From biology to machine learning and back: understanding transposable element methylation and its phenotypic effects

Katia Antonenko

[✉️] *ekaterina.antonenko@minesparis.psl.eu*

Paris Postdocs Seminar — Institut Imagine — 21 January 2026

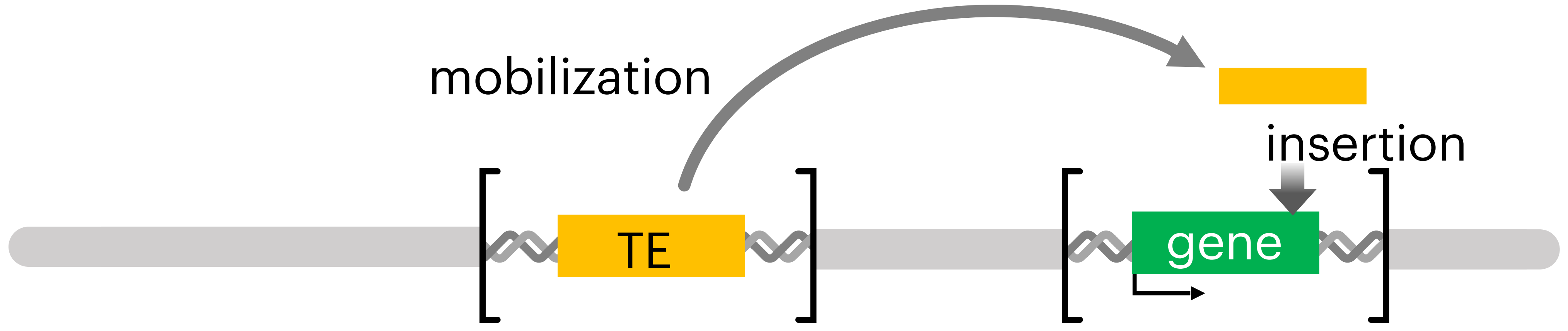
PÉpiTE team

Contents

- Background: transposable elements and methylation
- Part I: analysis of our TE cohort
- Part II: understanding methylation
- Part III: associations with gene expression
- Conclusions

Transposable Elements

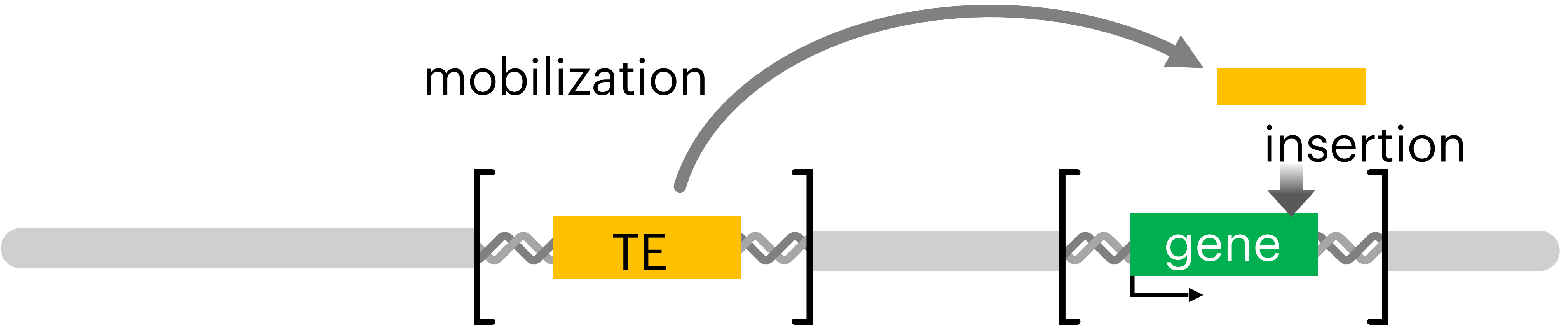
- Transposable Elements (TEs, "jumping genes") are an important source of mutations



