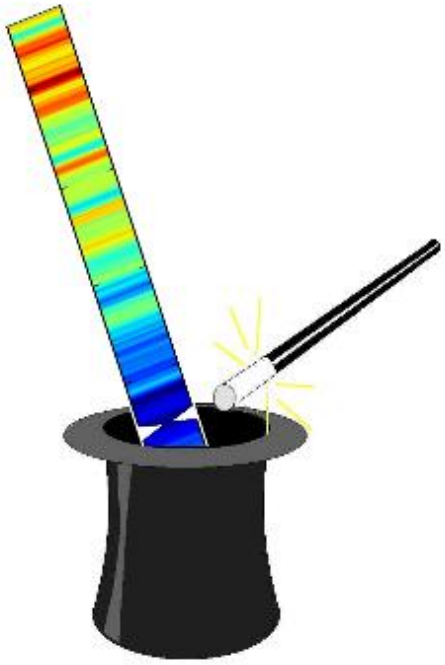
chromoWIZ

Visualisation of chromosomal structure



Authors:
Heidrun Gundlach
Mihaela Martis
Thomas Nussbaumer

July 2010

Contents

1 Introduction to *chromoWIZ*

1.1 What is *chromoWIZ*?

chromoWIZ is a software tool to visualize and compare the chromosomal architecture of sequenced genomes. It calculates the content of specified genomic elements and intrinsic sequence features in a sliding window approach along chromosomes.
chromoWIZ requires a multiple sequence *FASTA* file for the sequence input and a *GFF3* file as annotation input. A graphical interface guides the user through the creation of customizable heatmaps (Fig. 2), linecharts (Fig. 4) and barcharts (Fig. 5), leading to the output of high resolution picture files.

1.2 System Requirements

chromoWIZ is implemented in *Python* (version 2.5) and uses a *TkInter GUI*. It is tested under *Linux*, but it should also work under *Windows*. All *Python* modules are included in the *Python* standard installation. Only the *Python* Image Library (*PIL*) is needed additionally for the export of high resolution graphic files. If *PIL* is installed, the export of heatmaps, linecharts and barcharts in a higher quality (*PDF*, *JPEG*, *PNG*) is possible.

2 Installation and Usage

2.1 Installation

1. unzip *chromoWIZ* package:
   $ unzip chromoWIZ.zip -d /usr/local/data

2. navigate to the *src* package:
   $ cd /usr/local/data/chromoWIZ/src

2.2 Usage

1. run the data extraction using:
   $ python ./extract_data.py chromoWIZ.conf

2. start the *TkInter GUI* using:
   $ python ./visualize_data.py

   - choose a database (*e.g. Bd.db*) via "*File⇒Open*"
   - choose a density table (*e.g. density_Bd*)
   - choose a calculation identifier (*e.g. 500000_win_100000__shift_Gene_CDS*)

The following steps are necessary to visualize the data:

- create a configuration file (see 4).
- run the data extraction.
- start the *GUI* to adjust and export heatmaps, barcharts or linecharts.

3 Workflow

4 Configuration file

Apart from an annotation file in *GFF3* format, a multiple *FASTA* file is required. These two files plus additional parameter have to be specified in a single configuration file. It consists of three sections:

1. General parameters
2. Parameters for data extraction
3. Parameters for data postprocessing

The first section is mandatory. The second and third sections can be activated by enabling the parameter "*extract_data*" and "*calc_densities*".
The second section stores the annotated elements into the database, the third section executes a postprocessing step.

4.1 General parameter

- *genome_id (e.g. Bd)*
  The "*genome_id*" represents the name of the database and is equal to the postfix of the density table.

- *workspace (e.g. /usr/local/data/chromoWIZ/Bd)*
  Absolute path to the working directory including all directories and files created by *chromoWIZ* . If it does not exist yet, it will be created automatically. If a database already exists in the specified "*workspace*", the current calculation will be stored there.

- *seq_file (e.g. /usr/local/data/chromoWIZ/seq/brachy1.0_wholegenome_unmasked.mfa)*
  Absolute path to the genome sequence multiple *FASTA* file. The "*seq_file*" must contain all sequences for a certain genome.

