Methylation of Transposable Elements and Gene Expression in *Arabidopsis Thaliana*

Katia Antonenko[1,2,3], Marie Dogo[1,2,3], Jérémy Cohen[1,2,3], Sylvain Caillaud[1,2,3], Louna De Oliveira[4], Aurélien Petit[4], Vincent Colot[4], Chloé-Agathe Azencott[1,2,3], Pierre Baduel[4]

1. CBIO-Centre for Computational Biology, Mines Paris, PSL Research University 2. Institut Curie, PSL Research University 3. U1331 Institut National de la Santé et de la Recherche Médicale - INSERM 4. Institut de Biologie de l'École Normale Supérieure (IBENS) ENS, CNRS UMR8197, Inserm U1024, Paris

27 November 2025

Contents

- Background: transposable elements and methylation
- Motivation: to explain GWAS findings
- Model: to understand methylation spreading
- Conclusions

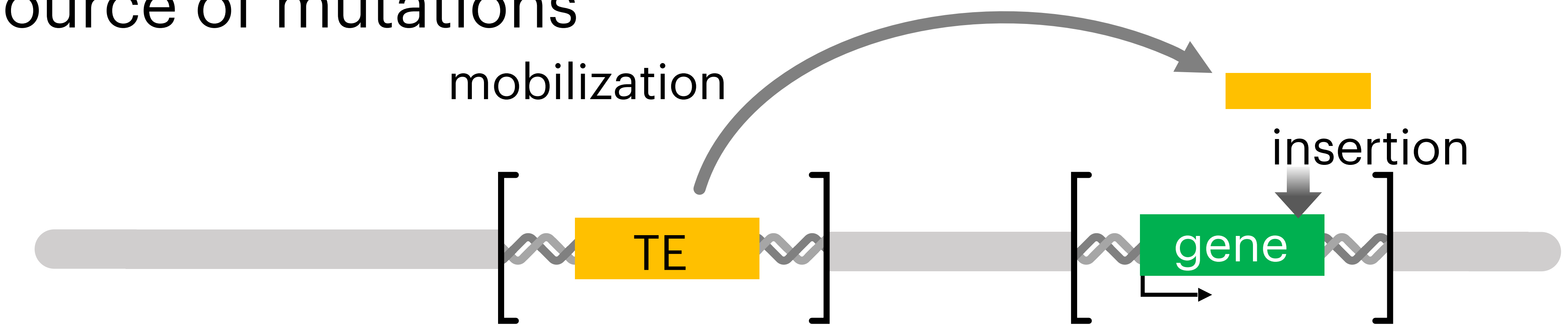
Transposable Elements

Transposable Elements

- Transposable Elements (TEs, “jumping genes”) are an important source of mutations

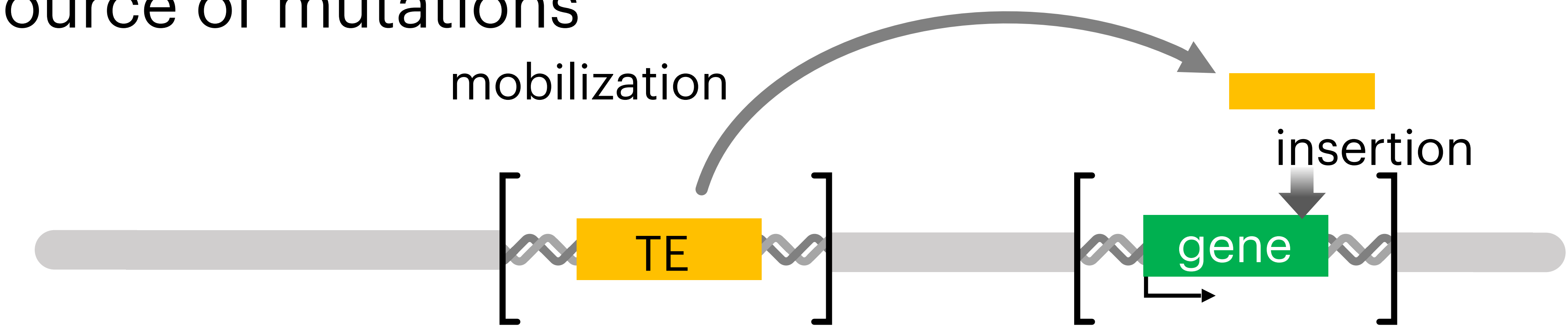
- TEs transpose by cut-and-paste or copy-and-paste mechanisms

- BUT: most TEs are degraded and do not transpose

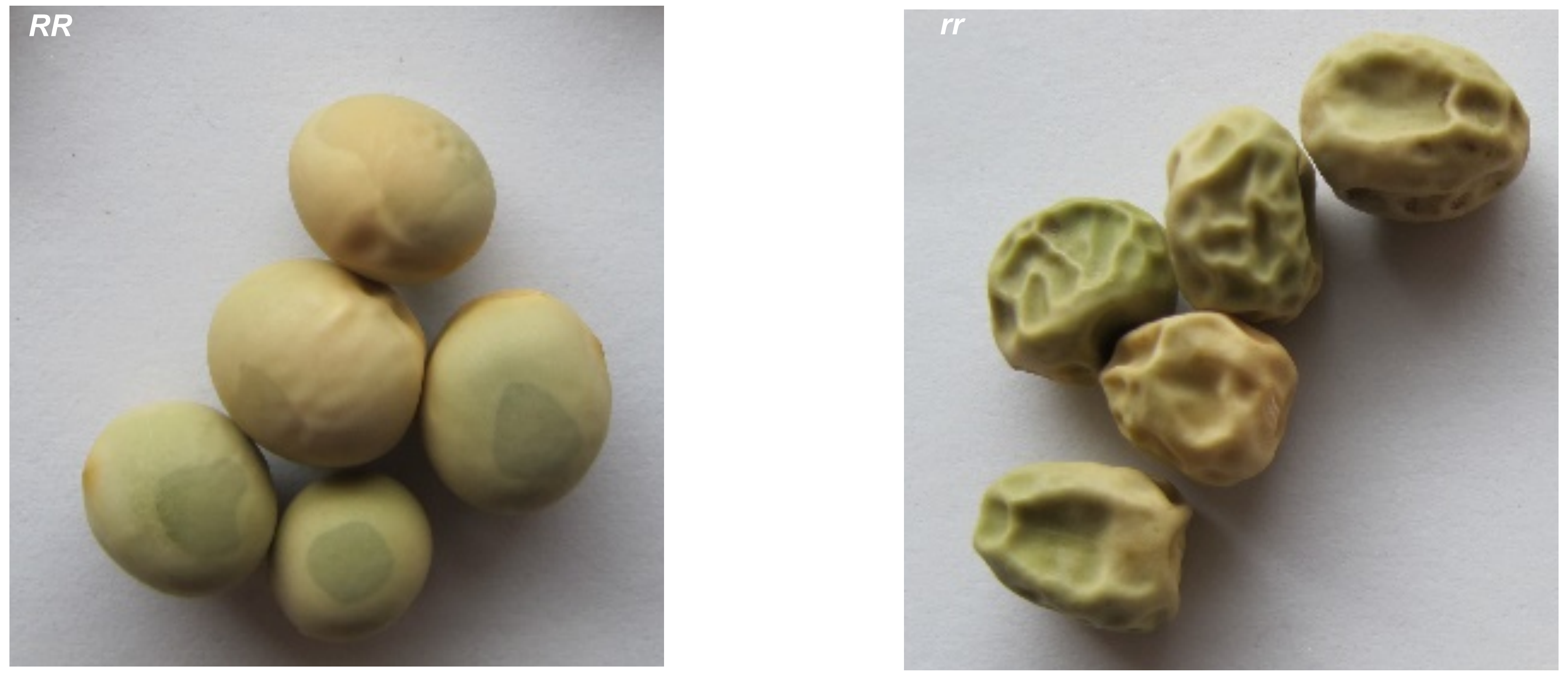


Transposable Elements

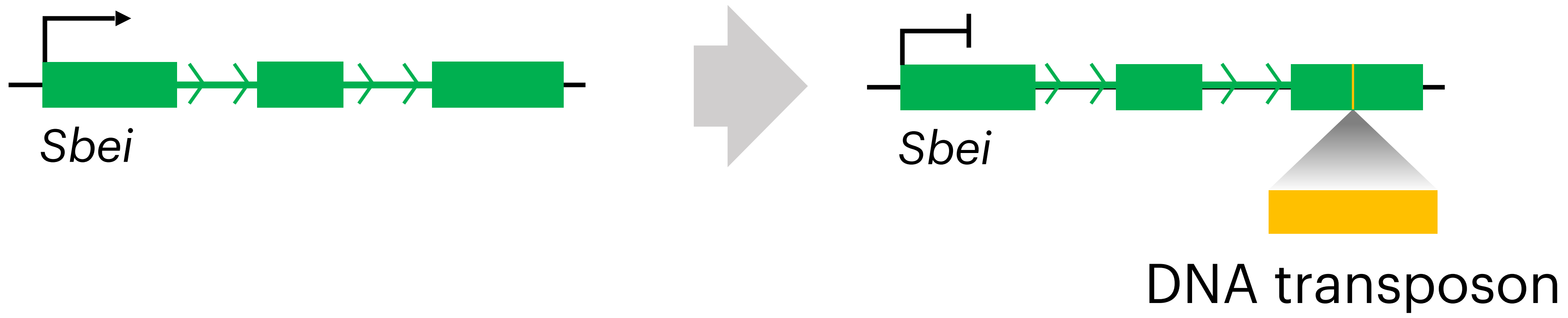
- Transposable Elements (TEs, “jumping genes”) are an important source of mutations
- TEs transpose by cut-and-paste or copy-and-paste mechanisms
- BUT: most TEs are degraded and do not transpose



Mutations may be **deleterious**...

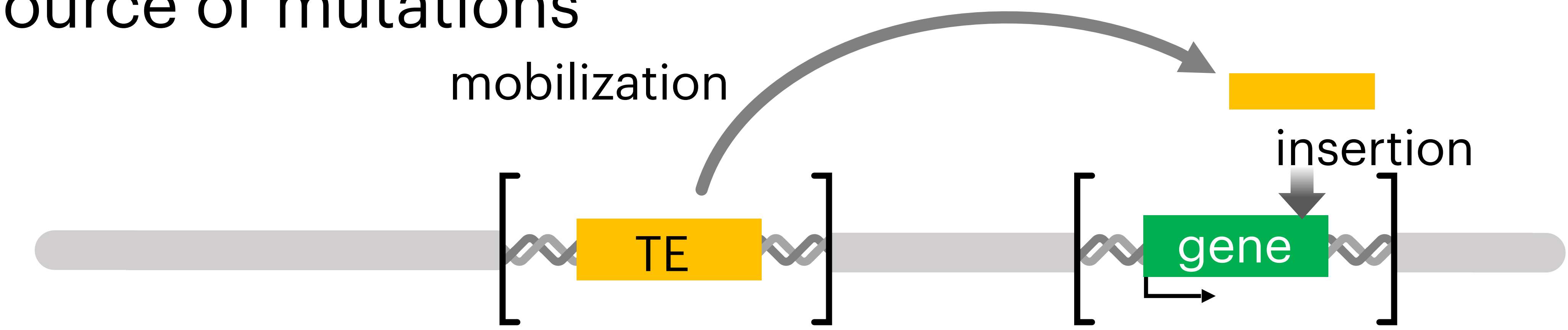


Bhattacharyya *et al. Cell* 1990

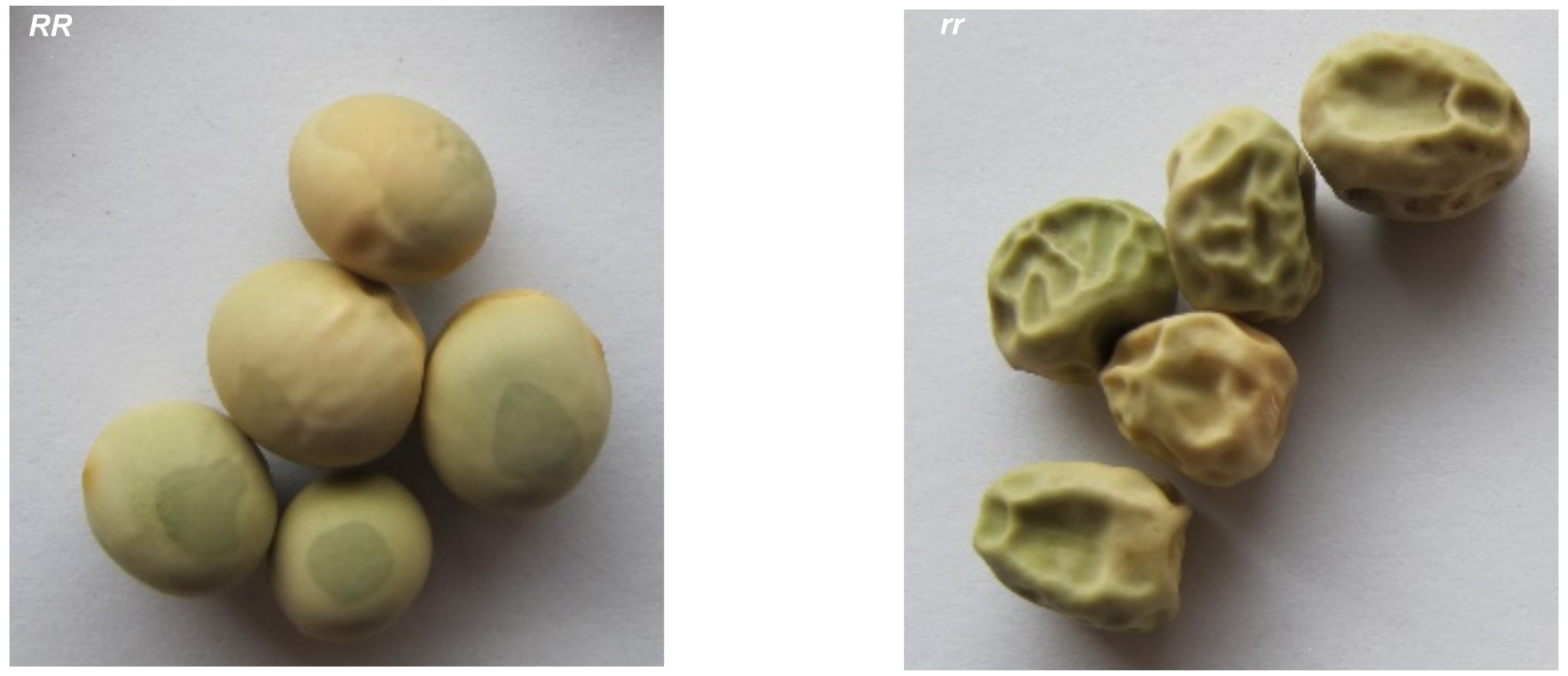


Transposable Elements

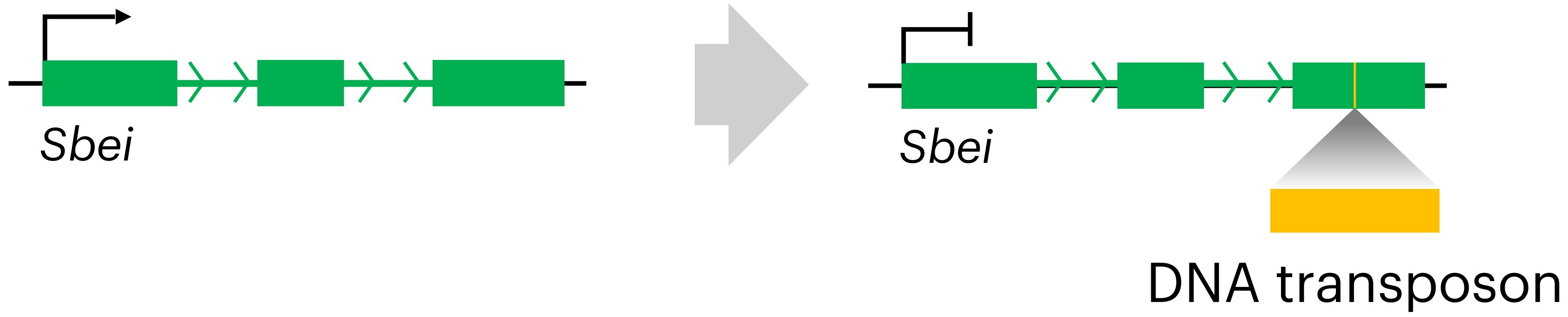
- Transposable Elements (TEs, “jumping genes”) are an important source of mutations
- TEs transpose by cut-and-paste or copy-and-paste mechanisms
- BUT: most TEs are degraded and do not transpose



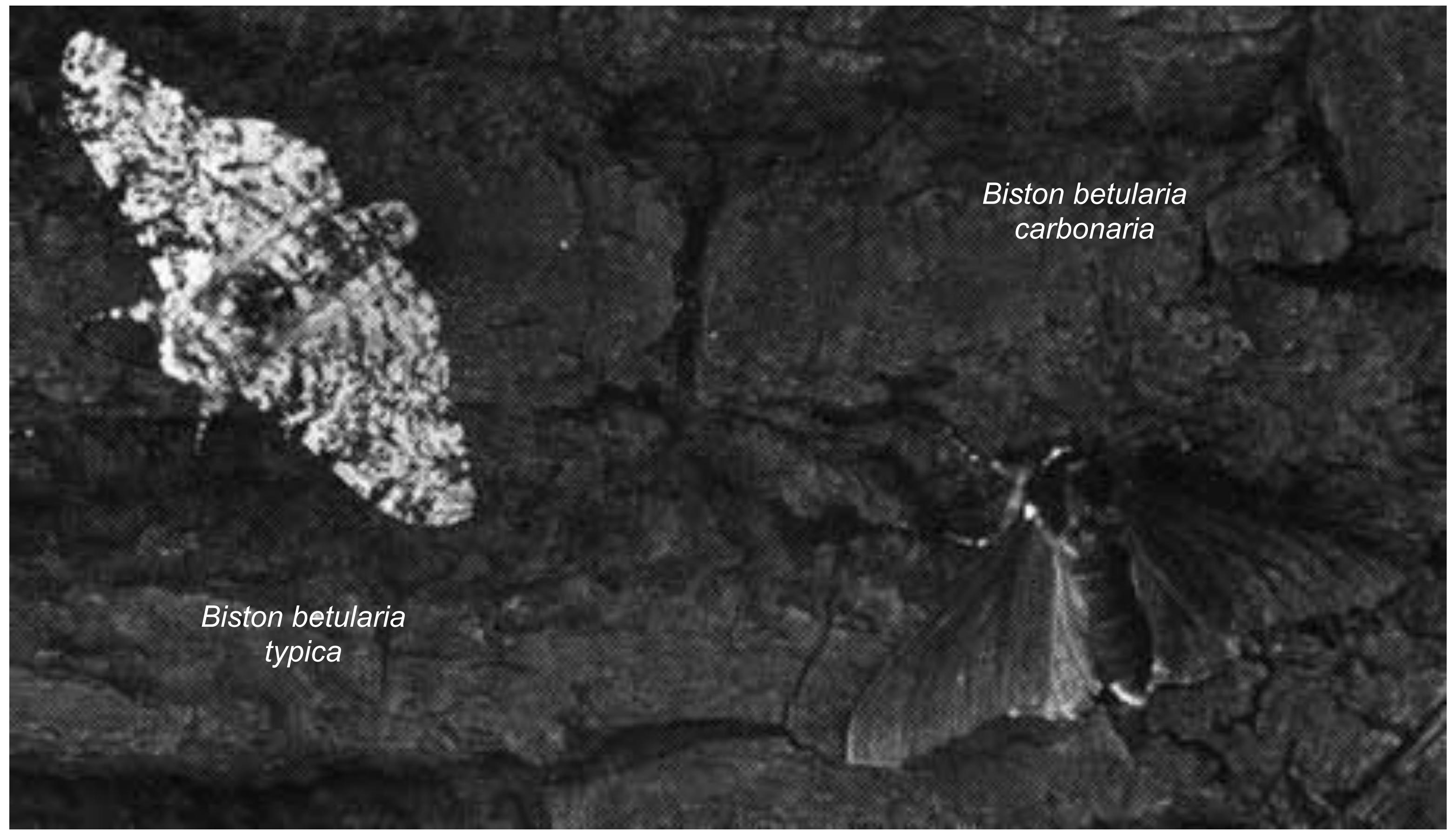
Mutations may be **deleterious**...



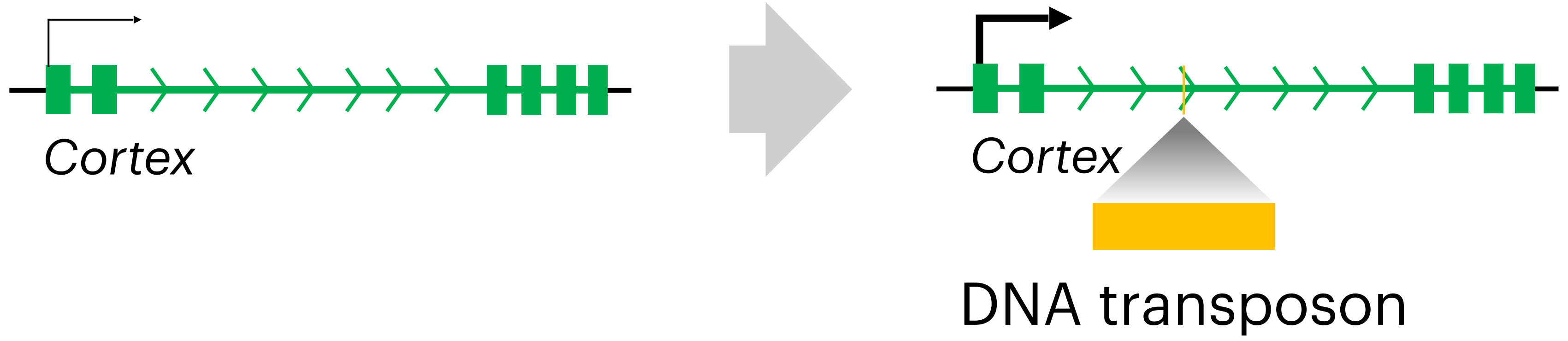
Bhattacharyya *et al. Cell* 1990



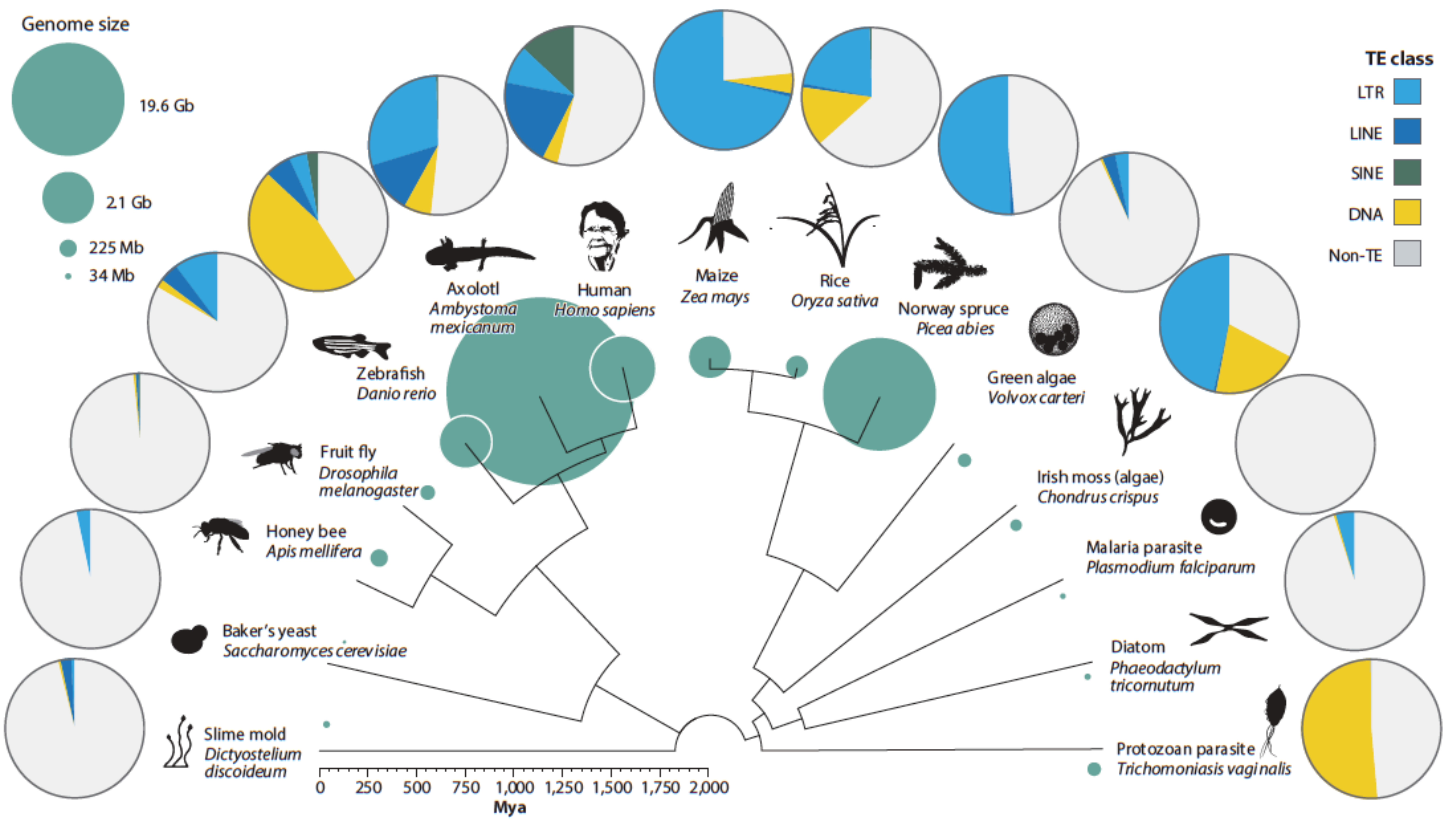
...yet sometimes **adaptive**



Kettelwell. *Heredity* 1956; van’t Hof *et al. Nature* 2016

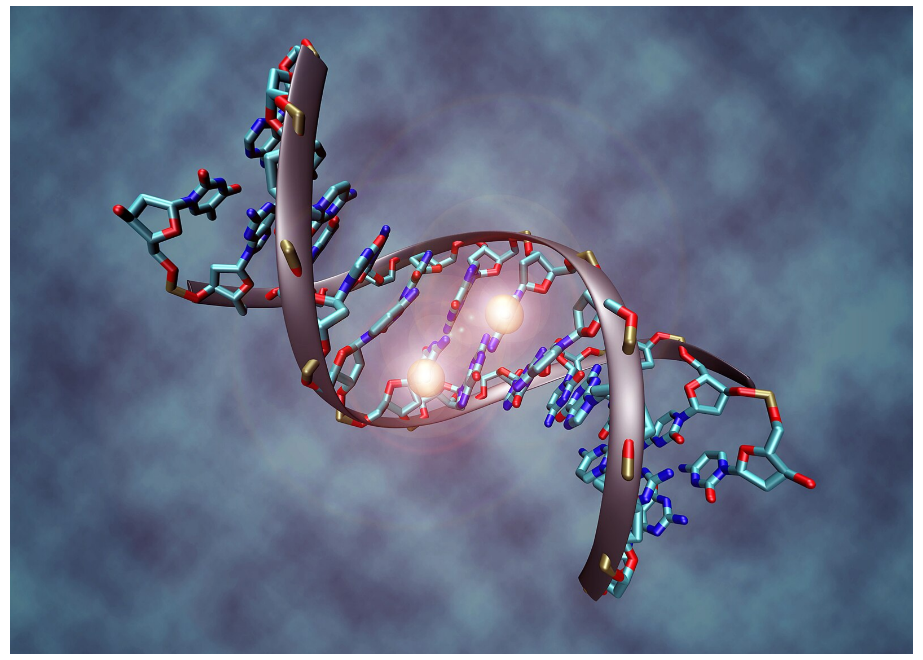


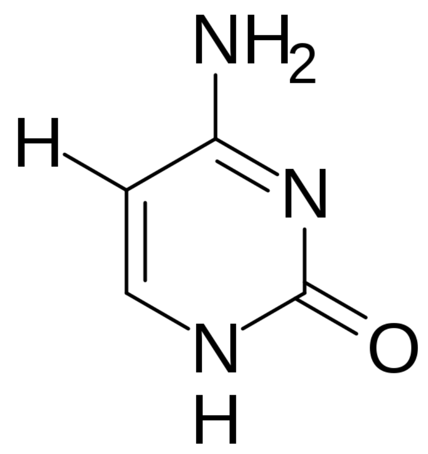
Transposable Elements



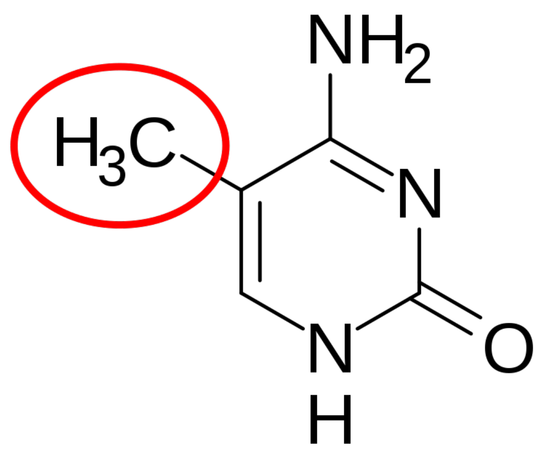
Wells & Feschotte, *Annual Review of Genetics 2020*

