Genotoul Biostat Bioinfo day
Toulouse, Dec 13, 2018

GeneNeighborhood

an R package to explore the direct neighbors of your favorite gene set

Pascal GP Martin

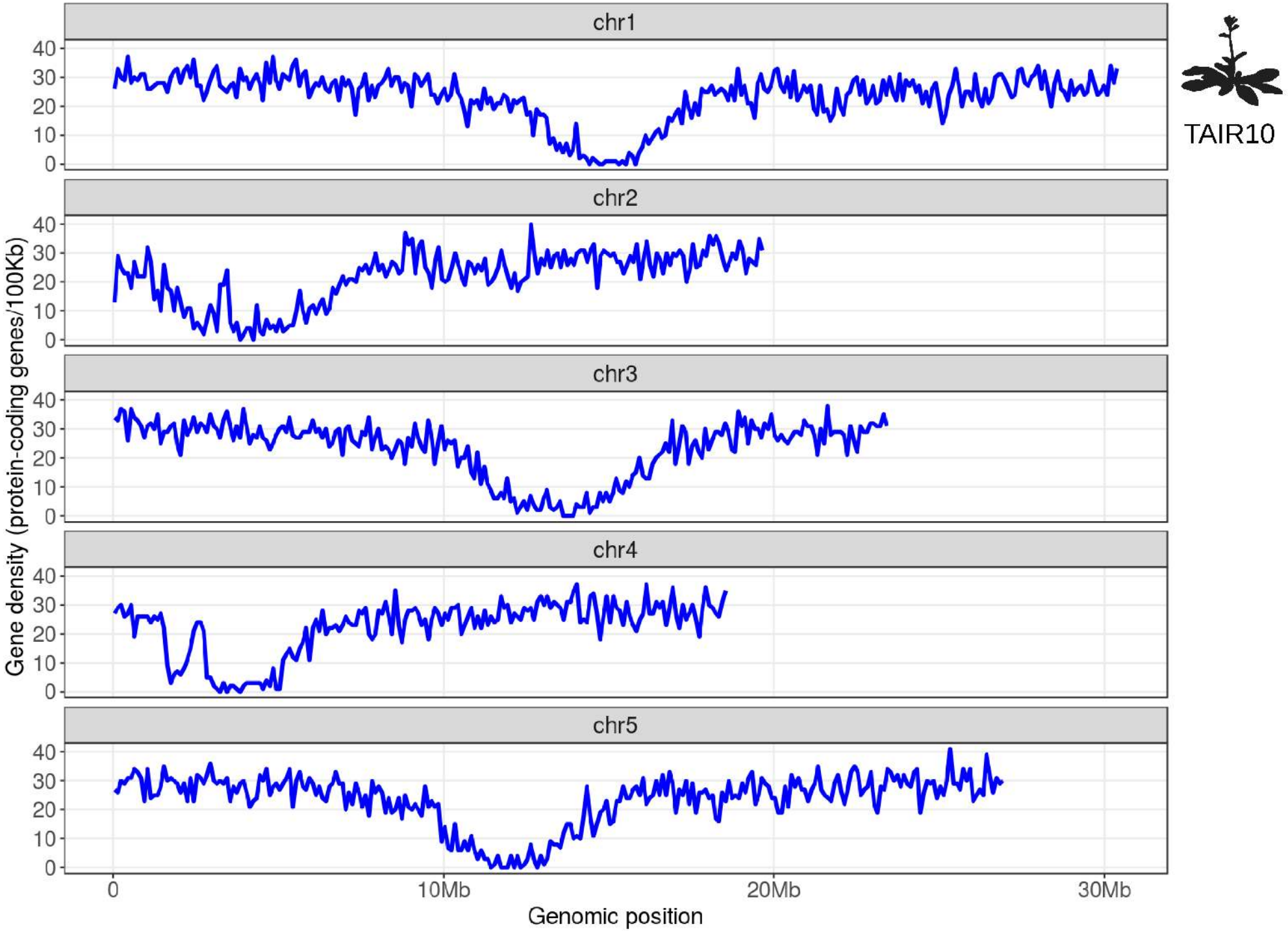
Pascal.Martin@inra.fr

github.com/pgpmartin

@PgpMartin

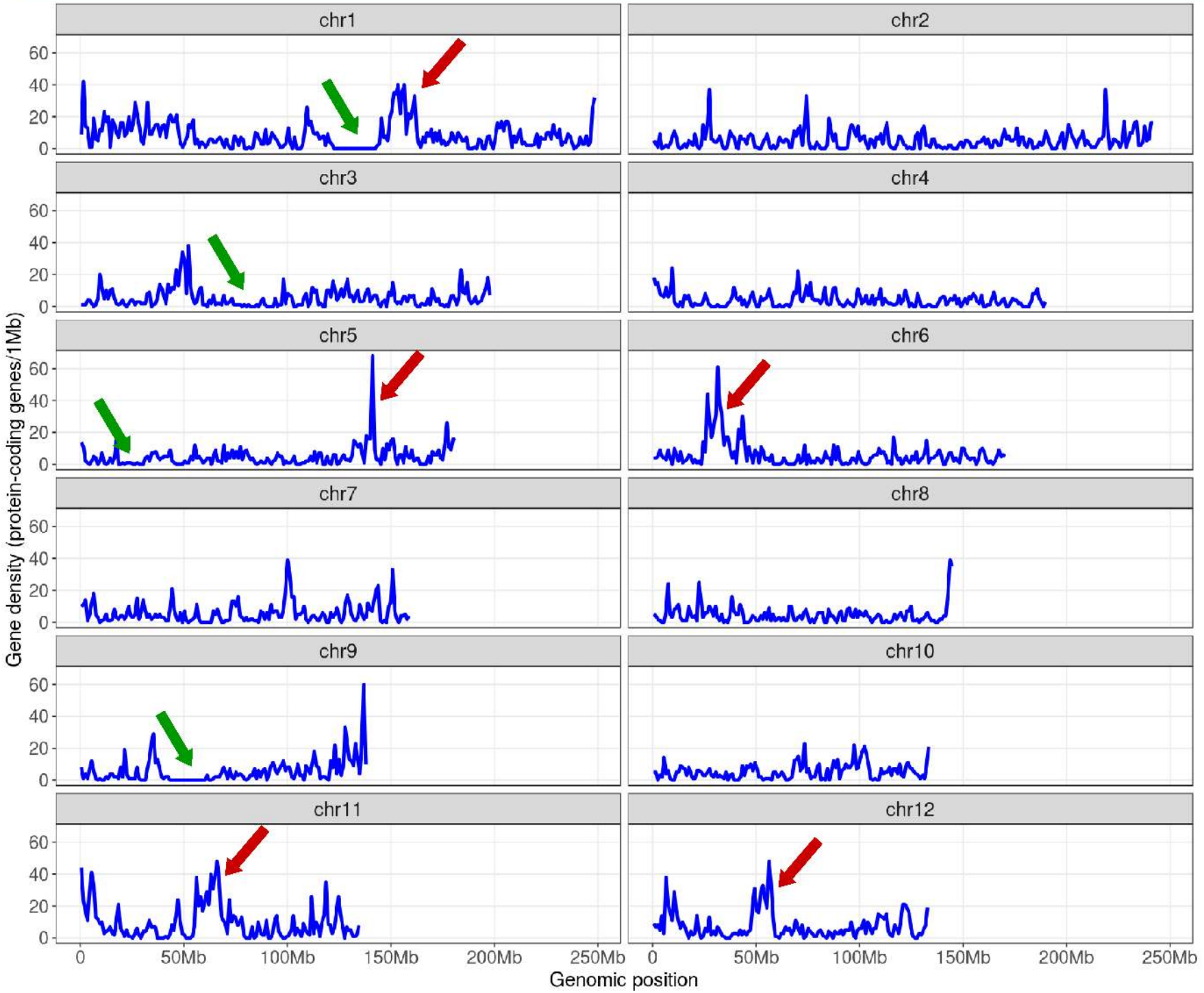
perso.pgpmartin.fr

Genes are not randomly distributed in genomes

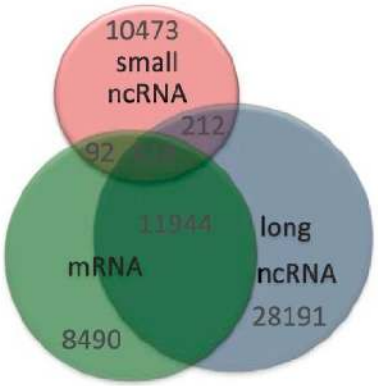


Genes are not randomly distributed in genomes

