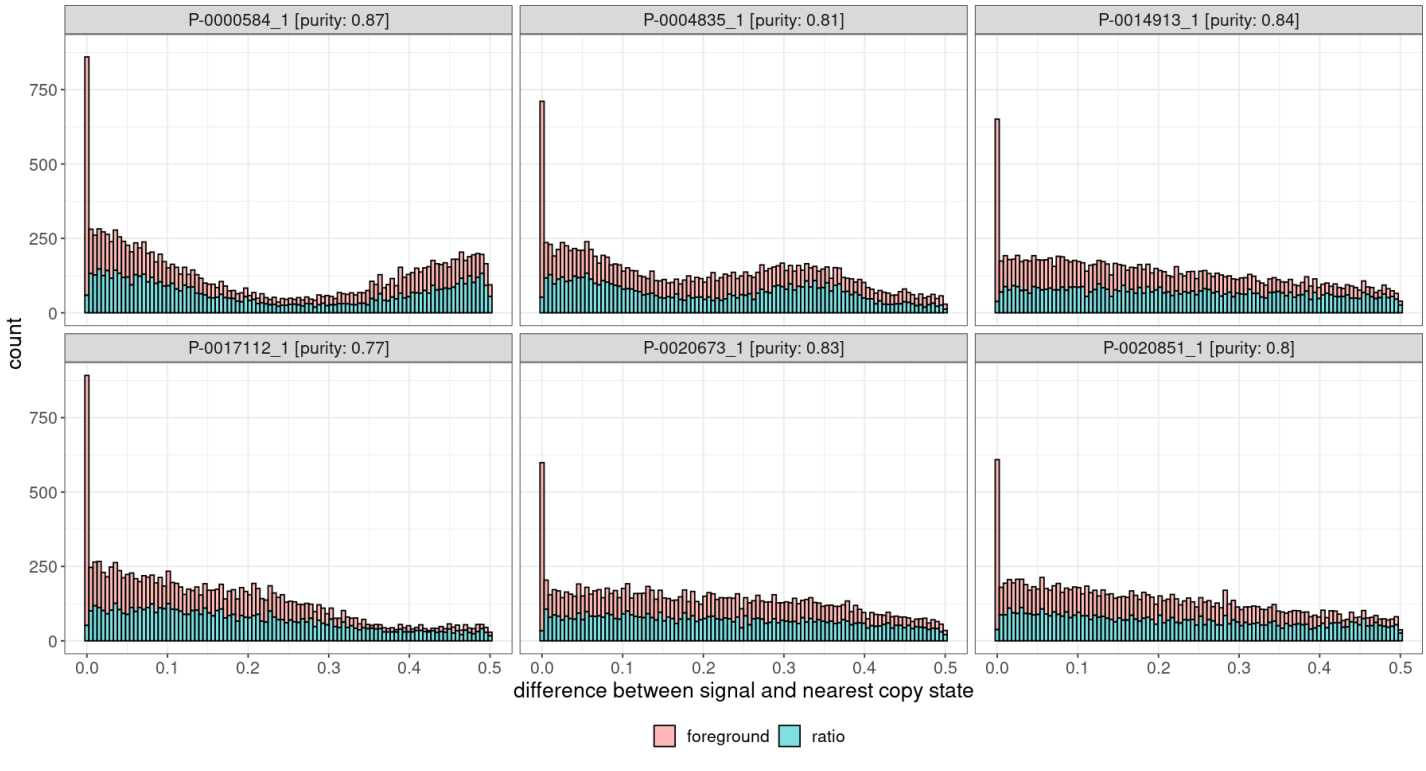
In [66]:

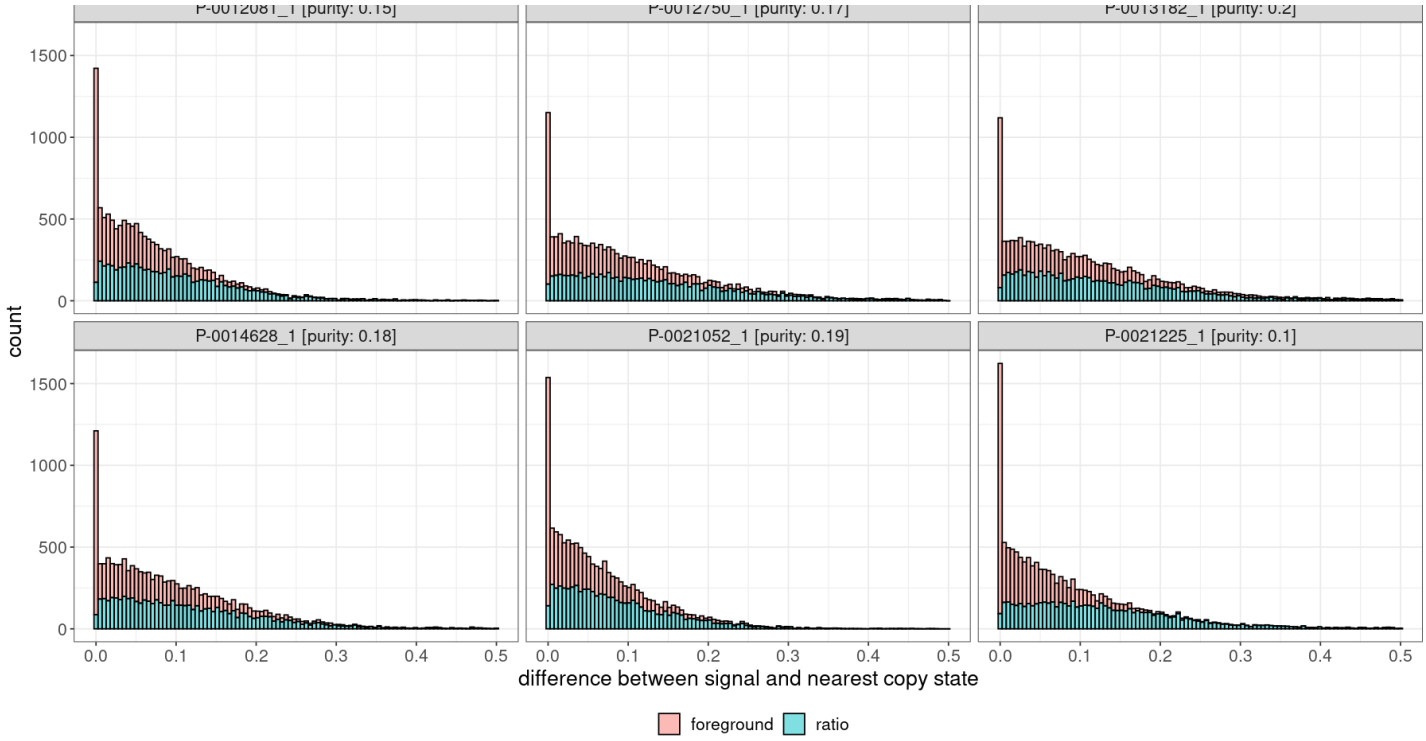
```
sample_stats = NULL
for (idx in 1:dim(metadata_updated)[1]){
    sample <- get_abs_collapsed(pair_name=metadata_updated[idx,]$pair) %>%
```

```
        select(seqnames, start, ratio.abs, foreground.abs, ratio, foreground) %>%
        gather(signal, val, -seqnames, -start) %>%
        mutate(drycleaned=str_detect(signal, 'foreground')) %>%
        mutate(diff_val=abs(val-round(val)), sample=metadata_updated[idx,]$pair)
    if(is.null(sample_stats)){
        sample_stats <- sample
    } else {
        sample_stats <- sample_stats %>% full_join(sample, by = c("seqnames", "start", "
signal", "val",
                                                                  "drycleaned", "diff_v
al", "sample"))
    }
}
```

In [68]:

```
options(repr.plot.width=15, repr.plot.height=8)
sample_purity <- metadata_updated %>% pull(purity, pair)

tibble(sample=names(sort(sample_purity, decreasing = TRUE)[1:6]), purity=sort(sample_pur
ity, decreasing = TRUE)[1:6]) %>%
    left_join(sample_stats %>% filter(!grepl('.abs', signal)), by = "sample") %>%
    ggplot() +
        geom_histogram(aes(x=diff_val,fill=signal, group=signal),
                       color="black", alpha=0.5,  bins=100) +
        #geom_line(aes(x=diff_val, group=signal), stat="density", size=1) +
        #facet_wrap(drycleaned~., nrow=1) +
        theme_bw() +
        theme(legend.position="bottom", text = element_text(size=16)) +
        labs(x="difference between signal and nearest copy state", fill="",
            title="High-purity samples [collapsed]") +
        facet_wrap(paste0(sample, " [purity: ",purity, "]")~.)