4.2 Parameters for data extraction

- *gff3_file (e.g. /usr/local/data/chromoWIZ/gff3/Bd_gene_1.2_CDS.gff3)*
  Absolute path to the *GFF3* annotation file.

- *gff3_type (e.g. CDS,gene)*
  Parent-child relationship of parsed genetic elements.

- *anno_id (e.g. CDS_gene)*
  Unique identifier within one genome for the *GFF3* annotation.

The "*gff3_types*" have to seperated by a comma and must represent a *1:n* relationship (*e.g.* a gene can be composed of several Exons).

4.3 Parameters for data postprocessing

- *win_size (e.g. 500000)*
  Size of window in *bp*.

- *shift (e.g. 100000)*
  Shift of sliding window in *bp*.

- *min_chromosome_length (20000000)*
  Minimal sequence length in *bp*, all sequences in the "*seq_file*" shorter than the specified "*min_chromosome_length*" will not be used.

4.4 Config file creator

The *TkInter GUI* allows the creation of new configuration files and the modification of existing files. It is recommended to use this tool because all input parameter will be automatically validated. Existing configuration files can be included via "*File⇒Open*". The parameter will be inserted into the *GUI* and can be changed there.

To start the config file creator please execute the following command:
$ python ./config_file_creator.py

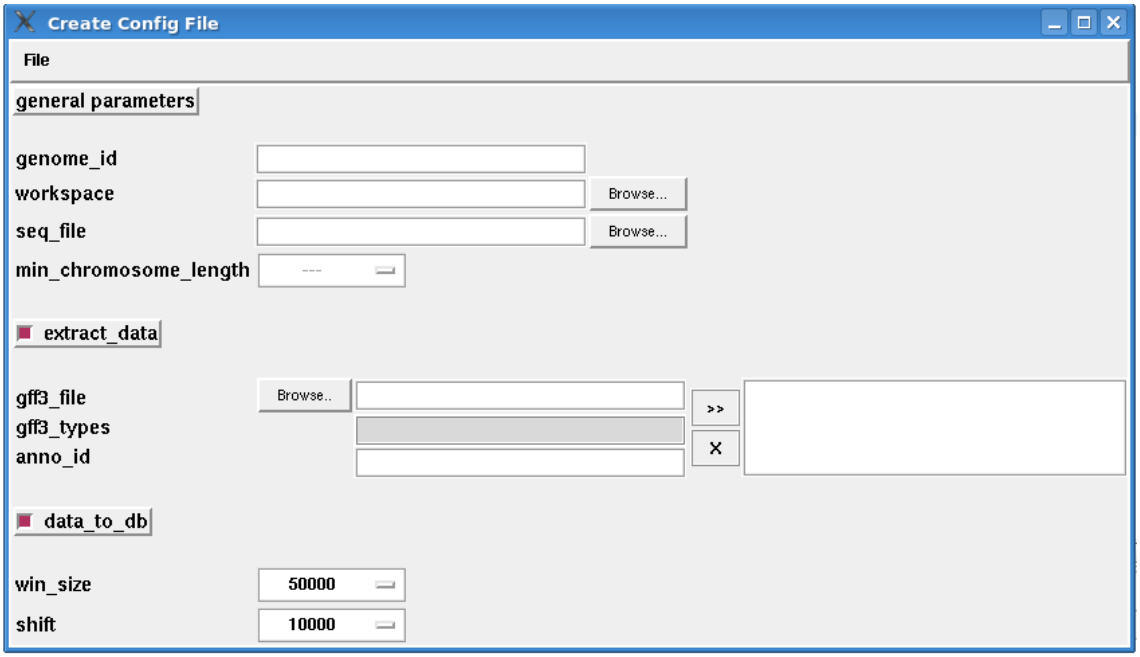


Figure 1: config file creator

4.5 Config file examples

4.5.1 single calculation in *GFF3* format

The configuration file represents an example for a single calculation. The file was generated with the Config file creator tool (see 4.4).

```
# generated with automatisation of chromoWIZ
genome_id::Bd
workspace::/usr/local/data/chromoWIZ/Bd
seq_file:: /usr/local/data/chromoWIZ/seq/brachy1.0_wholegenome_unmasked.mfa
min_chromosome_length::20000000


# data extraction
extract_data::yes
gff3_file::/usr/local/chromoWIZ/data/gff3/Bd_satellite_tandem_repeats.gff3
gff3_type::tandem_repeat,transposon_fragment
anno_id::Satellite

# density calculation
calc_densities::yes
win_size::500000
shift::100000
```

Listing 1: configuration file with one annotation elements

4.5.2 multiple calculations in *GFF3* format

In *chromoWIZ* the declaration of multiple *GFF3* type tupel is possible. In the following configuration file, four different tupel are specified (*Satellite*,*CDS_gene*,*DNA_TE*,*Retro_TE*) . "*win_size*" and "*shift*" values count for all *GFF3* type tupel. The "*min_chromosome_length*" has to identically for all calculations within one "*genome_id*".

```
# generated with automatisation of chromoWIZ
genome_id::Bd