Epigenetic Regulation of Transposable Elements





cytosine



methylated cytosine

DNA methylation:

- is an essential regulatory mechanism of TEs activity
- targets CG / CHG / CHH in plants
  [H = anything besides G]
- affects TE / gene expression (silencing)
- may spread to flanking regions
- example:

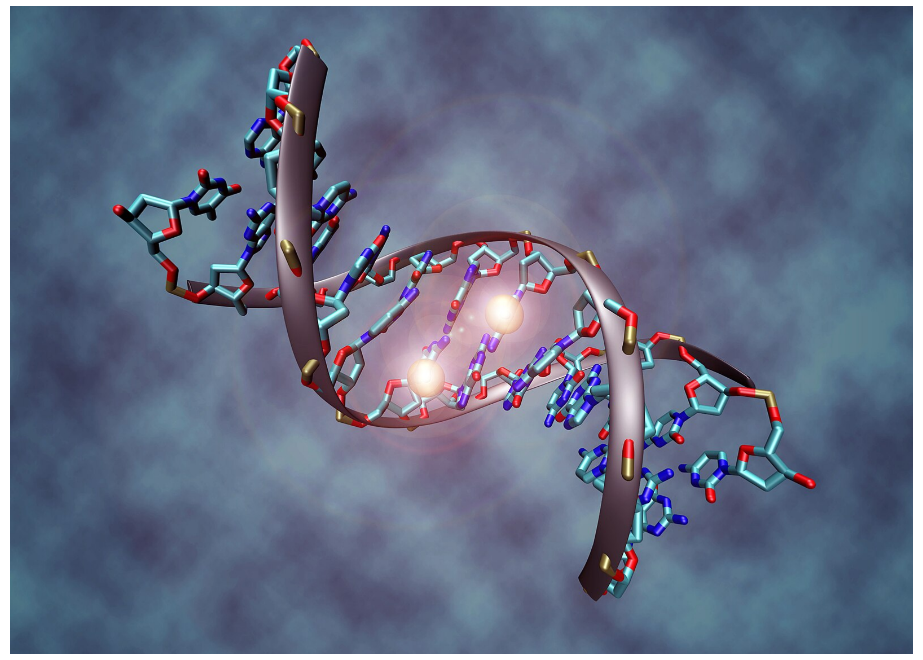
methylated promoter ⟹ no RNA ⟹
⟹ no protein ⟹ no function

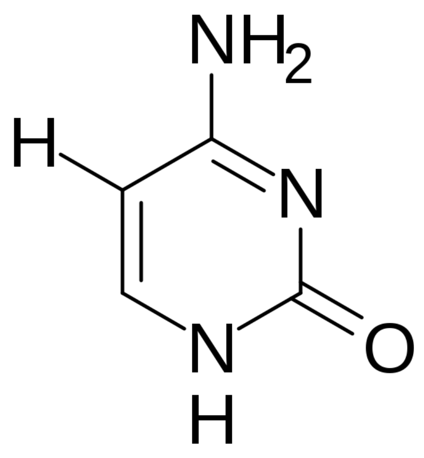
Epigenetic Regulation of Transposable Elements

DNA methylation:

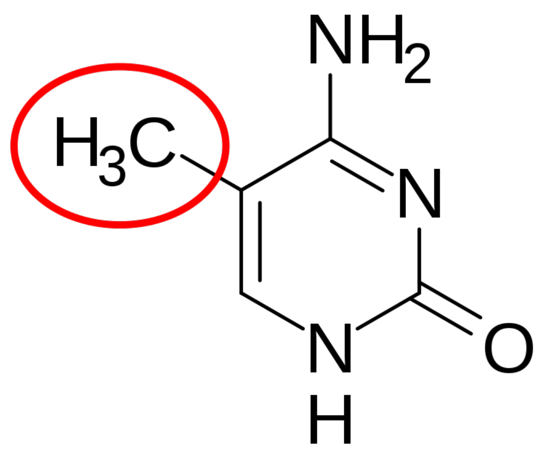
- is an essential regulatory mechanism of TEs activity
- targets CG / CHG / CHH in plants
  [H = anything besides G]
- affects TE / gene expression (silencing)
- may spread to flanking regions
- example:

  methylated promoter ⟹ no RNA ⟹
  ⟹ no protein ⟹ no function



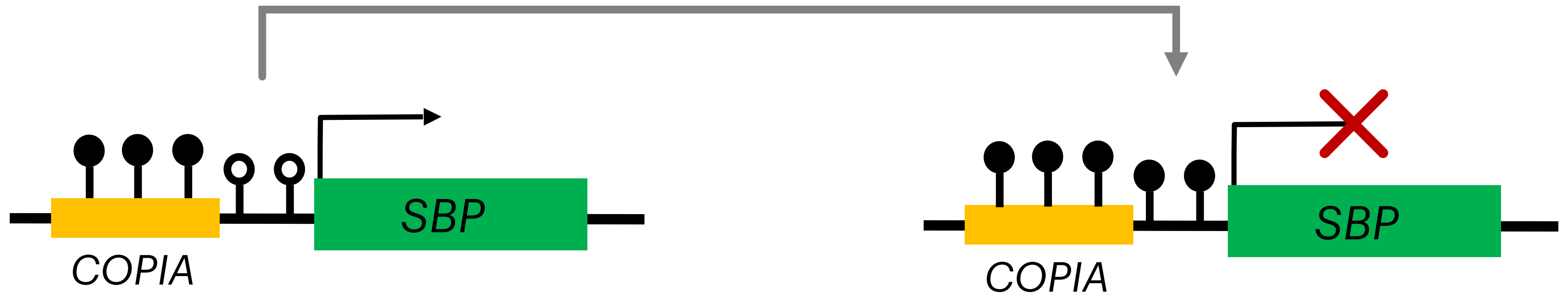


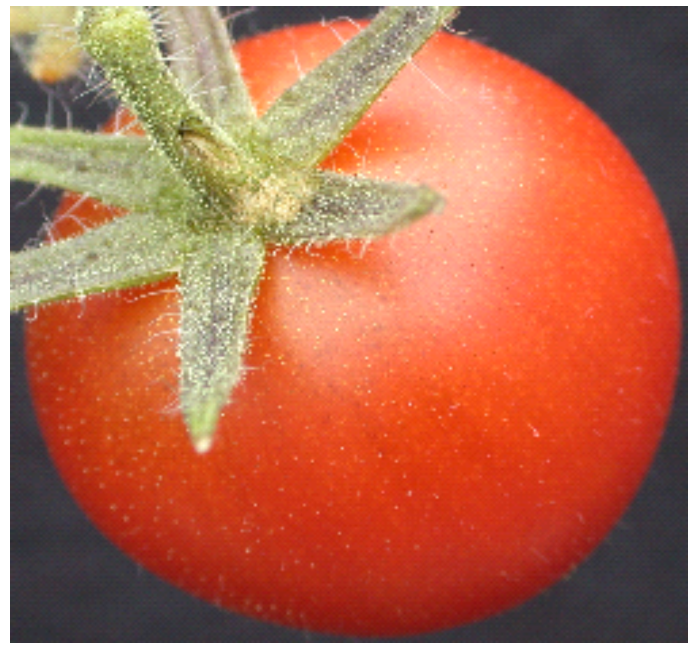
cytosine

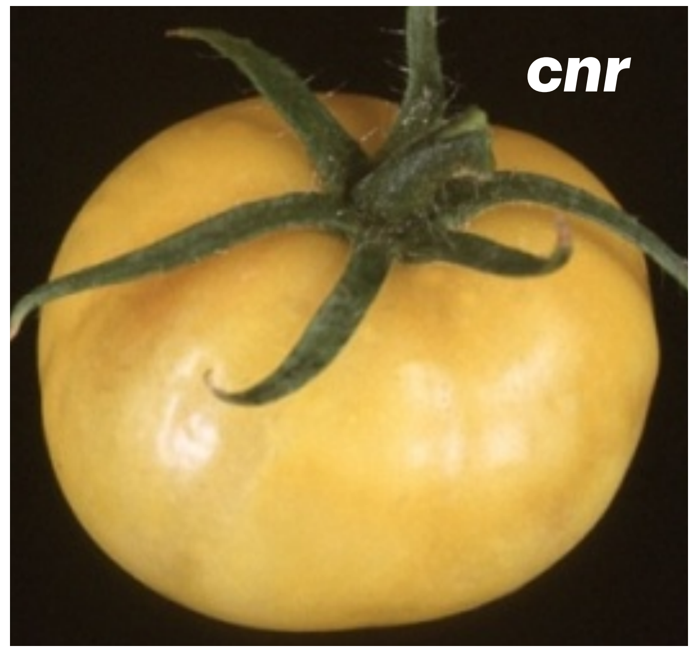


methylated cytosine

spontaneous epimutation





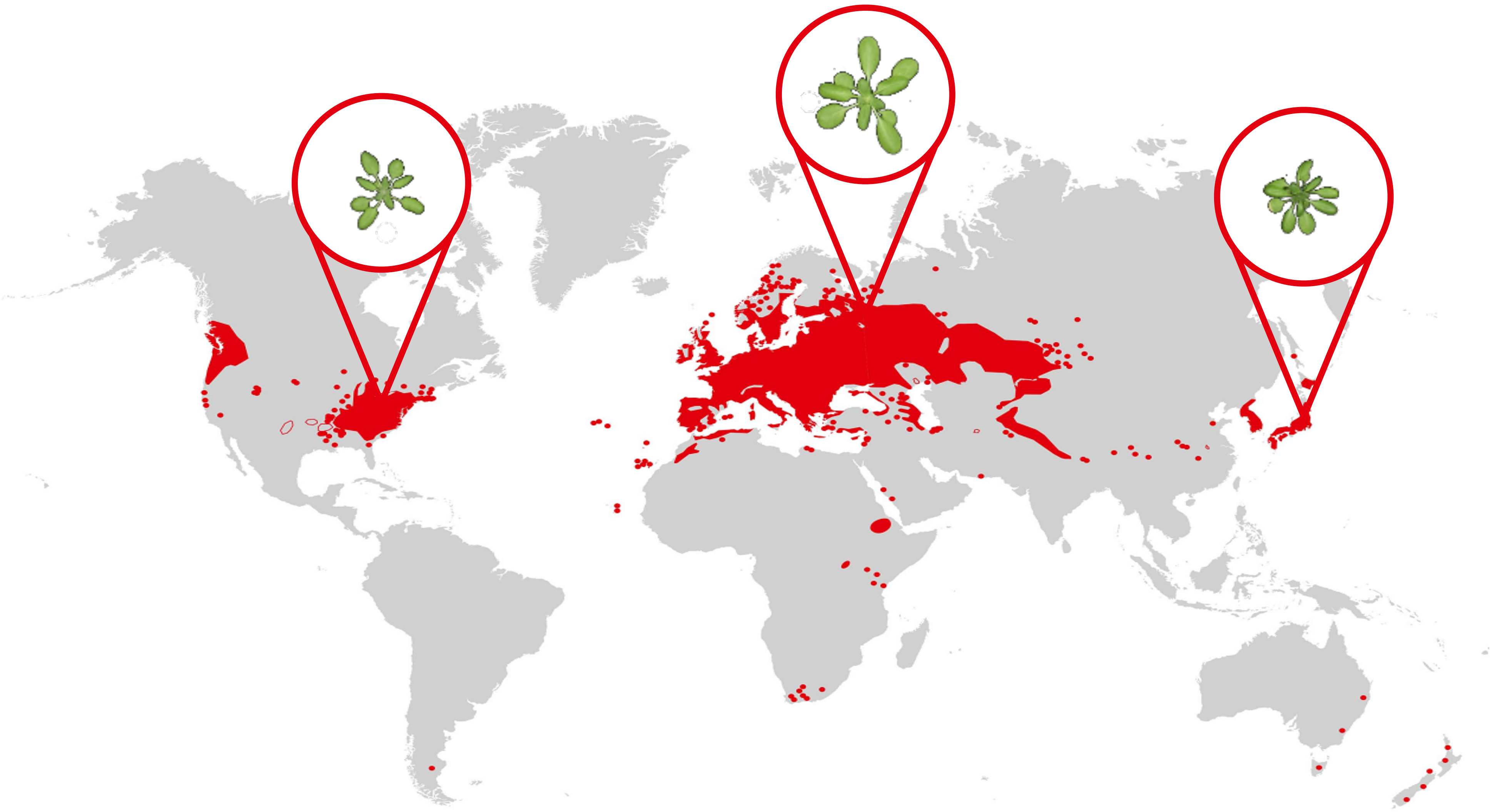


Manning et al., *Nat Genet 2006*

⟹ perfect Mendelian segregation though no DNA changes observed

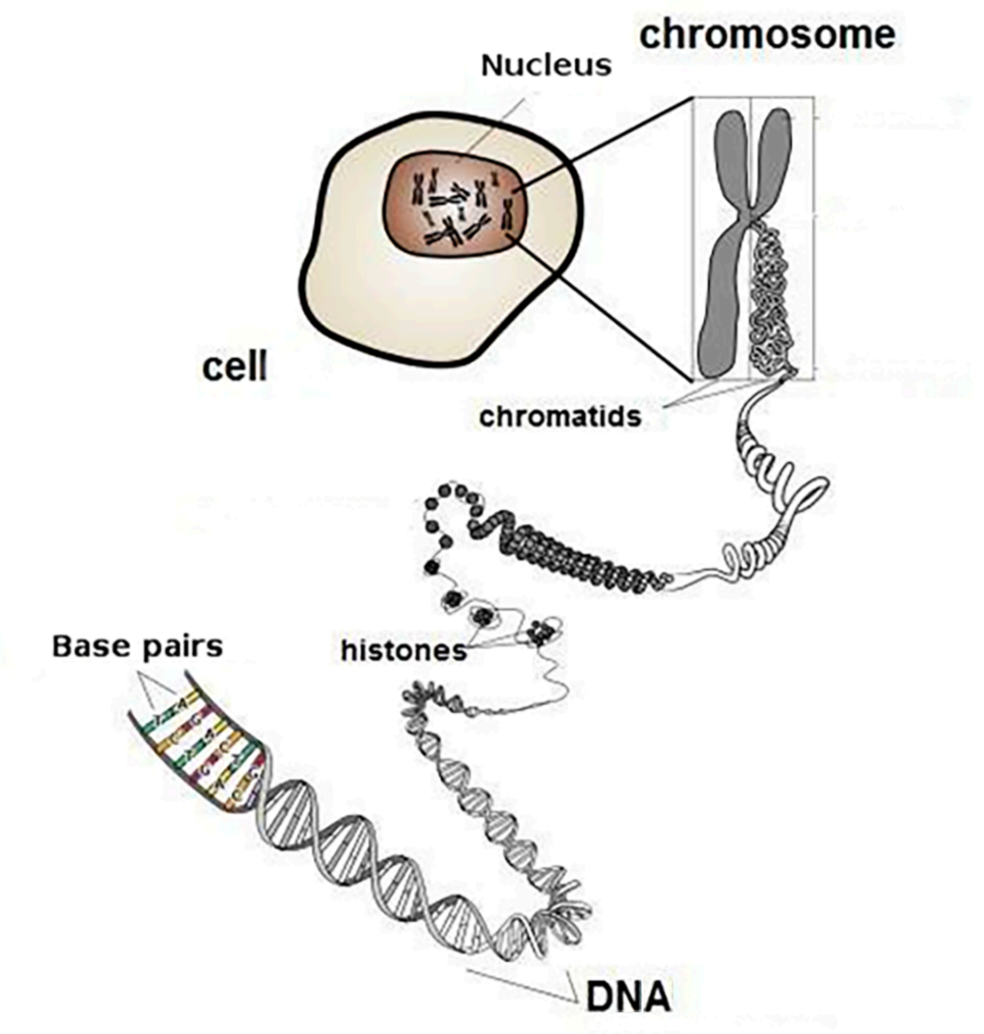
Our data: *Arabidopsis Thaliana*

- 89 strains from throughout the world, **sequenced with ultra-long reads (Nanopore)**
- **TE annotation + Full methylation profiles** (for all contexts CG, CHG, CHH)
- Gene expression data

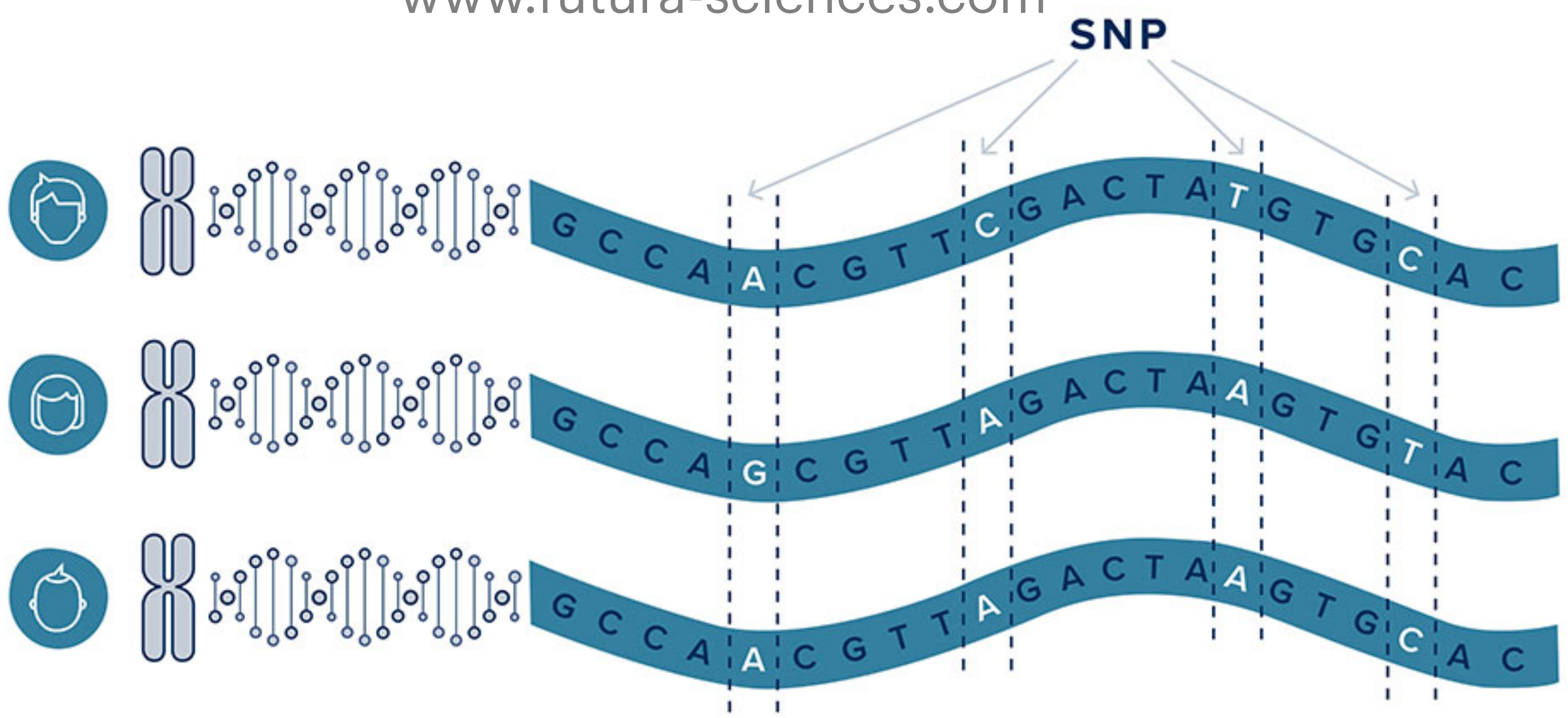


Kawakatsu *et al. Cell 2016*, Alonso-Blanco *et al. Cell 2016*, Quadrana *et al. eLife 2016*

From genotype to phenotype



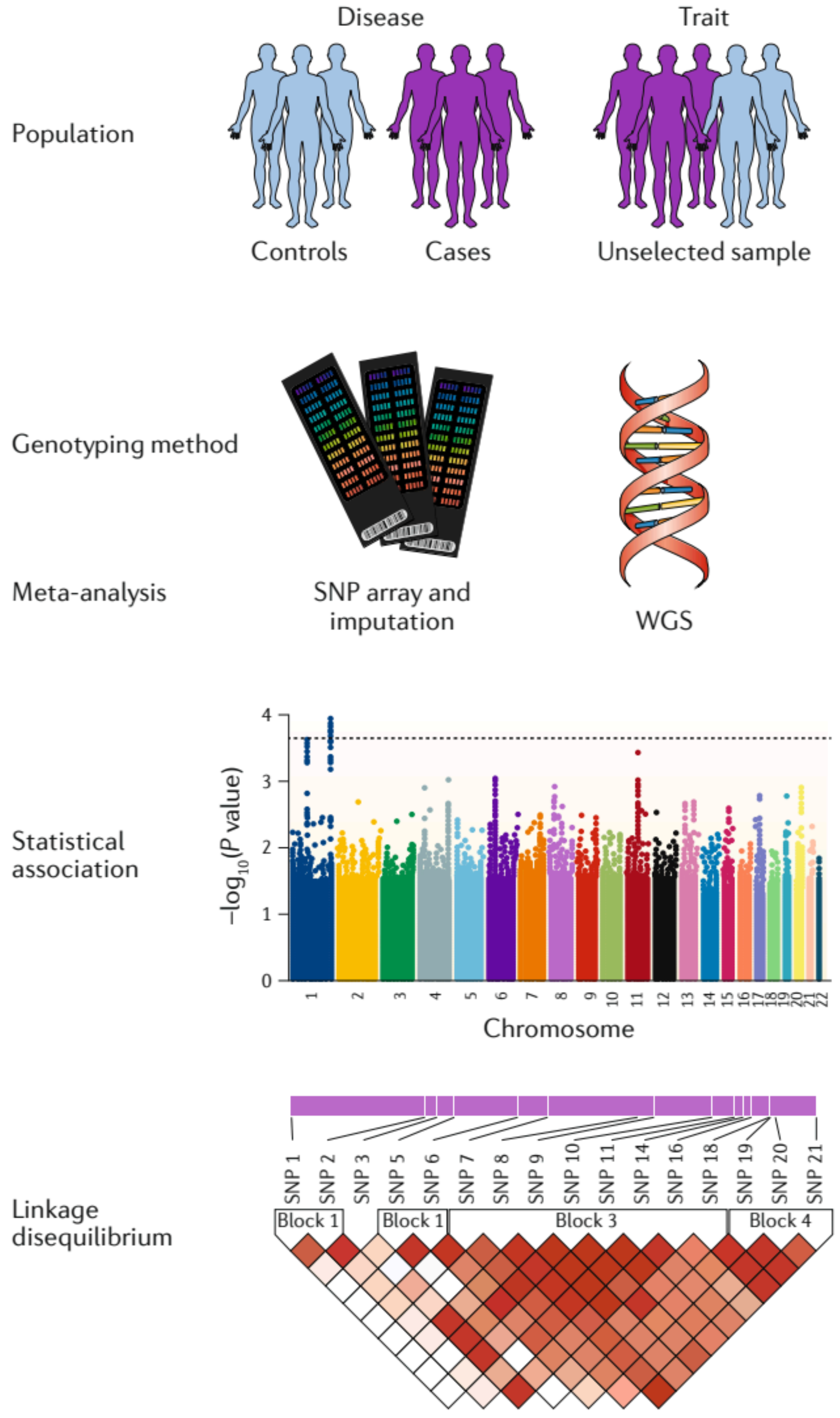
www.futura-sciences.com


SNP
G C C A A C G T T C G A C T A T G T G C A C
G C C A G C G T T A G A C T A A G T G T A C
G C C A A C G T T A G A C T A A G T G C A C