tibble(sample=names(sort(sample_purity)[1:6]), purity=sort(sample_purity)[1:6]) %>%
    left_join(sample_stats %>% filter(!grepl('.abs', signal)), by = "sample") %>%
    ggplot() +
        geom_histogram(aes(x=diff_val,fill=signal, group=signal),
                       color="black", alpha=0.5,  bins=100) +
        #geom_line(aes(x=diff_val, group=signal), stat="density", size=1) +
        #facet_wrap(drycleaned~., nrow=1) +
        theme_bw() +
        theme(legend.position="bottom", text = element_text(size=16)) +
        labs(x="difference between signal and nearest copy state", fill="",
            title="Low-purity samples [collapsed]") +
        facet_wrap(paste0(sample, " [purity: ",purity, "]")~.)
```





In [90]:

```
sample_stats_before = NULL
for (idx in 1:dim(metadata)[1]){
    drcln_sample <- readRDS(metadata[idx,]$dryclean) %>% as_tibble() %>% select(seqnames,
start, end, foreground)
    ratio_sample <- readRDS(metadata[idx,]$cov) %>% as_tibble() %>% select(seqnames, star
t, end, ratio)

    sample <- drcln_sample %>%
        full_join(ratio_sample, by = c("seqnames", "start", "end")) %>%
        gather(signal, val, -seqnames, -start, -end) %>%
        mutate(drycleaned=str_detect(signal, 'foreground')) %>%
        mutate(diff_val=abs(val-round(val)), sample=metadata[idx,]$pair)

    if(is.null(sample_stats_before)){
        sample_stats_before <- sample
    } else {
        sample_stats_before <- sample_stats_before %>% full_join(sample, by = c("seqname
s", "start", "end", "signal",
                                                                                 "val",
"drycleaned", "diff_val", "sample"))
    }
}
```

In [91]:

```
options(repr.plot.width=15, repr.plot.height=8)
sample_purity <- metadata %>% pull(purity, pair)

tibble(sample=names(sort(sample_purity, decreasing = TRUE)[1:6]), purity=sort(sample_pur
ity, decreasing = TRUE)[1:6]) %>%
    left_join(sample_stats_before %>% filter(!grepl('.abs', signal)), by = "sample") %>%
    ggplot() +
        geom_histogram(aes(x=diff_val,fill=signal, group=signal),
                       color="black", alpha=0.5,  bins=100) +
        #geom_line(aes(x=diff_val, group=signal), stat="density", size=1) +
        #facet_wrap(drycleaned~., nrow=1) +
        theme_bw() +
        theme(legend.position="bottom", text = element_text(size=16)) +
        labs(x="difference between signal and nearest copy state", fill="",
            title="High-purity samples [before]"
            ) +
        facet_wrap(paste0(sample, " [purity: ",purity, "]")~.)

tibble(sample=names(sort(sample_purity)[1:6]), purity=sort(sample_purity)[1:6]) %>%
```

```
    left_join(sample_stats_before %>% filter(!grepl('.abs', signal)), by = "sample") %>%
    ggplot() +
        geom_histogram(aes(x=diff_val,fill=signal, group=signal),
                       color="black", alpha=0.5,  bins=100) +
        #geom_line(aes(x=diff_val, group=signal), stat="density", size=1) +
        #facet_wrap(drycleaned~., nrow=1) +
        theme_bw() +
        theme(legend.position="bottom", text = element_text(size=16)) +
        labs(x="difference between signal and nearest copy state", fill="",
            title="Low-purity samples [before]"
            ) +
        facet_wrap(paste0(sample, " [purity: ",purity, "]")~.)
```