workspace::/usr/local/data/chromoWIZ/Bd
seq_file:: /usr/local/data/chromoWIZ/seq/brachy1.0_wholegenome_unmasked.mfa
min_chromosome_length::20000000

# data extraction
extract_data::yes
gff3_file::/usr/local/chromoWIZ/data/gff3/Bd_satellite_tandem_repeats.gff3
gff3_type::tandem_repeat,transposon_fragment
anno_id::Satellite


gff3_file::/usr/local/chromoWIZ/data/gff3/Bd_genes_1.2_CDS_one_splice_var.gff3
gff3_type::CDS,gene
anno_id::CDS_gene


gff3_file::/usr/local/chromoWIZ/data/gff3/Bd_TEs_v2.2__DNA-TEs.gff3
gff3_type::transposable_element,transposon_fragment
anno_id::DNA_TE


gff3_file::/usr/local/chromoWIZ/data/gff3/Bd_TEs_v2.2__Retro-TEs.gff3
```

```
gff3_type::transposable_element,transposon_fragment
anno_id::Retro_TE


# density calculation
calc_densities::yes
win_size::500000
shift::100000
```

Listing 2: configuration file with four annotation elements

4.5.3 single calculation in tab format

In cases where annotation in *GFF3* format is not available, offers an alternative input in tab format. In these files only *chromoWIZ* relevant information is stored.

```
genome_id :: brachy
workspace :: /nfs/plant/data/repeats/heatmaps_Thomas/ChromoWIZ_Test/TAB/run_tab
seq_file :: /nfs/plant/data/repeats/ANGELA/current/Brachy/sequences/mtfa/Bd_all.tfa

extract_data :: yes

tab_file :: /nfs/plant/databases/COORDs/Brachy_1.2_coords.fa
tab_type :: exon
anno_id :: CDS_genes

# calculate densities
calc_densities :: yes
win_size :: 500000
shift :: 100000
min_chromosome_length :: 20000000
```

Listing 3: configuration file in tab format

A subset of the Brachypodium distachyon annotation file in tab format can be found in the figure below. A line starting with “>” declares a parent tag, followed by parent and sequence identifier. Start and stop position of a parent tag are seperate entries. Child elements are represented by additional lines. One child element is composed of element type start and stopposition. An element is minus stranded when the stop position is higher than the start position.

```
>Bradi1g00200.1 chr01_pseudomolecule
exon    10581   10850
start   10581
exon    11252   11638
stop    11638
>Bradi1g00210.1 chr01_pseudomolecule
exon    18479   18608
start   18479
exon    18706   18753
exon    19307   19426
exon    19730   19875
exon    20010   20157
exon    20579   20662
exon    20767   20882
exon    21145   21405
exon    21495   21718
exon    22787   23039
stop    23039
```

Listing 4: annotation in tab format

5 GUI for display and adjustments

The *GUI* offers methods for modifying and exporting heatmaps, barcharts and linecharts.