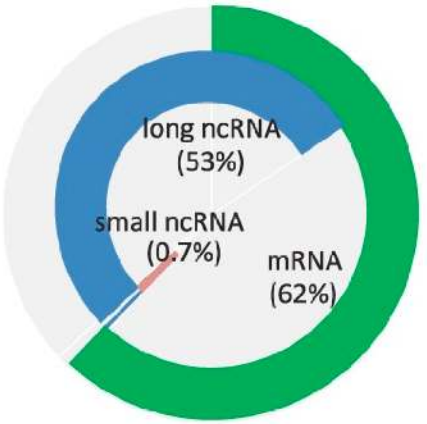
Gencode v28



A large fraction of the genome is transcribed

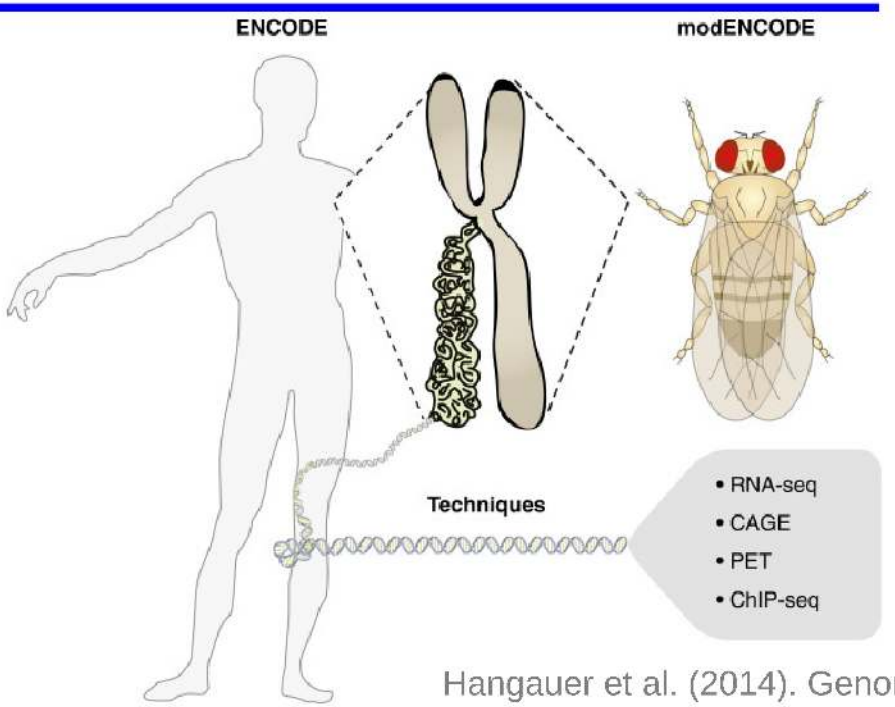


Number of loci



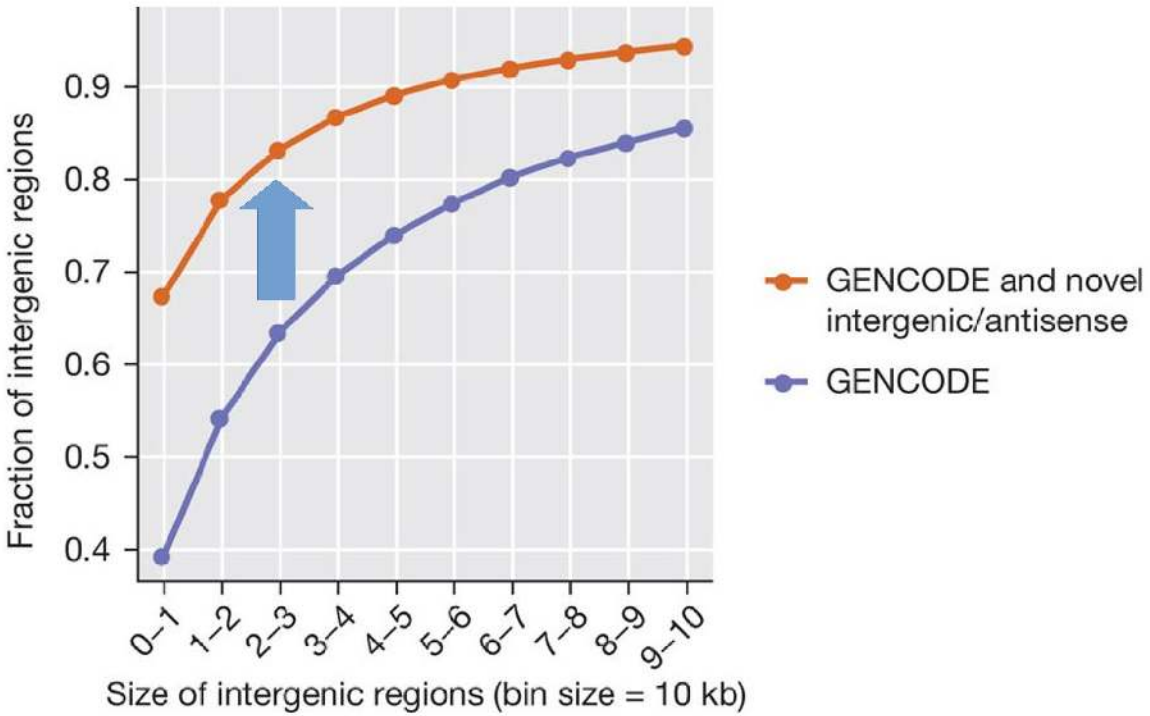
bp coverage of
the transcriptome

Pertea M. (2012). The human transcriptome. Genes.



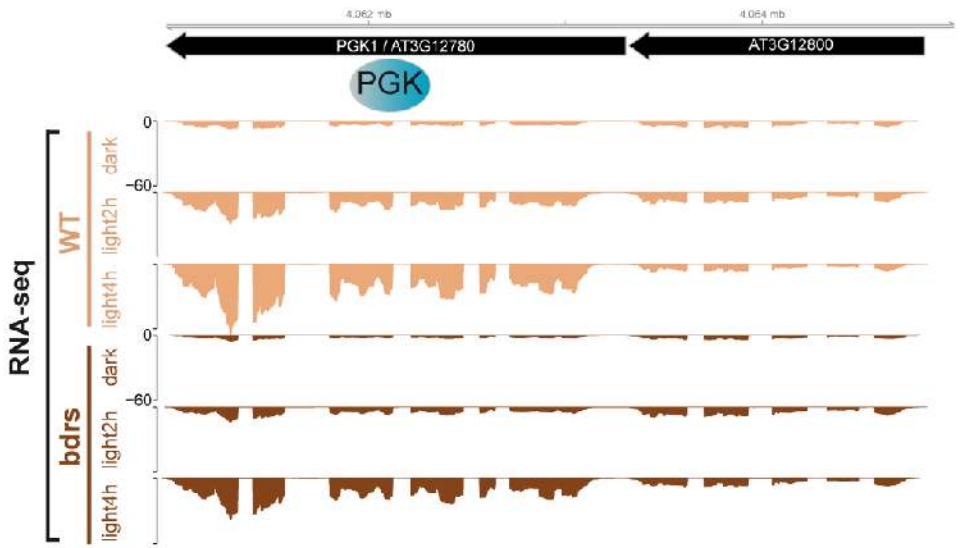
Hangauer et al. (2014). Genome Biol.

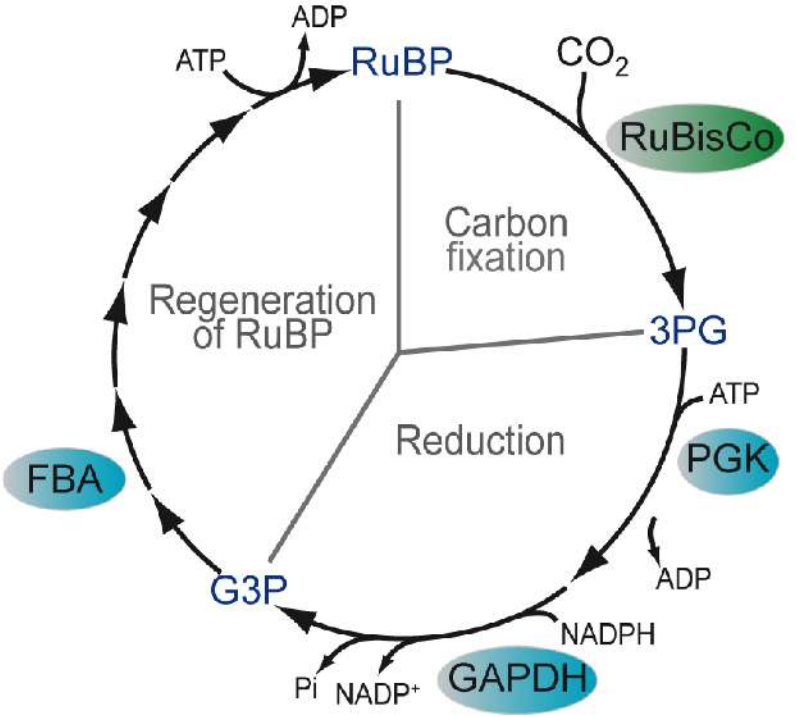
- ✔ enhancers
- ✔ promoters
- ✔ repetitive elements
- ✔ . . .