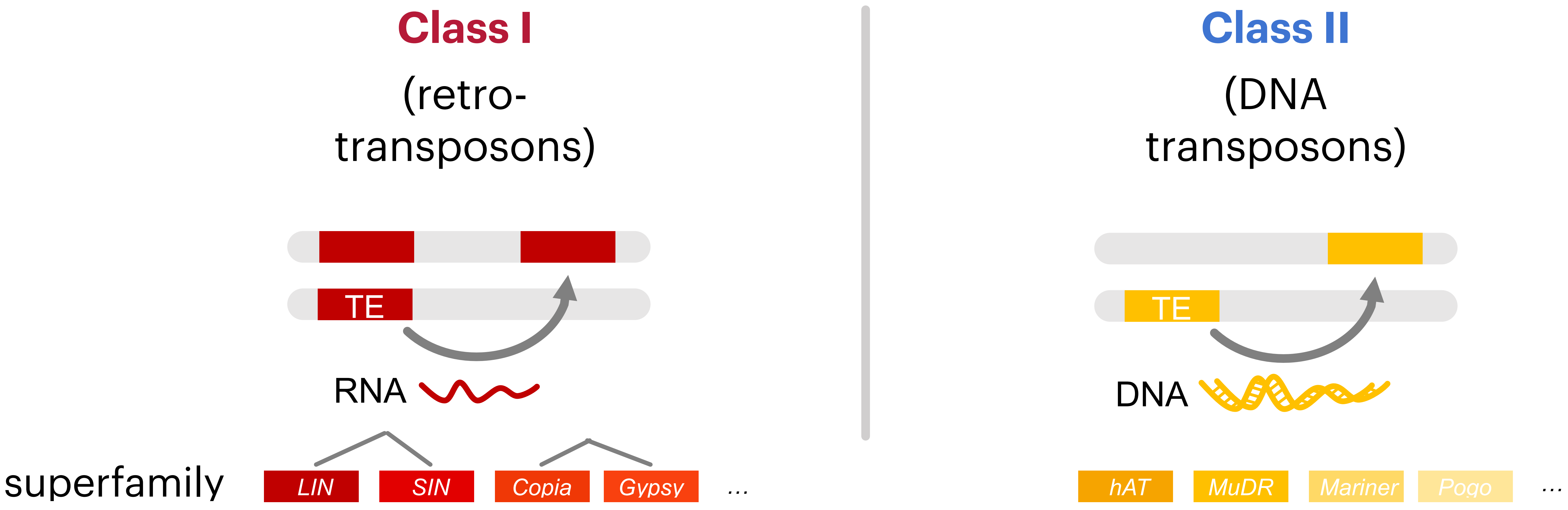
Barbara McClintock
Nobel prize 1983

Transposable Elements

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



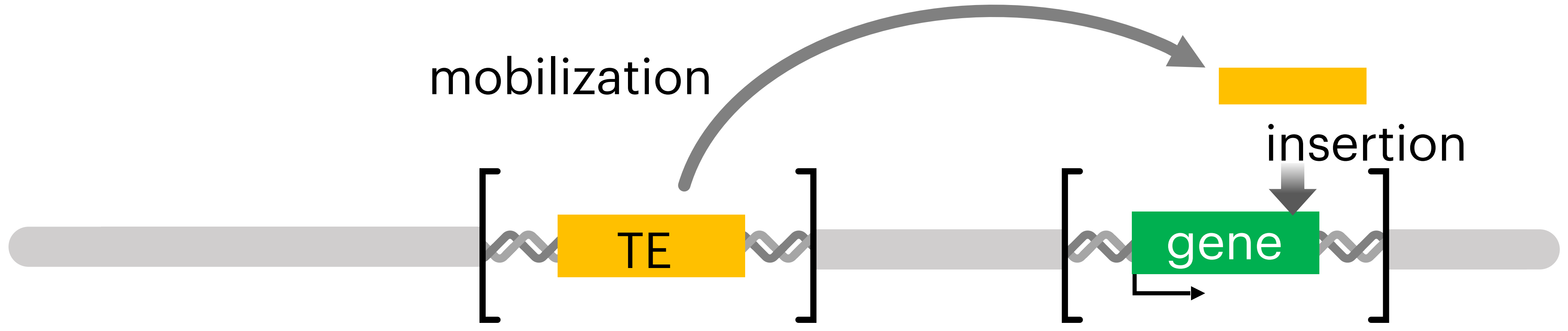
- TEs transpose by cut-and-paste or copy-and-paste mechanisms