5.1 display options

- *heatmaps*
  A heatmap allows the visualisation of genomic elements in transistions between red and blue. Red areas are very element rich, blue corresponds to element poor areas (Fig. 2).

- *barcharts*
  In stacked heatmaps one or more heatmap tracks can be arranged among themselves (Fig. 3).

- *linecharts*
  In linecharts density values are represented as colored lines (Fig. 4).

- *stacked barcharts*
  In stacked barcharts density values of chosen calculation identifier are piled up (Fig. 5).

5.2 command line options

If a lot of visualisation elements should be calculated, modifying and exporting them within the *GUI* may take too much time. For that reason, *chromoWIZ* offers a possibility to create and export heatmaps, barcharts and linecharts via command line calls.

5.2.1 commands

- *heatmaps*
  $ python ./visualize_data.py -c <s> -d <s> -m <c> -o <s> -s <s> -t <s> -v heatmap -x <i>

- *stacked barcharts*
  $ python ./visualize_data.py -c <s> -d <s> -o <s> -s <s> -t <s> -v barchart

- *linecharts*
  $ python ./visualize_data.py -c <s> -d <s> -m <c> -o <s> -s <s> -t <s> -v linechart -x=<i>

- *stacked heatmaps*
  $ python ./visualize_data.py -c <s> -d <s> -m <c> -o <s> -s <s> -t <s> -v stacked

<s> . . . string
<c> . . . character
<i> . . . number

5.2.2 input parameter

- *-c*
  calculation identifier(s) to use, must be seperated by a comma. "all" if every calculation identifier should be taken
  (*e.g. "-c 500000_win_100000_shift__GC_percent","-c all",*
  "-c 500000_win_100000_shift__GC_percent,500000_win_100000_shift__CDS_gene").
- *-d*
  absolute path to the sqlite database file (*e.g. "-d /usr/local/chromoWIZ/Bd/Bd.db"*).
- *-m*
  percent (p) or absolute mode (a) (*e.g. "-m a" or "-m p"*).
- *-o*
  absolute path to the directory where results should be stored
  (*e.g. "-o /usr/local/data/chromoWIZ/Bd"*).
- *-s*
  "*seq_id(s)*" to use, must be seperated by a comma, "all" if every "*seq_id*" should be taken (*e.g. "-s Bd1,Bd2,Bd3,Bd4" or "-s all"*).
- *-t*
  density_table of sqlite database. prefix "density_" + "*genome_id*" (*e.g. "-t density_Bd"*).
- *-v*
  possible values: heatmap, barchart, linechart, stacked (*"-v=heatmap" or "-v barchart" or "-v linechart" or "-v stacked"*).
- *-x*
  maximum intensity for all calculations. "-1" indicates that no max parameter should be set (*e.g. "-x 100" or "-x -1"*).

5.2.3 supported parameter in display options

	-c	*-d*	*-m*	*-o*	*-s*	*-t*	*-v*	*-x*
heatmap	✓	✓	✓	✓	✓	✓	✓	✓
barchart	✓	✓		✓	✓	✓	✓	
linechart	✓	✓	✓	✓	✓	✓	✓	✓
stacked	✓	✓	✓	✓	✓	✓	✓	

Table 1: supported parameter in heatmap, stacked_heatmaps, linecharts and barcharts

