WGDI Documentation

version 0.4.4

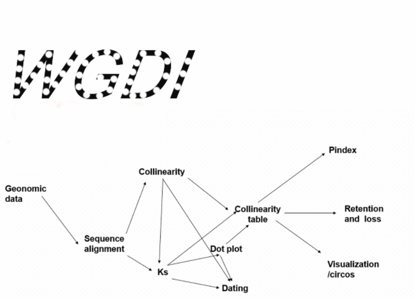
Pengchuan Sun

■■ 28, 2021

Contents

Welcome to WGDI's documentation!

Description

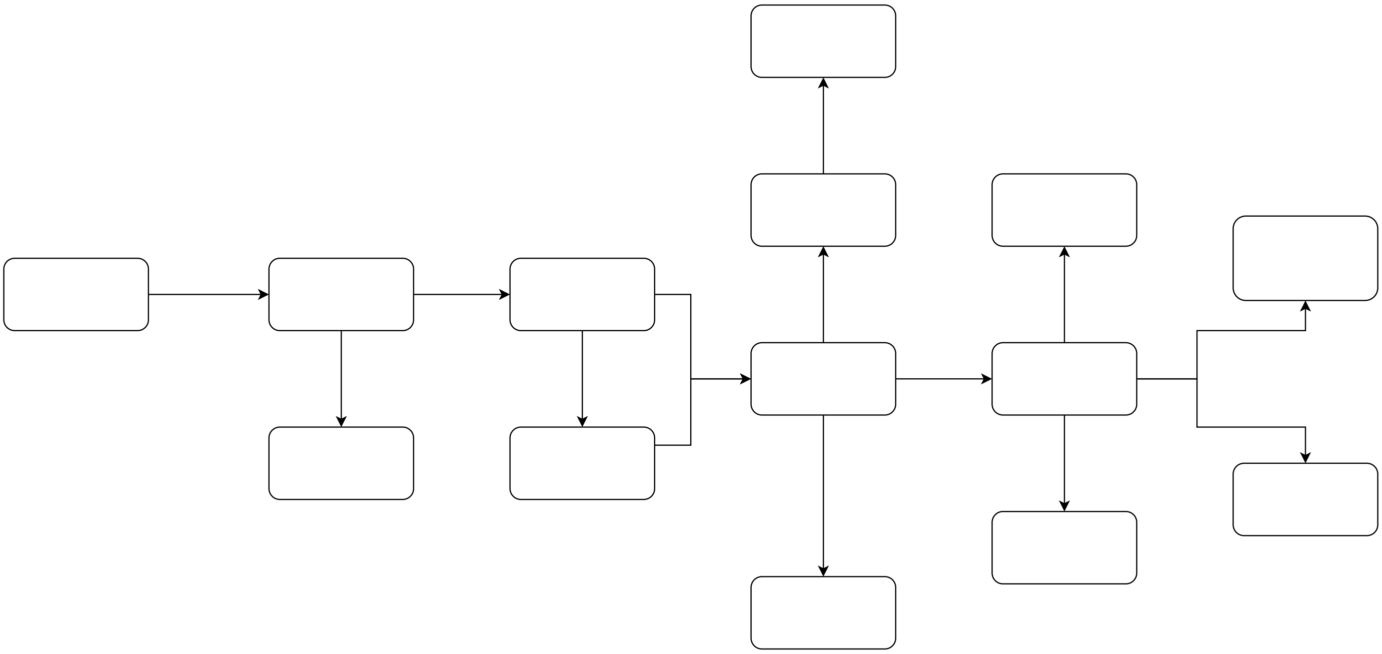


This is a gold standard for complex genomic analysis, including the construction of genomic homology maps, event-related collinear gene mapping, repeated gene classification, molecular evolution distance estimation, and the determination and correction of evolution rate differences, etc. Finely identify the whole genome duplication events and generate the genomic homology tables.

Table of Contents

Introduction

This is a gold standard for complex genomic analysis, including the construction of genomic homology maps, event-related collinear gene mapping, repeated gene classification, molecular evolution distance estimation, and the determination and correction of evolution rate differences, etc.



Installation

Python package and command line interface (IDLE) for the analysis of whole genome duplications (WGDI). The environment required for installation is python3.