Scientific DX GmbH, 2020

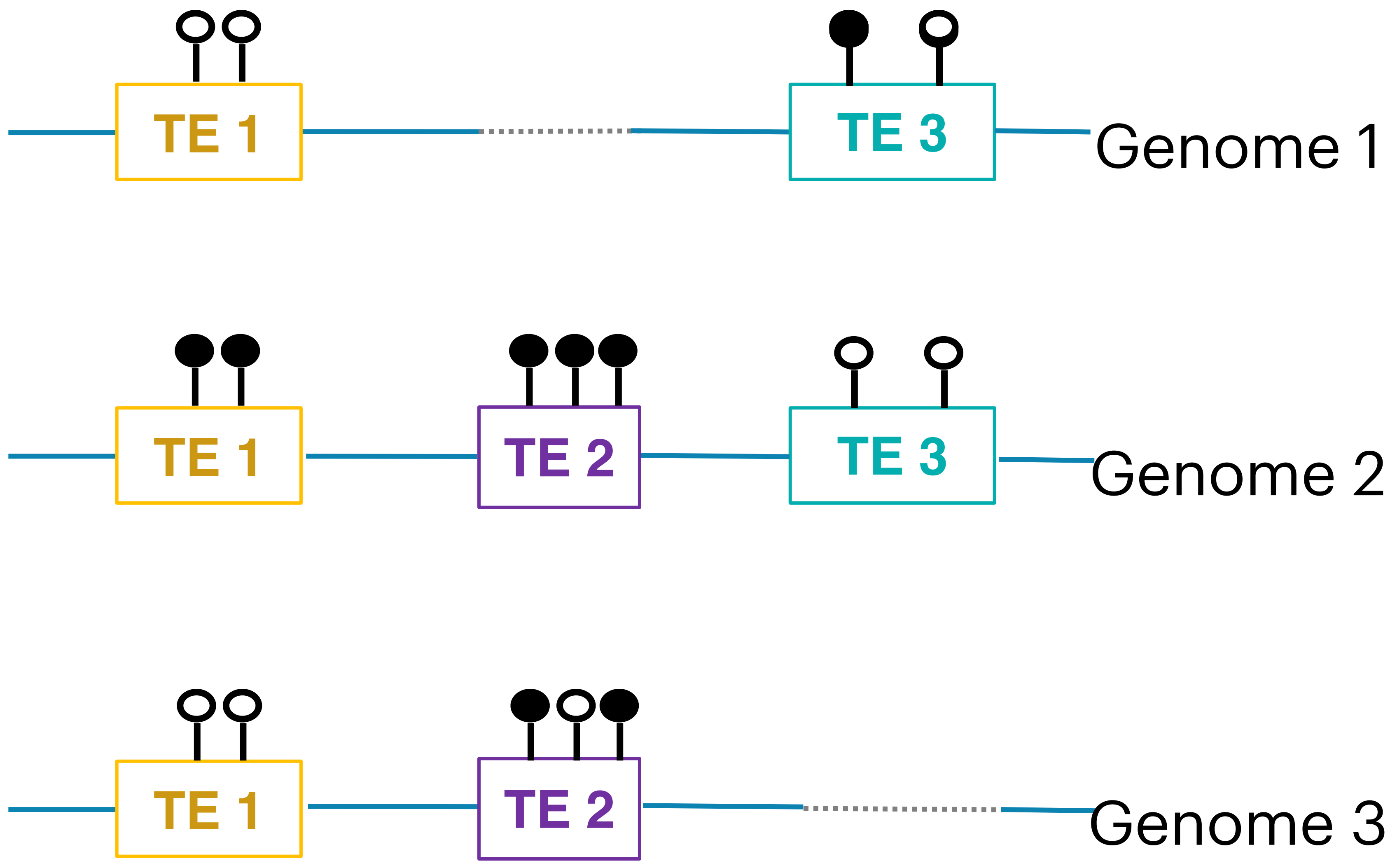
Genome-Wide Association Study

⟹



Tam et al., *Nature Reviews Genetics 2019*

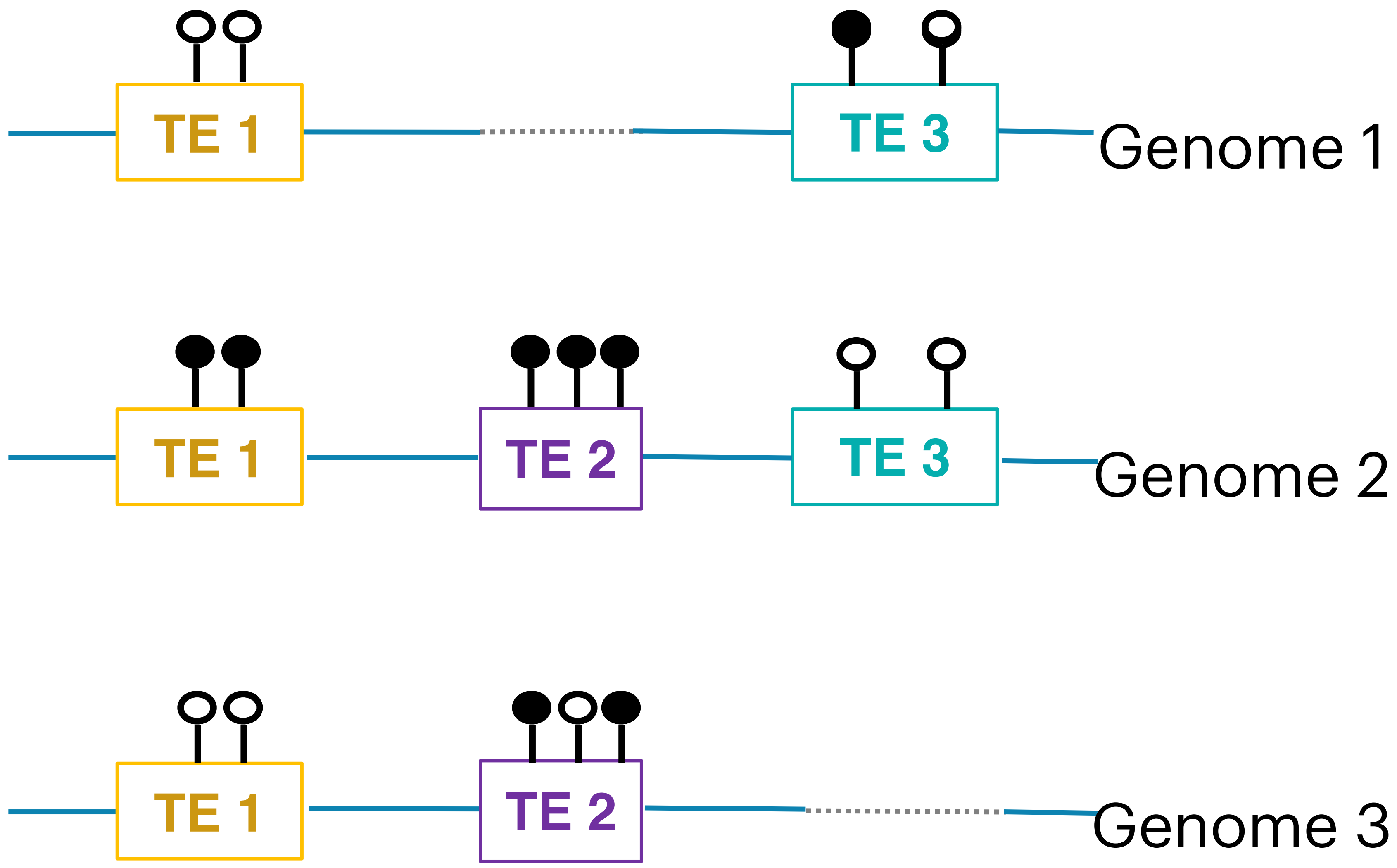
From epi-genotype to phenotype



Genome-Wide Association Study ⟹

	Gene A	Gene B	Gene C
Genome 1			
Genome 2			
Genome 3			

From epi-genotype to phenotype



Genome-Wide Association Study

⟹

3 groups:

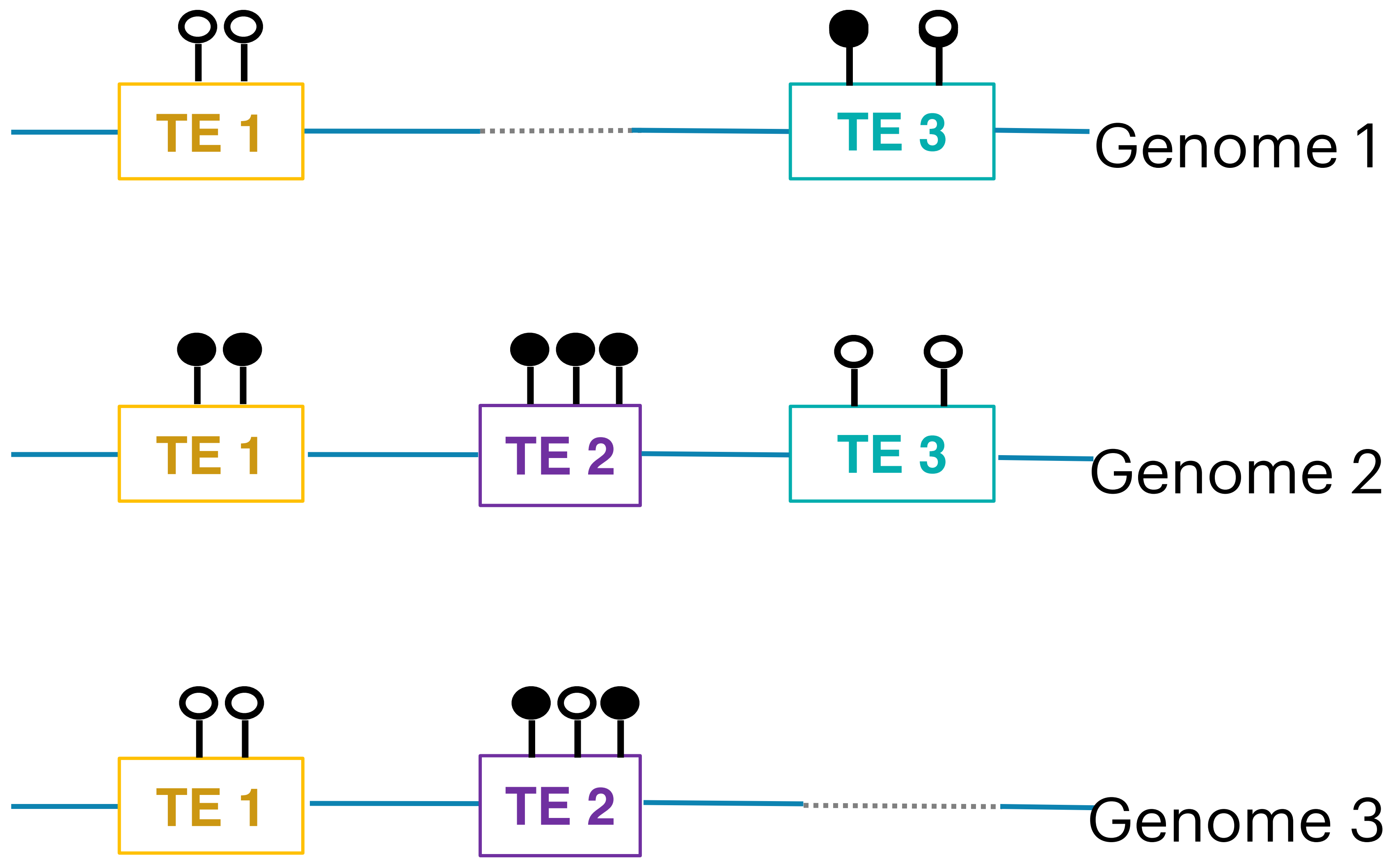
0 = absent

1 = present and not methylated (< 5%)

2 = present and methylated (> 5%)

	Gene A	Gene B	Gene C
Genome 1			
Genome 2			
Genome 3			

From epi-genotype to phenotype



Genome-Wide Association Study ⟹

3 groups:

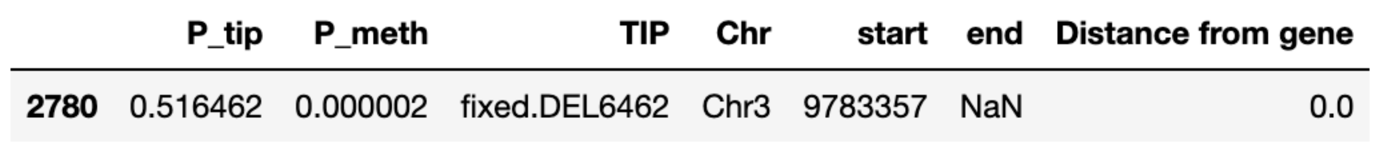
0 = absent

1 = present and not methylated (< 5%)

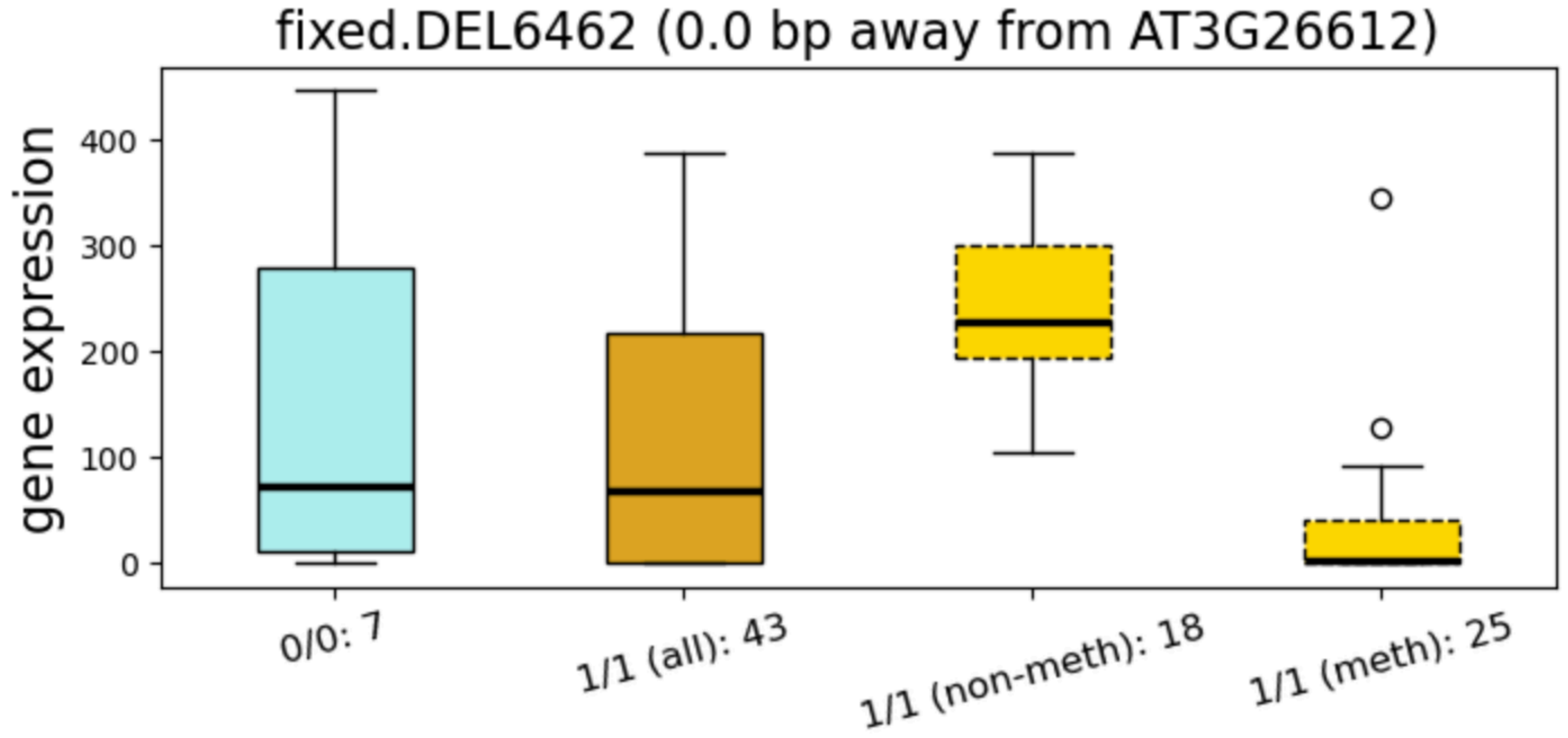
2 = present and methylated (> 5%)

	Gene A	Gene B	Gene C
Genome 1			
Genome 2			
Genome 3			

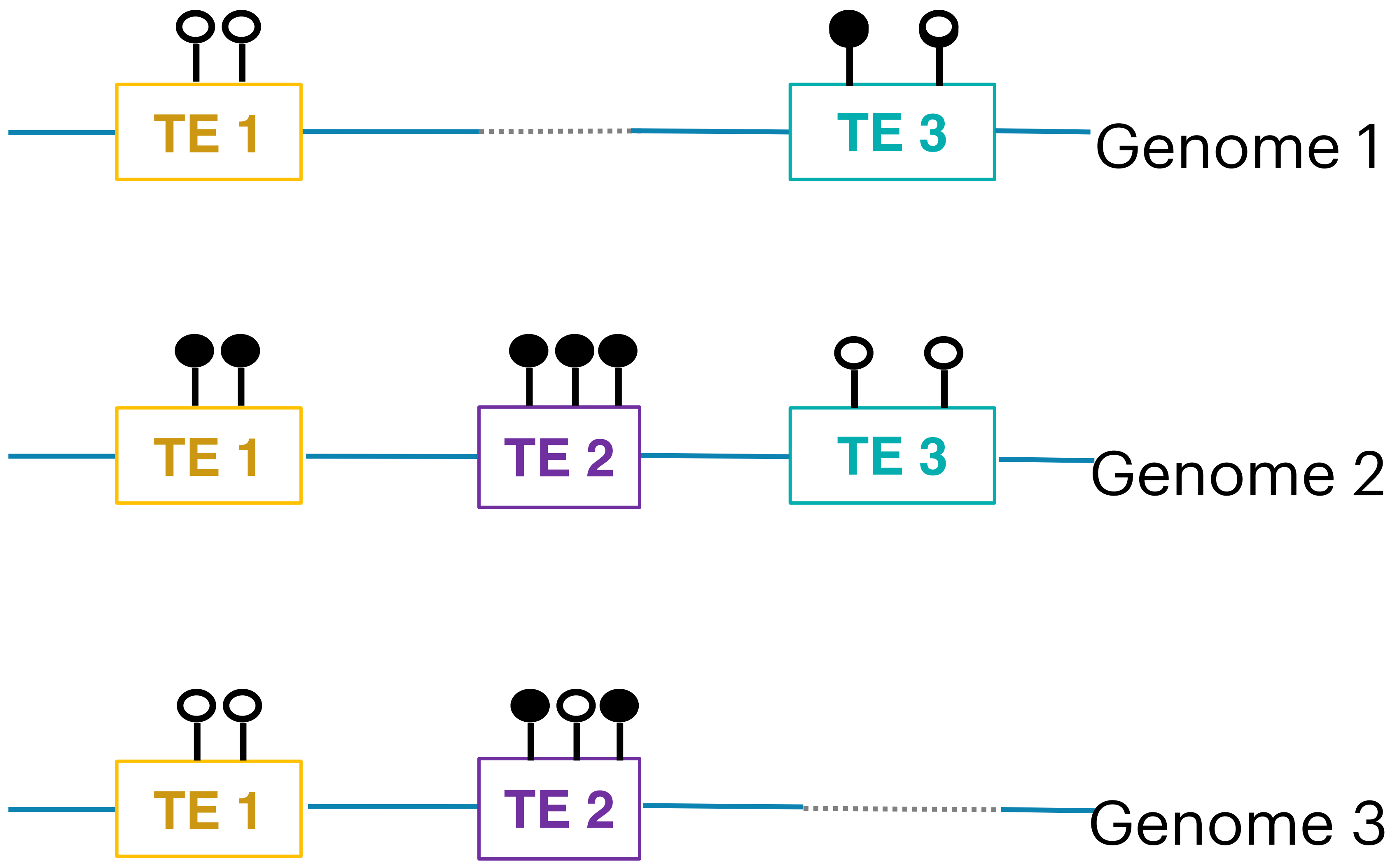
For the moment, cis- effects only:



	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
2780	0.516462	0.000002	fixed.DEL6462	Chr3	9783357	NaN	0.0



From epi-genotype to phenotype



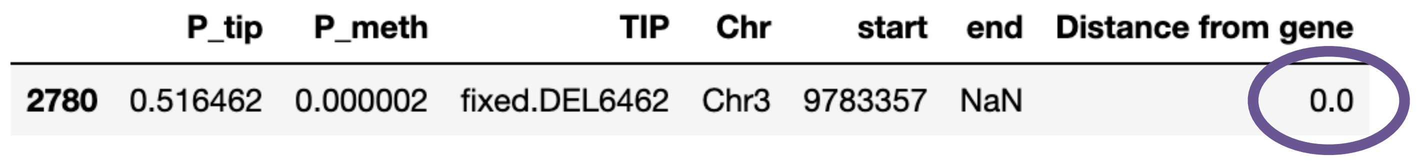
Genome-Wide Association Study

⇒

3 groups:
0 = absent
1 = present and not methylated (< 5%)
2 = present and methylated (> 5%)

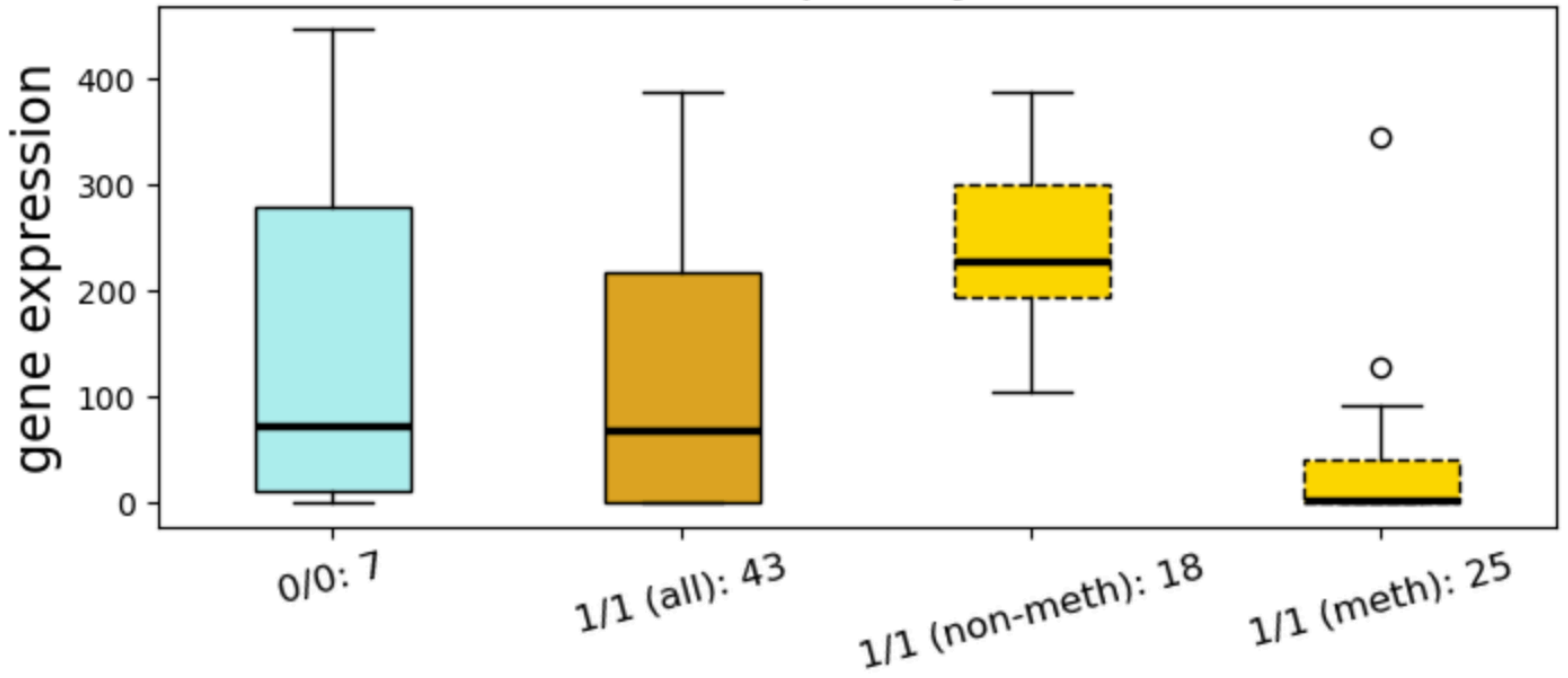
	Gene A	Gene B	Gene C
Genome 1			
Genome 2			
Genome 3			

For the moment, cis- effects only:

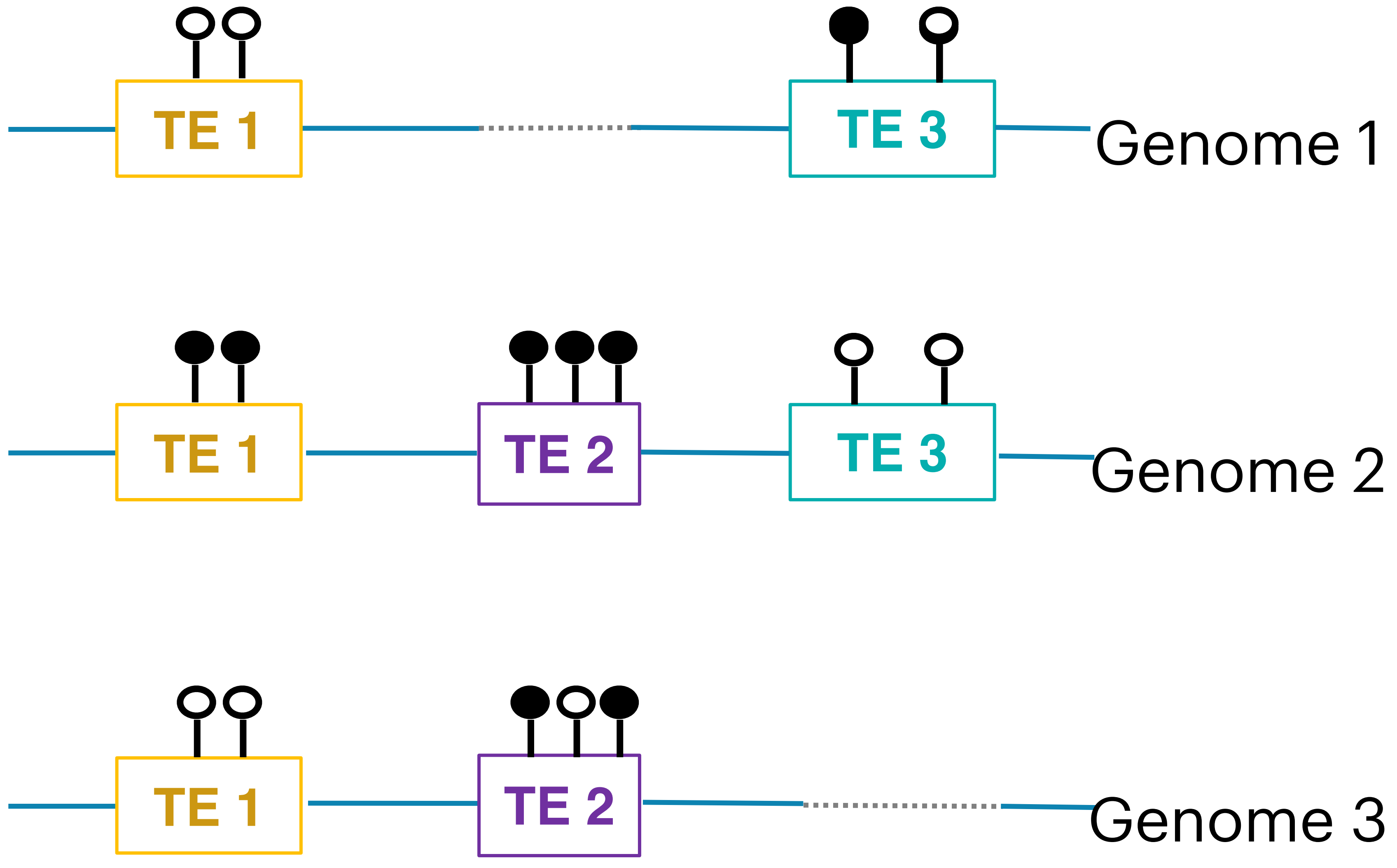


	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
2780	0.516462	0.000002	fixed.DEL6462	Chr3	9783357	NaN	0.0

fixed.DEL6462 (0.0 bp away from AT3G26612)



From epi-genotype to phenotype



Genome-Wide Association Study

⟹

3 groups:

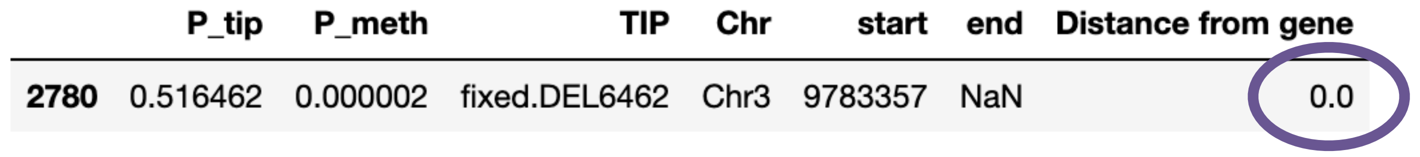
0 = absent

1 = present and not methylated (< 5%)