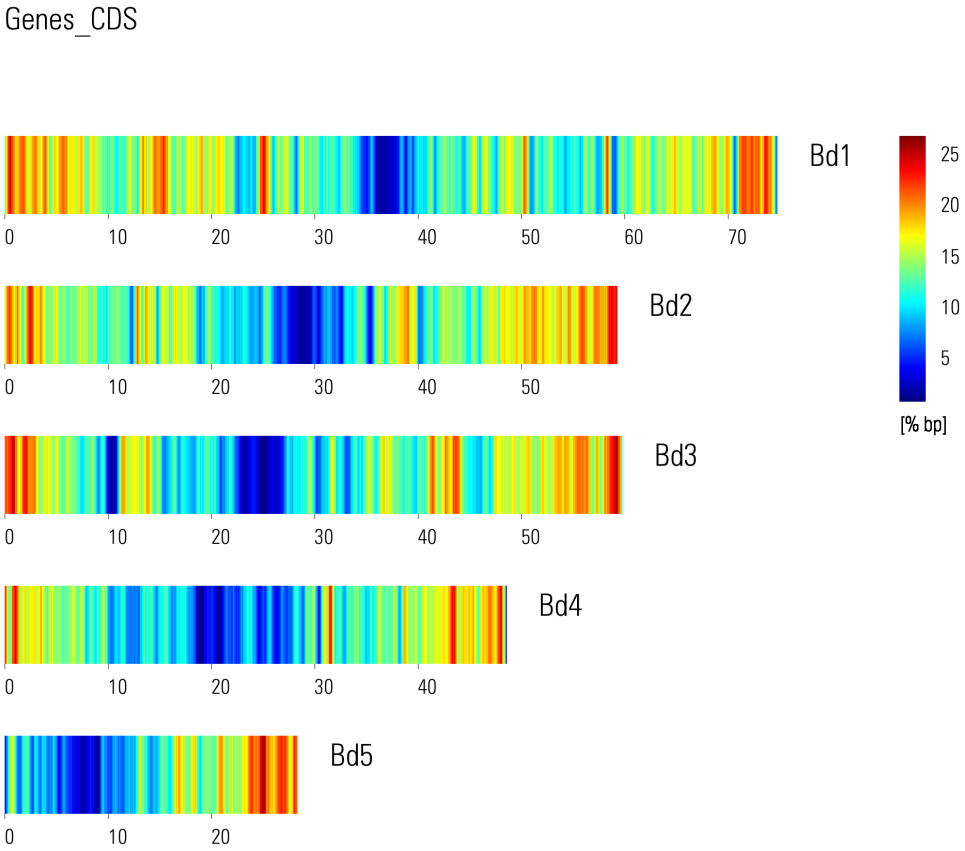


Figure 2: heatmap

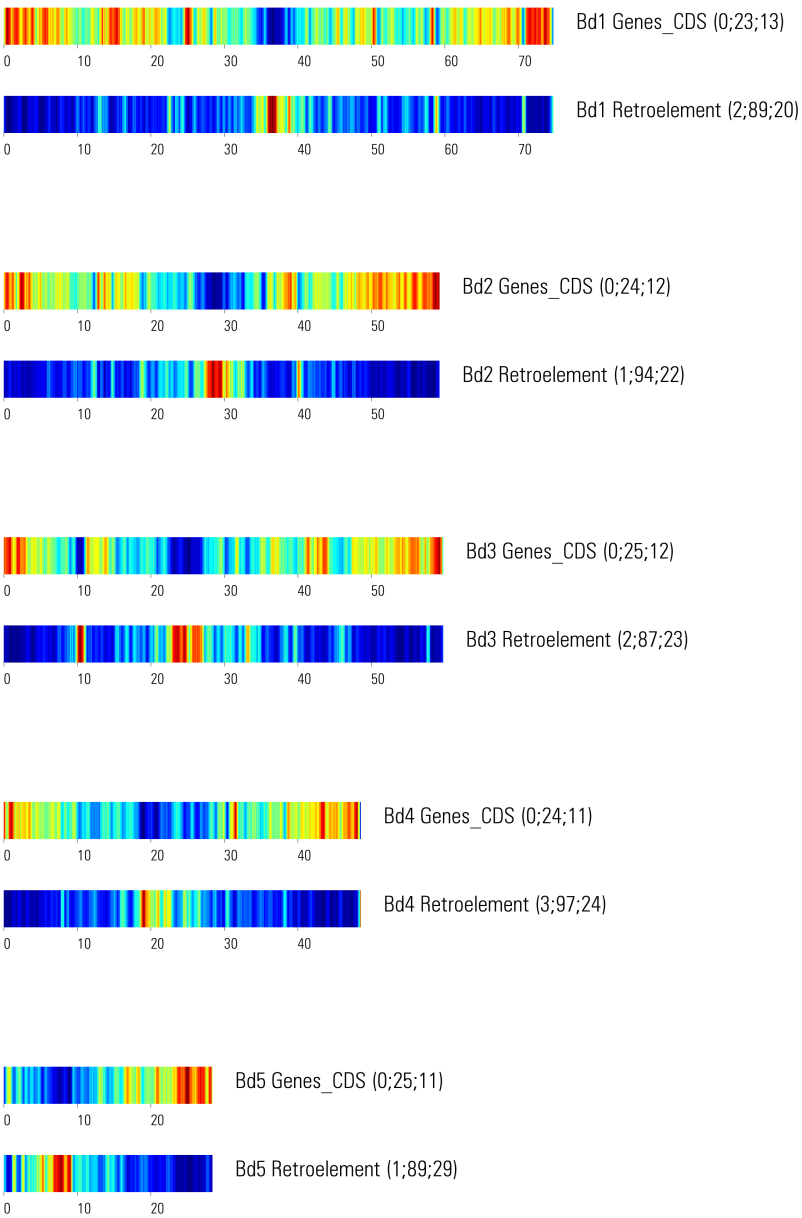


Figure 3: stacked heatmaps

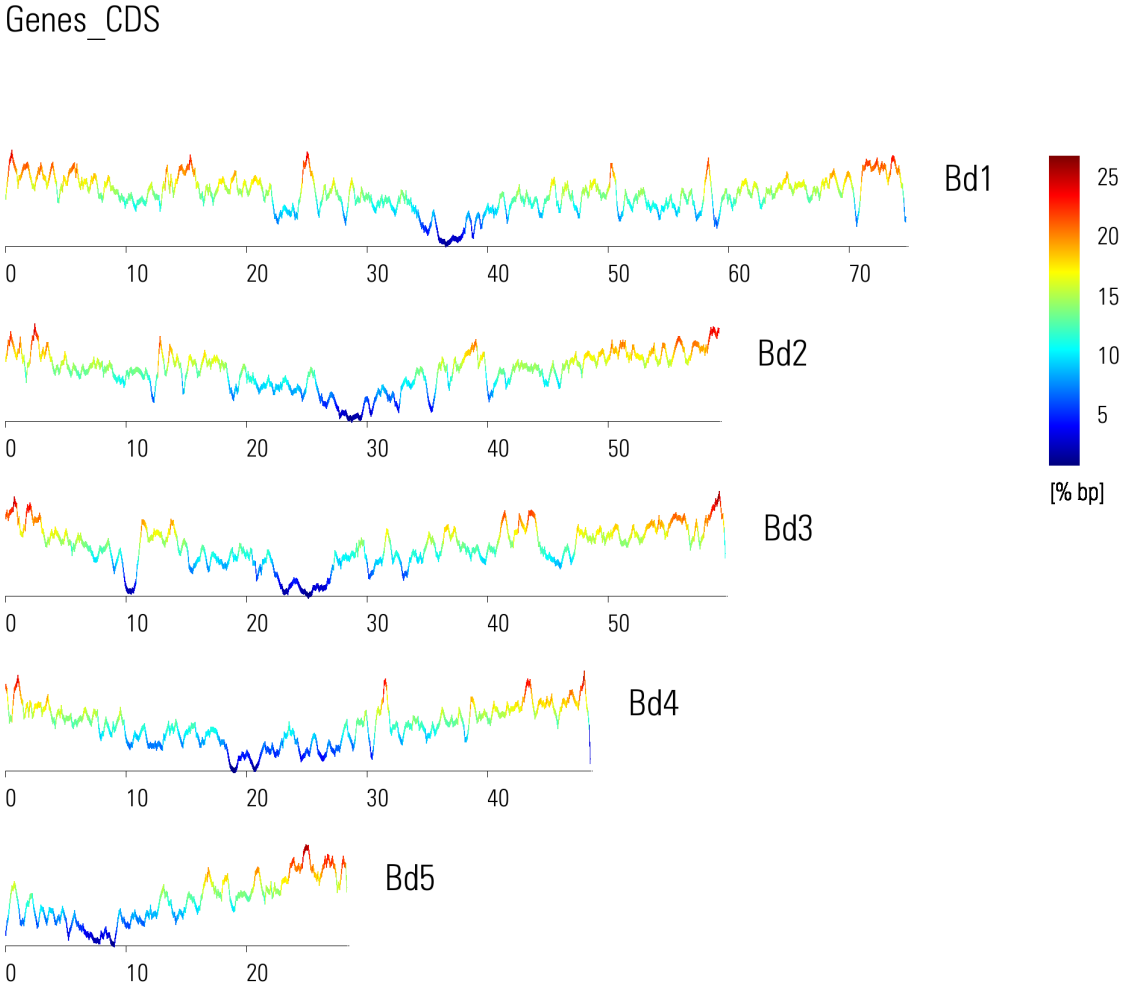