Method

Bioconda

```
conda install -c bioconda  wgdi
```

Pypi

```
pip install wgdi
```

Github

```
git clone https://github.com/SunPengChuan/wgdi.git
cd wgdi
python setup.py install
```

Third party software

Some parts of WGDI use the following additional python libraries:

paml

mafft

muscle

pal2nal

After you download and install the above package. You can run the following command to configure the path of the existing software.

```
wgdi -conf help > conf.ini
```

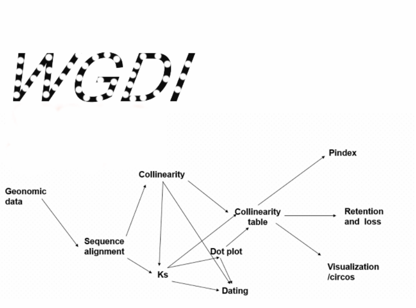
conf.ini:

```
[ini]
mafft_path = C:\bio\mafft-win\mafft.bat
pal2nal_path = C:\bio\[pal2nal.v14\pal2nal.pl
yn00_path = C:\bio\paml4.9j\bin\yn00.exe
muscle_path = C:\bio\muscle3.8.31_i86win32.exe
```

Uninstall

If you don't need `wgdi`, you can uninstall with `pip uninstall wgdi` or `conda remove wgdi`.

usage



Point open the **Centents** on the left.

We support the use of **WGDI** to complete the work on the icon number.

Common file

- conf

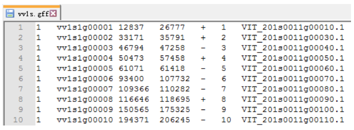
The .conf file contains the parameters required for the corresponding operation, which are read when wGDI is performed. Using `wgdi -* ? > *.conf` to generate needs to be in the same directory as the file mentioned in the parameter. And total.conf contains all the parameters, using `wgdi -conf ? > total.conf` is generated.

In the **conf file**: **gff1**, **lens1**, **gff2**, and **lens2** represent the files of species 1 and 2, respectively.

We will not explain in detail when we explain the parameters.

genome1_name and **genome1_name** represent the names of species 1 and 2, respectively. These parameters will be used to label the picture for your convenience.

- gff

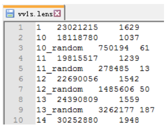


vv1s.gff

```
1   vv1s1g00001 12837   26777   +   1   VIT_201s0011g00010.1
1   vv1s1g00002 33171   35791   +   2   VIT_201s0011g00030.1
1   vv1s1g00003 46794   47258   -   3   VIT_201s0011g00040.1
1   vv1s1g00004 50473   57458   +   4   VIT_201s0011g00050.1
1   vv1s1g00005 61071   61418   -   5   VIT_201s0011g00060.1
1   vv1s1g00006 93400   107732  -   6   VIT_201s0011g00070.1
1   vv1s1g00007 109366  110282  -   7   VIT_201s0011g00080.1
1   vv1s1g00008 116646  118695  +   8   VIT_201s0011g00090.1
1   vv1s1g00009 150565  175325  -   9   VIT_201s0011g00100.1
1   vv1s1g00010 194371  206245  -   10  VIT_201s0011g00110.1
```

Column	Information	Explanation
1	Chr	Chromosome number
2	ID	Gene name
3	Strat	The location of the gene
4	End	Gene ending position
5	Direction	Direction of the gene sequence
6	Order name	Full name

- lens



vv1s.lens

```
1   23021215    1629
10  18118780    1037
10_random   750194  61
11  19815517    1239
11_random   278485  13
12  22690056    1542
12_random   1485606 50
13  24390809    1559
13_random   3262177 187
14  30252880    1948
```

Column	Information	Explanation
1		Chromosome number
2	Chr lens	Number of chromosome sequences
3		Number of chromosome genes
*_random		Not slicing the genes on the chromosomes

- The explosion is the output file of the blast+ ,available in the -6 and m-8 formats..

Contents

dotplot

dotplot is show homologous gene dotplot.