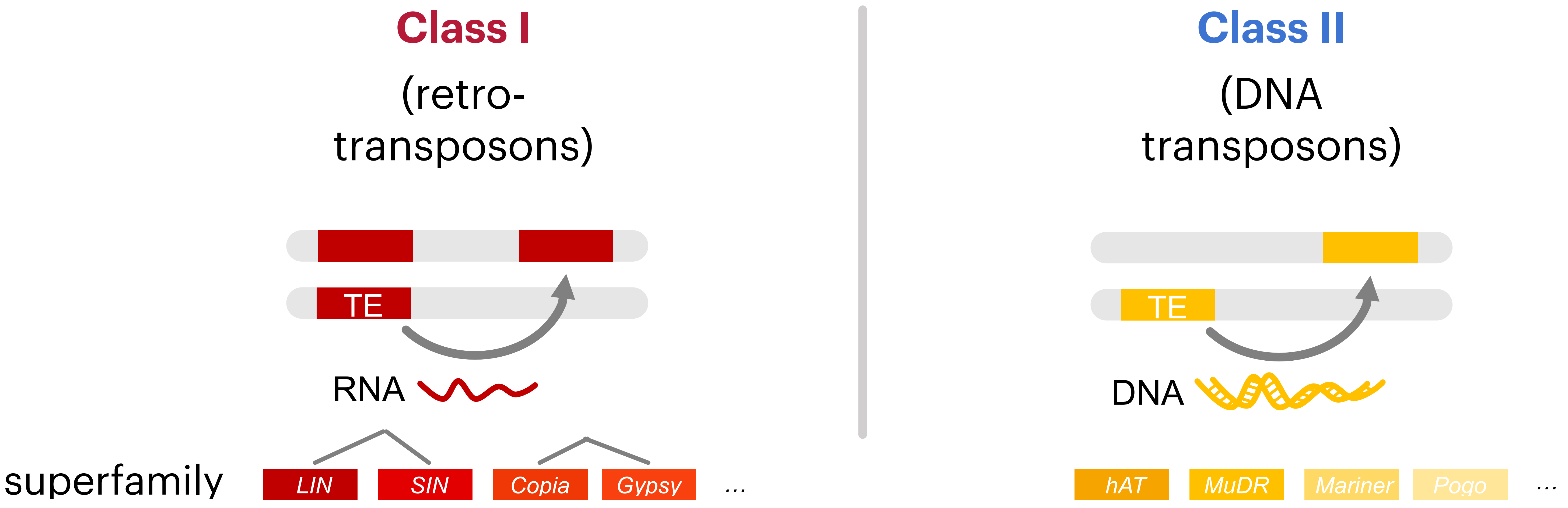
Barbara McClintock
Nobel prize 1983

Transposable Elements

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



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



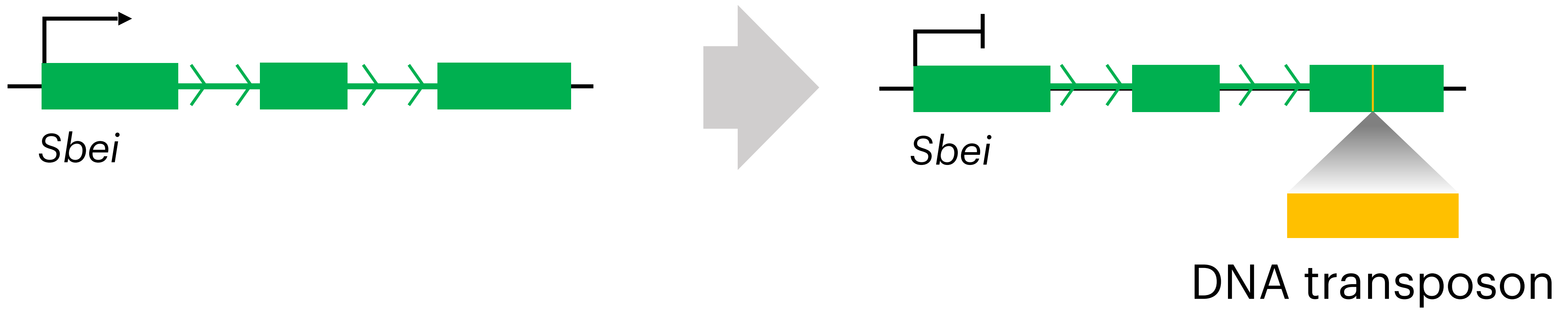
Barbara McClintock
Nobel prize 1983

- BUT: most TEs are degraded and do not transpose

Transposable Elements

Mutations may be **deleterious**...