Figure 4: linechart

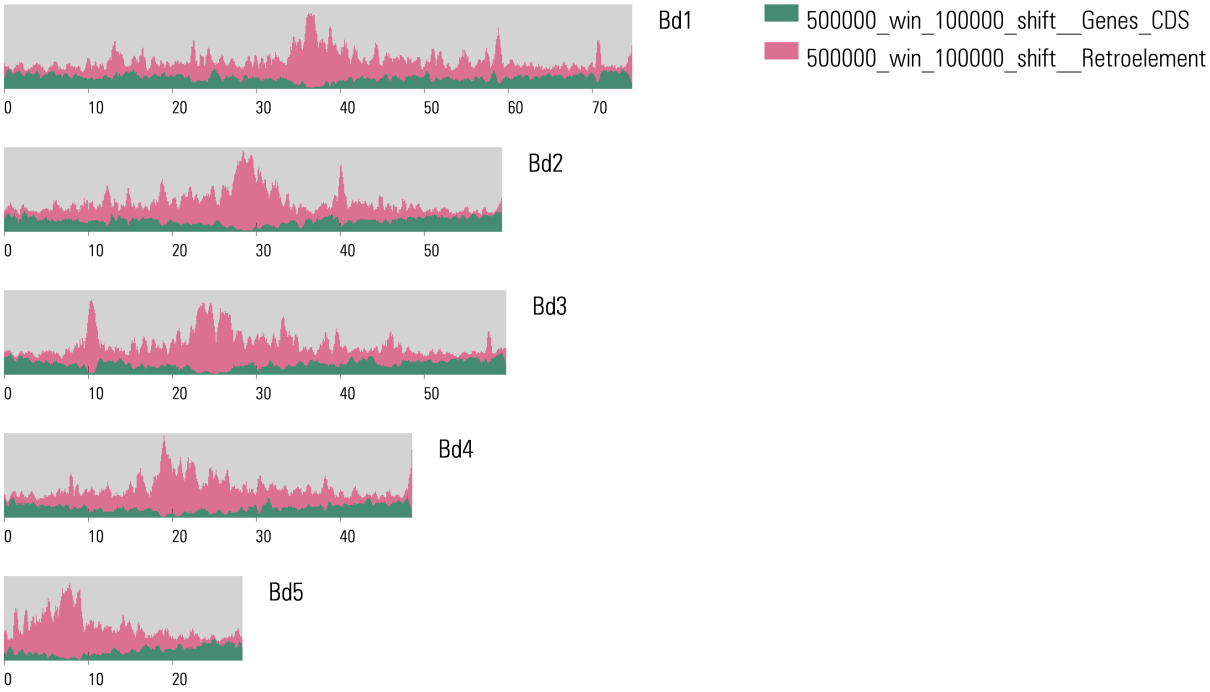


Figure 5: stacked barcharts

6 Feedback

If you have questions concerning *chromoWIZ* you can write an email to the following address:

chromoWIZ@helmholtz-muenchen.de

If you provide us additional ideas for *chromoWIZ* or if you found some bugs in the program, please contact us.