Parameters

Use command to enter the folder `wgdi -d ? > dotplot.conf` Take out the parameter file.

```
[dotplot]
blast= blast file
gffl = gff1 file
gff2 = gfi2 file
lens1 = lens1 file
lens2 = lens2 file
genome1_ name =  Genome1 name
genome2_ name =  Genome2 name
multiple  = 1
score = 100
evalue = 1e-5
repeat_number = 20
position = order
markersize = 0.5
figsize = 10,10
savefile = savefile(.png, .pdf)
```

Parameters	Standards and instructions
multiple	Type: int Default: 1 The best number of homologous genes.
score	Type: int Default: 100 Score value in the blast results.
evalue	Type: float Default: 1e-5 Evalue value in the blast result.
repeat_number	Type: int Default: 20 The maximum number of homologous genes is allowed to be copied, the rest removed.
position	Type: {start,order,end} Default: order The position of the gene corresponds to the gff file.
markersiz	Type: float Default: 0.5 The size of the point in the plot.
figsize	Type: int,int Default: 10,10 Control the proportion of the size of the saved picture.
savefile	Type: {*.png,*.pdf} Default: *.png Save pictures support png, pdf, svg formats.

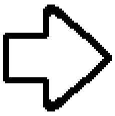
Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi -d dotplot.conf` to run the parameter file and output the results you want.



colinearscan

colinearscan is a simple way to run ColinearScan.

Parameters

Use command to enter the folder `wgdi -cl ? > colinearscan.conf` Take out the parameter file.:

```
[colinearscan]
gff1 = gff1 file
gff2= gff2 file
lens1 = lens1 file
lens2 = lens2 file
blast = blast file
dir = Output file
evalue = 1e-5
score = 100
mg = 50,50
repeat_number = 20
positon = order
```

Parameters	Standards and instructions

dir	Type: str Default: - The directory of the generated file.
evalue	Type: float Default: 1e-5 Evalue in the blast result.
score	Type: int Default: 100 Score value in the blast results.
mg	Type: int,int Default: 50,50 The maximum clearance length (mg) is an important parameter for detecting collinear regions.
repeat_number	Type: int Default: 20 The maximum number of homologous genes is allowed to be removed more than part of the population.
position	Type: {start,order,end} Default: order The position of the gene corresponds to the gff file.

Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi -cl colinearscan.conf` to run the parameter file and output the results you want.

ks

ks is calculate Ka/Ks for homologous gene pairs by Comdel.

Parameters

Use command to enter the folder `wgdi -ks ? > ks.conf` Take out the parameter file.:

```
[ks]
cds_file = cds file
pep_file = pep file
align software = muscle
pairs_file = gene  pairs file
ks_file = ks result
```

Parameters	Standards and instructions
cds_file	Type: str Default: - A cds file of one or more genomes.
pep_file	Type:str Default:- A protein file for one or more genomes. non-essential files, if you do not write this parameter, Then this file will be translated through the biopython module cds-file.
align_software	Type:{muscle,mafft} Default: muscle Multi-sequence comparison software.

pairs_file	Type:str Default: - The same gene pairs of ks need to be calculated, either by pressing, or separating the list, or as the output of ColinearScan.
ks_file	Type:str Default: - The output file name of ks.

Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi -ks ks.conf` to run the parameter file and output the results you want.

align

align is show event-related genomic alignment in a dotplot.

Parameters

Use command to enter the folder `wgdi -a ? > align.conf` Take out the parameter file.:

```
[alignment]
gff1 =  gff1 file
gff2 =  gff2 file
lens1 = lens1 file
lens2 = lens2 file
genome1_ name =  Genome1 name
genome2_ name =  Genome2 name
markersize = 0.5
position = order
colors = red,blue,green
figsize = 10,10
savefile = savefile(.csv)
savefig= savefig(.png,.pdf)
block_list = 1.txt
blockinfo = block information file
```

Parameters	Standards and instructions
markersize	Type: str Default: 0.5 The size of the control point.
position	Type: str Default: order The position of the gene corresponds to the gff file.
colors	Type: {red,blue,green} Default: red,blue,green Set multiple sets of colors based on grouping, split with a comma.
figsize	Type: int,int Default: 10,10 Control the proportion of the size of the saved picture.
savefile	Type: str Default: *.csv A resulting collinear list.

savefig	Type: {.*png,*.pdf} Default: *.png Save pictures support png, PDF formats.
block_list	Type: str Default: - Add a class column to the blockinfo file to group and express different groups with different numbers.
blockinfo	Type:str Default:- Integrate collinear and ks files.

Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi -a align.conf` to run the parameter file and output the results you want.

blockks

blockks is show Ks of blocks in a dotplot.