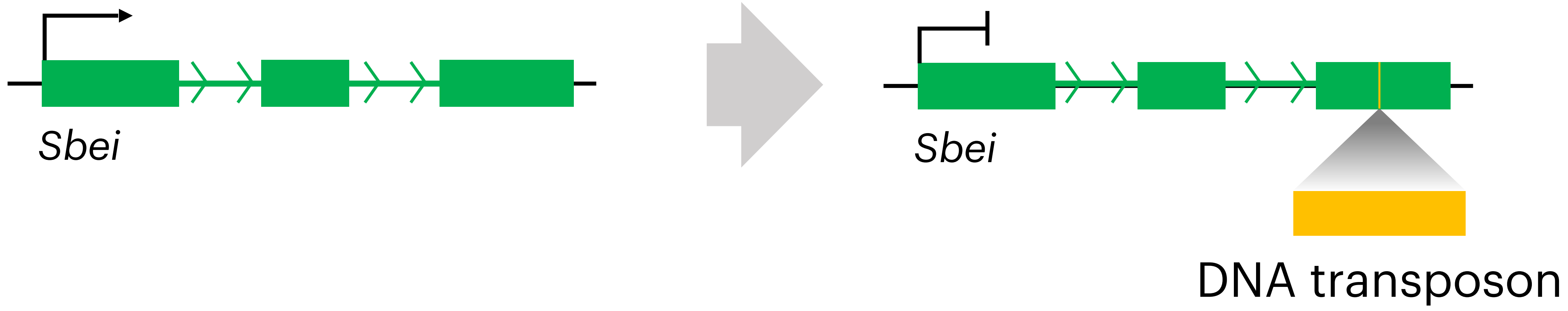
Bhattacharyya *et al. Cell* 1990



Transposable Elements

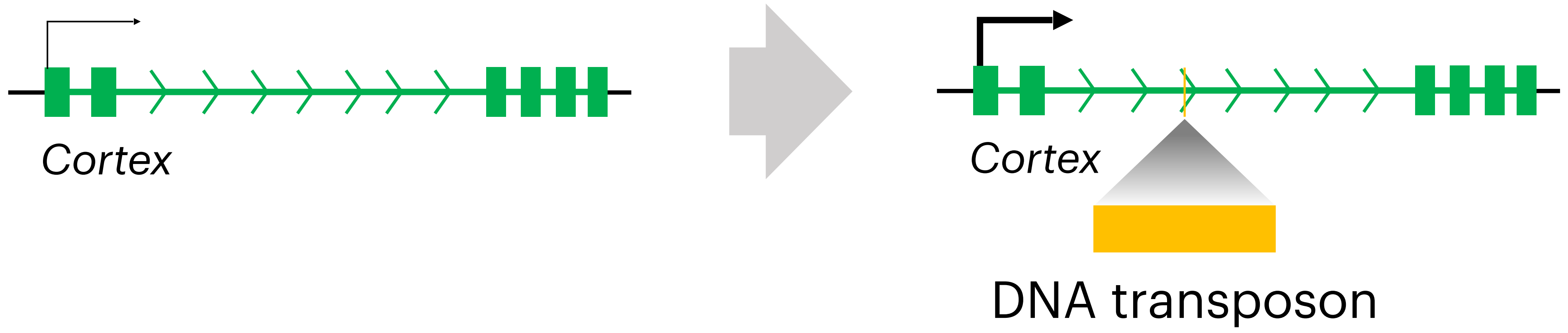
Mutations may be **deleterious**...