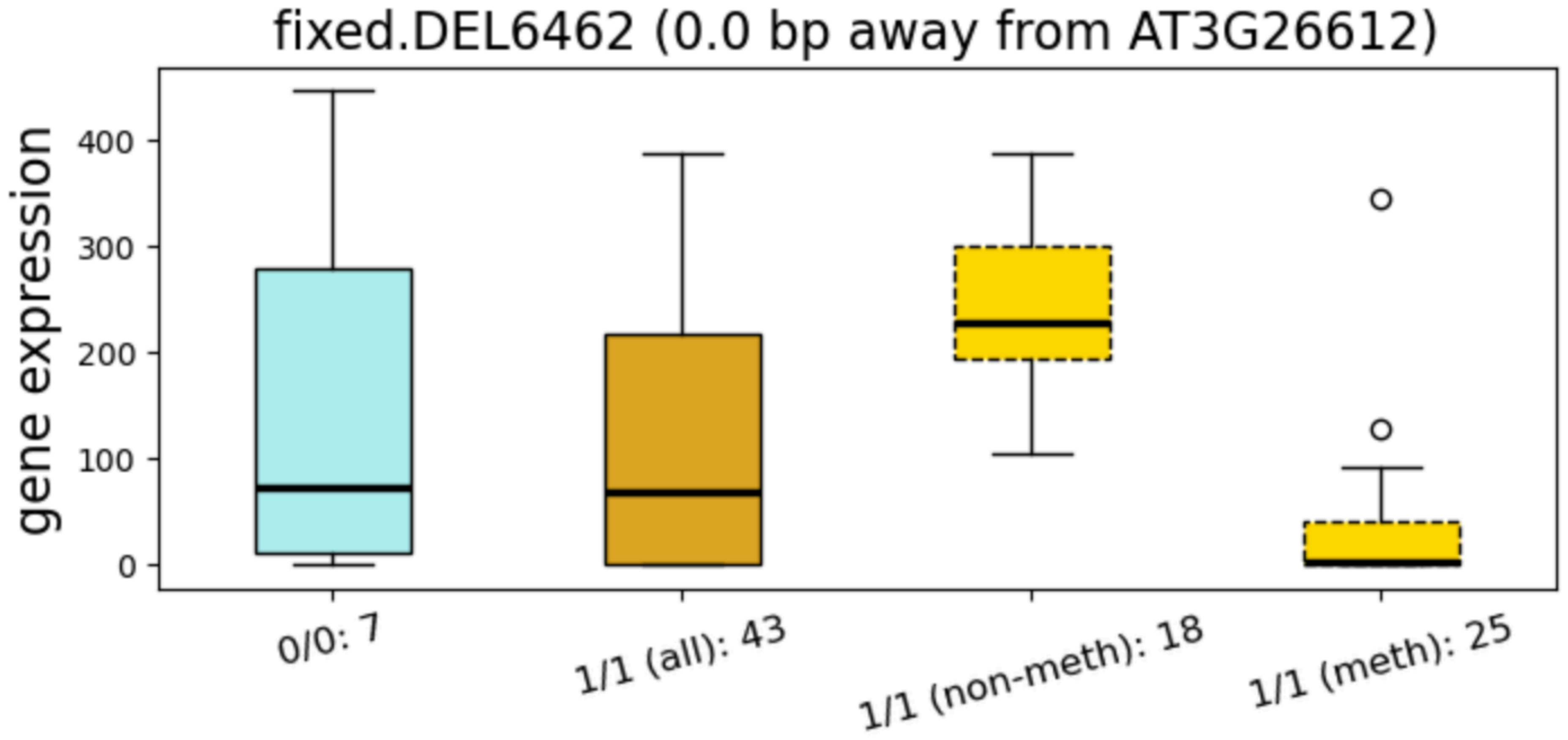
2 = present and methylated (> 5%)

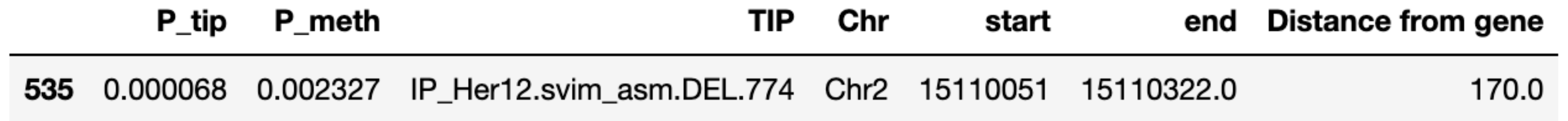
	Gene A	Gene B	Gene C
Genome 1			
Genome 2			
Genome 3			

For the moment, cis- effects only:

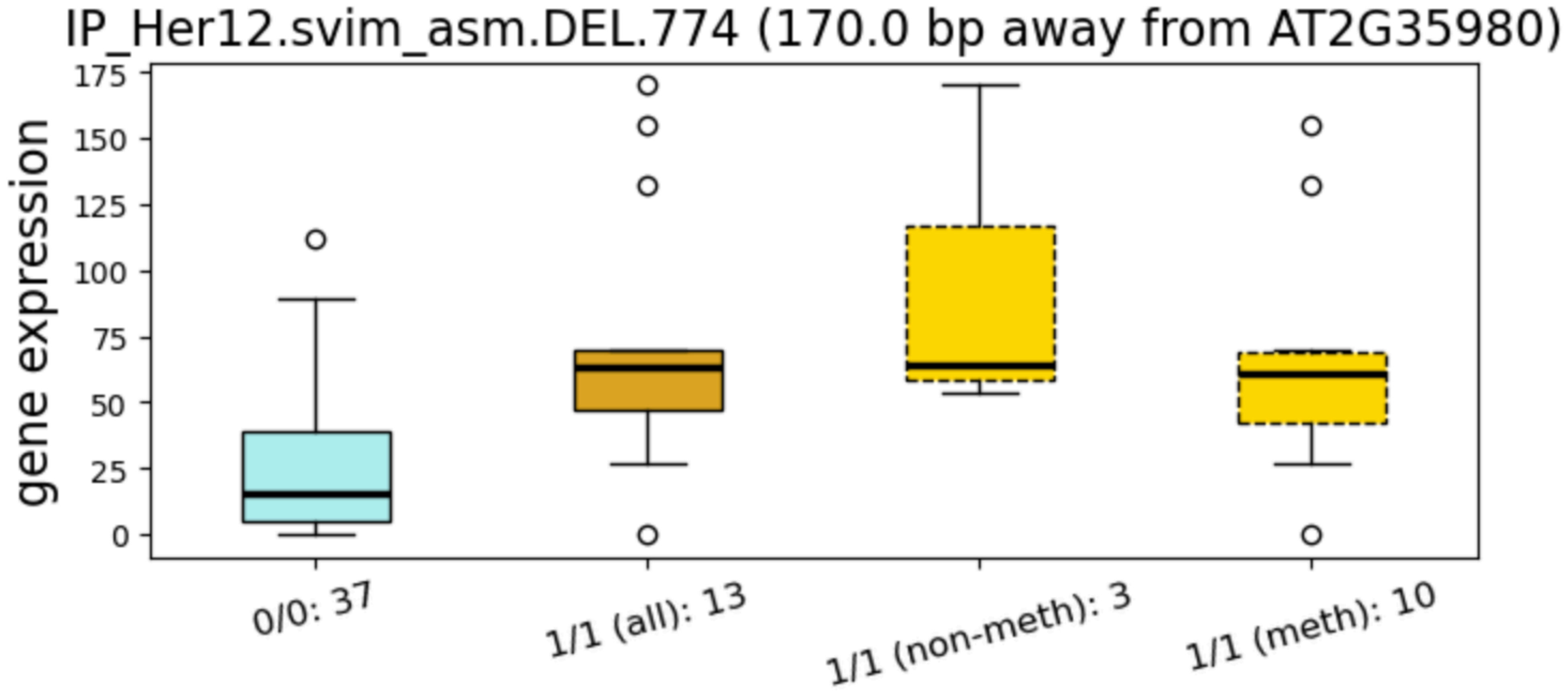


	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
2780	0.516462	0.000002	fixed.DEL6462	Chr3	9783357	NaN	0.0

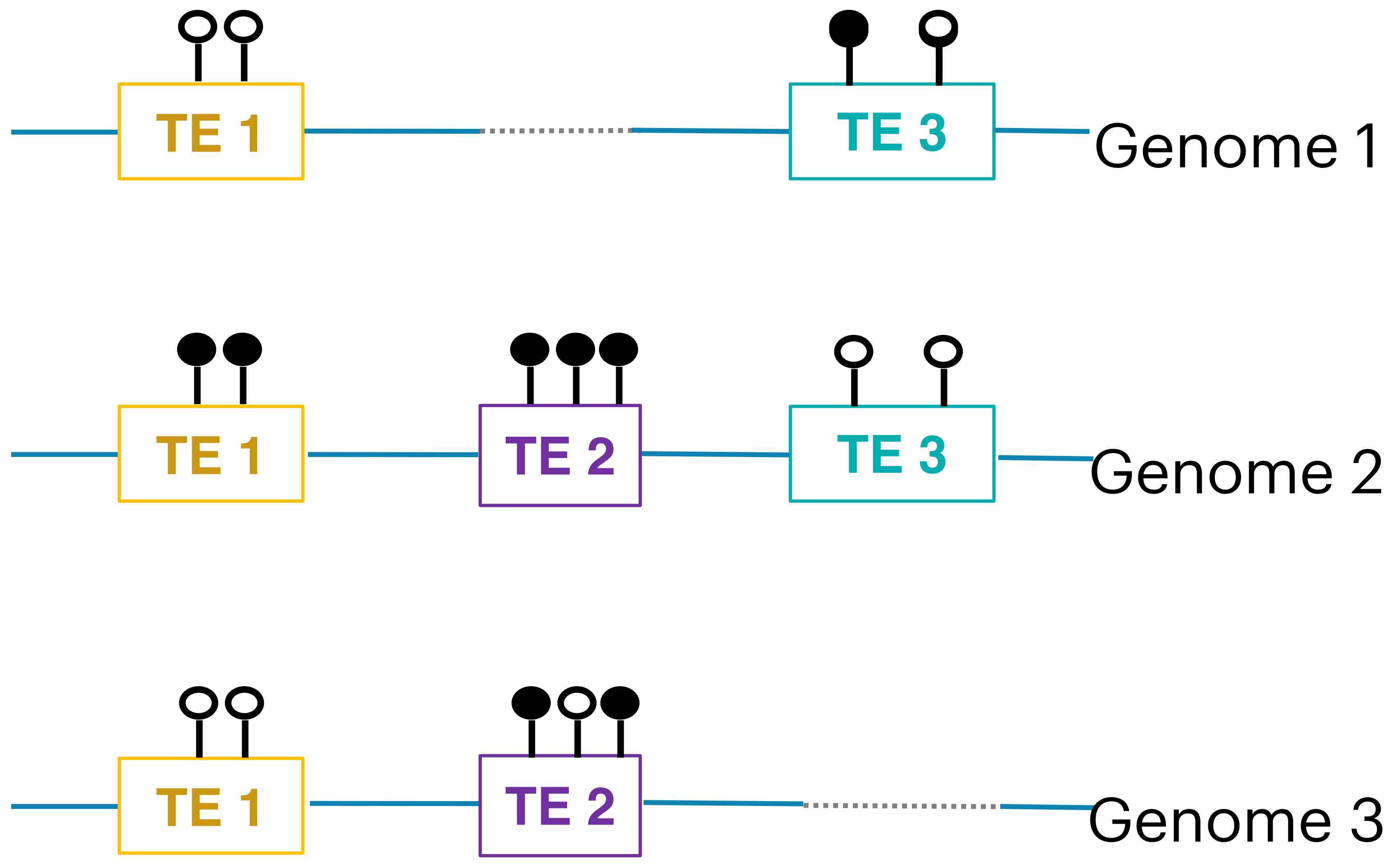




	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
535	0.000068	0.002327	IP_Her12.svim_asm.DEL.774	Chr2	15110051	15110322.0	170.0



From epi-genotype to phenotype



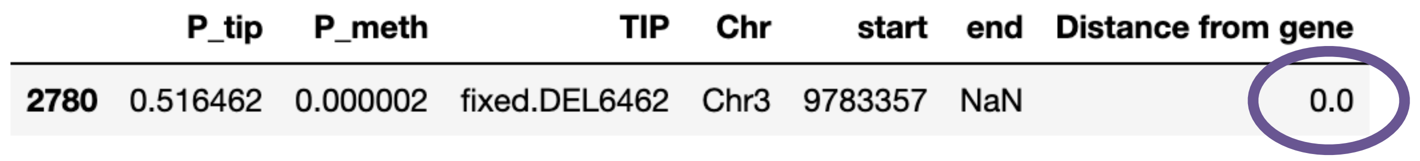
Genome-Wide Association Study

3 groups:

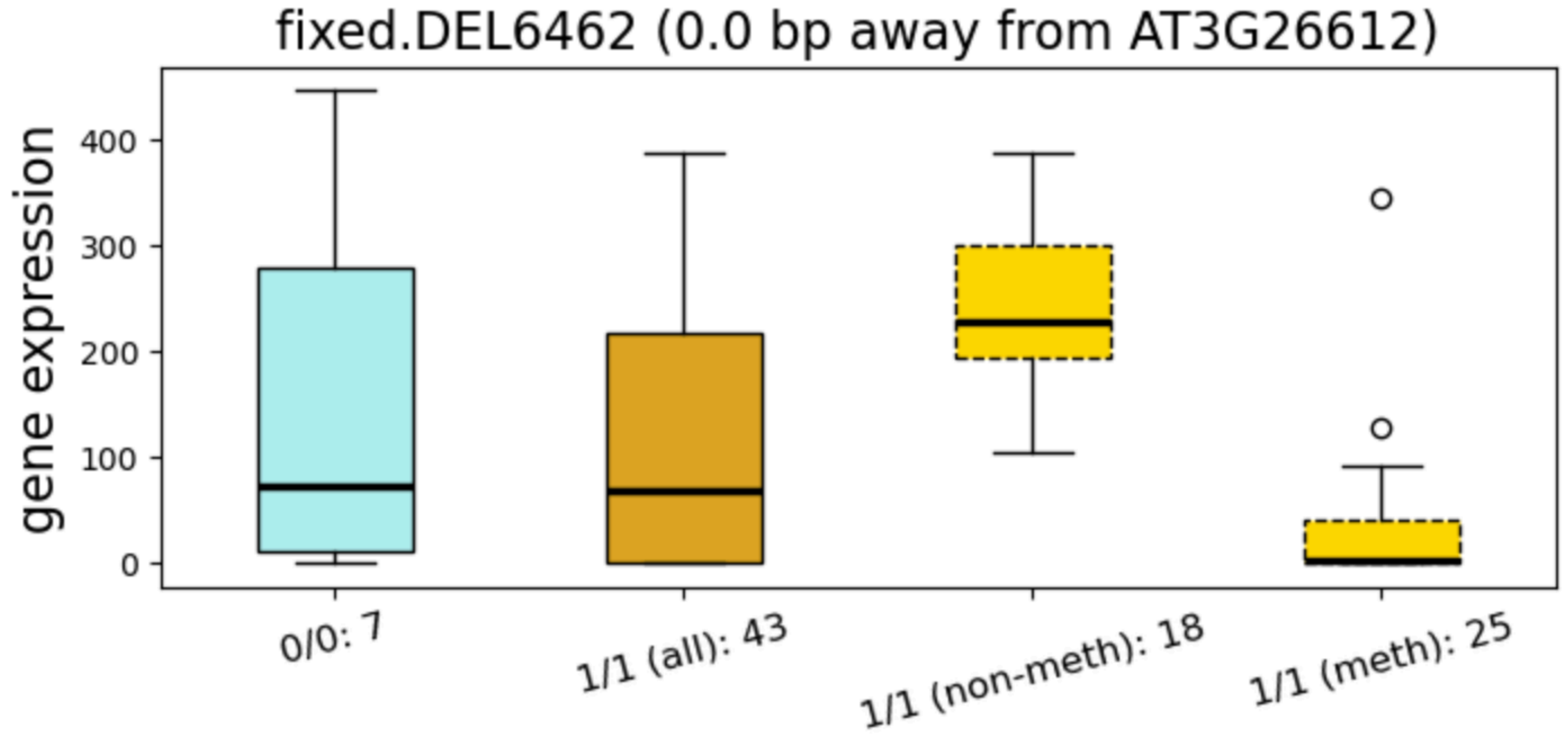
0 = absent
1 = present and not methylated (< 5%)
2 = present and methylated (> 5%)

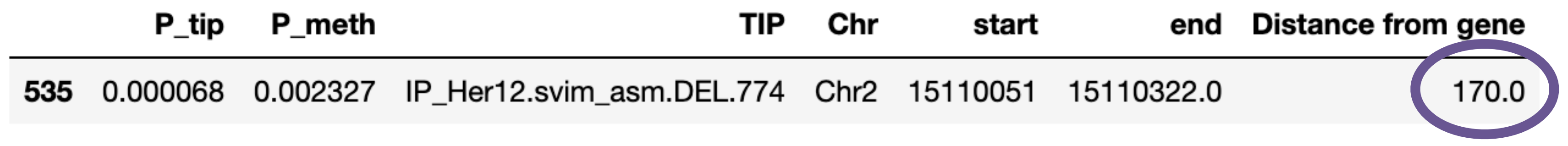
	Gene A	Gene B	Gene C
Genome 1			
Genome 2			
Genome 3			

For the moment, cis- effects only:

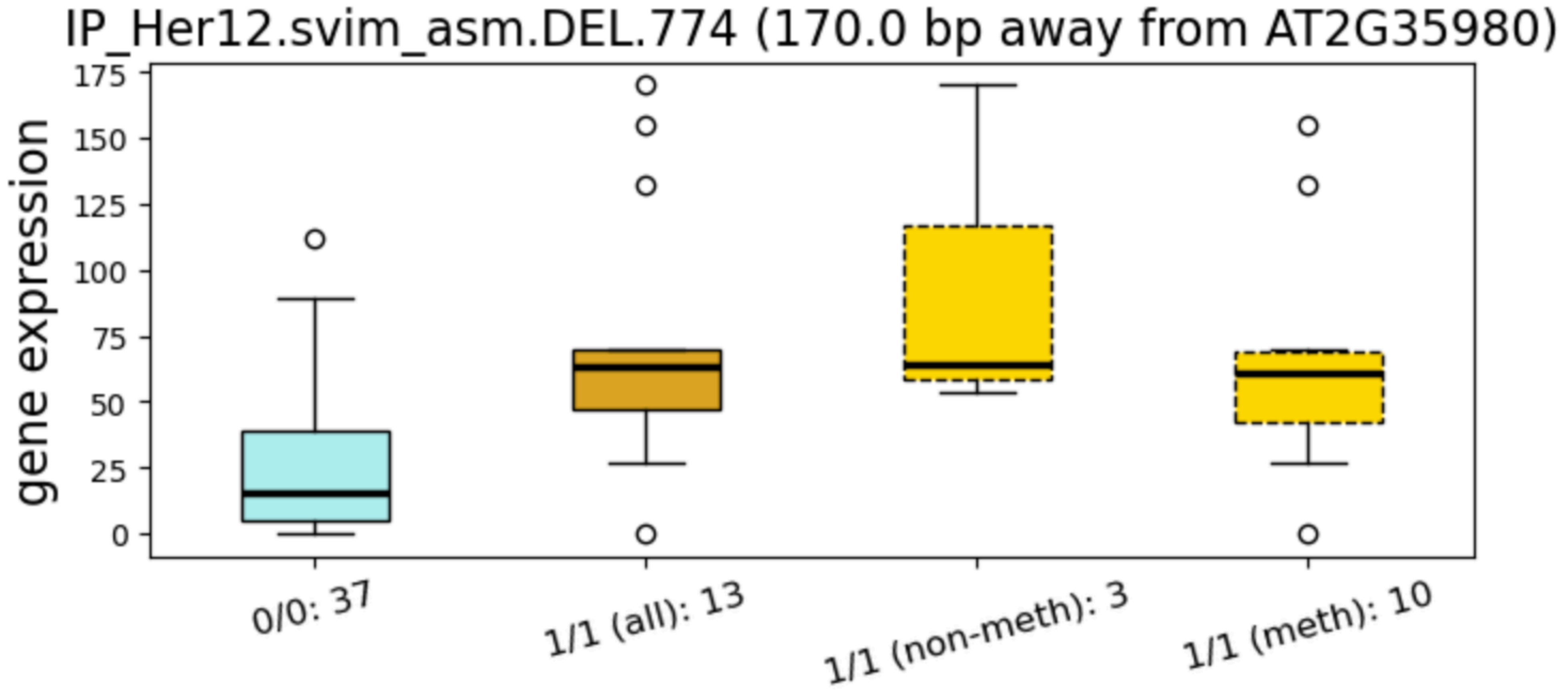


	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
2780	0.516462	0.000002	fixed.DEL6462	Chr3	9783357	NaN	0.0

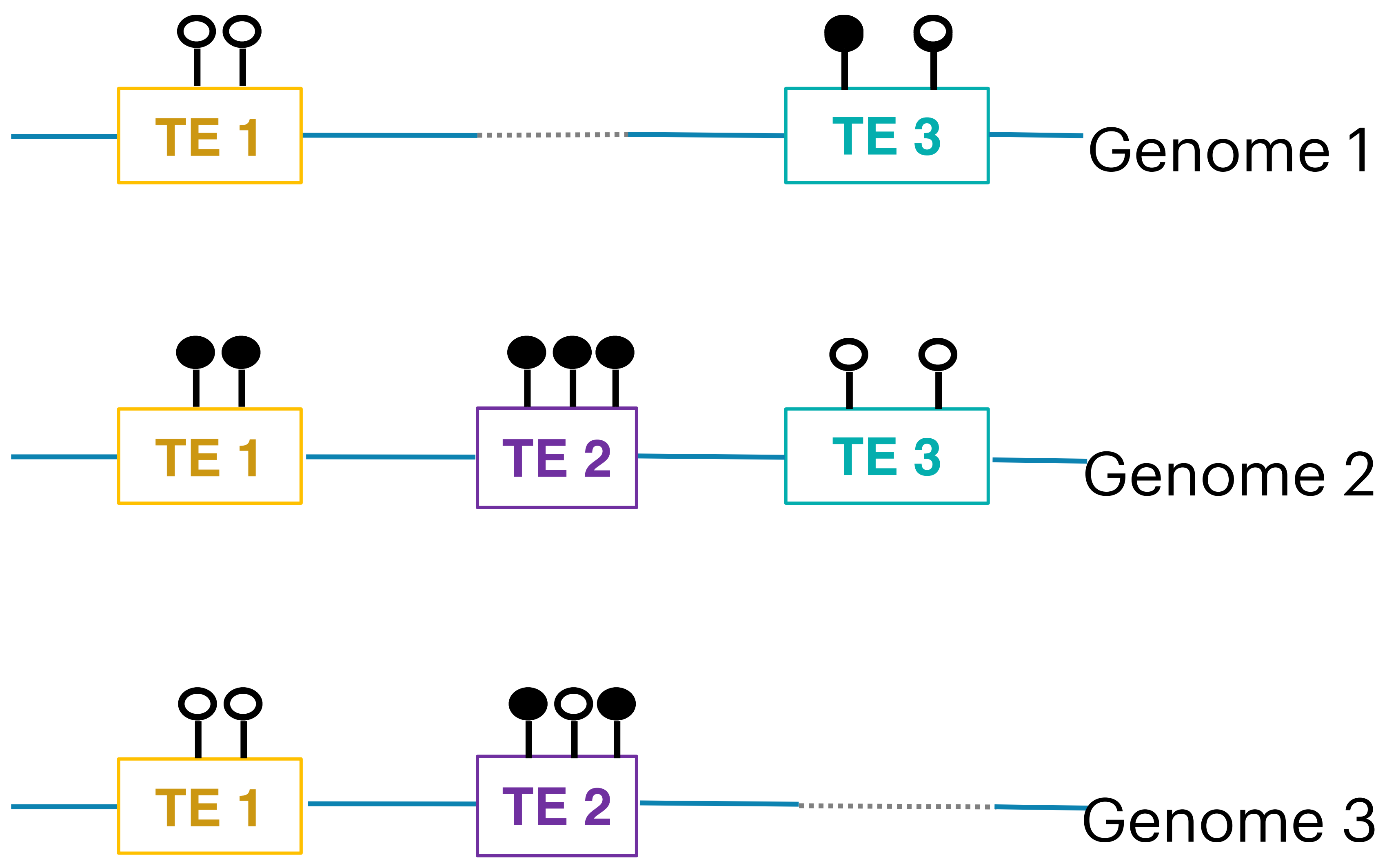

fixed.DEL6462 (0.0 bp away from AT3G26612)



	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
535	0.000068	0.002327	IP_Her12.svim_asm.DEL.774	Chr2	15110051	15110322.0	170.0


IP_Her12.svim_asm.DEL.774 (170.0 bp away from AT2G35980)

From epi-genotype to phenotype



Genome-Wide Association Study

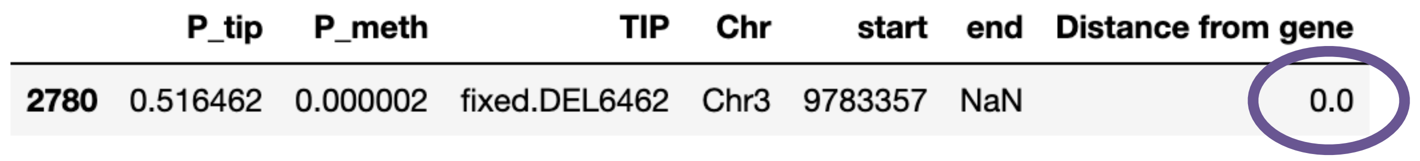
⟹

3 groups:
0 = absent
1 = present and not methylated (< 5%)
2 = present and methylated (> 5%)

	Gene A	Gene B	Gene C
Genome 1			
Genome 2			
Genome 3			

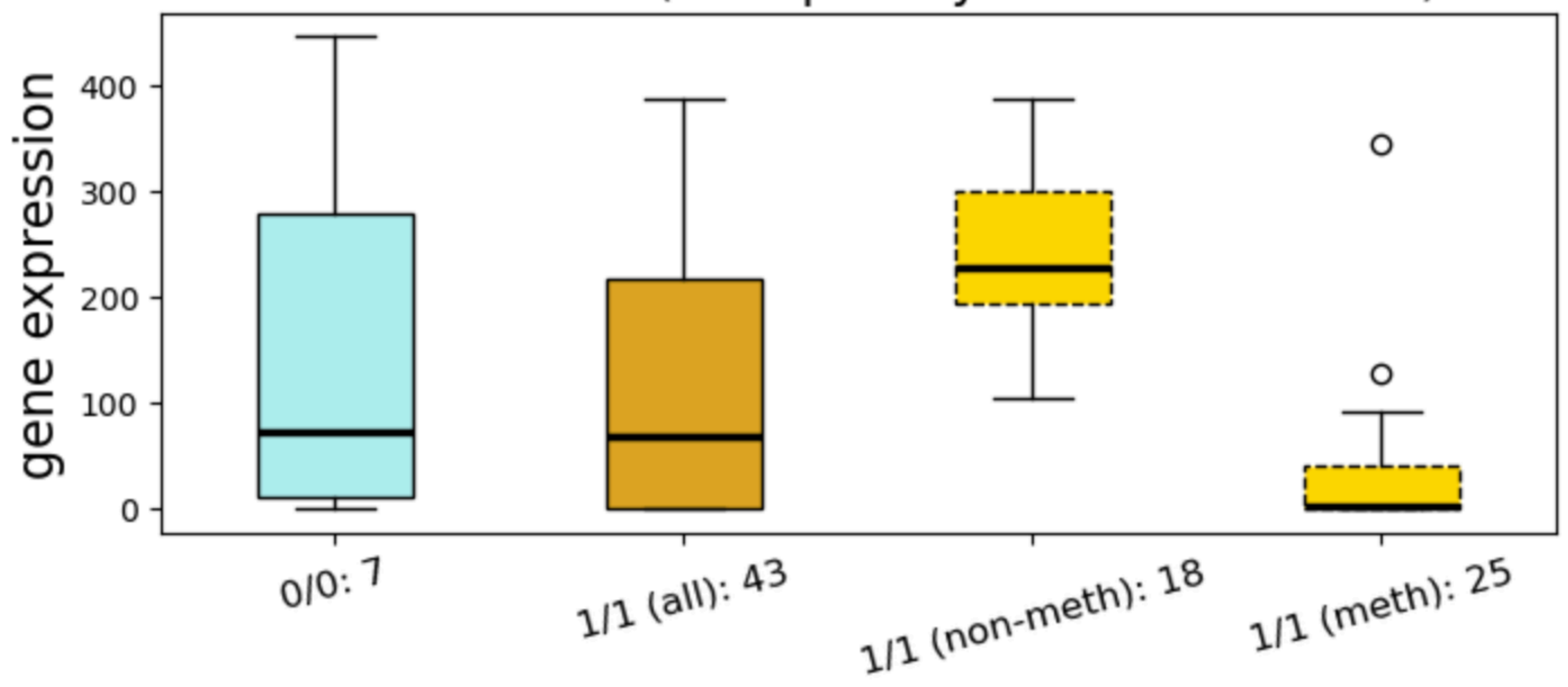
Spreading ?