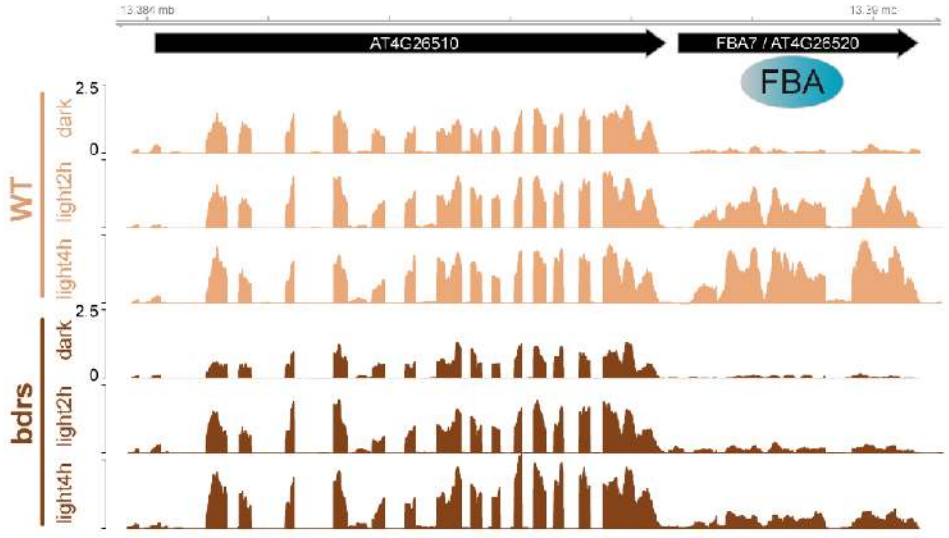


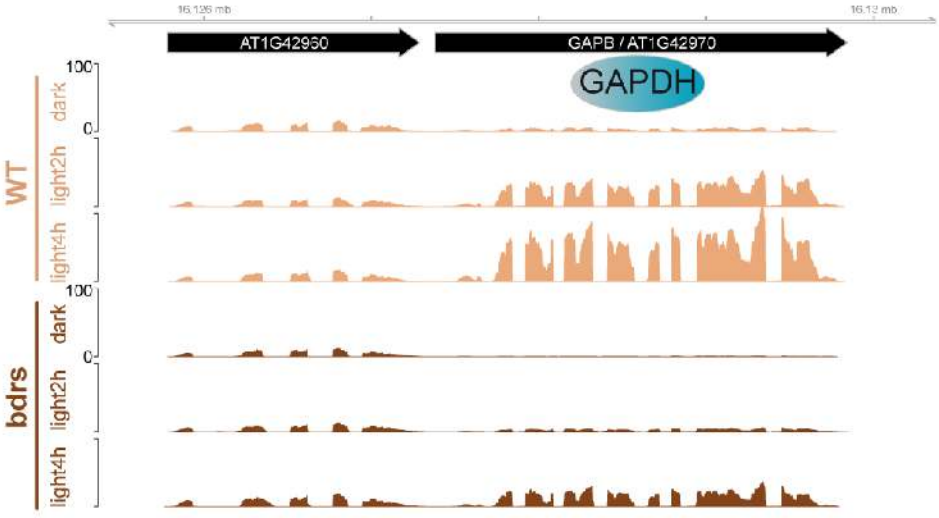
Djebali S et al. (2012) Landscape of transcription in human cells. Nature

with some functional consequences...