Bhattacharyya *et al.* *Cell* 1990

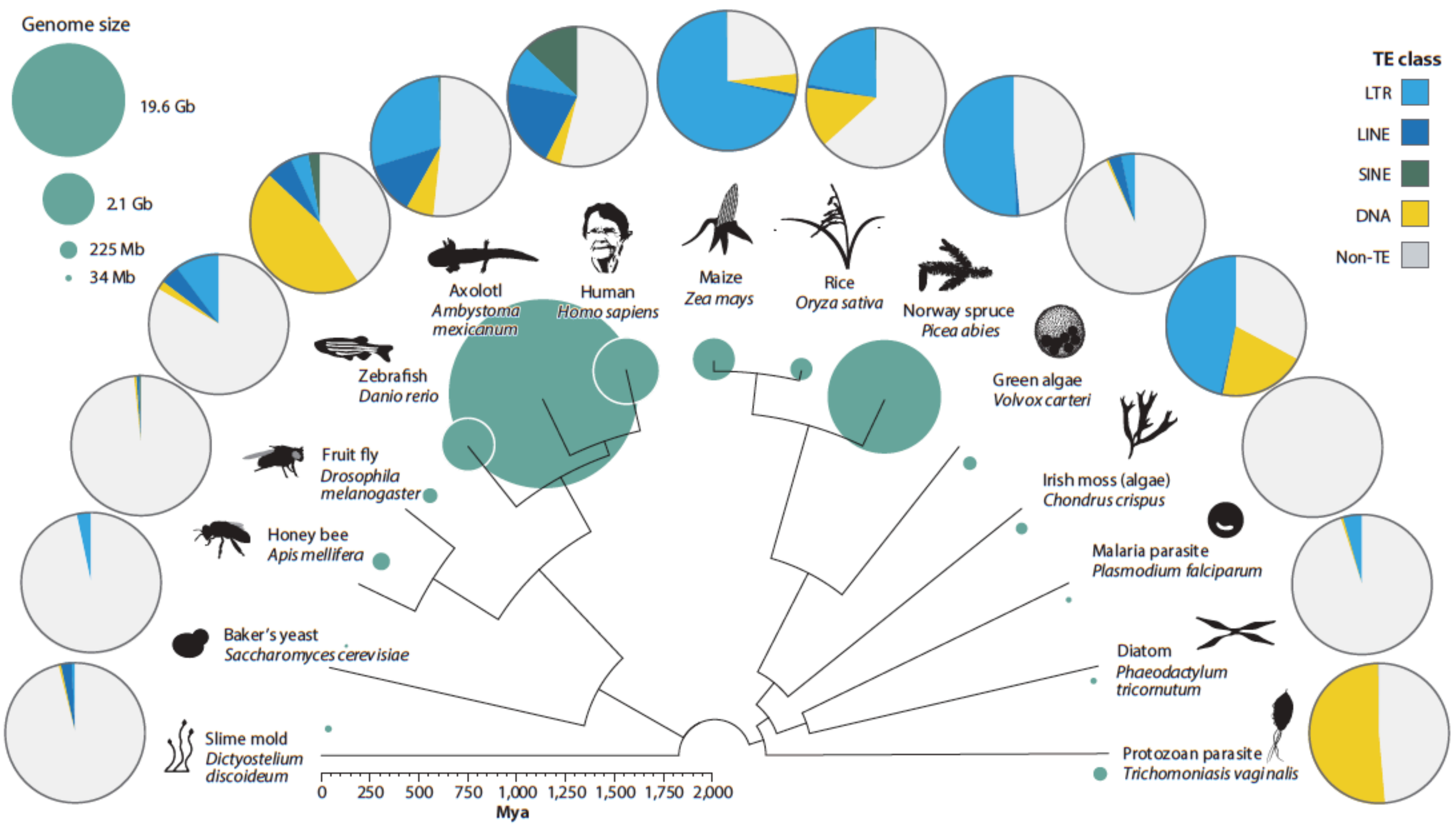


...yet sometimes **adaptive**

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



Transposable Elements



Wells & Feschotte, *Annual Review of Genetics 2020*

Epigenetic Regulation of Transposable Elements

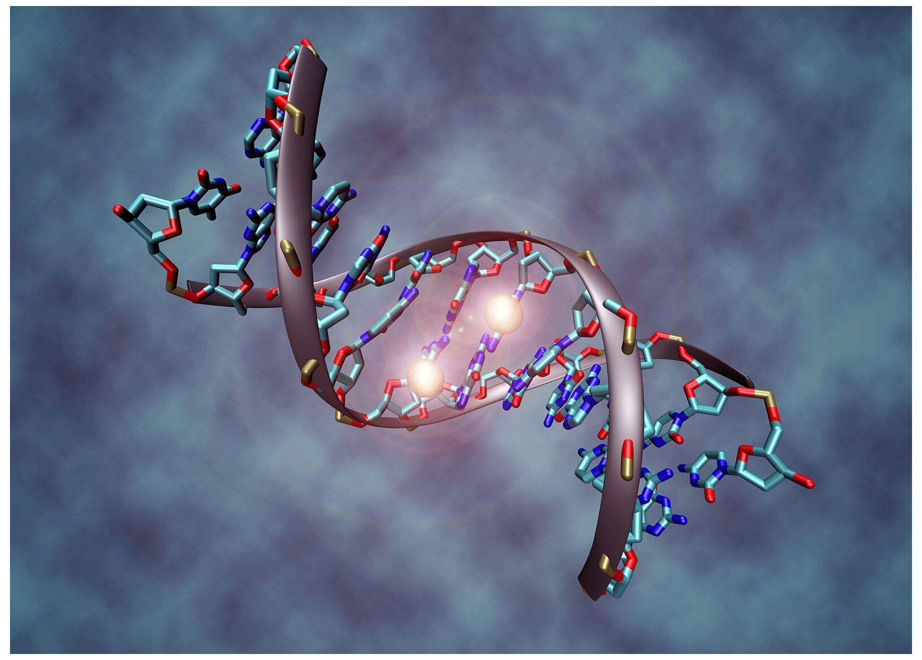
DNA methylation:

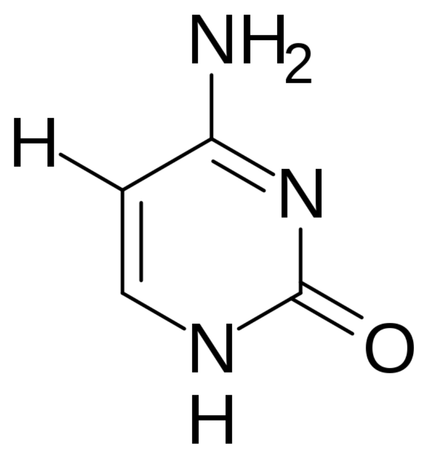
- is an essential **regulatory mechanism** of TEs activity
- targets **CG / CHG / CHH** in plants
  [H = anything besides G]
- is regulated by multiple pathways
- **affects TE / gene expression** (~ silencing)
- may **spread** to flanking regions
- example:

  methylated promoter $\Longrightarrow$ no RNA $\Longrightarrow$

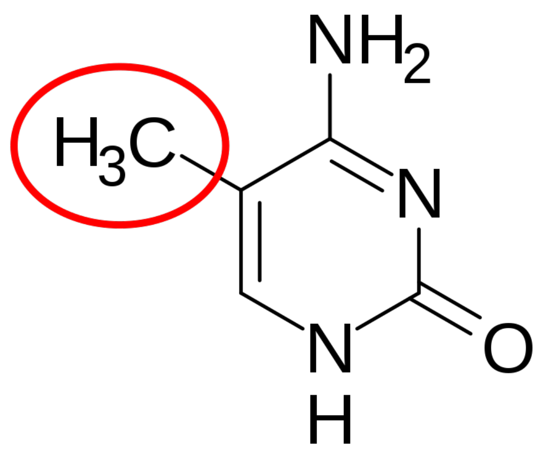
  $\Longrightarrow$ no protein $\Longrightarrow$ no function

- sometimes may be **inherited**





cytosine



methylated cytosine

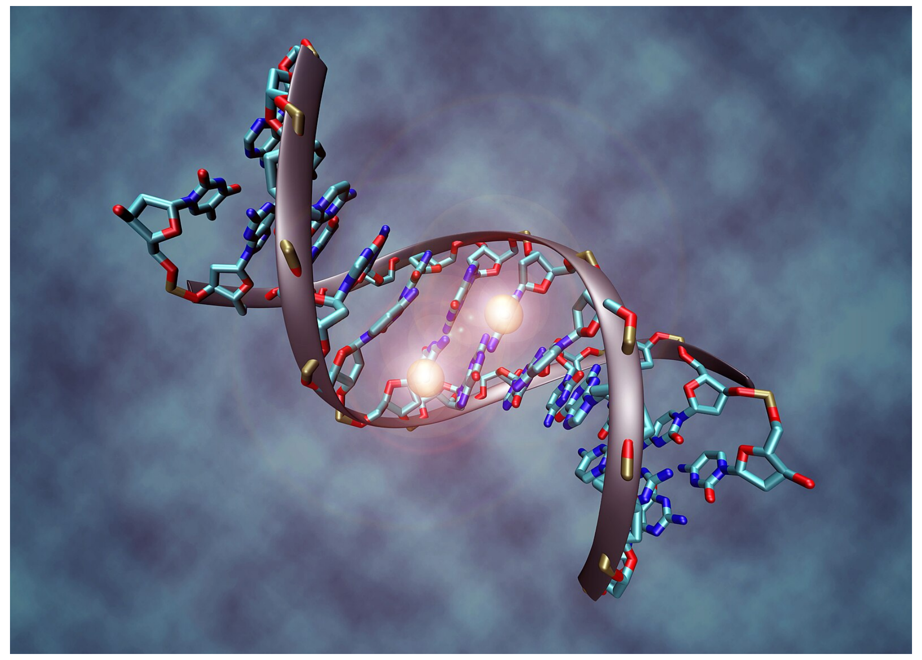
Epigenetic Regulation of Transposable Elements