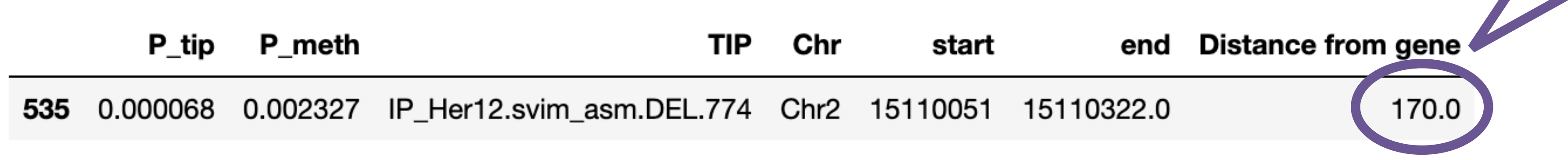
For the moment, cis- effects only:



	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
2780	0.516462	0.000002	fixed.DEL6462	Chr3	9783357	NaN	0.0

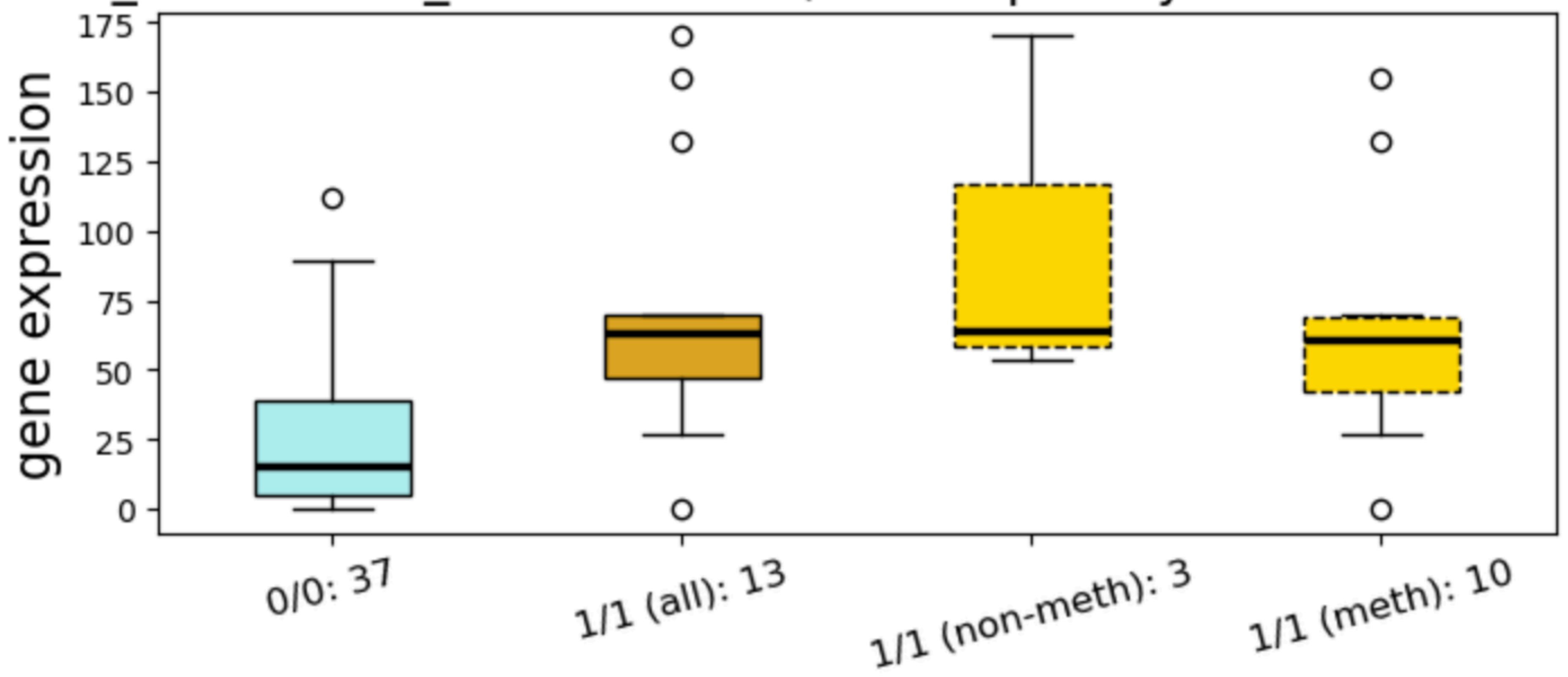
fixed.DEL6462 (0.0 bp away from AT3G26612)



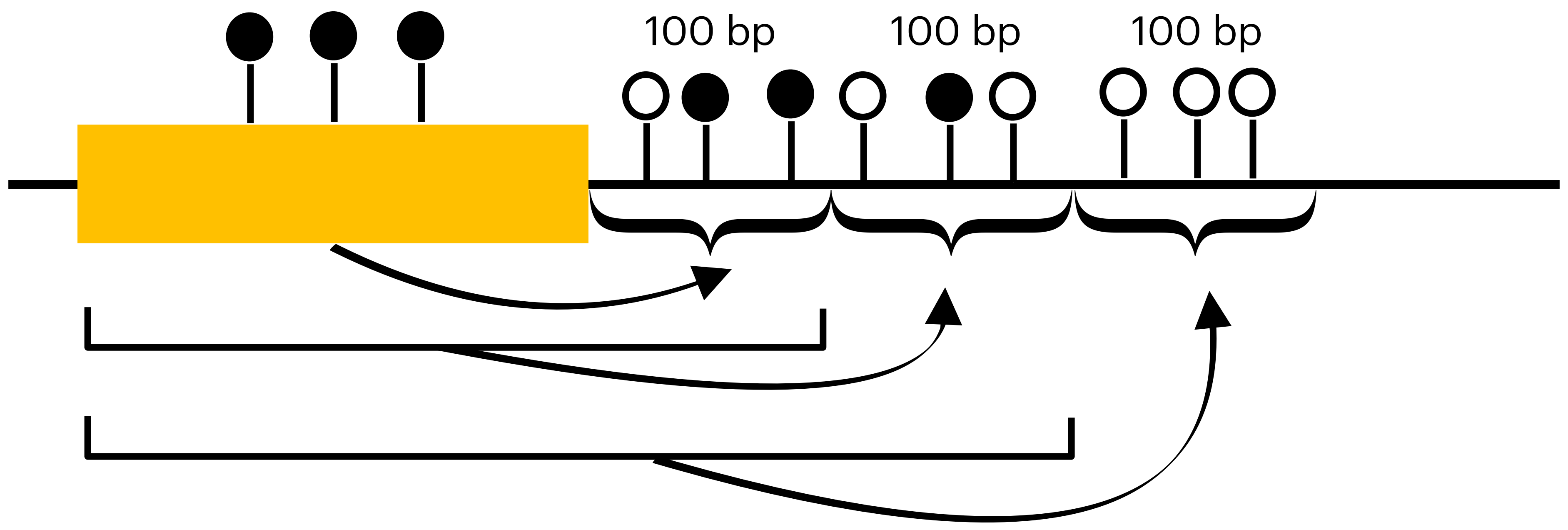


	P_tip	P_meth	TIP	Chr	start	end	Distance from gene
535	0.000068	0.002327	IP_Her12.svim_asm.DEL.774	Chr2	15110051	15110322.0	170.0

IP_Her12.svim_asm.DEL.774 (170.0 bp away from AT2G35980)

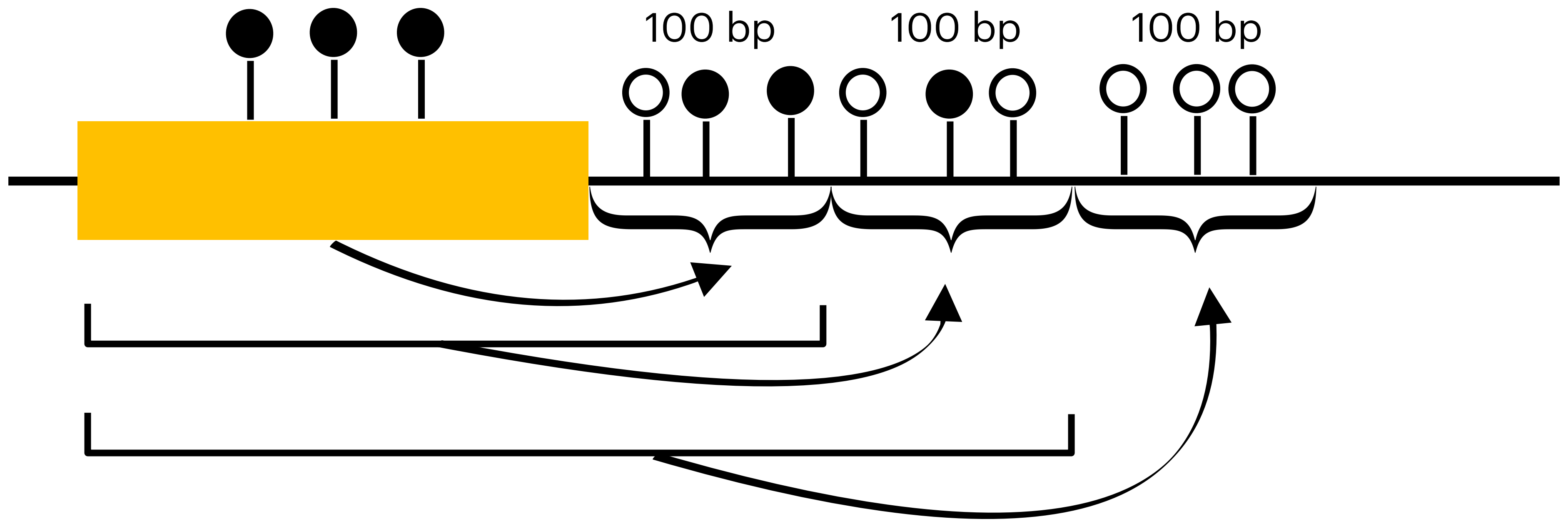


Prediction of methylation spreading



Model:
Random Forest

Prediction of methylation spreading

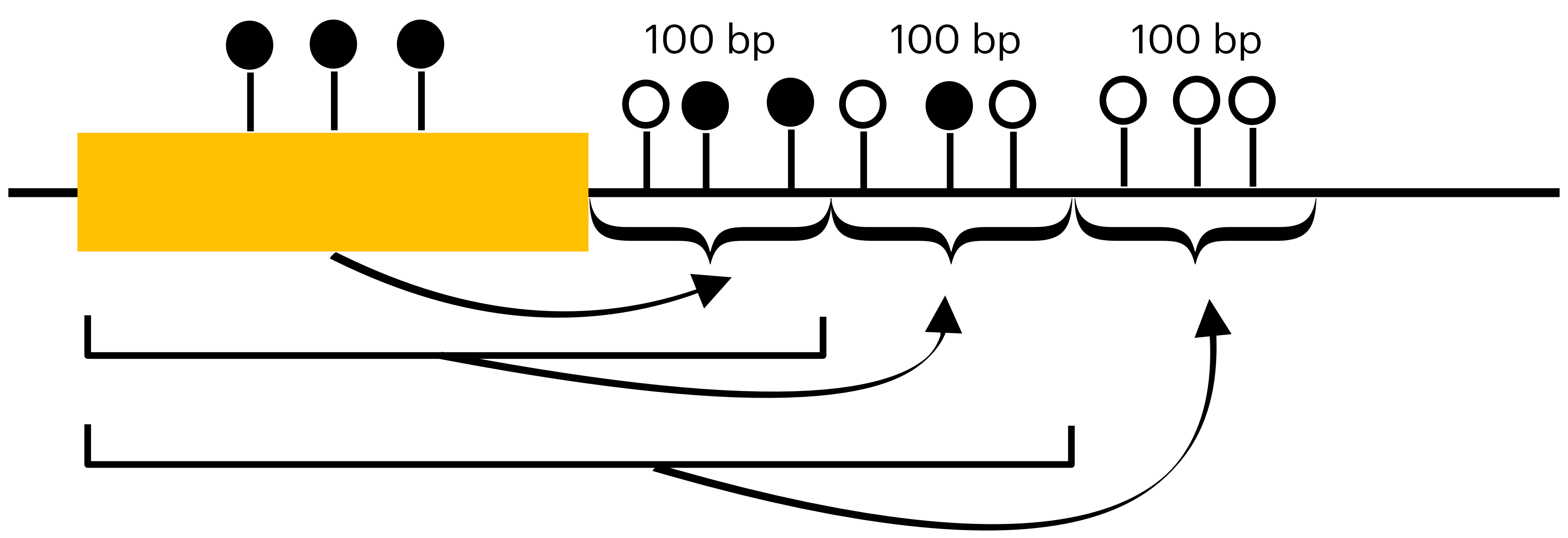


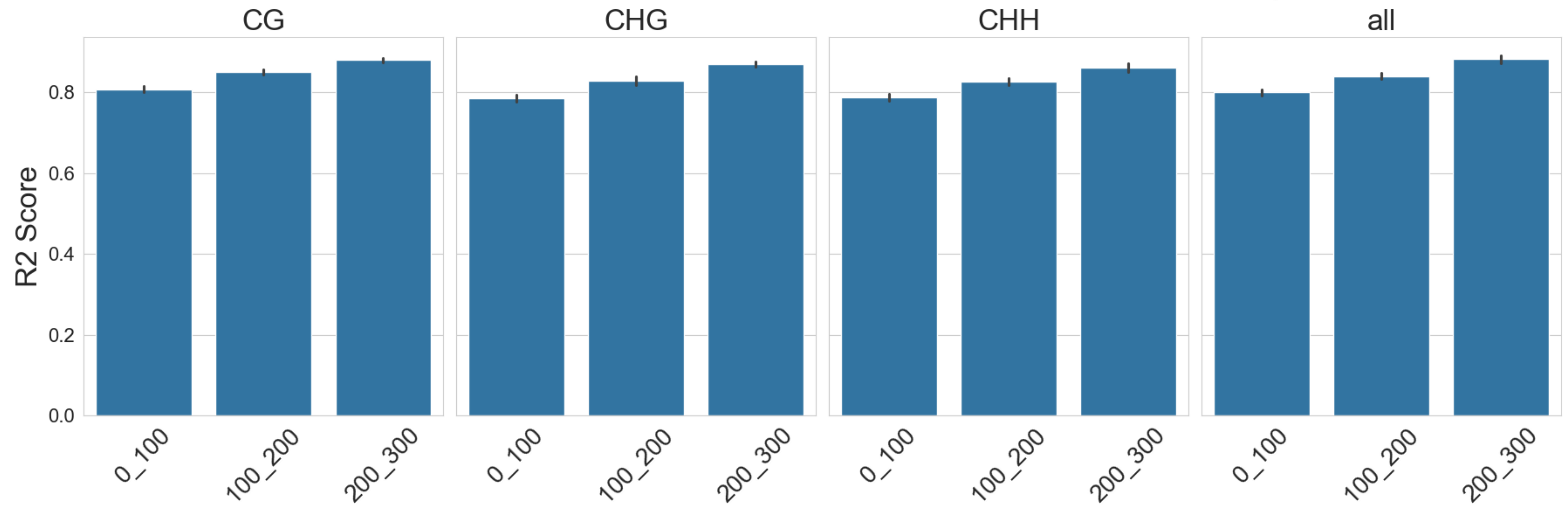
Model:
Random Forest

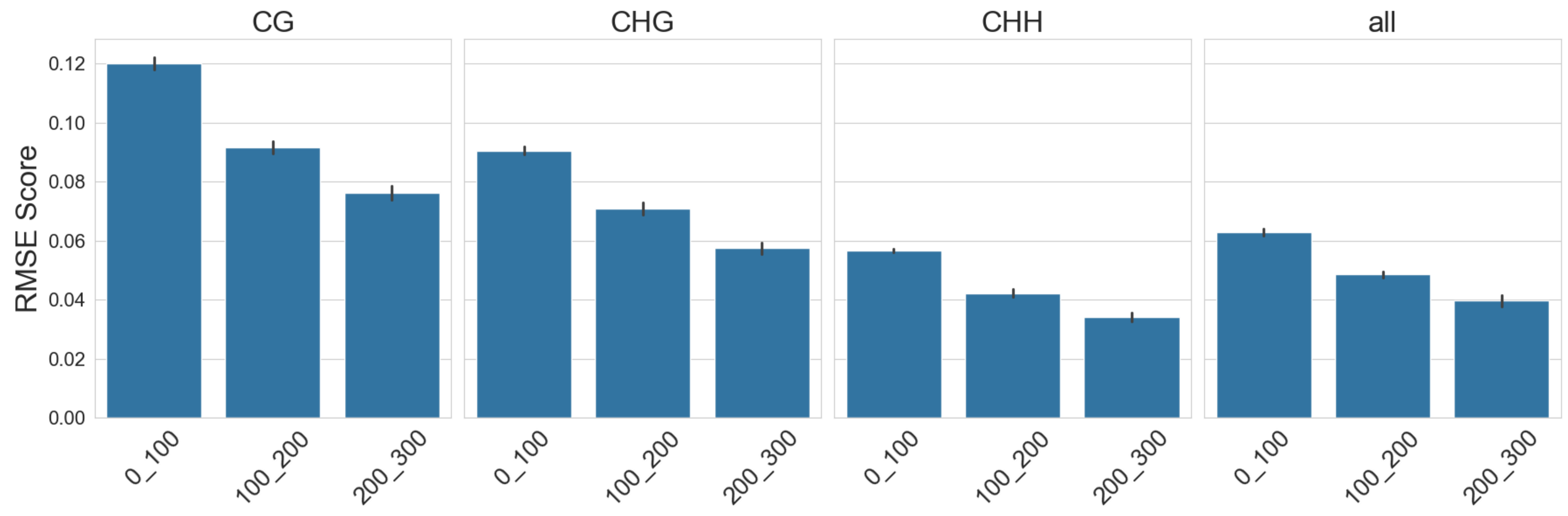
Features:

- **TE** (length, distance to pericentromere, superfamily, insertion frequency, divergence, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Methylation** in CG, CHG, CHH contexts (average genome-wide, TE, on the ends of TE, previous windows)
- **Densities** of CG, CHG, CHH contexts

Prediction of methylation spreading





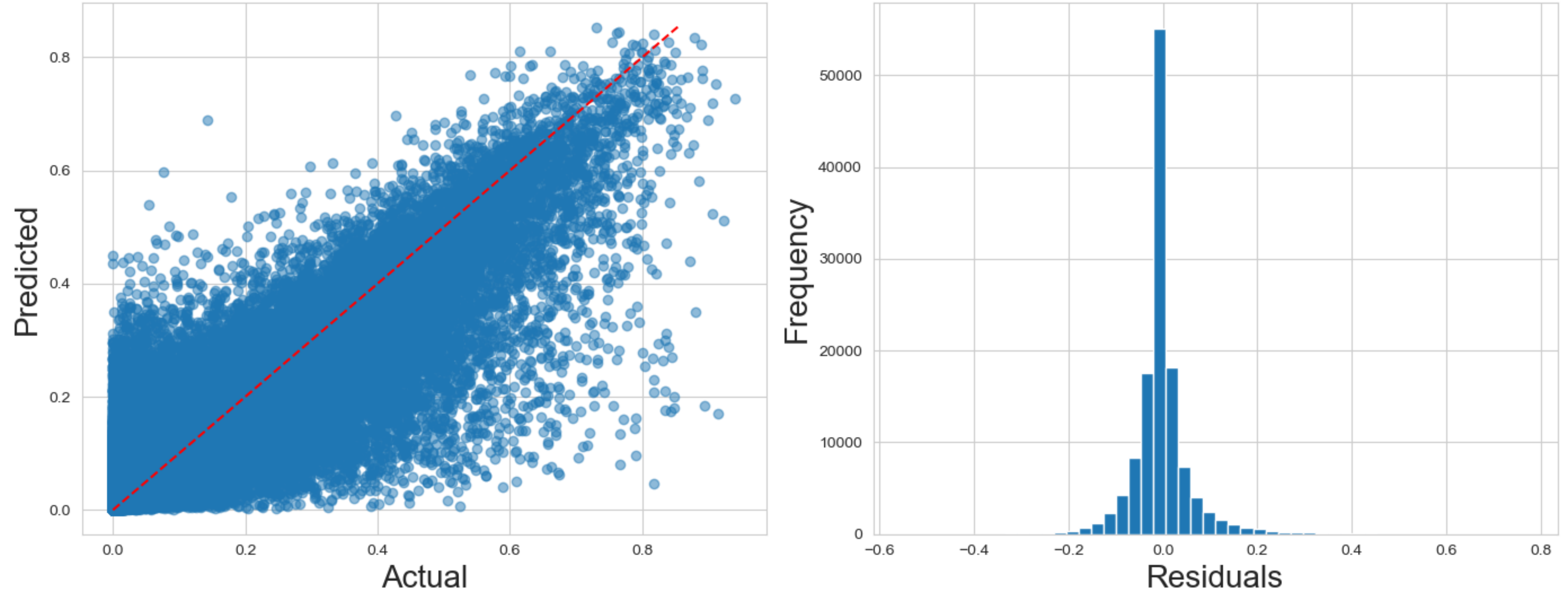


Model:
Random Forest

Features:

- **TE** (length, distance to pericentromere, superfamily, insertion frequency, divergence, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Methylation** in CG, CHG, CHH contexts (average genome-wide, TE, on the ends of TE, previous windows)
- **Densities** of CG, CHG, CHH contexts

all methylation in 0_100 window



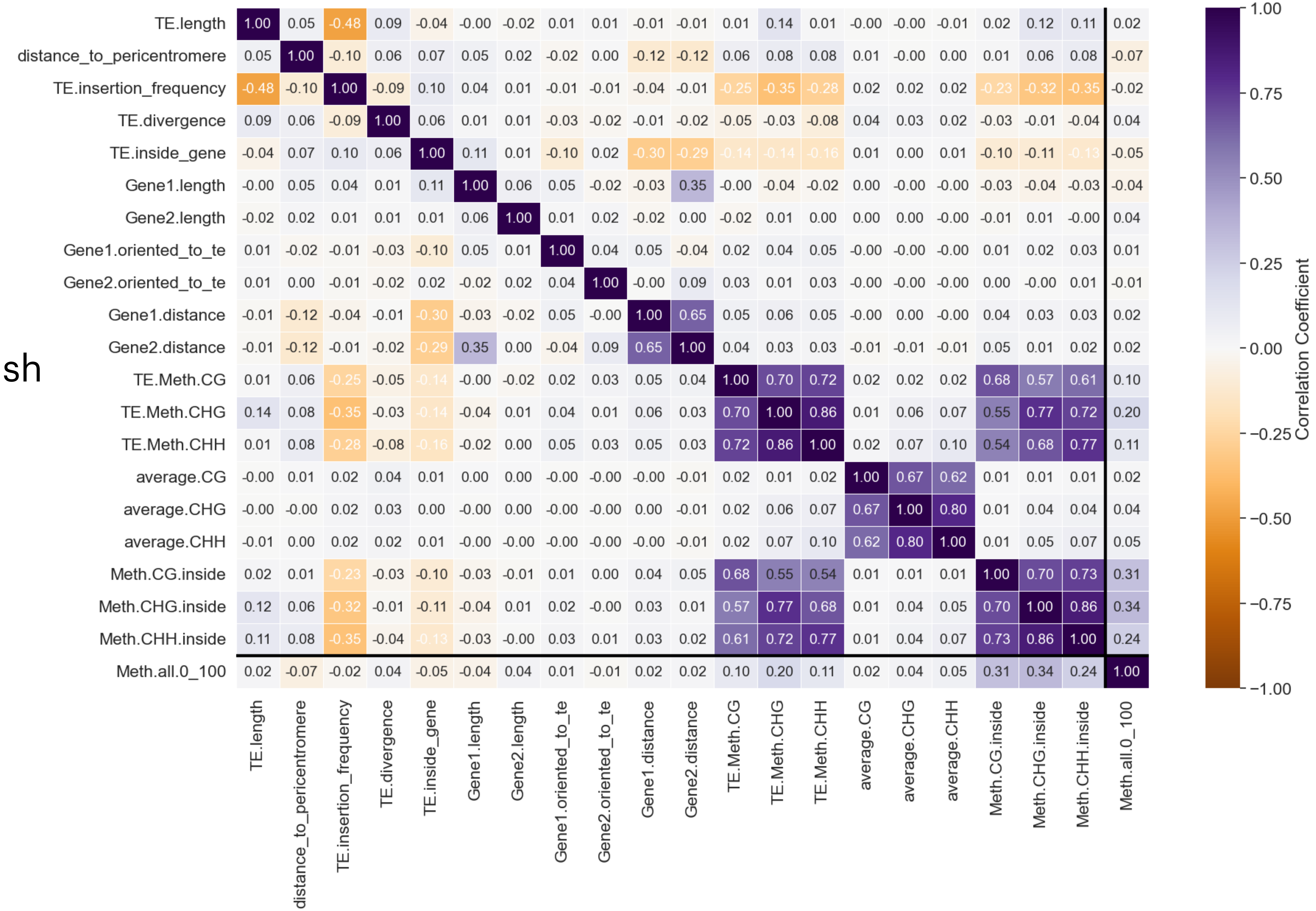
Prediction of methylation spreading

- The model predicts well, but which features define methylation level?

Prediction of methylation spreading

- The model predicts well, but which features define methylation level?