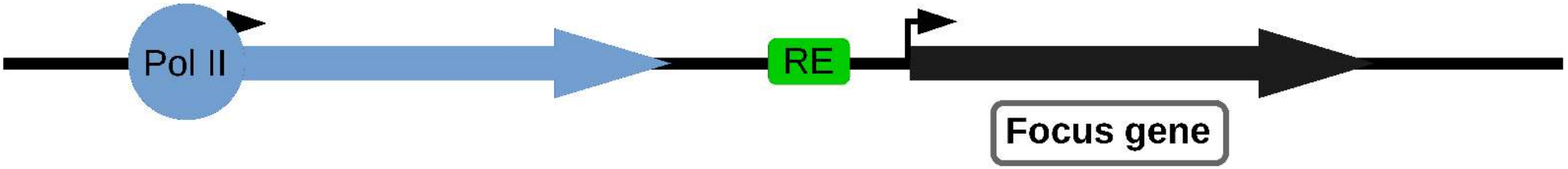






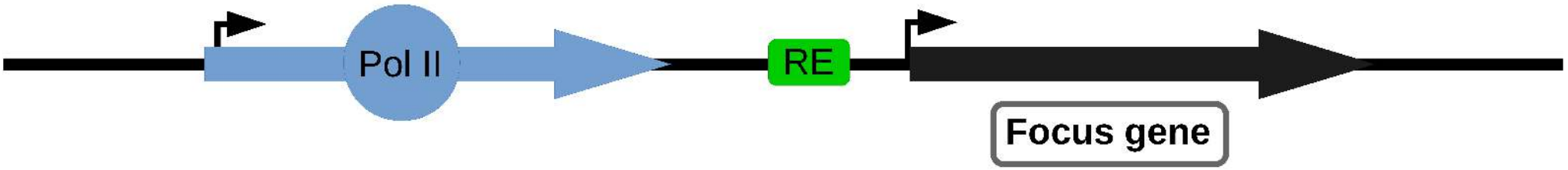
Transcriptional interferences

Shearwin et al., Trends Genet 2005



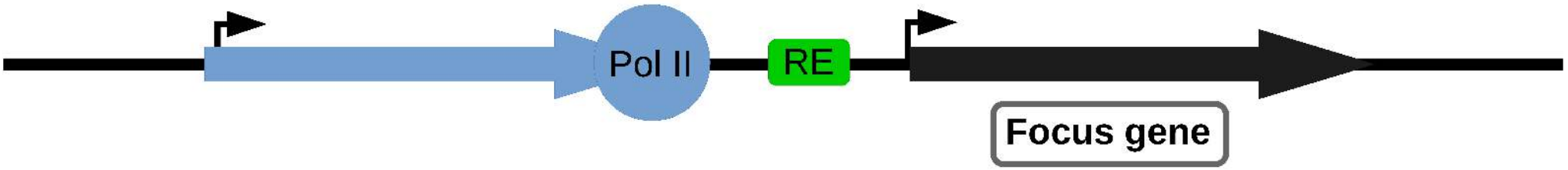
Transcriptional interferences

Shearwin et al., Trends Genet 2005



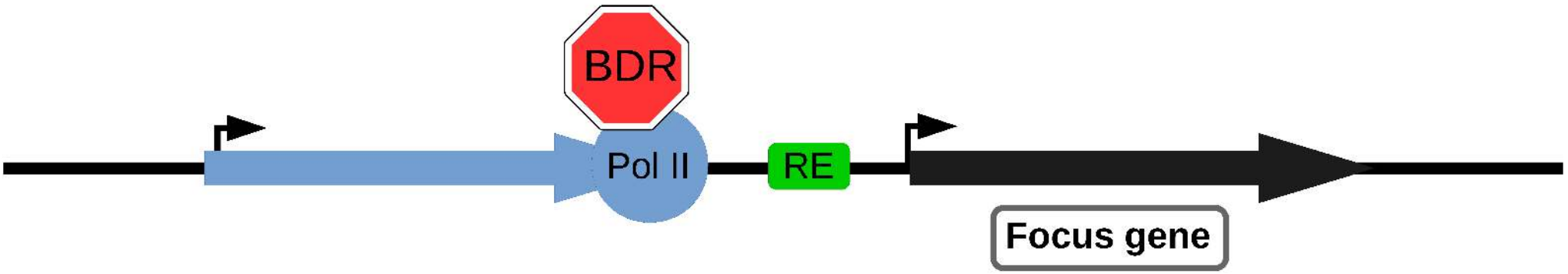
Transcriptional interferences

Shearwin et al., Trends Genet 2005



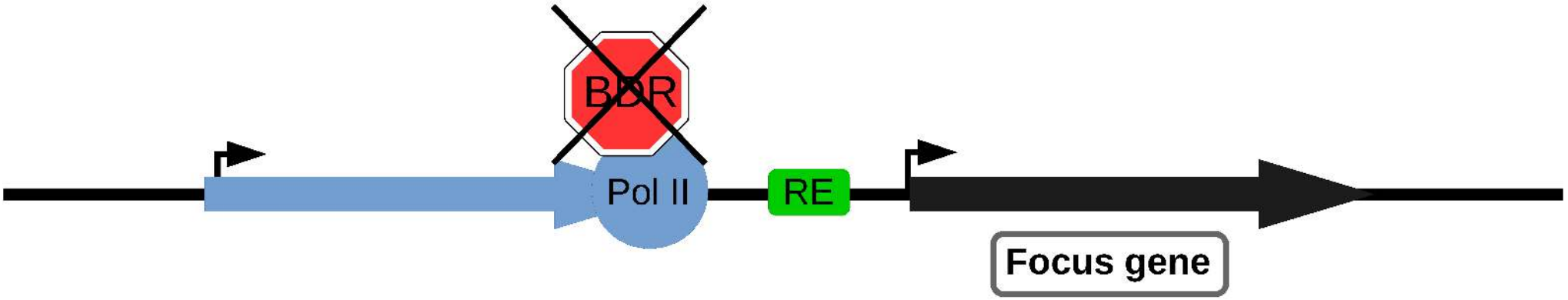
Transcriptional interferences

Shearwin et al., Trends Genet 2005



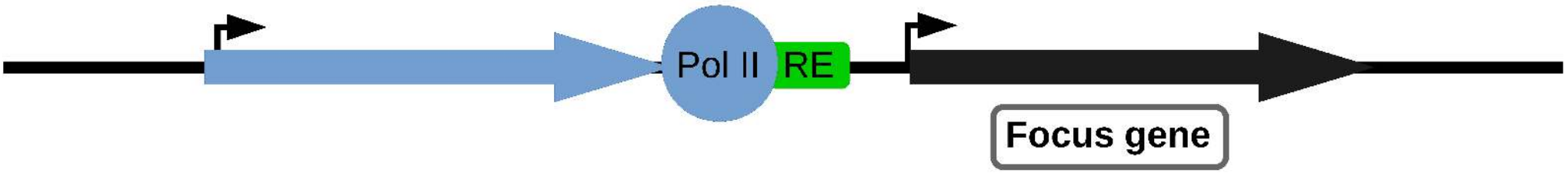
Transcriptional interferences

Shearwin et al., Trends Genet 2005



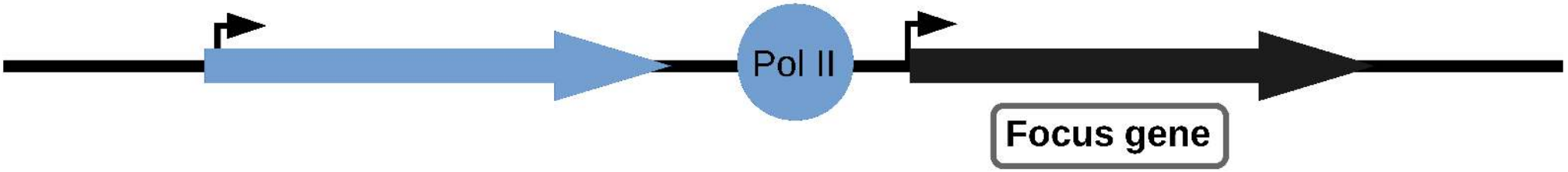
Transcriptional interferences

Shearwin et al., Trends Genet 2005



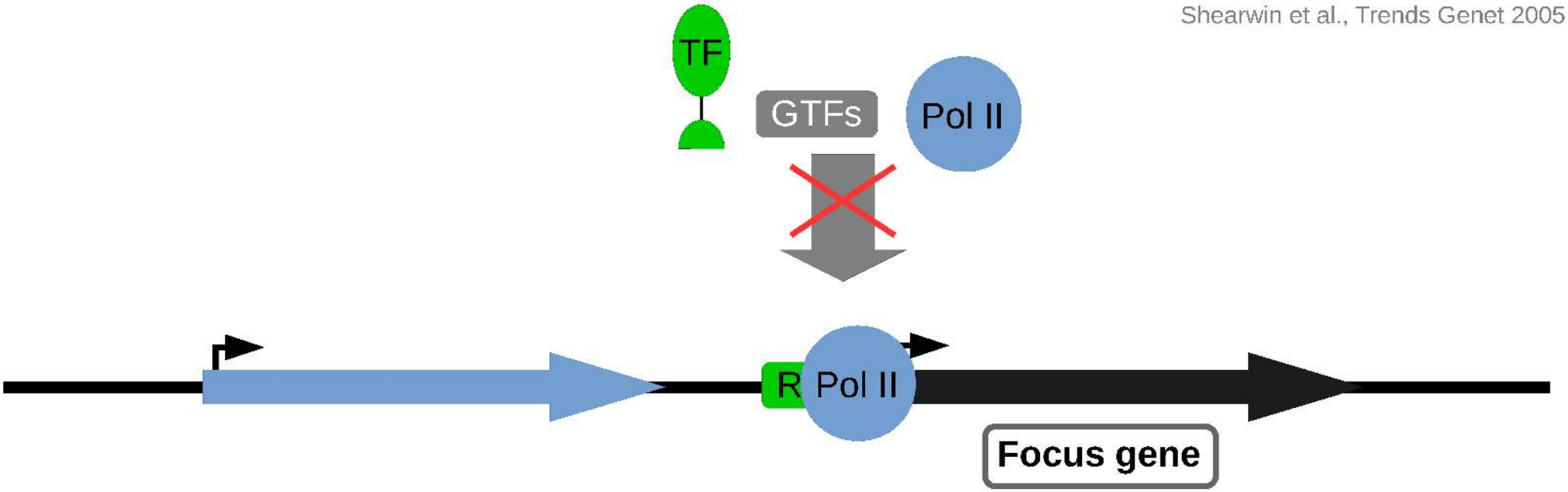
Transcriptional interferences

Shearwin et al., Trends Genet 2005



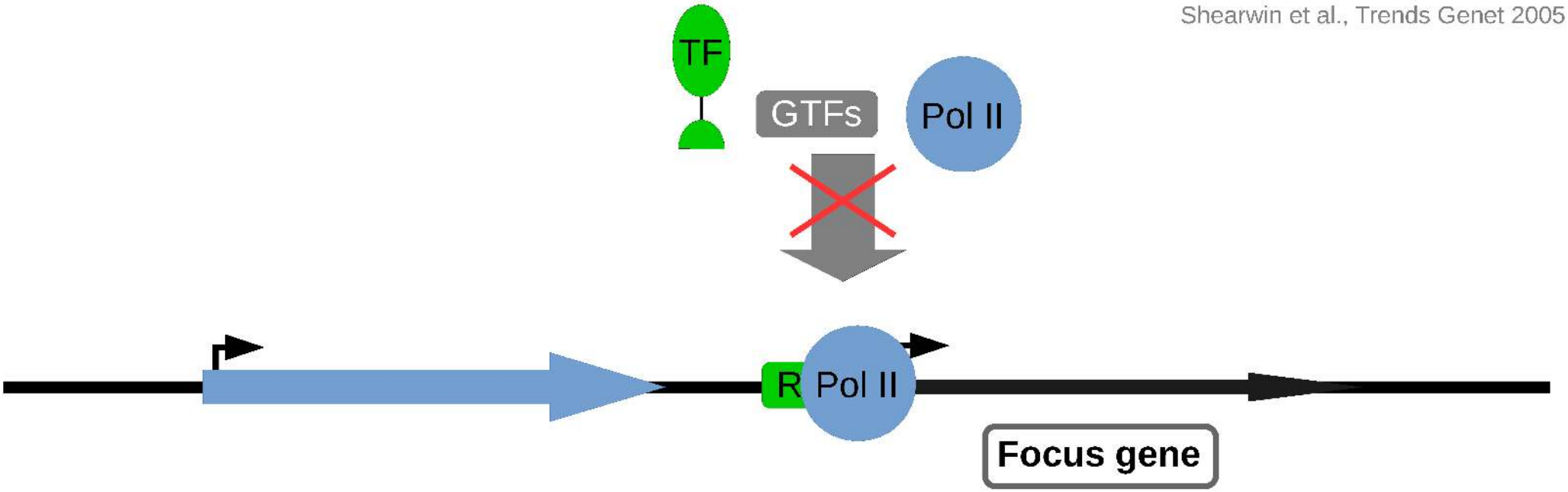
Transcriptional interferences

Shearwin et al., Trends Genet 2005

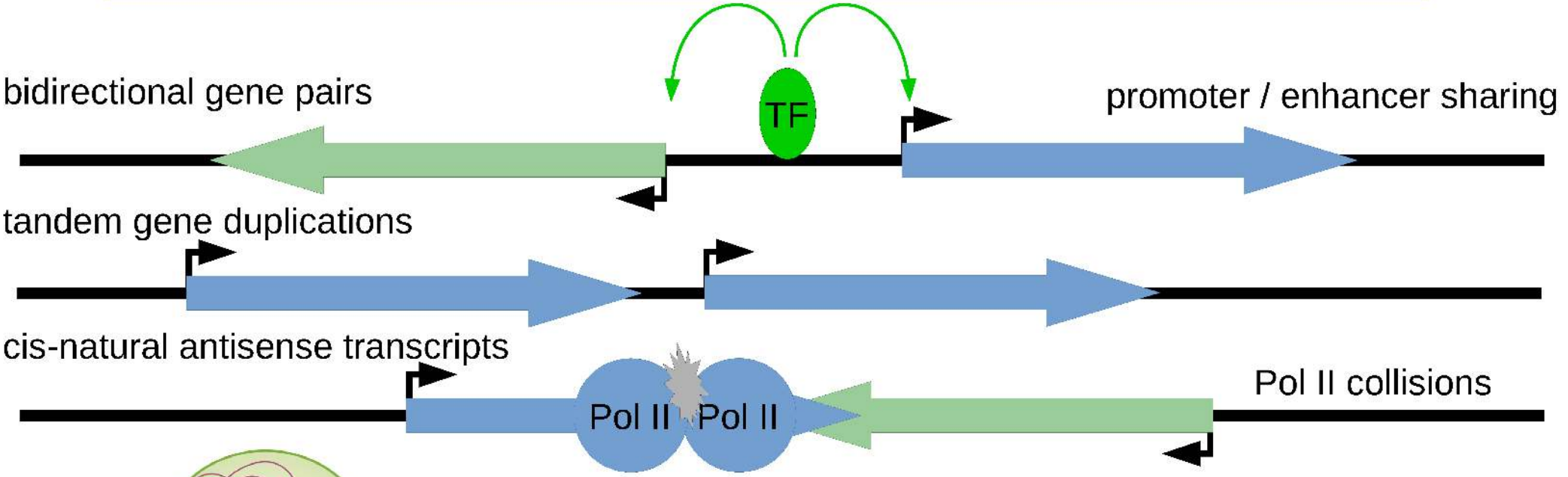


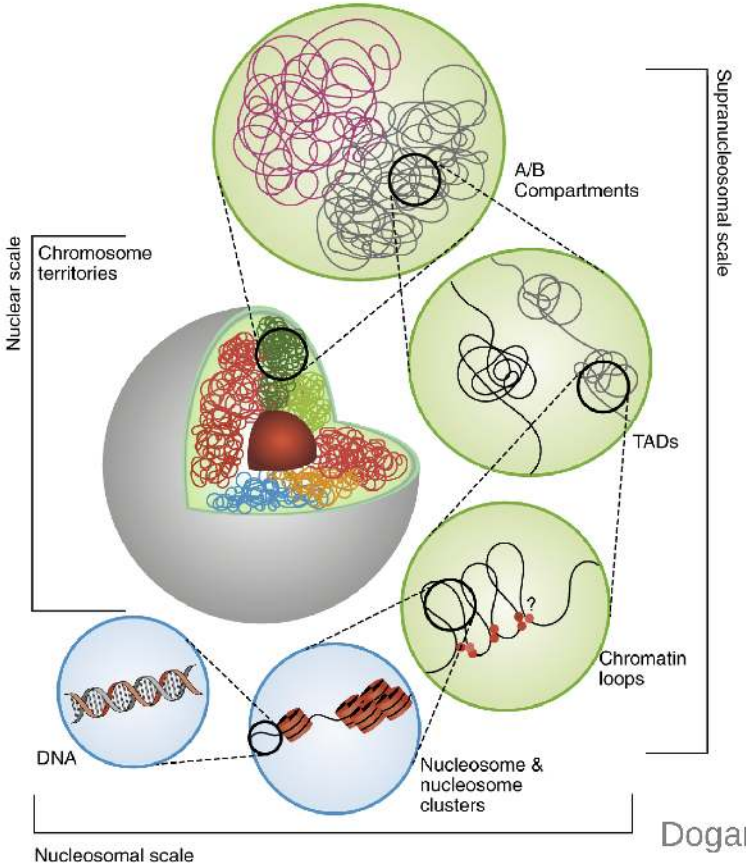
Transcriptional interferences

Shearwin et al., Trends Genet 2005