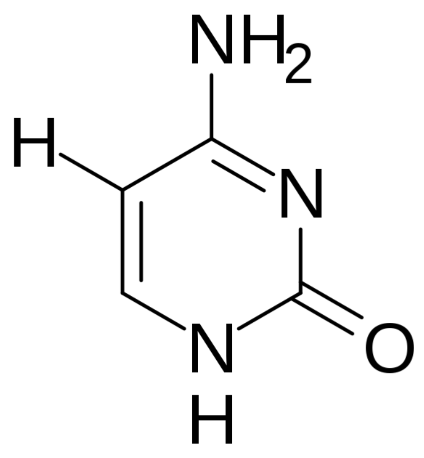
DNA methylation:

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

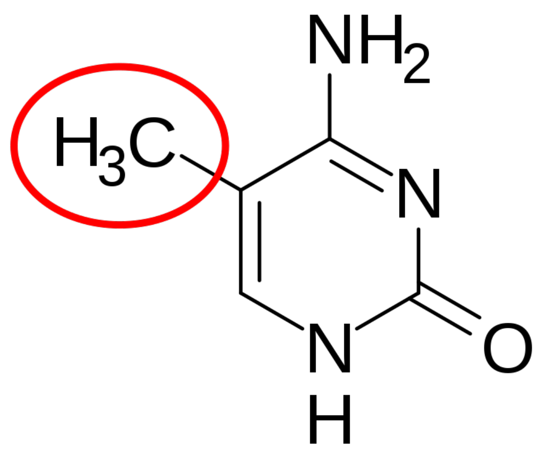
  methylated promoter $\Longrightarrow$ no RNA $\Longrightarrow$
  $\Longrightarrow$ no protein $\Longrightarrow$ no function