- Some features are highly correlated ⟹ hard to distinguish between them

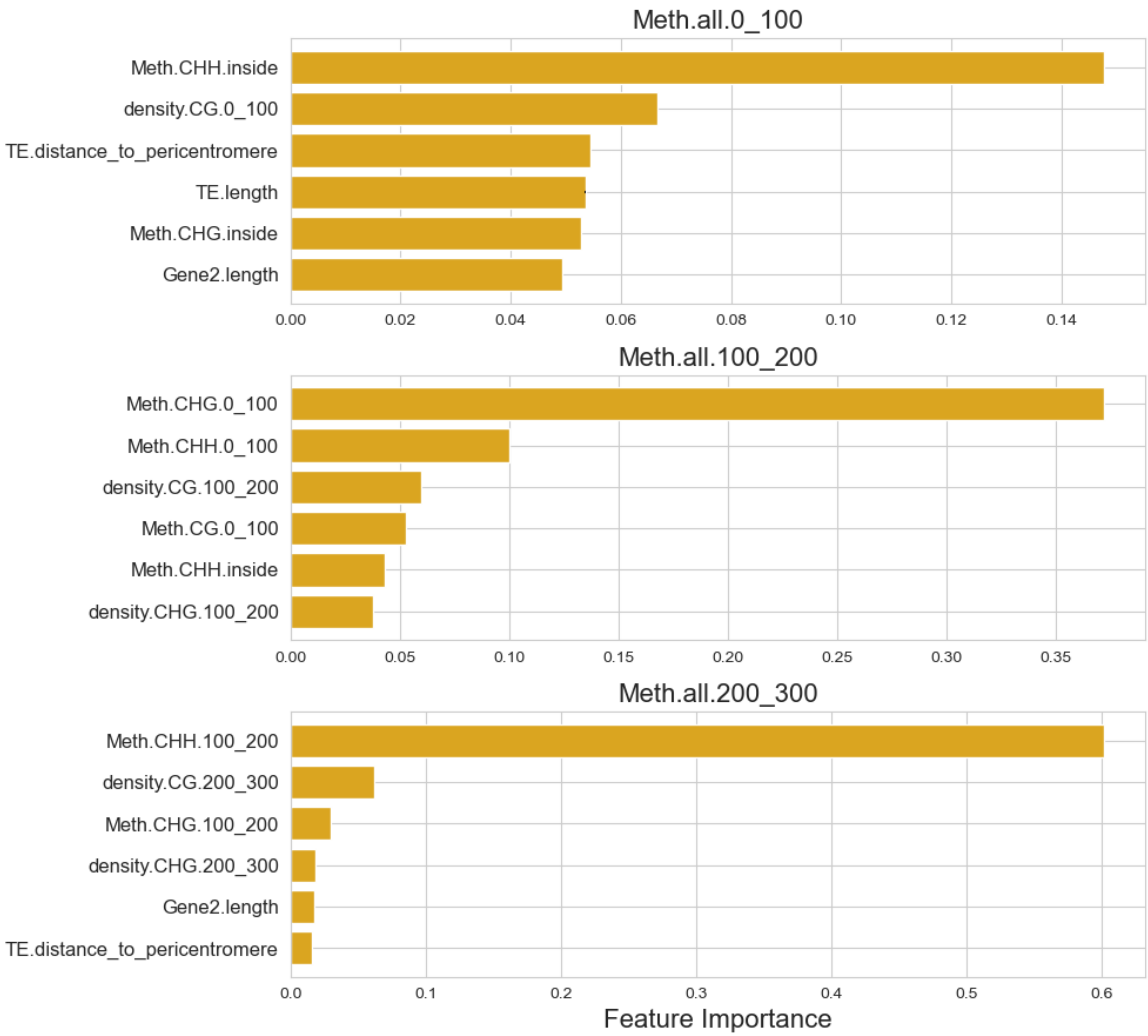


	TE.length	distance_to_pericentromere	TE.insertion_frequency	TE.divergence	TE.inside_gene	Gene1.length	Gene2.length	Gene1.oriented_to_te	Gene2.oriented_to_te	Gene1.distance	Gene2.distance	TE.Meth.CG	TE.Meth.CHG	TE.Meth.CHH	average.CG	average.CHG	average.CHH	Meth.CG.inside	Meth.CHG.inside	Meth.CHH.inside	Meth.all.0_100
TE.length	1.00	0.05	-0.48	0.09	-0.04	-0.00	-0.02	0.01	0.01	-0.01	-0.01	0.01	0.14	0.01	-0.00	-0.00	-0.01	0.02	0.12	0.11	0.02
distance_to_pericentromere	0.05	1.00	-0.10	0.06	0.07	0.05	0.02	-0.02	0.00	-0.12	-0.12	0.06	0.08	0.08	0.01	-0.00	0.00	0.01	0.06	0.08	-0.07
TE.insertion_frequency	-0.48	-0.10	1.00	-0.09	0.10	0.04	0.01	-0.01	-0.01	-0.04	-0.01	-0.25	-0.35	-0.28	0.02	0.02	0.02	-0.23	-0.32	-0.35	-0.02
TE.divergence	0.09	0.06	-0.09	1.00	0.06	0.01	0.01	-0.03	-0.02	-0.01	-0.02	-0.05	-0.03	-0.08	0.04	0.03	0.02	-0.03	-0.01	-0.04	0.04
TE.inside_gene	-0.04	0.07	0.10	0.06	1.00	0.11	0.01	-0.10	0.02	-0.30	-0.29	-0.14	-0.14	-0.16	0.01	0.00	0.01	-0.10	-0.11	-0.13	-0.05
Gene1.length	-0.00	0.05	0.04	0.01	0.11	1.00	0.06	0.05	-0.02	-0.03	0.35	-0.00	-0.04	-0.02	0.00	-0.00	-0.00	-0.03	-0.04	-0.03	-0.04
Gene2.length	-0.02	0.02	0.01	0.01	0.01	0.06	1.00	0.01	0.02	-0.02	0.00	-0.02	0.01	0.00	0.00	0.00	-0.00	-0.01	0.01	-0.00	0.04
Gene1.oriented_to_te	0.01	-0.02	-0.01	-0.03	-0.10	0.05	0.01	1.00	0.04	0.05	-0.04	0.02	0.04	0.05	-0.00	-0.00	-0.00	0.01	0.02	0.03	0.01
Gene2.oriented_to_te	0.01	0.00	-0.01	-0.02	0.02	-0.02	0.02	0.04	1.00	-0.00	0.09	0.03	0.01	0.03	-0.00	-0.00	-0.00	0.00	-0.00	0.01	-0.01
Gene1.distance	-0.01	-0.12	-0.04	-0.01	-0.30	-0.03	-0.02	0.05	-0.00	1.00	0.65	0.05	0.06	0.05	-0.00	0.00	-0.00	0.04	0.03	0.03	0.02
Gene2.distance	-0.01	-0.12	-0.01	-0.02	-0.29	0.35	0.00	-0.04	0.09	0.65	1.00	0.04	0.03	0.03	-0.01	-0.01	-0.01	0.05	0.01	0.02	0.02
TE.Meth.CG	0.01	0.06	-0.25	-0.05	-0.14	-0.00	-0.02	0.02	0.03	0.05	0.04	1.00	0.70	0.72	0.02	0.02	0.02	0.68	0.57	0.61	0.10
TE.Meth.CHG	0.14	0.08	-0.35	-0.03	-0.14	-0.04	0.01	0.04	0.01	0.06	0.03	0.70	1.00	0.86	0.01	0.06	0.07	0.55	0.77	0.72	0.20
TE.Meth.CHH	0.01	0.08	-0.28	-0.08	-0.16	-0.02	0.00	0.05	0.03	0.05	0.03	0.72	0.86	1.00	0.02	0.07	0.10	0.54	0.68	0.77	0.11
average.CG	-0.00	0.01	0.02	0.04	0.01	0.00	0.00	-0.00	-0.00	-0.00	-0.01	0.02	0.01	0.02	1.00	0.67	0.62	0.01	0.01	0.01	0.02
average.CHG	-0.00	-0.00	0.02	0.03	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.01	0.02	0.06	0.07	0.67	1.00	0.80	0.01	0.04	0.04	0.04
average.CHH	-0.01	0.00	0.02	0.02	0.01	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	0.02	0.07	0.10	0.62	0.80	1.00	0.01	0.05	0.07	0.05
Meth.CG.inside	0.02	0.01	-0.23	-0.03	-0.10	-0.03	-0.01	0.01	0.00	0.04	0.05	0.68	0.55	0.54	0.01	0.01	0.01	1.00	0.70	0.73	0.31
Meth.CHG.inside	0.12	0.06	-0.32	-0.01	-0.11	-0.04	0.01	0.02	-0.00	0.03	0.01	0.57	0.77	0.68	0.01	0.04	0.05	0.70	1.00	0.86	0.34
Meth.CHH.inside	0.11	0.08	-0.35	-0.04	-0.13	-0.03	-0.00	0.03	0.01	0.03	0.02	0.61	0.72	0.77	0.01	0.04	0.07	0.73	0.86	1.00	0.24
Meth.all.0_100	0.02	-0.07	-0.02	0.04	-0.05	-0.04	0.04	0.01	-0.01	0.02	0.02	0.10	0.20	0.11	0.02	0.04	0.05	0.31	0.34	0.24	1.00

Prediction of methylation spreading

Impurity-based feature importances

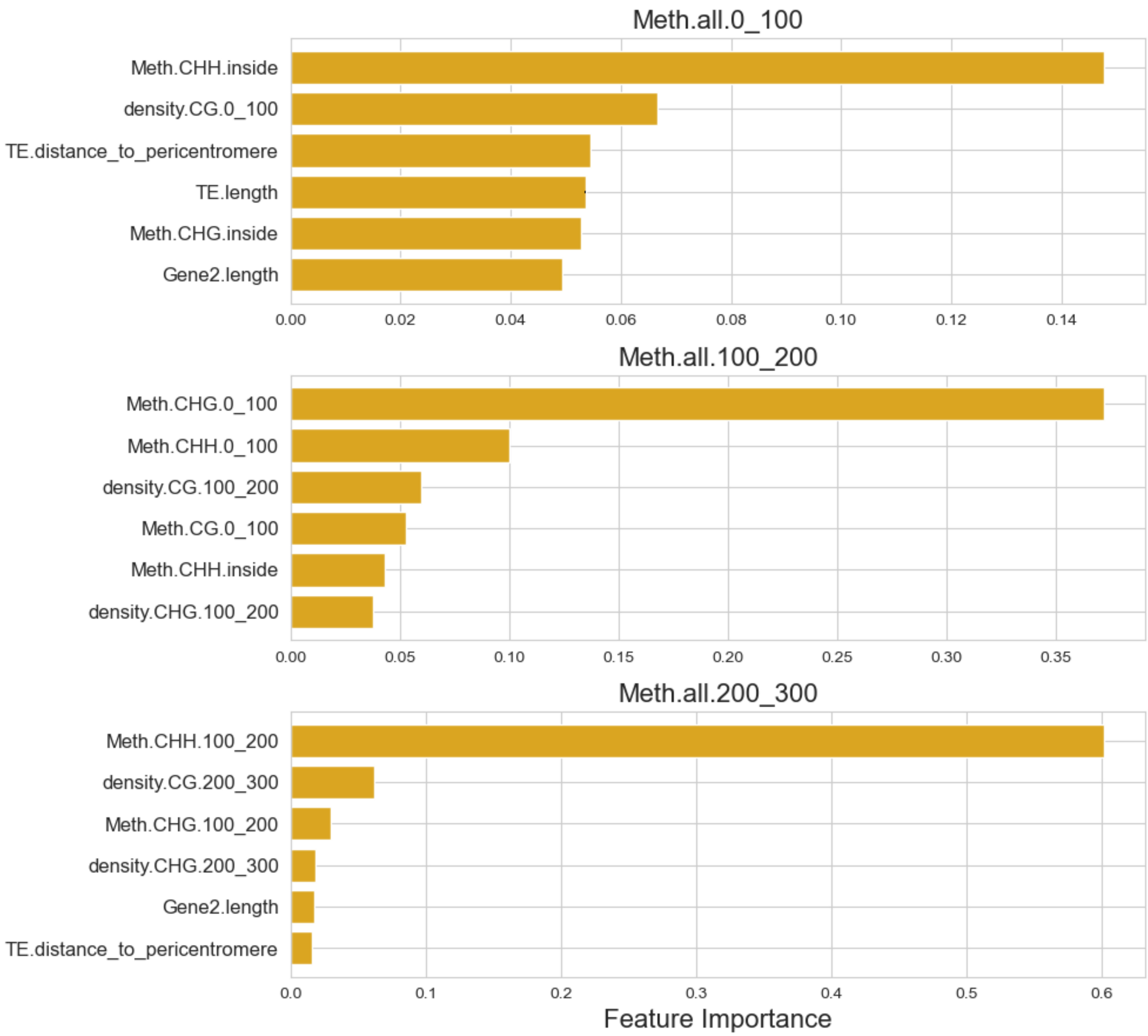
* 10 independent runs with different random seeds



Prediction of methylation spreading

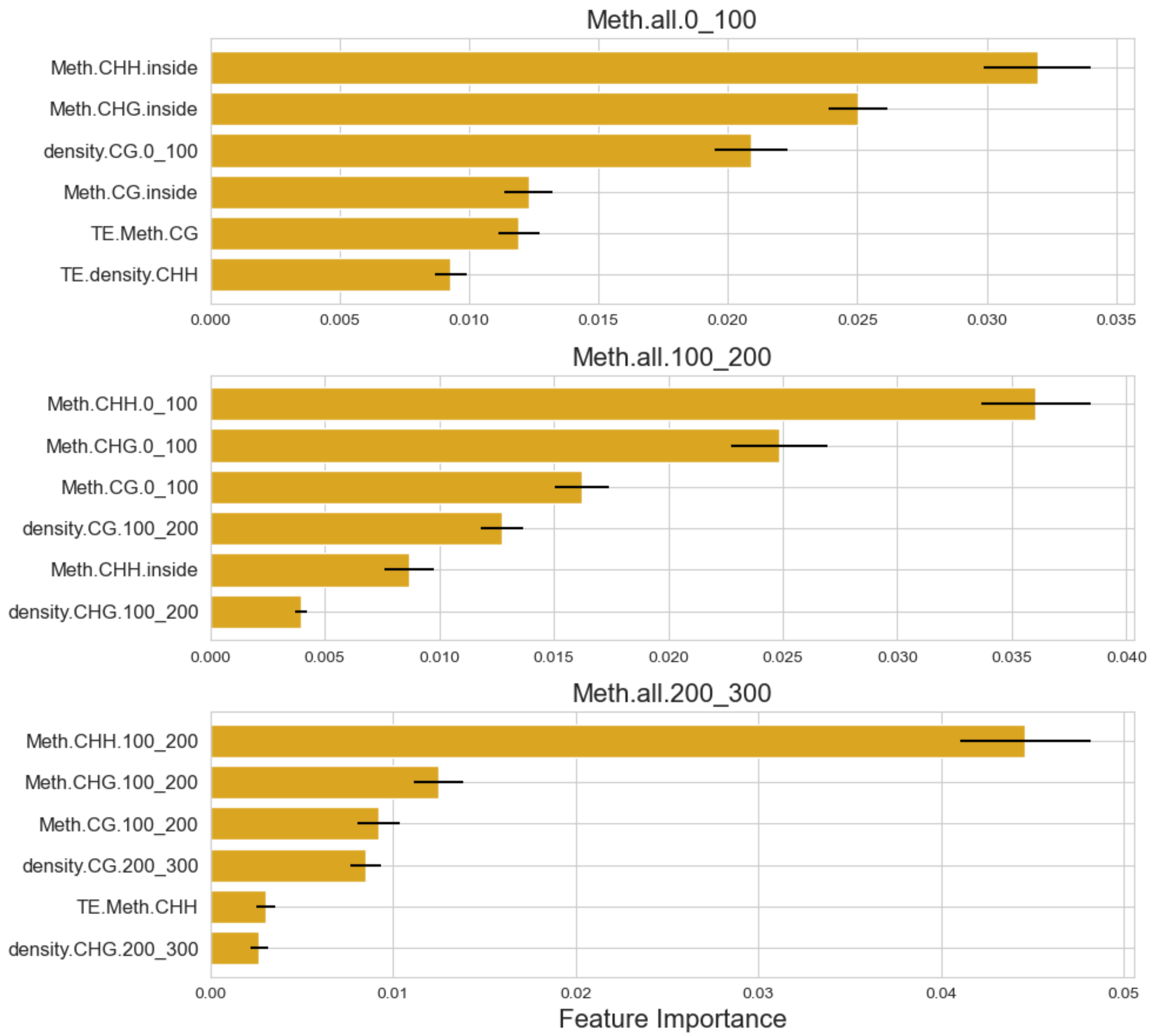
Impurity-based feature importances

* 10 independent runs with different random seeds



SHAP values

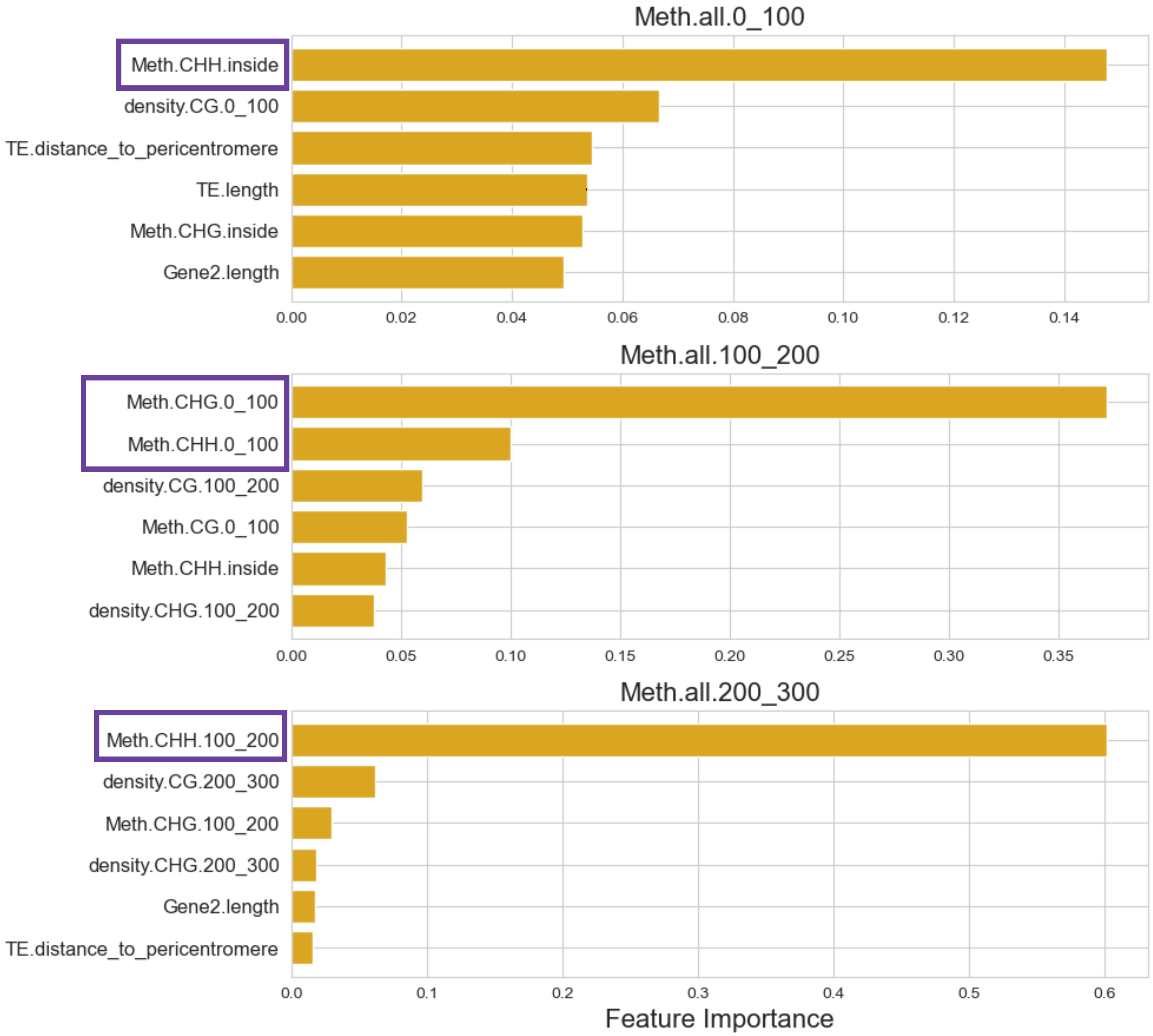
* 10 folds in a cross-validation manner



Prediction of methylation spreading

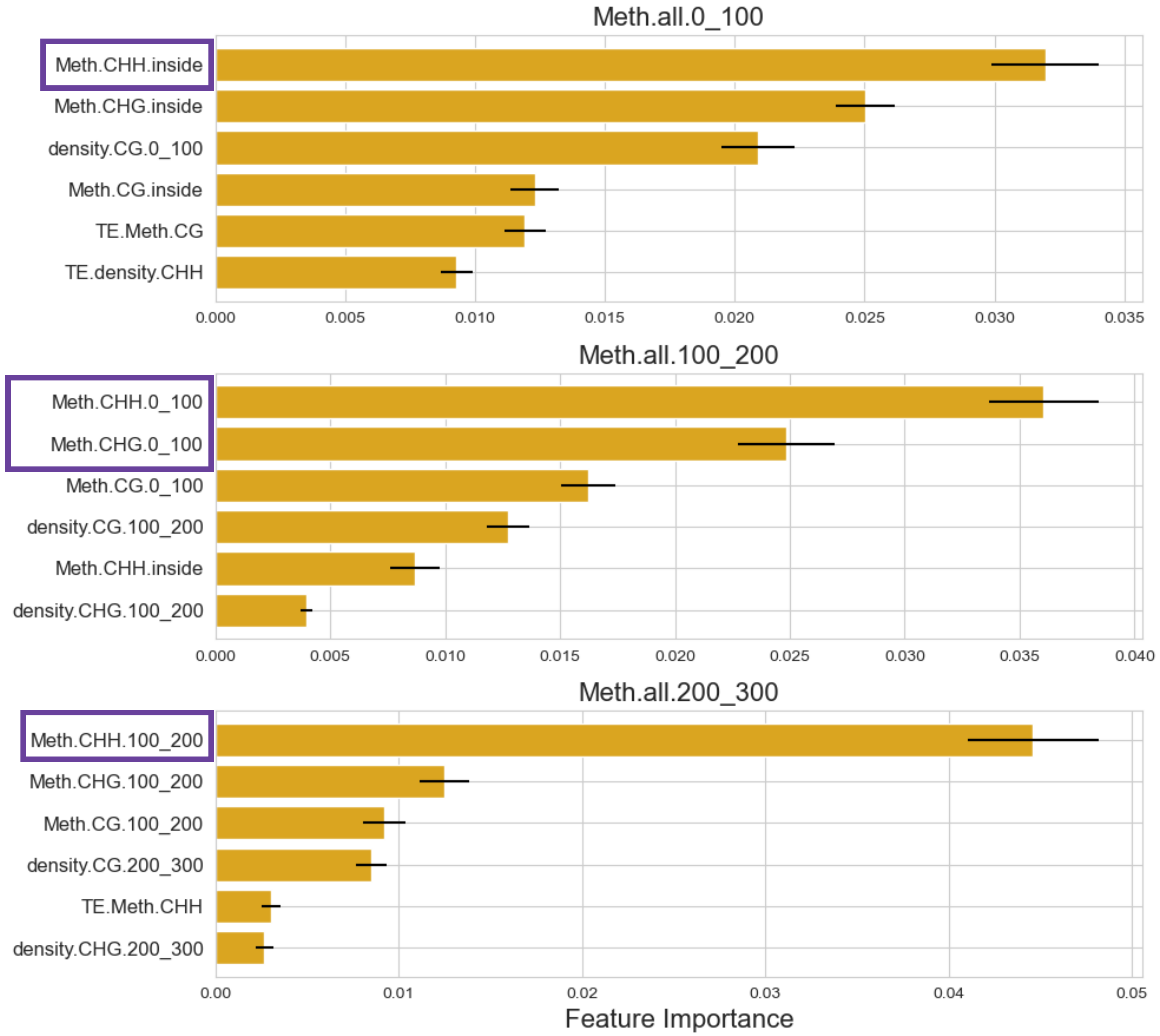
Impurity-based feature importances

* 10 independent runs with different random seeds



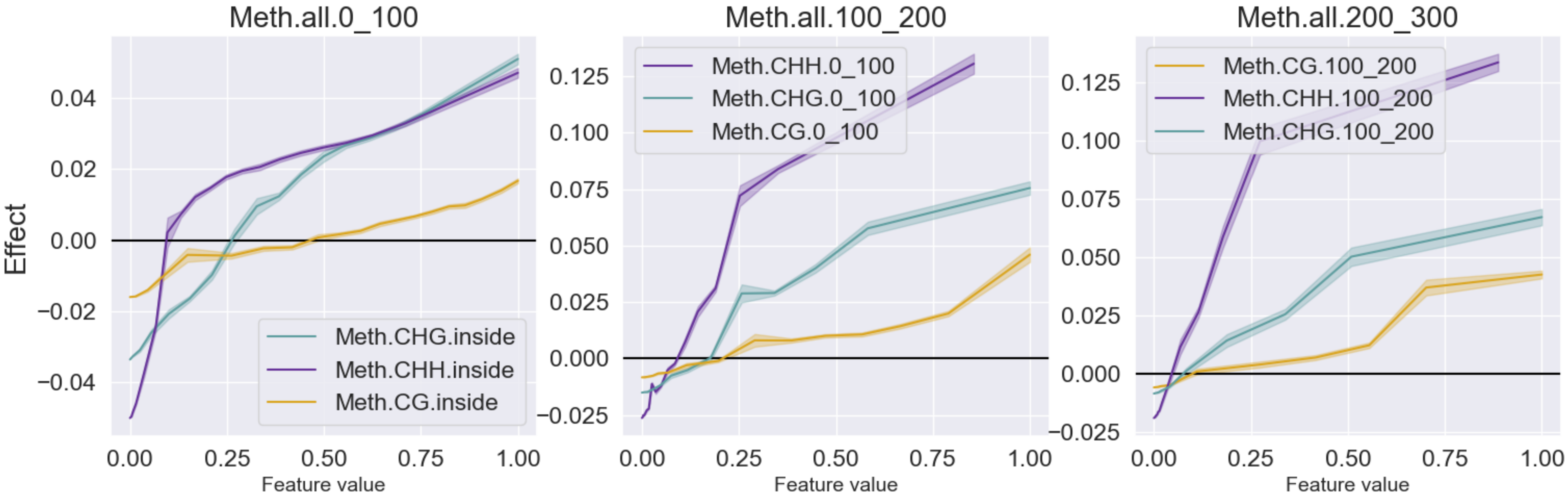
SHAP values

* 10 folds in a cross-validation manner



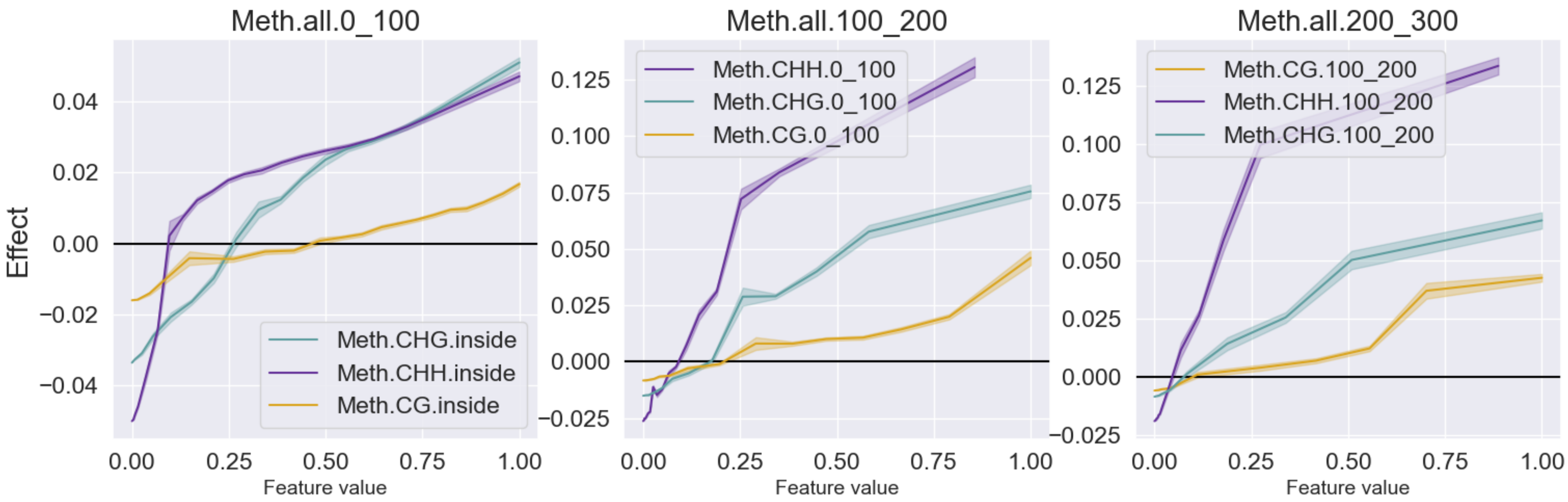
Prediction of methylation spreading

Accumulated Local Effects (ALE)



Prediction of methylation spreading

Accumulated Local Effects (ALE)

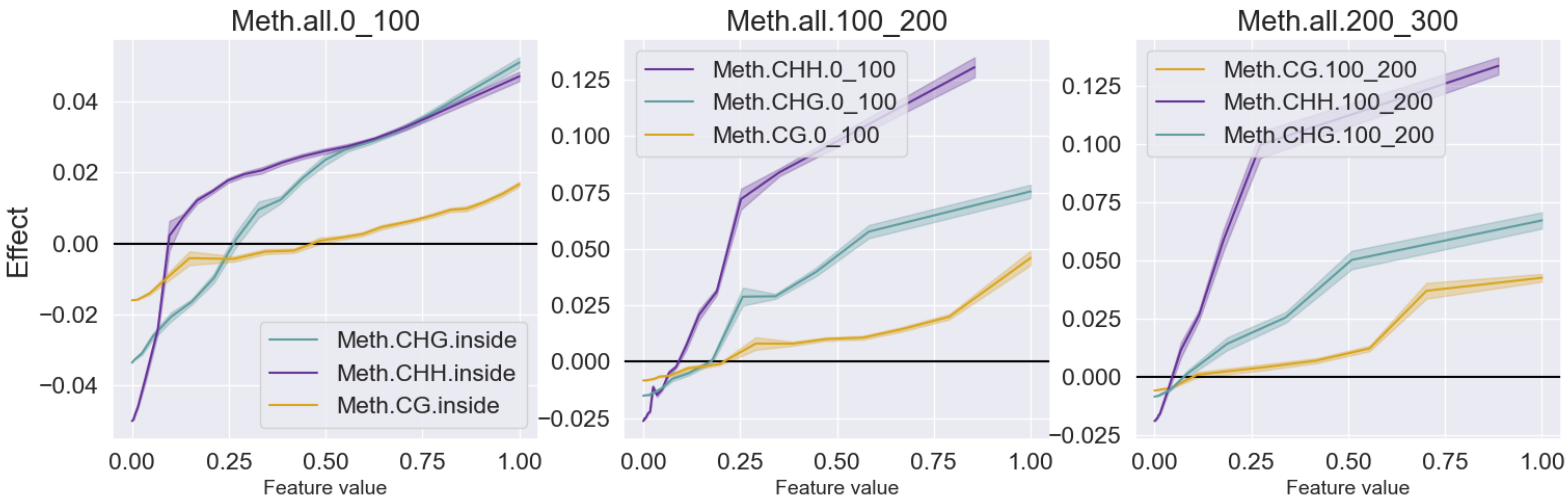


Conclusion:

- **Methylation of the TE inside ends** consistently comes as the most important feature with monotonous effect increase
- TE is **methylated on the ends** $\Longrightarrow$ more **likely to spread**

Prediction of methylation spreading

Accumulated Local Effects (ALE)



Conclusion:

- **Methylation of the TE inside ends** consistently comes as the most important feature with monotonous effect increase
- TE is **methylated on the ends** ⟹ more **likely to spread**

Question:

- What defines the **methylation of the TE inside ends**?