Parameters

Use command to enter the folder `wgdi -bk ? > blockks.conf` Take out the parameter file.

```
[blockks]
lens1 = lens1 file
lens2 = lens2 file
genome1_ name = Genome1 name
genome2_ name = Genome2 name
blockinfo = block information (*.csv)
pvalue = 0.05
tandem = true
markersize = 1
area = 0,1
block_length = int number
figsize = 10,10
savefile = save_image(.png,.pdf,.svg)
```

Parameters	Standards and instructions
colinearity	Type: str Default: - Colinscan results file.
ks	Type: str Default: - ks calculation results.
markersize	Type: str Default: 1 The size of the control point.
area	Type: str Default: 0,1 Show the range of ks.
block_length	Type: str Default: int number Show the minimum length of a collinear block.

position	Type: str Default: order The position of the gene corresponds to the gff file.
figsize	Type: int,int Default: 10,10 Control the proportion of the size of the saved picture.
savefile	Type: {*.png, *.pdf,*.svg} Default: *.png Save pictures support png, pdf■svg formats.

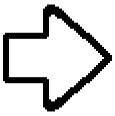
Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi -bk blockks.conf` to run the parameter file and output the results you want.



circos

circos is a simple way to run circos.

Parameters

Use command to enter the folder `wgdi -ci ? > circos.conf` Take out the parameter file.:

```
[circos]
gff =  gff file
lens = lens file
radius = 0.2
angle_gap = 0.05
ring_width= 0.015
colors  = color confige(chr:color,chr:color)
position = end
alignment = text.txt
chr_label =
figsize = 10,10
savefig = saving image(.png,.pdf)
```

Parameters	Standards and instructions
radius	Type: float Default: 0.2 Radius, value between 0-1.
angle_gap	Type: float Default: 0.05 Gap between chromosomes.
ring_width	Type: float Default: 0.015 The width of the ring.
colors	Type: str Default: - Set multiple sets of colors based on grouping, split with a comma.
position	Type: {start,order,end} Default: order The position of the gene corresponds to the gff file.
alignment	Type: str Default: - Colinear List.

chr_label	Type: str Default: Shorthand A shorthand for chromosomes.
figsize	Type: int,int Default: 10,10 The size ratio of the image.
savefile	Type: {*.png, *.pdf} Default: *.png Save pictures support png, pdf formats.

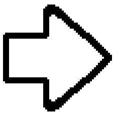
Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi –ci circos.conf` to run the parameter file and output the results you want.



kspeaks

kspeaks is a simple way to get ks peaks.

Parameters

Use command to enter the folder `wgdi -kp ? > kp.conf` Take out the parameter file.:

```
[kspeaks]

 blockinfo = block information (*.csv)
 pvalue = 0.05
 tandem = true
 block_ length = int number
 ks_area = 0,10
 multiple = 1
 homo = 0,1
 fontsize = 9
 area = 0,3
 figsize = 10,6.18
 savefig = saving image(.png,.pdf)
 savefile = ks medain savefile
```

Parameters	Standards and instructions
blockinfo	Type: str Default: - Integrate collinear and ks files.
pvalue	Type:str Default: 0.05 P-value in collinear results.
tandem	Type:str Default: true The criterion is that there are no more than 200 genes with a difference in genetic location.
block_length	Type:str Default: int number Minimum length of collinear blocks.

ks_area	Type:str Default: 0,10 Show the range of ks.
multiple	Type:str Default: 1 The optimal number of homologous genes.
homo	Type:str Default: 0,1 Collinear fragments favor the best matching values, with a range of -1, 1.
fontsize	Type:str Default: 9 The size of the font.
area	Type:str Default: 0,3 The extent of the drawing display.
figsize	Type:str Default: 10,6.18 The size ratio of the image
savefig	Type:{*.png, *.pdf} Default: *.png Save pictures support png, PDF formats.
savefile	Type:*.csv Default: *.csv Save pictures support csv formats.

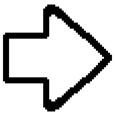
Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi -kp kp.conf` to run the parameter file and output the results you want.



retain

retain is show subgenomes in gene retention or genome fractionation.