Other mechanisms associated with co-expression or interference





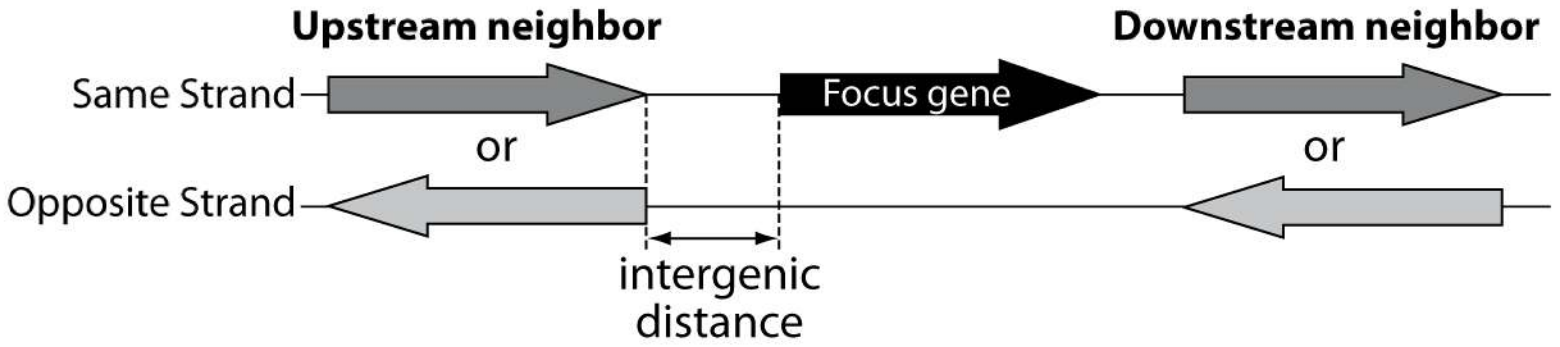
Dogan & Liu (2018) Nat Plants

Massively Parallel Reporter Assays

Akhtar et al. (2013) Cell

- enhancers / transcriptional activity
- chromatin state
- 3D position effects

Studying gene neighbors



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Gene sets:

- Differentially expressed
- Bound by a specific protein
- Present in a specific chromatin context
- Involved in a specific function
- . . .

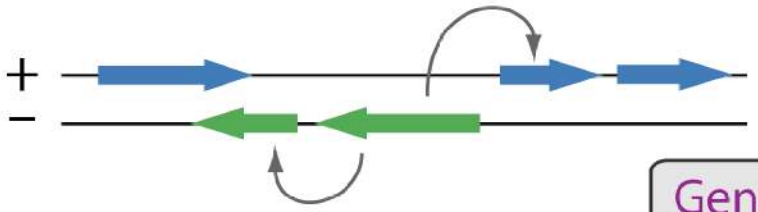
Specific neighborhood ?

- Orientation (incl. overlap) ?
- Specific proximity / distance ?

subgroups
Function / expression... of the neighbors ?

Generate hypothesis on how neighborhing genes interact

Package overview



Gene annotations

`getGeneNeighborhood()`

Gene neighbors annotations

Gene set(s)

DE analysis
TF binding
⋮

Neighbors orientation ?

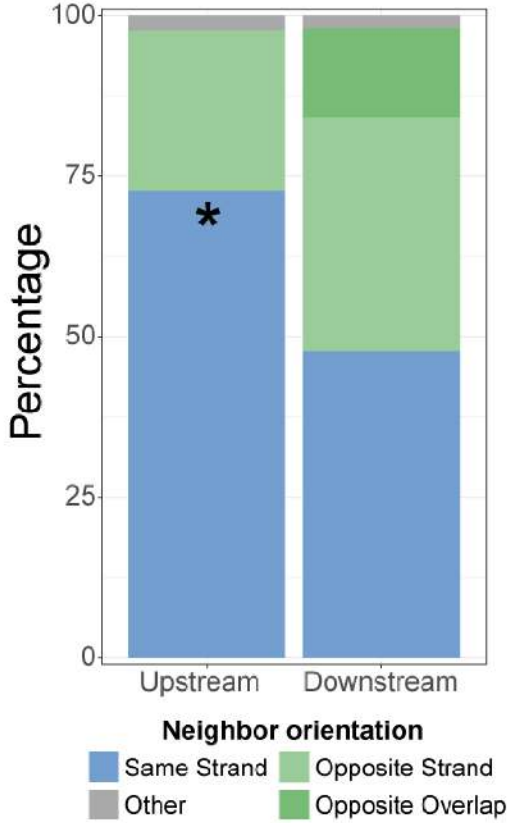
`analyzeNeighborsOrientation()`
`plotNeighborsOrientation()`

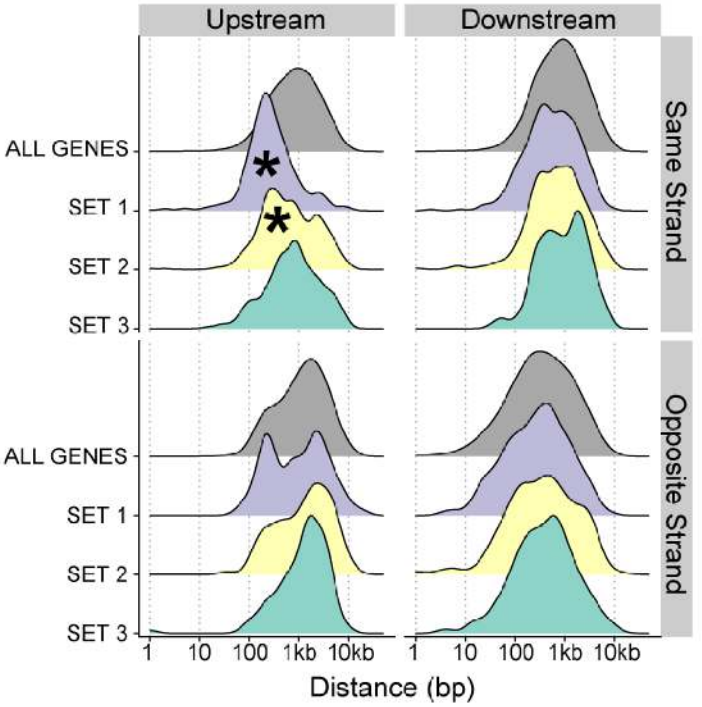
Neighbors distance ?