Prediction of inside methylation

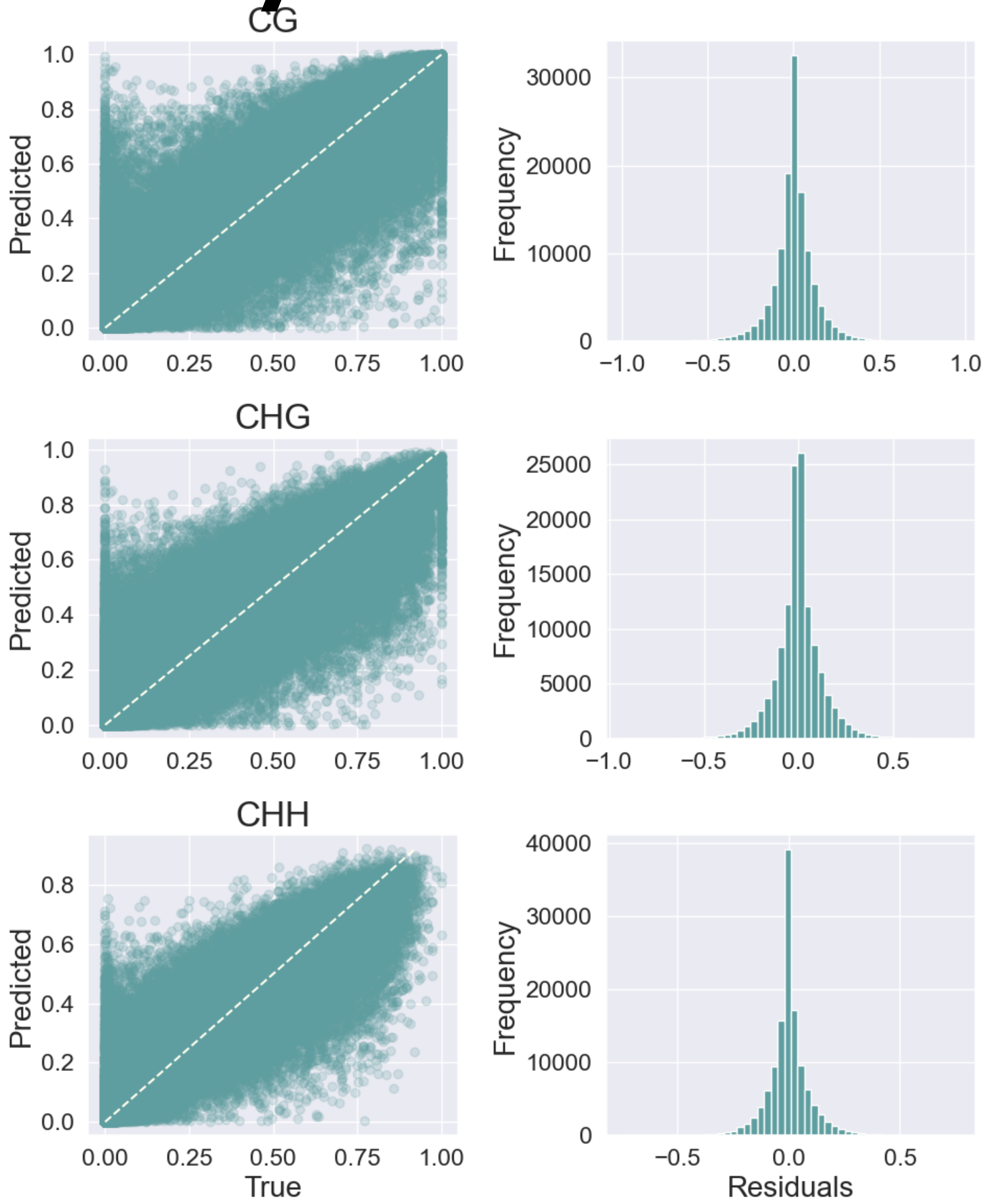
Features:

- **TE** (length, distance to pericentromere, superfamily, insertion frequency, divergence, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Methylation** in CG, CHG, CHH contexts (average genome-wide)
- **Densities** of CG, CHG, CHH contexts

Prediction of inside methylation

Features:

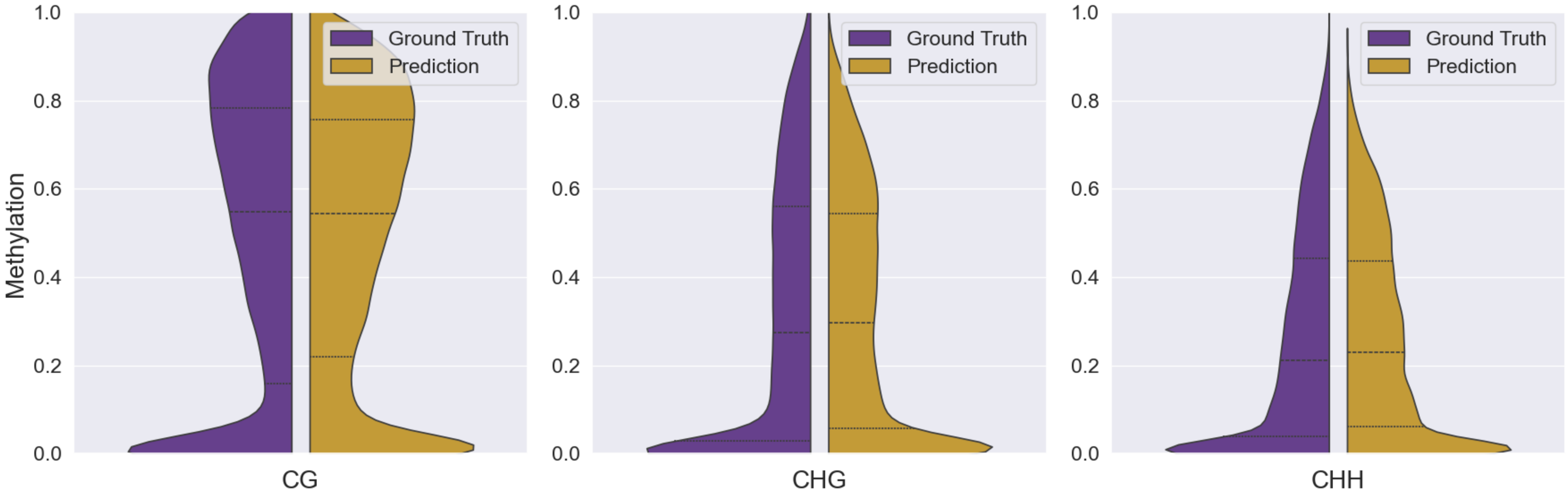
- **TE** (length, distance to pericentromere, superfamily, insertion frequency, divergence, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Methylation** in CG, CHG, CHH contexts (average genome-wide)
- **Densities** of CG, CHG, CHH contexts

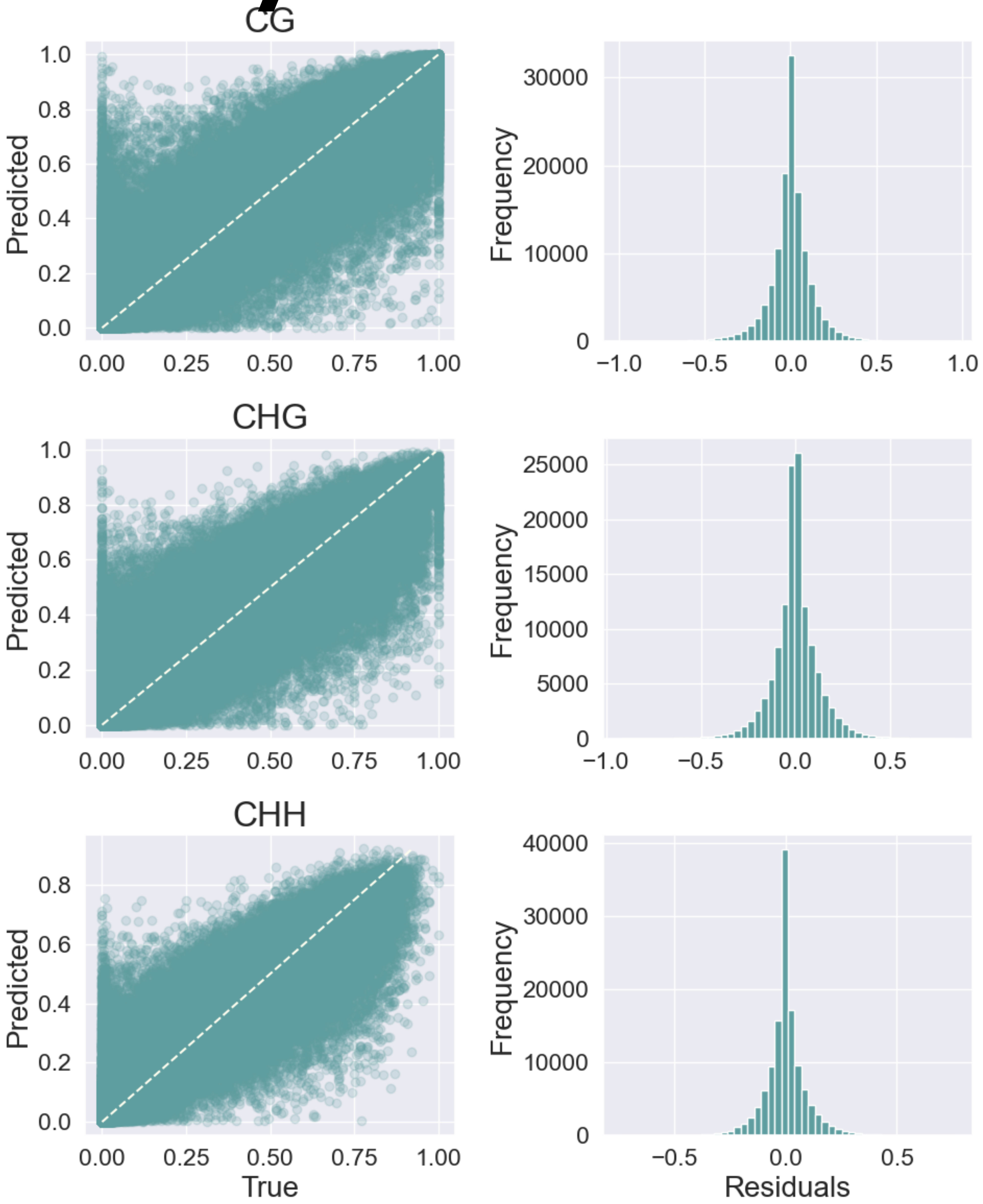


Prediction of inside methylation

Features:

- **TE** (length, distance to pericentromere, superfamily, insertion frequency, divergence, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Methylation** in CG, CHG, CHH contexts (average genome-wide)
- **Densities** of CG, CHG, CHH contexts

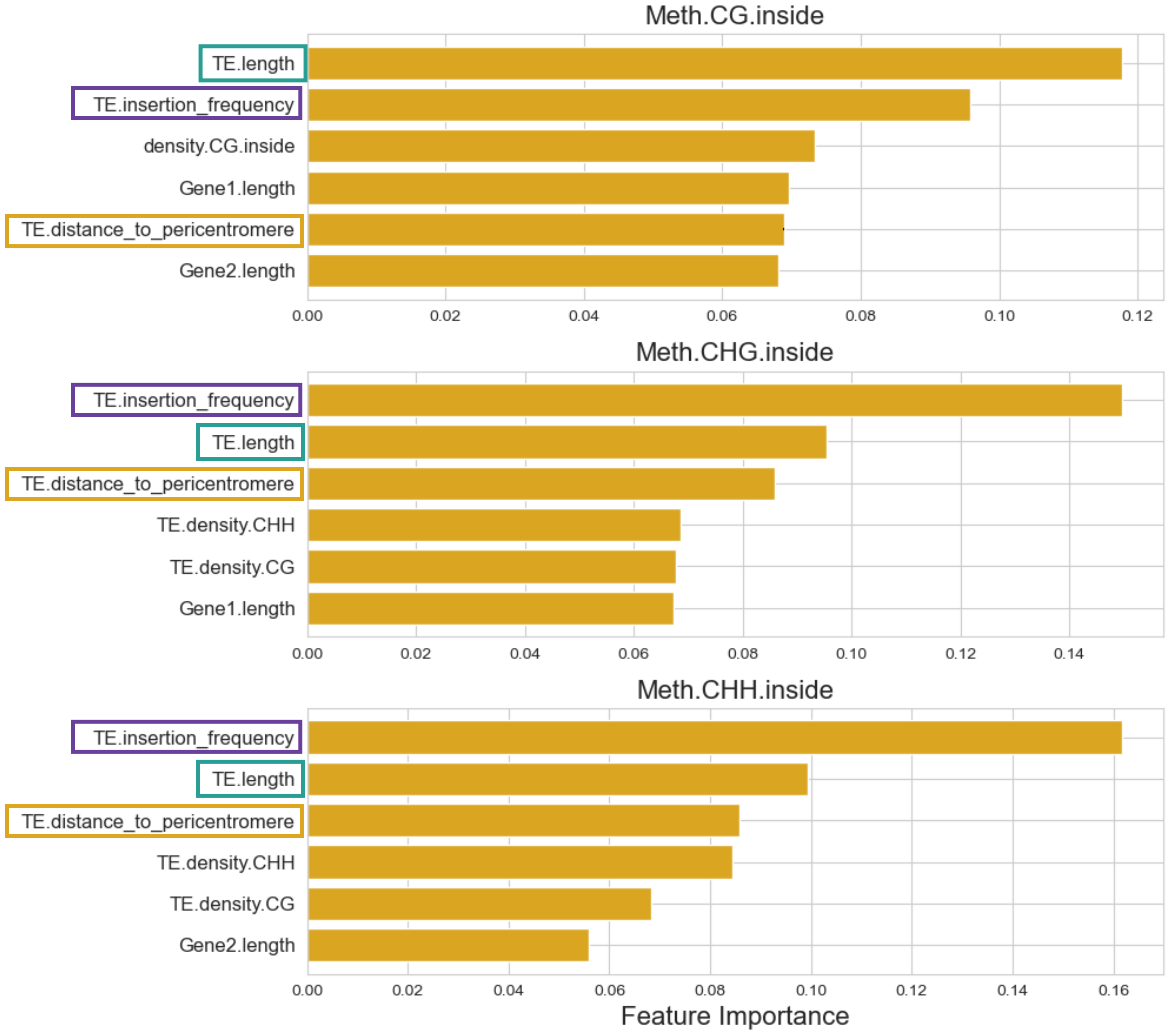




Prediction of inside methylation

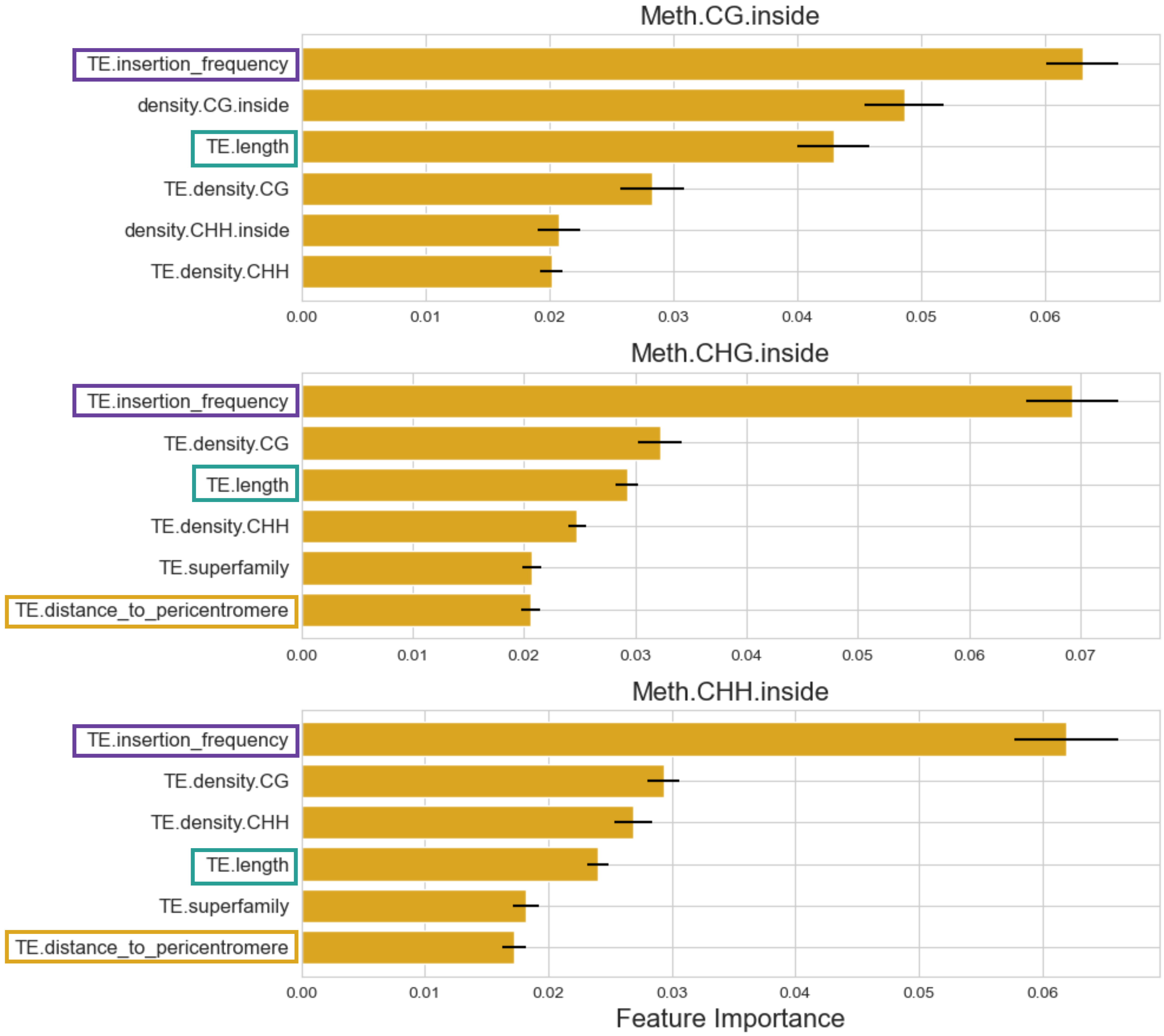
Impurity-based feature importances

* 10 independent runs with different random seeds



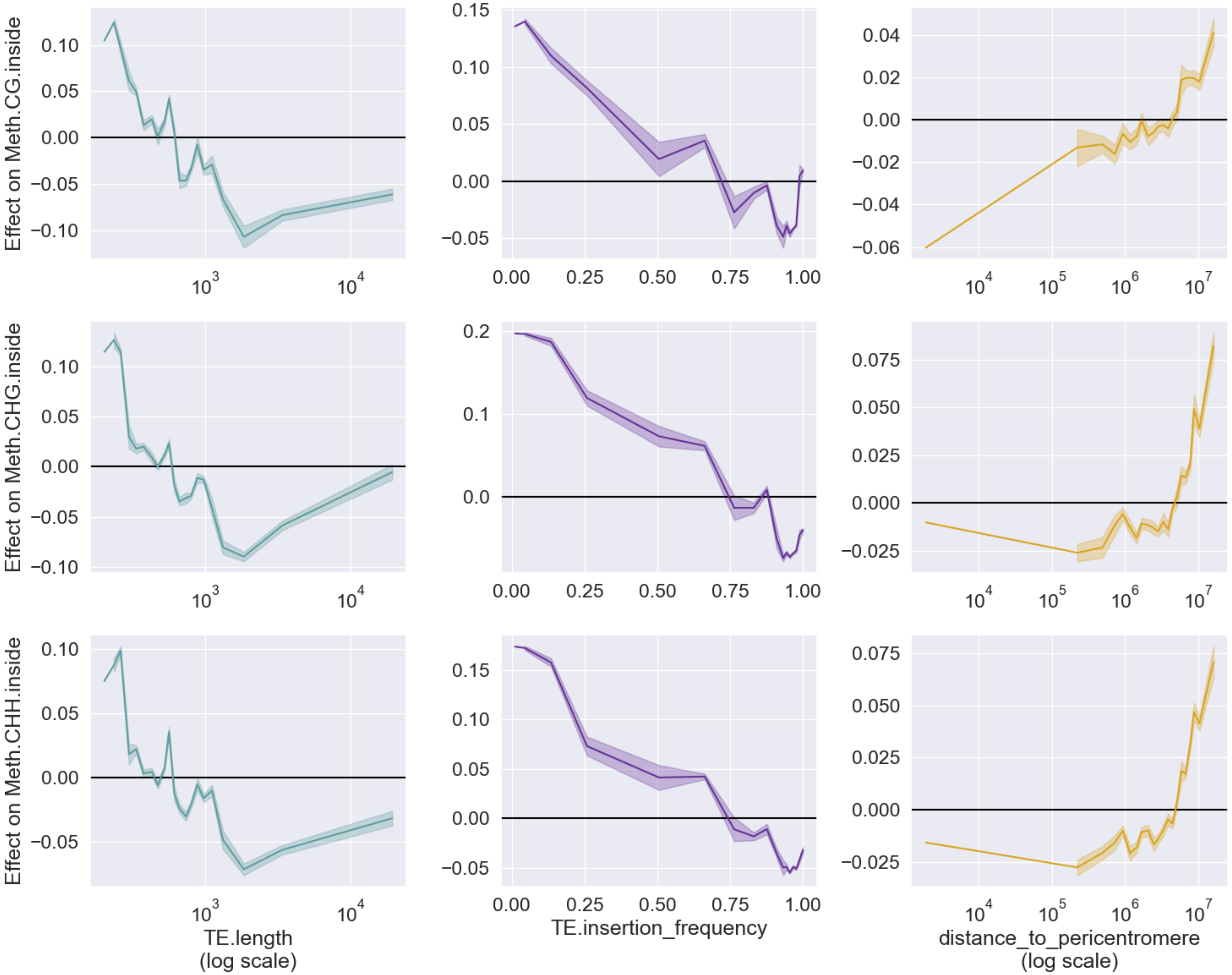
SHAP values

* 10 folds in a cross-validation manner



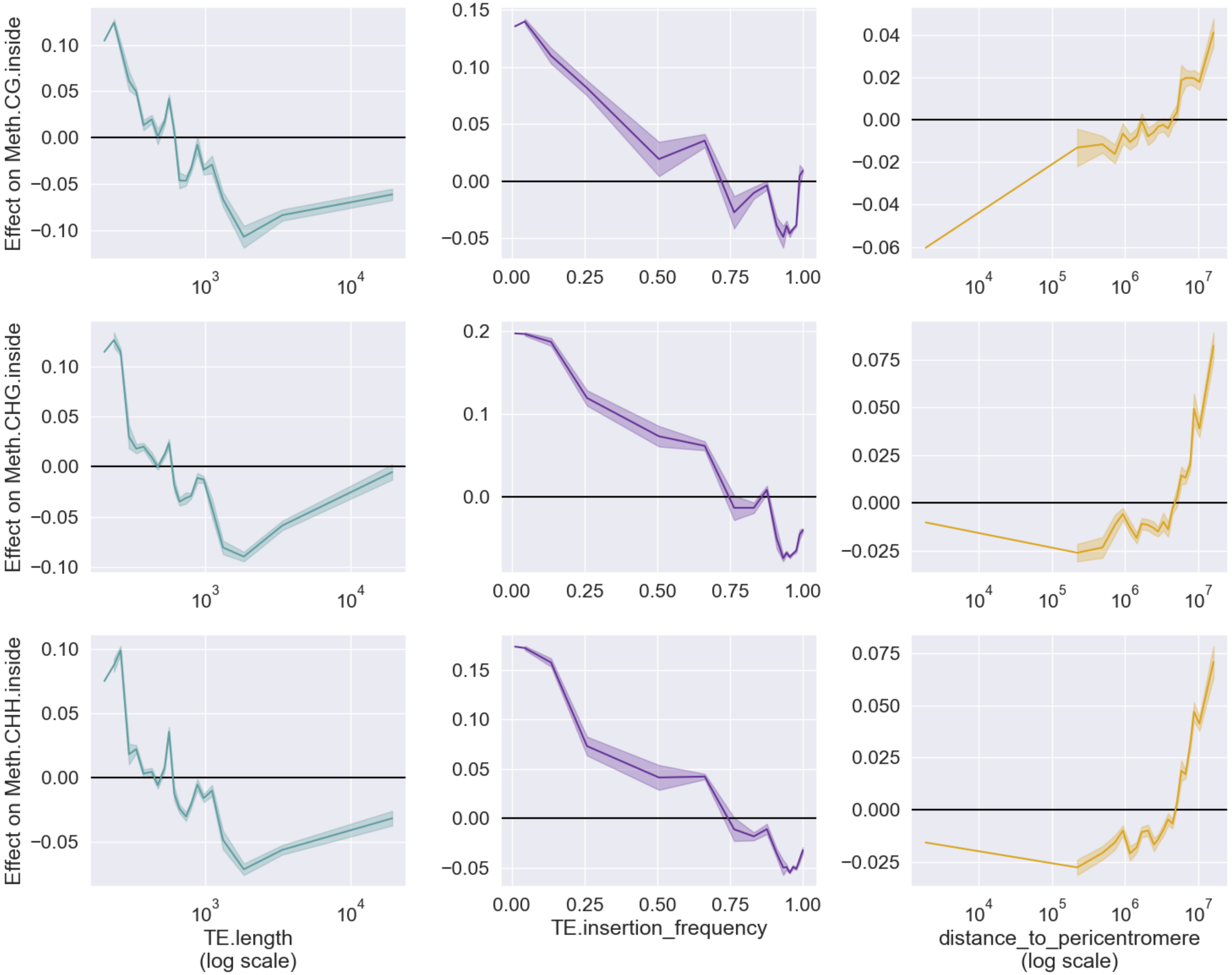
Prediction of inside methylation

Accumulated Local Effects



Prediction of inside methylation

Accumulated Local Effects



Conclusions

- Length of TE is unlikely to be a driving factor

  (rather a confounder)

- **Insertion frequency**:

  Rare (= new) TEs are targeted by methylation machinery

- **Distance to pericentromere**:

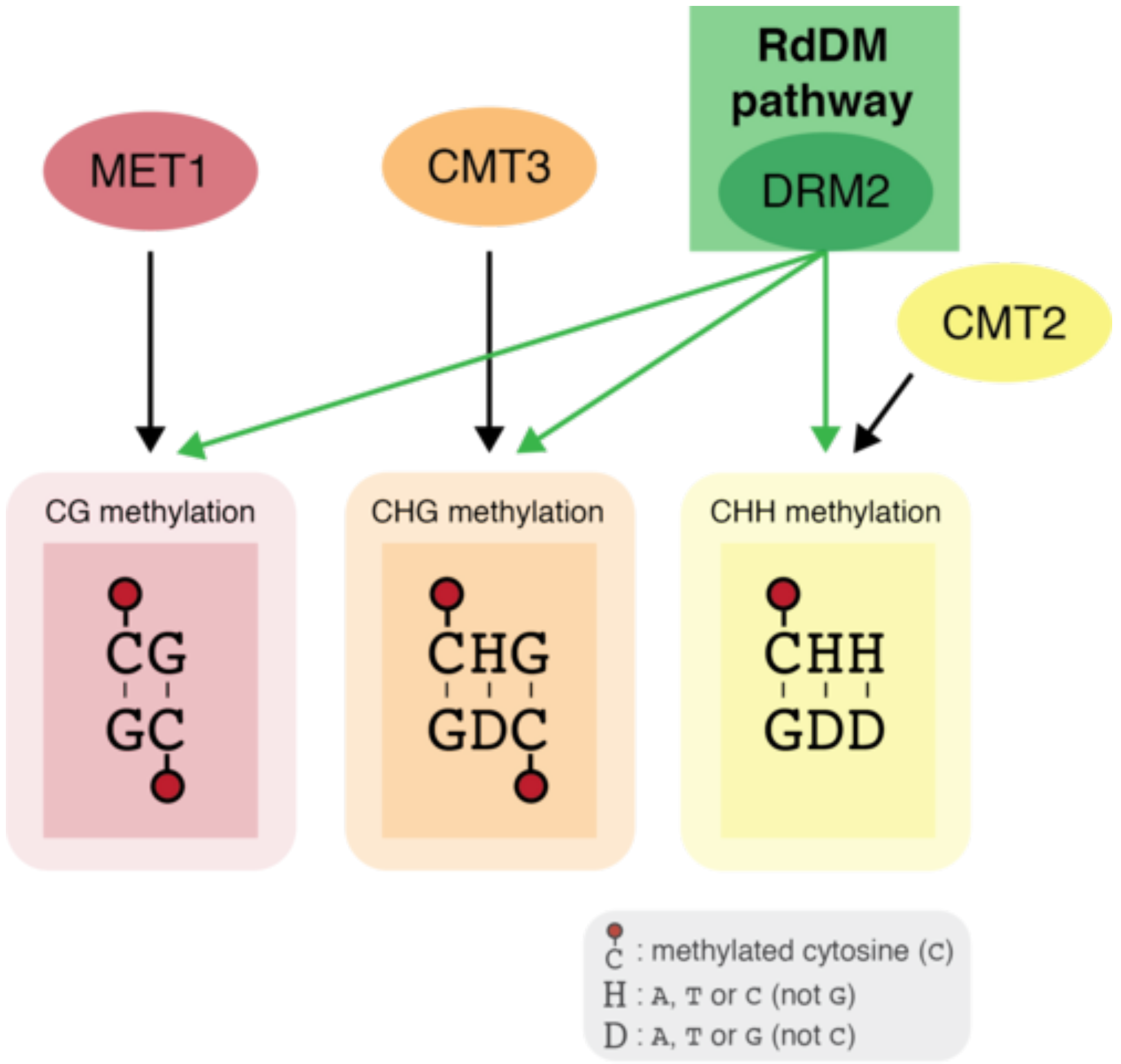
  More distant TEs are more likely to be targeted

Biological conclusions

- **The most important factors:**
  - methylation of insides in the CHG and CHH contexts
  - insertion frequency
  - distance to pericentromere

Biological conclusions

- **The most important factors:**
  - methylation of insides in the CHG and CHH contexts
  - insertion frequency
  - distance to pericentromere

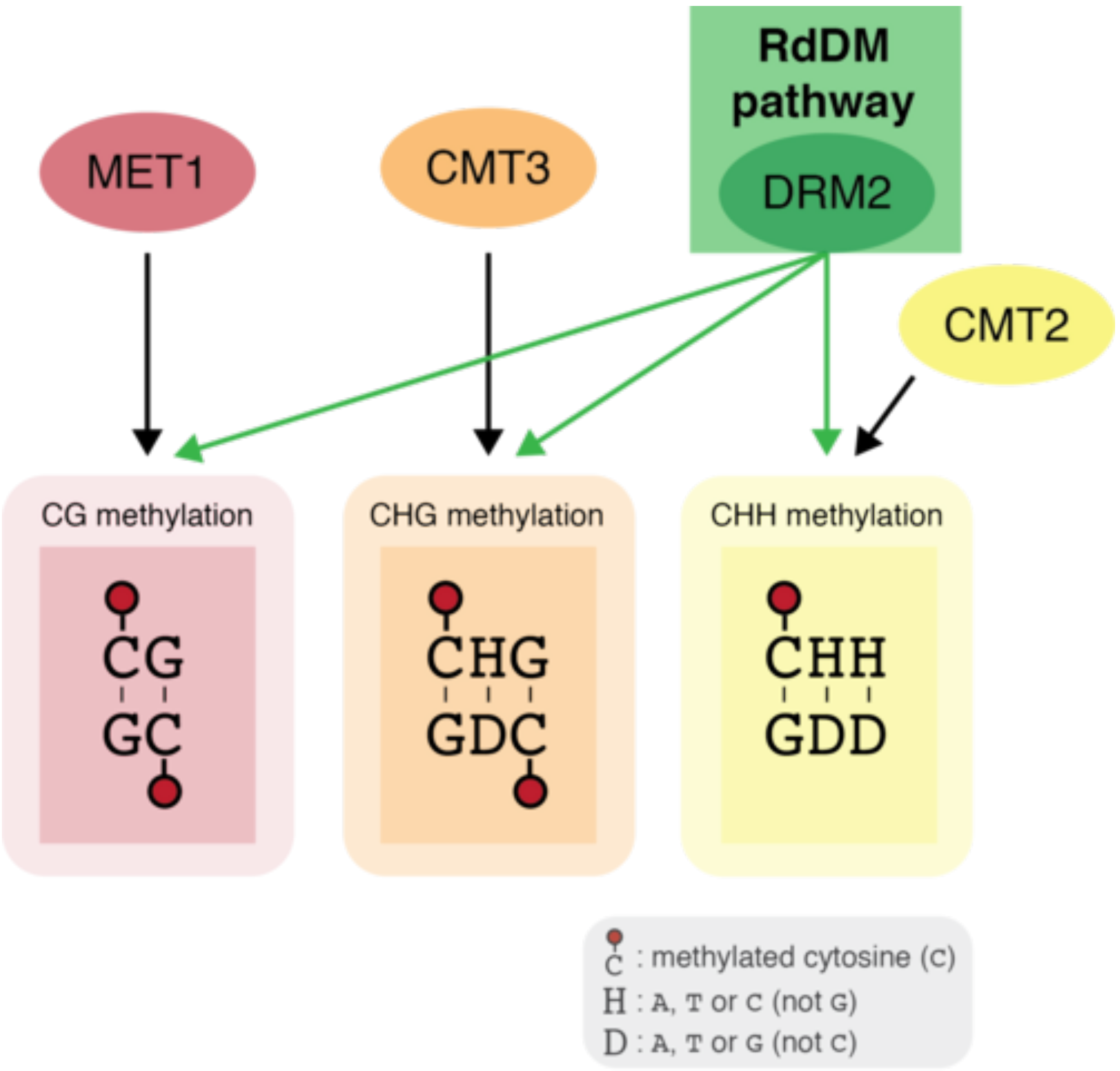


Wikipedia

- **Hypothesis:** the **RdDM** machinery is responsible for spreading
  - targets all contexts
  - the <u>only</u> pathway capable of adding DNA methylation *de novo*
  - targets rather chromosome arms

Biological conclusions

- **The most important factors:**
  - methylation of insides in the CHG and CHH contexts
  - insertion frequency
  - distance to pericentromere



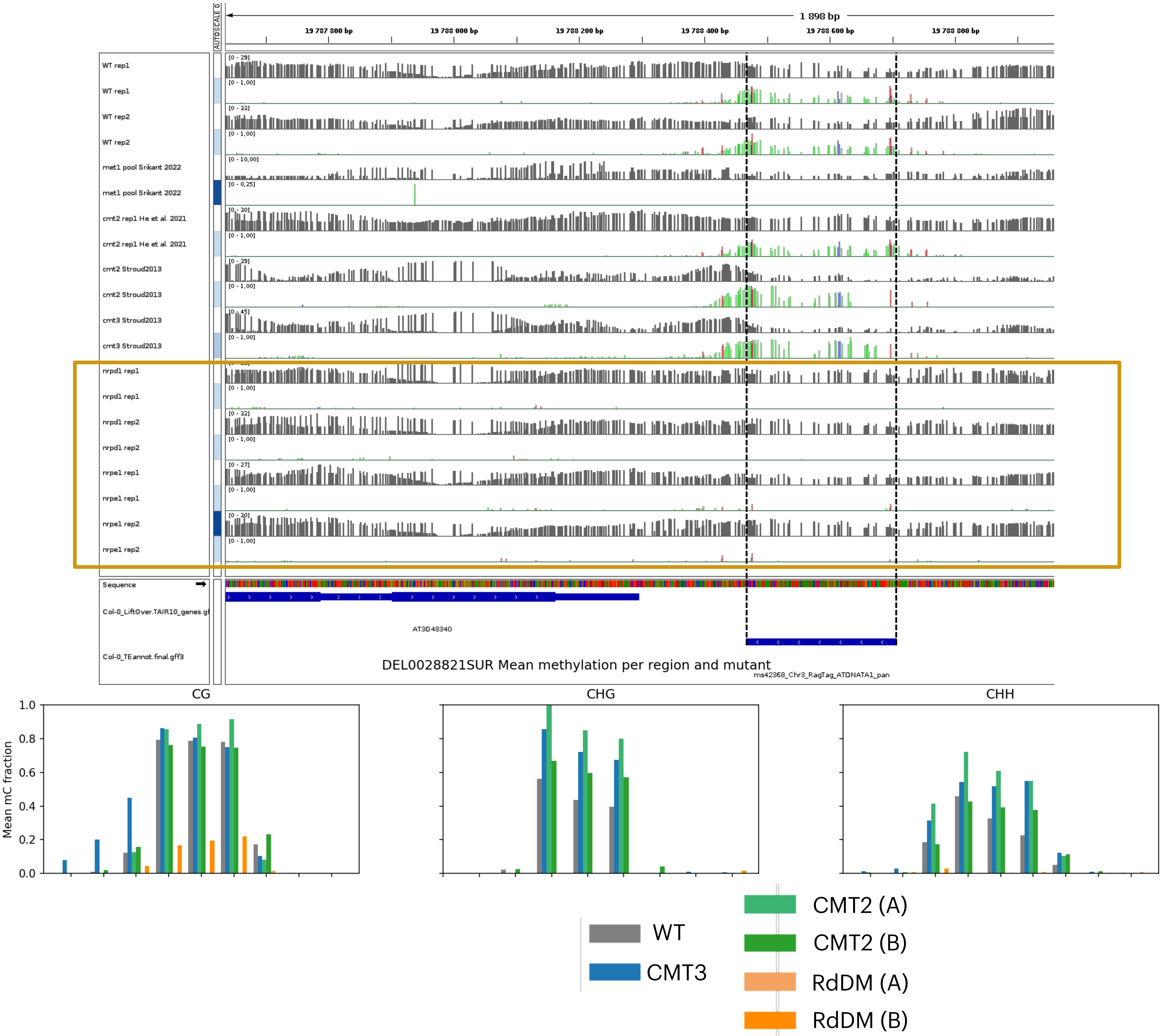
Wikipedia

- **Hypothesis:** the **RdDM** machinery is responsible for spreading
  - targets all contexts
  - the only pathway capable of adding DNA methylation *de novo*
  - targets rather chromosome arms

- **Test:** mutants of Col-0 strain of *A. Thaliana* where different methylation pathways are knocked out

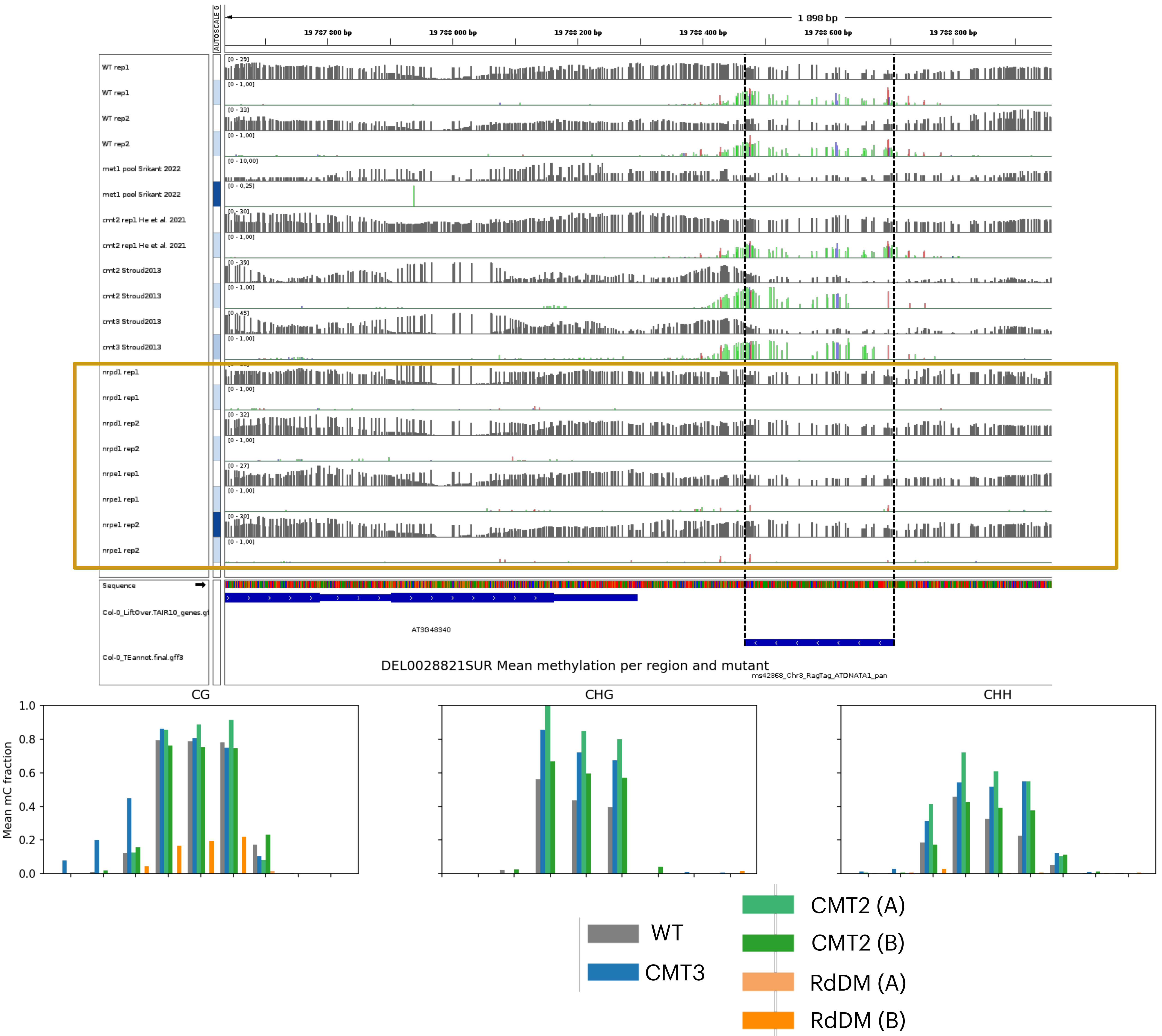
Biological confirmation

DEL0028821SUR

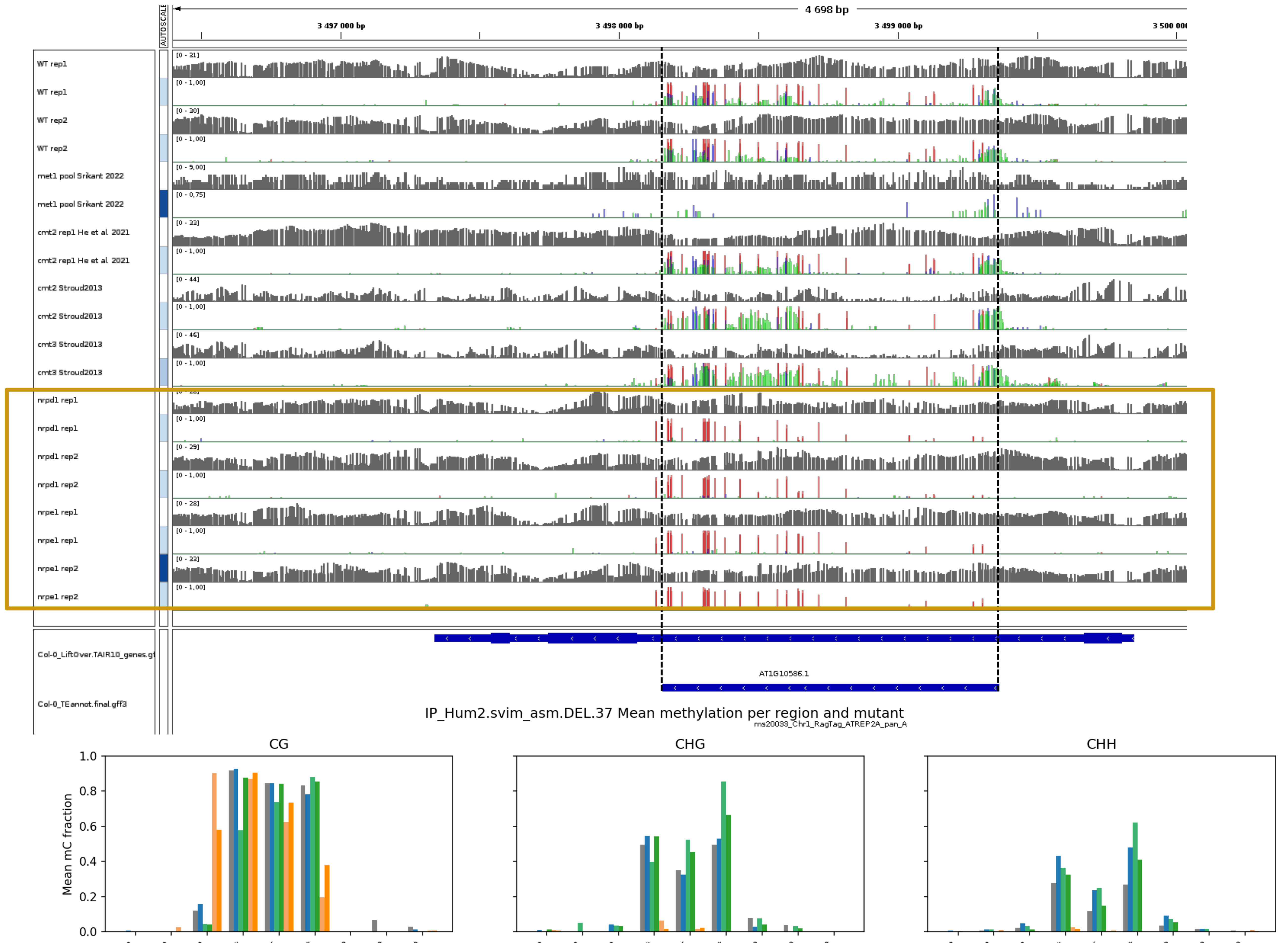


Biological confirmation

DEL0028821SUR

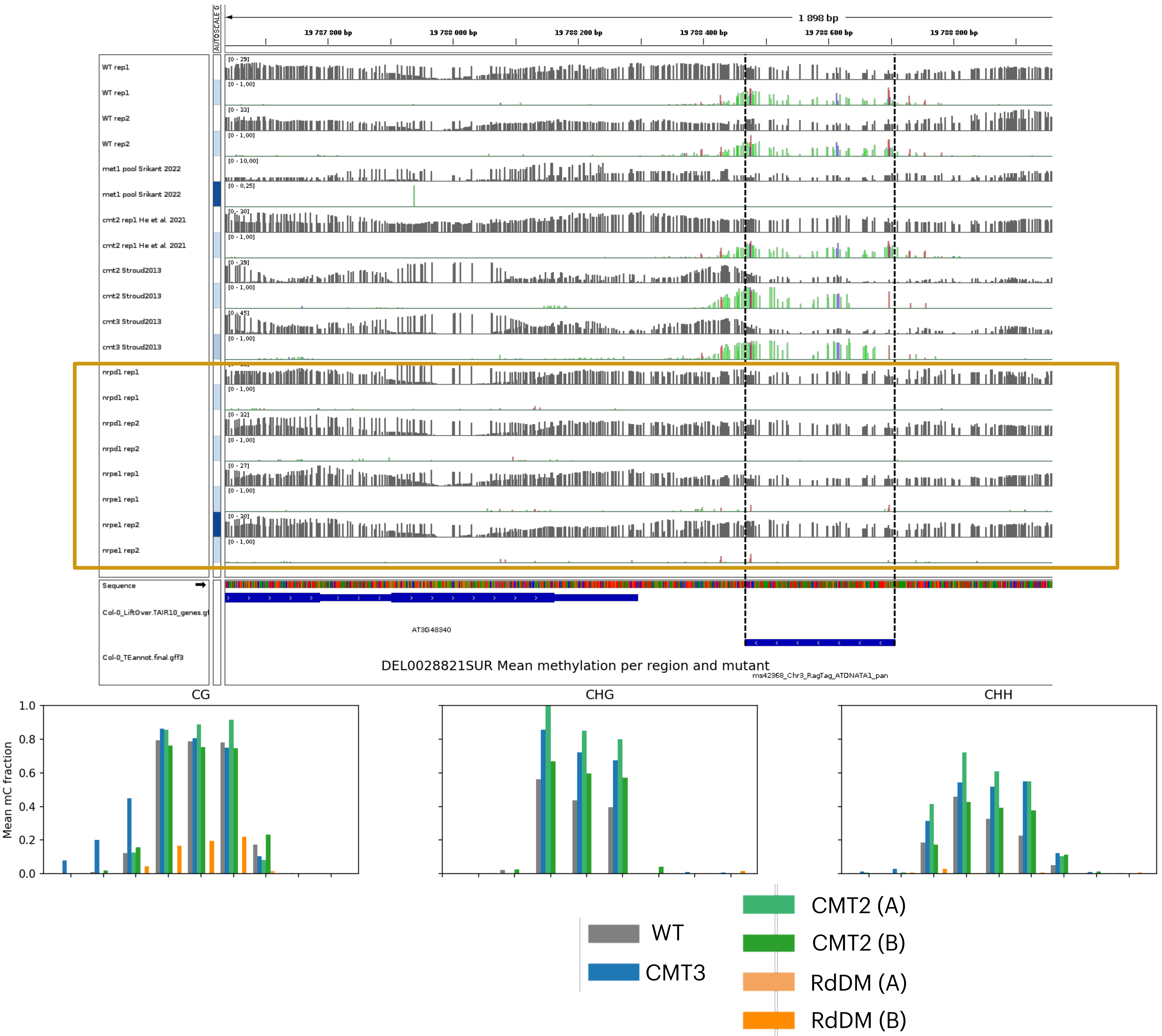


IP_Hum2.svim_asm.DEL.37

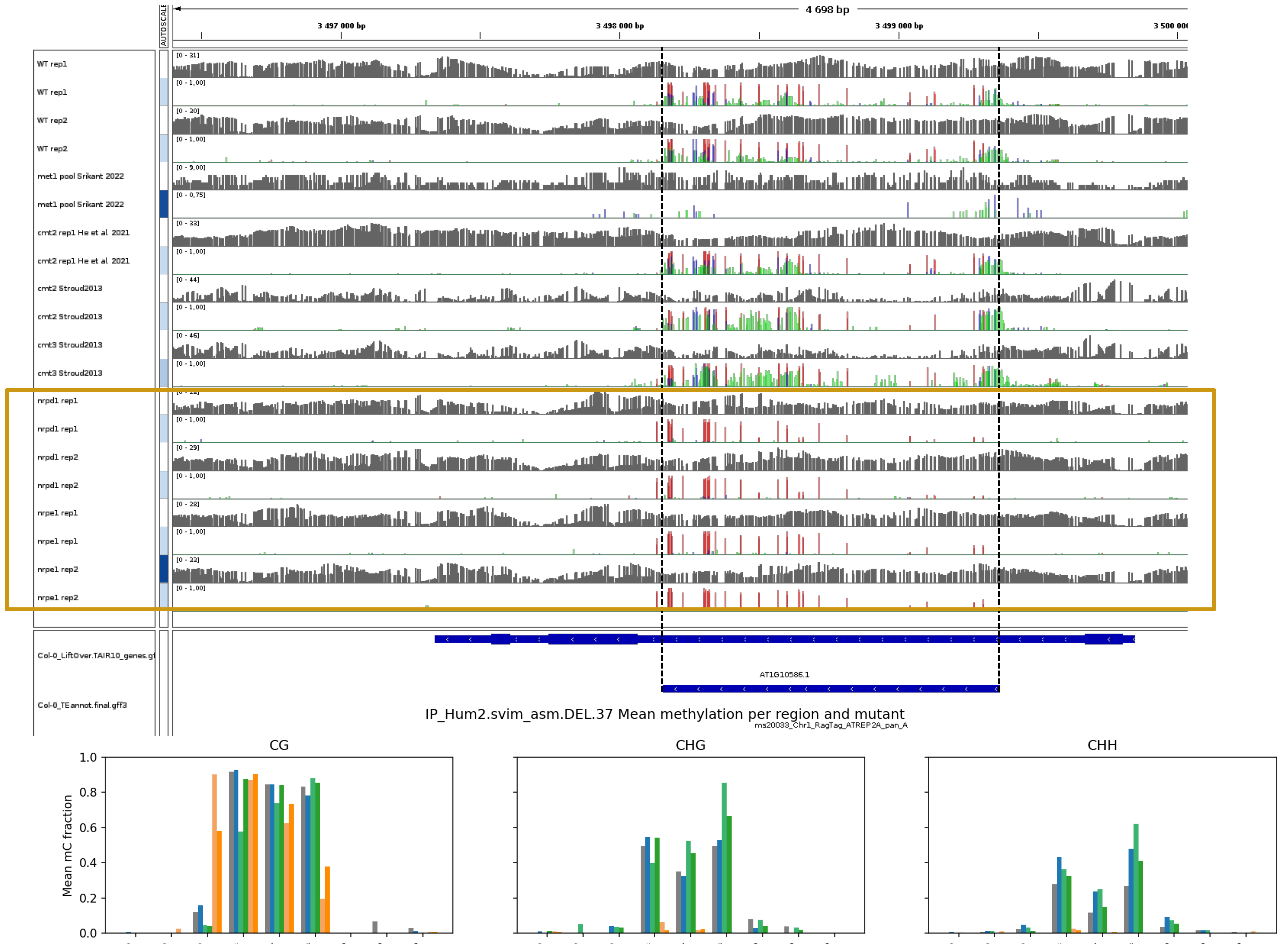


Biological confirmation

DEL0028821SUR



IP_Hum2.svim_asm.DEL.37



CHG and CHH methylation (and spreading!) disappear in RdDM mutants

Conclusions

- An example of the workflow:

biological phenomenon $\Longrightarrow$ machine learning model $\Longrightarrow$ explanations $\Longrightarrow$ real biological mechanisms

Conclusions

- An example of the workflow:

  biological phenomenon $\Longrightarrow$ **machine learning model** $\Longrightarrow$ **explanations** $\Longrightarrow$ **real biological mechanisms**

- Different **explainability tools** have been explored, and they provide **consistent conclusions**

Conclusions

- An example of the workflow:

  biological phenomenon ⟹ machine learning model ⟹ explanations ⟹ real biological mechanisms

- Different **explainability tools** have been explored, and they provide **consistent conclusions**

- Potential actor (**RdDM**) is identified

Conclusions

- An example of the workflow:

biological phenomenon $\Longrightarrow$ **machine learning model** $\Longrightarrow$ **explanations** $\Longrightarrow$ **real biological mechanisms**

- Different **explainability tools** have been explored, and they provide **consistent conclusions**

- Potential actor (**RdDM**) is identified

- We understand better one of the factors to **explain GWAS signals**

Acknowledgements

CBIO

Chloé-Agathe Azencott

Marie Dogo

Jérémy Cohen

Sylvain Cailloud

(and everyone else)

IBENS

Vincent Colot

Pierre Baduel

Louna De Oliveira

Aurélien Petit

(and everyone else)

MINES PARIS | PSL | CBIO

anr

IBENS Institut de Biologie de l'Ecole Normale Supérieure

ENS | PSL