Parameters

Use command to enter the folder `wgdi -r ? > retain.conf` Take out the parameter file.:

```
[retain]
alignment = alignment file
gff = gff file
colors = red,blue,green
refgenome = shorthand
figsize = 10,12
step = 50
ylabel = y label
savefile = retain file(result)
figurefile = result(.png,.pdf)
```

Parameters	Standards and instructions
alignment	Type:str Default: - Colinear List.

colors	Type:{red,blue,green} Default:- Set multiple sets of colors based on grouping, split with a comma.
refgenome	Type:str Default: - A short handbook of reference species.
figsize	Type:str Default: - The size ratio of the image.
step	Type:int Default: - The size of the sliding window.
ylabel	Type:str Default: - The y-axis label of the picture.
savefile	Type:str Default: - Results of the drawing.
figurefile	Type:{*.png, *.pdf} Default: *.png Save pictures support png, PDF formats.

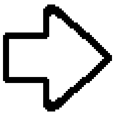
Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi -r retain.conf` to run the parameter file and output the results you want.



correspondence

correspondence is extract event-related genomic alignment.

Parameters

Use command to enter the folder `wgdi -c ? > correspondence.conf` Take out the parameter file.:

```
[correspondence]
blockinfo = blockinfo file(.csv)
lens1 = lens1 file
lens2 = lens2 file
tandem = (true/false)
pvalue = 0.05
block_length = 5
multiple = 1
homo = 0,1
savefile = savefile(.csv)
```

Parameters	Standards and instructions
blockinfo	Type: str Default:- Integrate collinear and ks files.
tandem	Type: {true,false} Default: true The criterion is that there are no more than 200 genes with a difference in genetic location.

pvalue	Type: str Default: 0.05 P-value in collinear results.
block_length	Type: str Default: int number Minimum length of collinear blocks.
multiple	Type: str Default: 1 The optimal number of homologous genes.
homo	Type: str Default: 0,1 Collinear fragments favor the best matching values, with a range of -1, 1.
savefile	Type: *.csv Default: *.csv Save pictures support csv formats.

Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi –c correspondence.conf` to run the parameter file and output the results you want.

pf

peaksfit is gaussian fitting of ks distribution.

Parameters

Use command to enter the folder `wgdi -pf ? > blockks.conf` Take out the parameter file.

Parameters	Standards and instructions
	Type: str Default: -
	Type: str Default: -
	Type: str Default: -
	Type: str Default: -
	Type: str Default: -
	Type: str Default: -
	Type: str Default: -
	Type: str Default: -

Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi –pf blockks.conf` to run the parameter file and output the results you want.

bi

bi is collinearity and Ks speculate whole genome duplication.

Parameters

Use command to enter the folder `wgdi -bk ? > blockks.conf` Take out the parameter file.:

```
[b1ockinfo]
blast = blast file
gff1 = gffl file
gff2 = gff2 file
lens1 = lens1 file
1ens2 = lens2 file
colinearity = colinearity file
score = 100
evalue = 1e-5
repeat_number = 30
position = order
ks = ks file
ks_col = ka_ NG86
savefile = block information (*.csv)
```

Parameters	Standards and instructions
colinearity	Type: str Default: - Colinscan results file.
score	Type: int Default: 100 Score value in the blast results.
evalue	Type: float Default: 1e-5 Evalue value in blast result.
repeat_number	Type: int Default: 20 The maximum number of homologous genes is allowed to be removed more than part of the population.
position	Type: {start,order,end} Default: order The position of the gene corresponds to the gff file.
ks	Type: str Default: - ks calculation results.
ks_col	Type: str Default: -
savefile	Type: *.csv Default: *.csv The resulting file.

Example

Modify

Modify the parameters that are right for you to run.

Begin

Use `wgdi –bi blockinfo.conf` to run the parameter file and output the results you want.

Convenient

- You can use `wgdi -conf ? > total.conf` generates a **total.conf** file with all parameters, and when you modify the parameters and run **WGDI**, **WGDI** will only read the parameters corresponding to the **total.conf** file to execute your command.
- We put in the example of **git**'s official website, where all parameters are in the **total.conf** file.
- When a folder runs **WGDI**, **WGDI** automatically generates results for you in the background, and you can exit the folder and go to the next folder to start working.
- **WGDI** performs the **.conf** file for the current folder, so you can copy the **.conf** file in bulk and make parameter modifications that apply to the target folder.

Help us

When you have used **WGDI**, you have good suggestions or ideas to email the PengChuan Sun's mailbox or submit changes on our github.