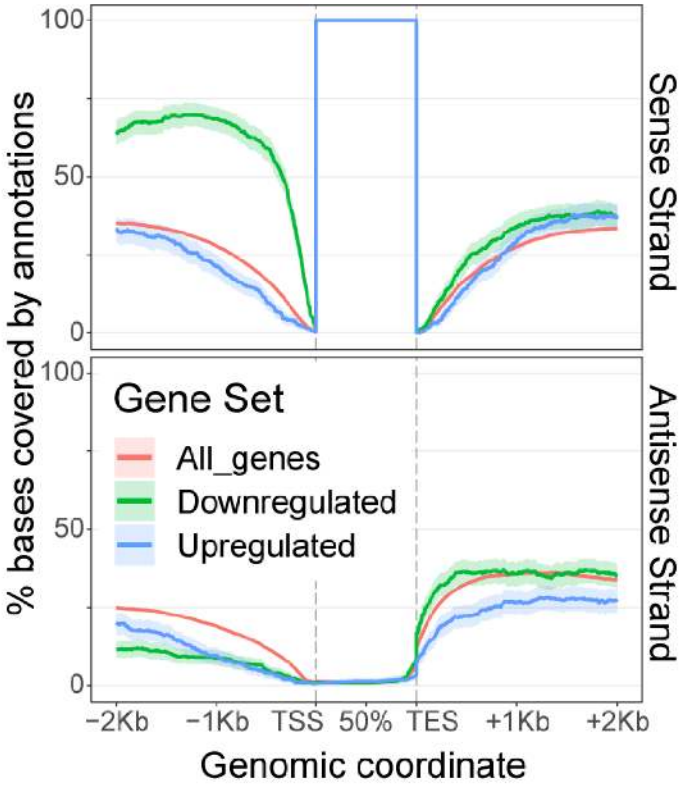
`analyzeNeighborsDistance()`
`plotDistanceDistrib()`

Metagene profiles

`annotationCoverageAroundFeatures()`
`getAvgProfileWithCI()`
`plotMetageneAnnotProfile()`







Defining Neighborhoods

Gene Neighborhood

Genes not overlapping with other genes:

UpstreamClass	Upstream	Focus gene	Downstream	DownstreamClass	Neighborclass
SameStrand	S		S	SameStrand	SS
SameStrand	S		O	OppositeStrand	SO
OppositeStrand	O		S	SameStrand	OS
OppositeStrand	O		O	OppositeStrand	OO

Genes overlapping with a single gene:

UPPER CASE: overlapping gene is on the same strand
lower case: overlapping gene is on the oppositre strand

UpstreamClass	Upstream	Focus gene	Downstream	DownstreamClass	Neighborclass	
SameOverlap	U		S or O	SameStrand or OppositeStrand	US / UO	U: upstream
OppositeOverlap	u		S or O	SameStrand or OppositeStrand	uS / uO	
SameStrand or OppositeStrand	S or O		D	SameOverlap	SD / OD	D: downstream
SameStrand or OppositeStrand	S or O		d	OppositeOverlap	Sd / Od	
SameStrand or OppositeStrand	S or O	I	S or O	SameStrand or OppositeStrand	I	I: identical (same start + same end)
SameStrand or OppositeStrand	S or O	i	S or O	SameStrand or OppositeStrand	i	
SameStrand or OppositeStrand	S or O	H	S or O	SameStrand or OppositeStrand	H	H: host
SameStrand or OppositeStrand	S or O	h	S or O	SameStrand or OppositeStrand	h	
other		G		other	G	G: guest
other		g		other	g	

Genes overlapping with a >1 gene:

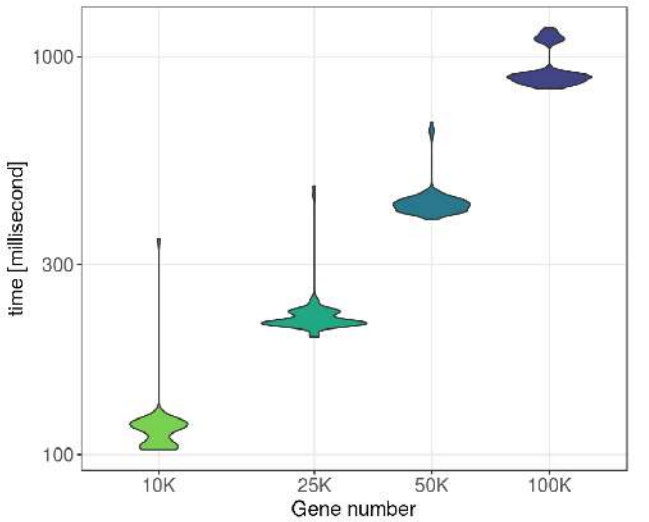
UpstreamClass	Upstream	Focus gene	Downstream	DownstreamClass	Neighborclass
other	or any other situation with multiple overlaps			other	c

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

TxDb objects
ensembldb
GenomicFeatures package

Filter annotations
(expressed genes,
gene types, ...)

`getGeneNeighborhood()`

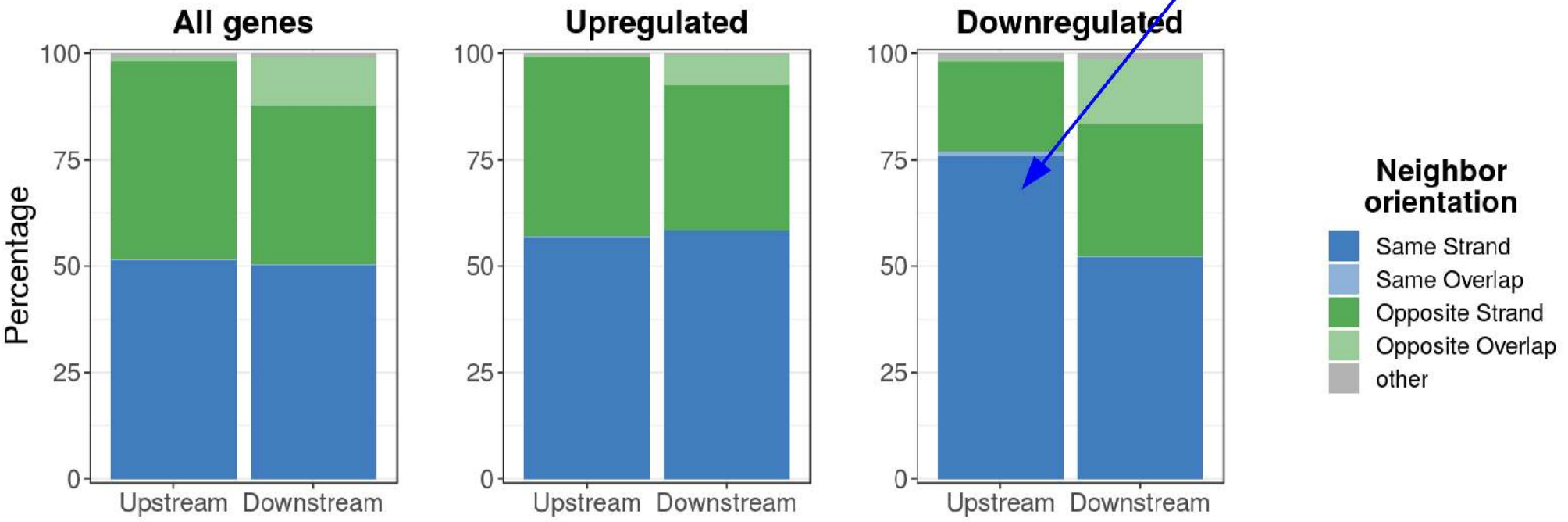


Analyzing neighbors orientations

`analyzeNeighborsOrientation()`

Side	Orientation	n	Percentage	n Universe	Percentage in Universe	p.value
Upstream	SameStrand	449	75.84	17068	51.23	**2,00E-35**
Upstream	SameOverlap	6	1.01	96	0.29	7,00E-03
Upstream	OppositeStrand	125	21.11	15267	45.82	1,00E+00
Upstream	OppositeOverlap	4	0.68	344	1.03	9,00E-01
Downstream	SameStrand	308	52.03	17099	51.32	4,00E-01
Downstream	SameOverlap	1	0.17	111	0.33	9,00E-01
Downstream	OppositeStrand	184	31.08	12708	38.14	1,00E+00
Downstream	OppositeOverlap	91	15.37	2857	8.57	**4,00E-08**

`plotNeighborsOrientation()`



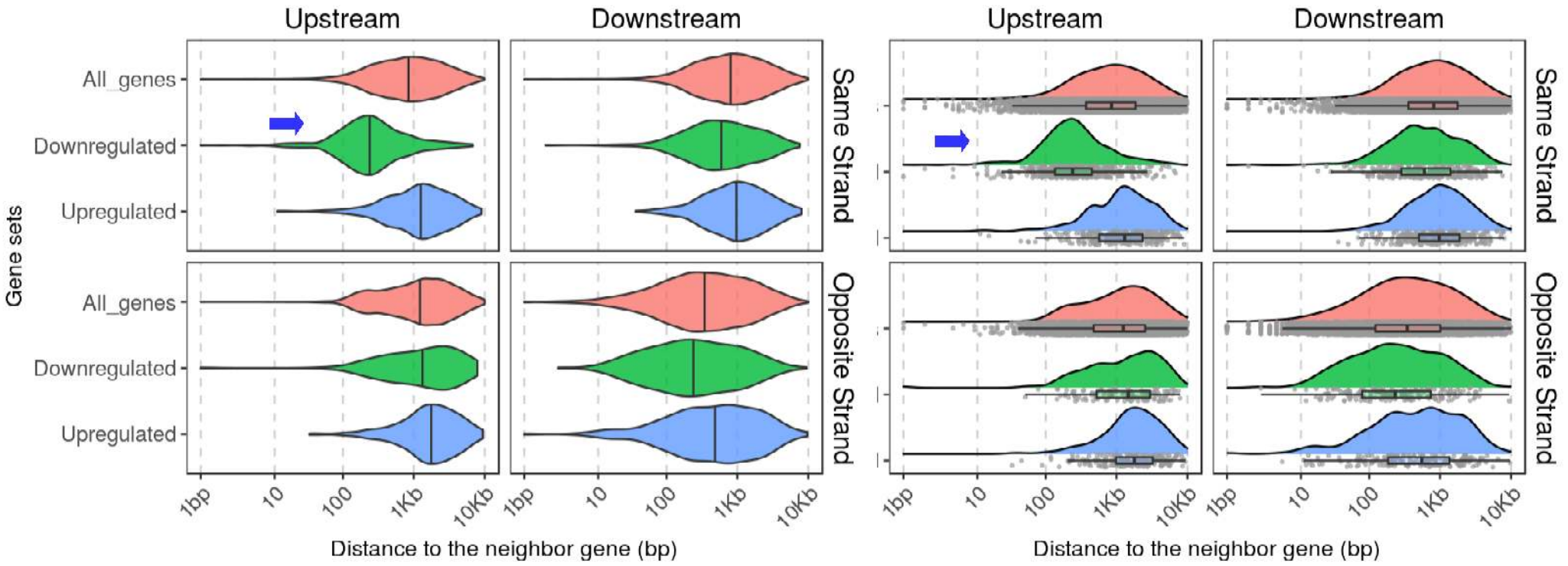
Analyzing neighbors distance

`analyzeNeighborsDistance()`

- Extract distances
- Descriptive statistics (mean, median, SD, 95 % CI, ...)
- Tests (KS, Mann-Whitney, permutation test of independence)

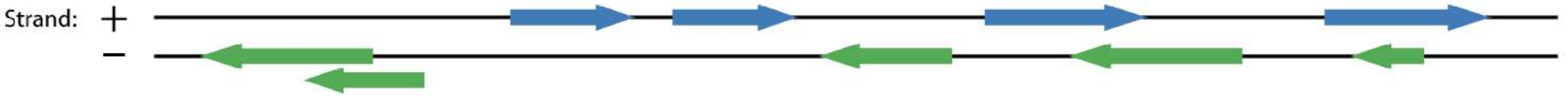
coin R package

`plotDistanceDistrib()`

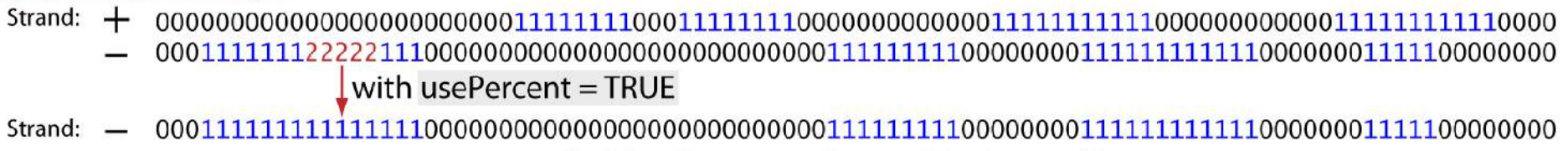


Metagene profiles

Gene annotations:

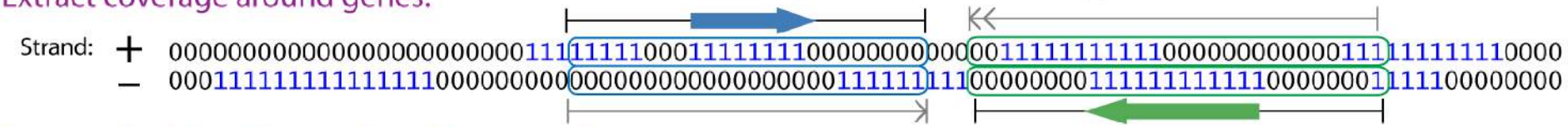


Annotation coverage:



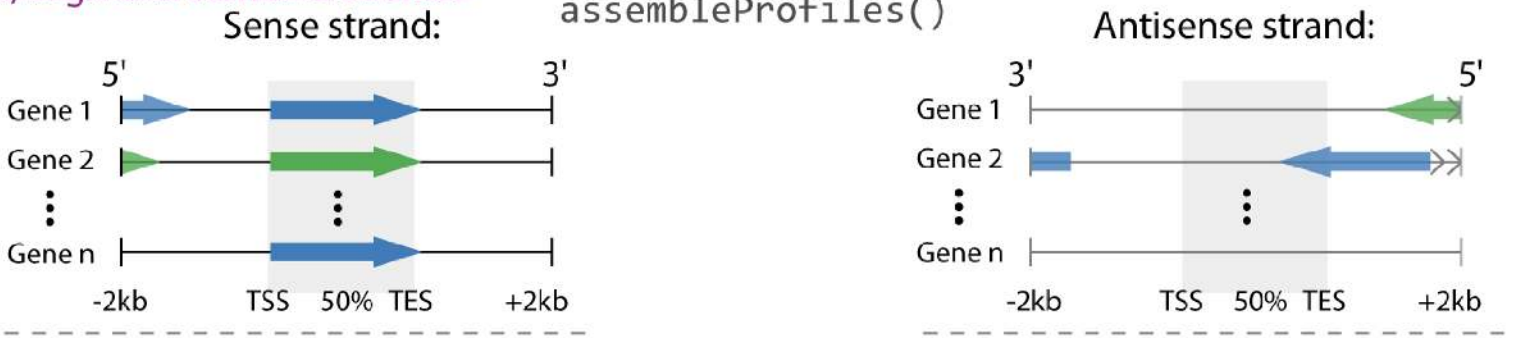
Extract coverage around genes:

annotationCoverageAroundFeatures()



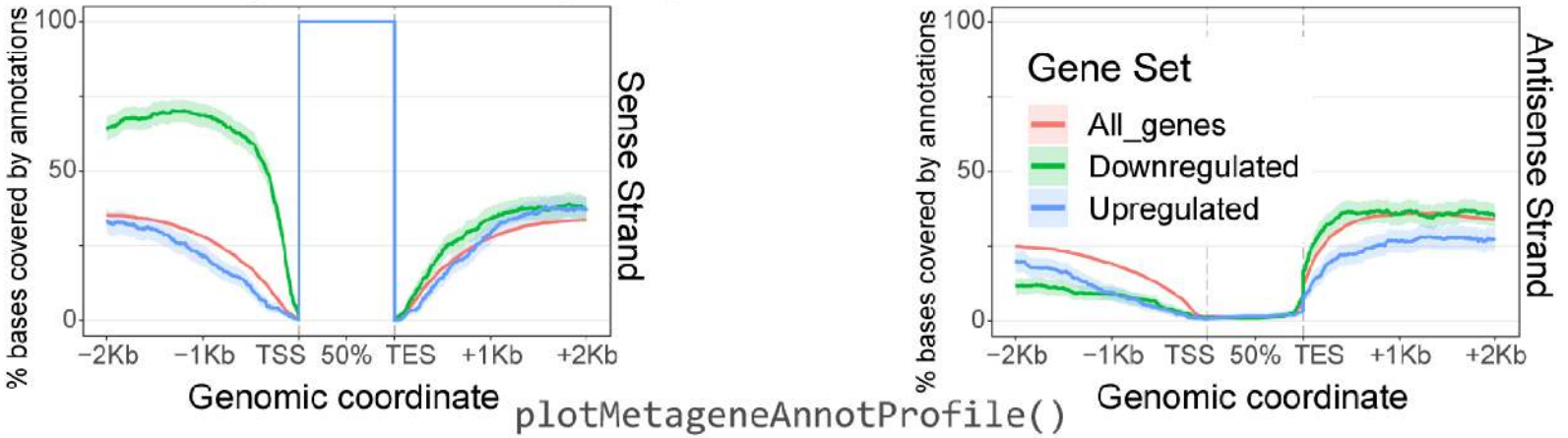
Bin gene bodies, align and combine strands:

assembleProfiles()



Average profiles for different gene sets:

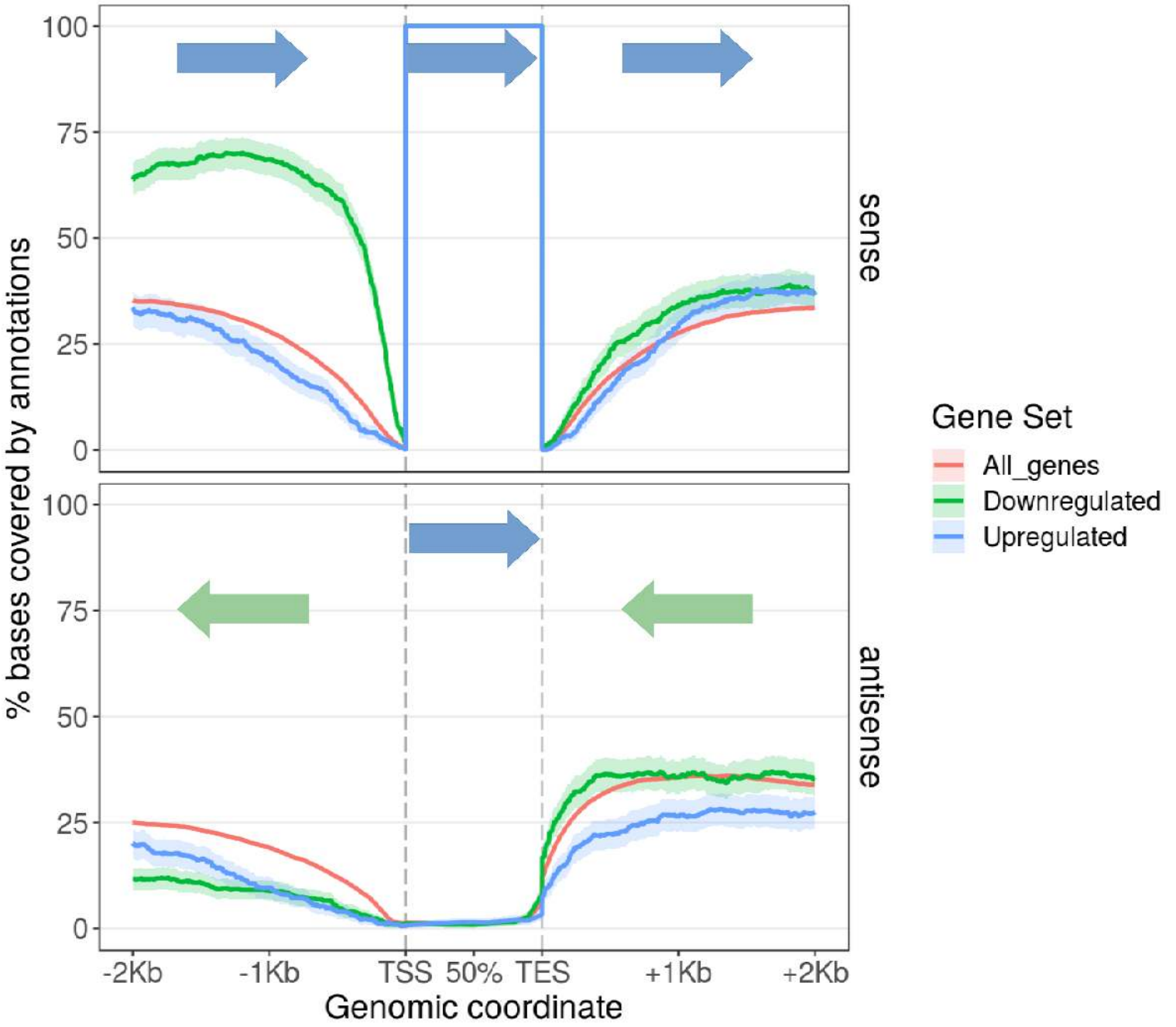
getAvgProfileWithCI()



plotMetageneAnnotProfile()

Metagene profiles

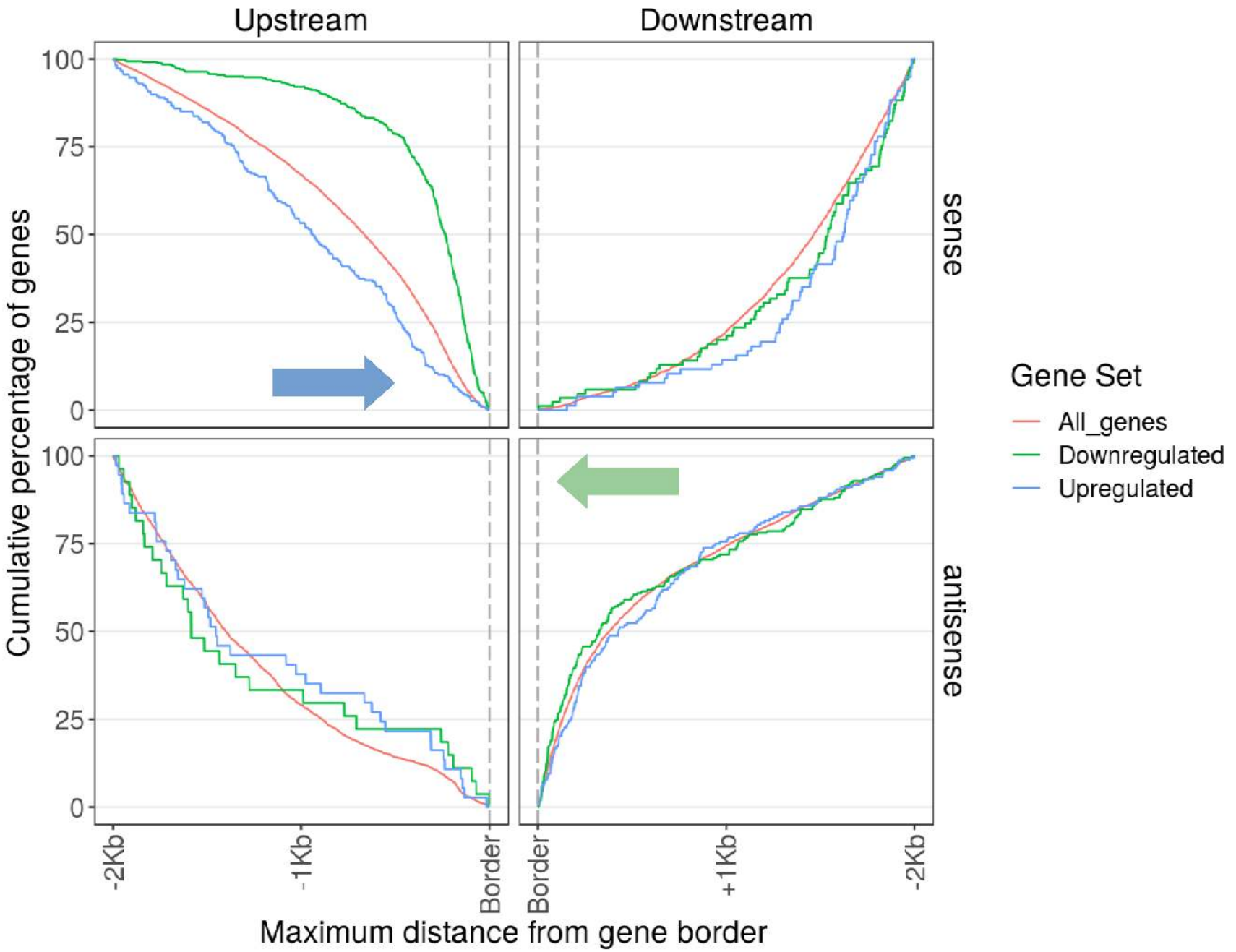
```
plotMetageneAnnotProfile()
```



Distances to single point features

`plotCumulPercentProfile()`

% of genes with a TES at less than X bp ?



Conclusion / future development

build passing build passing

github.com/pgpmartin/GeneNeighborhood

Neighbors orientation

- Facilitate the analysis of multiple gene sets

Neighbors distance

- Split analysis into 3 tasks :
  - extract distances
  - descriptive statistics
  - inference

Metagene profiles

- Add a wrapper to simplify the pipeline

+ Case studies in vignette → Submit to CRAN or BioC

Future development

- > 1 overlap
- Deal with a quantitative variable defining gene sets

Acknowledgements

INDIANA UNIVERSITY

Scott D Michaels

NIH

Xuhong Yu

Mathilde Gorieu

Funding:

Co-funded by the European Union

AgreenSkills plus

Pathways for inventive researchers

Community help / forums:

R

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Thank you for your attention !