- sometimes may be **inherited**

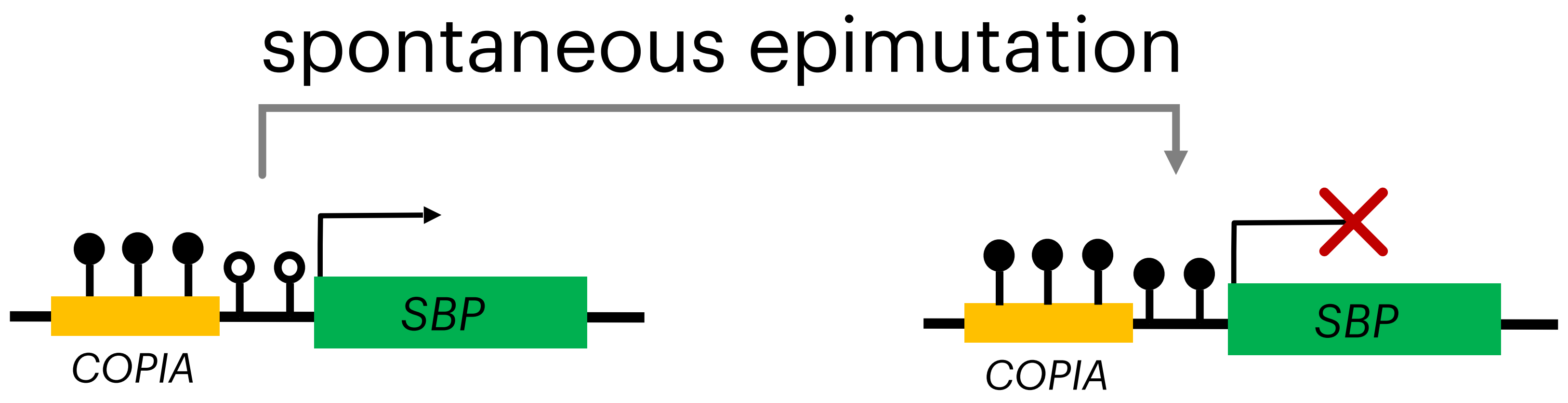




cytosine



methylated cytosine



Manning et al., *Nat Genet 2006*

$\Longrightarrow$ perfect Mendelian segregation though no DNA changes observed

Motivation

- Understanding better **methylation mechanisms** of TEs

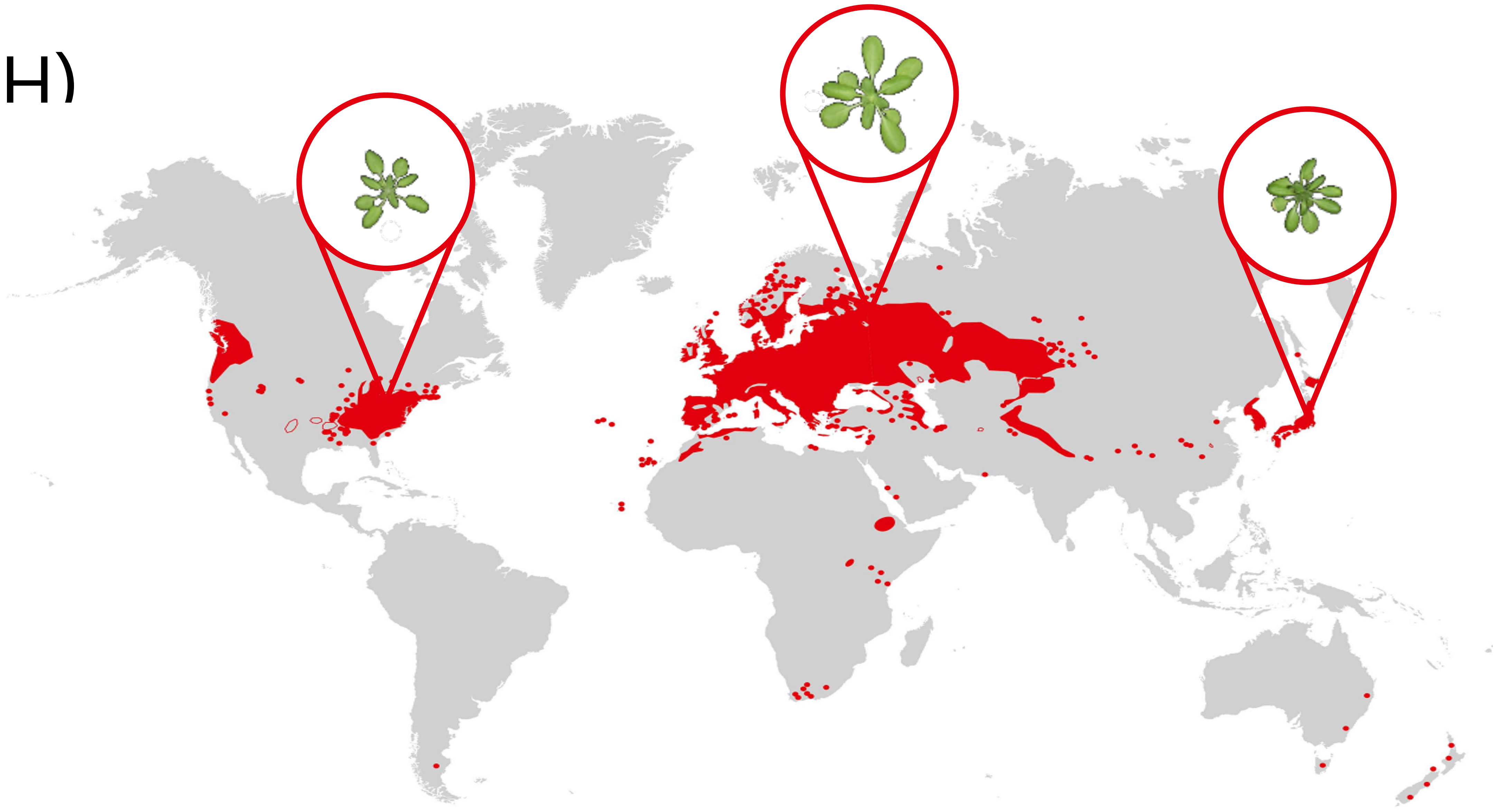
Motivation

- Understanding better **methylation mechanisms** of TEs
- Include **TEs and methylation variation** into **genotype-to-phenotype** studies

Part I: analysis of our TE cohort

Our data: *Arabidopsis Thaliana*

- 87 strains from throughout the world,
  sequenced with ultra-long reads (Nanopore)
- TE annotation (in-house pipeline: GraffiTE + Blast)
  = **Genotyping (same TE across all genomes) + exact positions**
- Full methylation profiles
  (for all contexts CG, CHG, CHH)
- Gene annotation
- SNP annotation
- Gene expression data



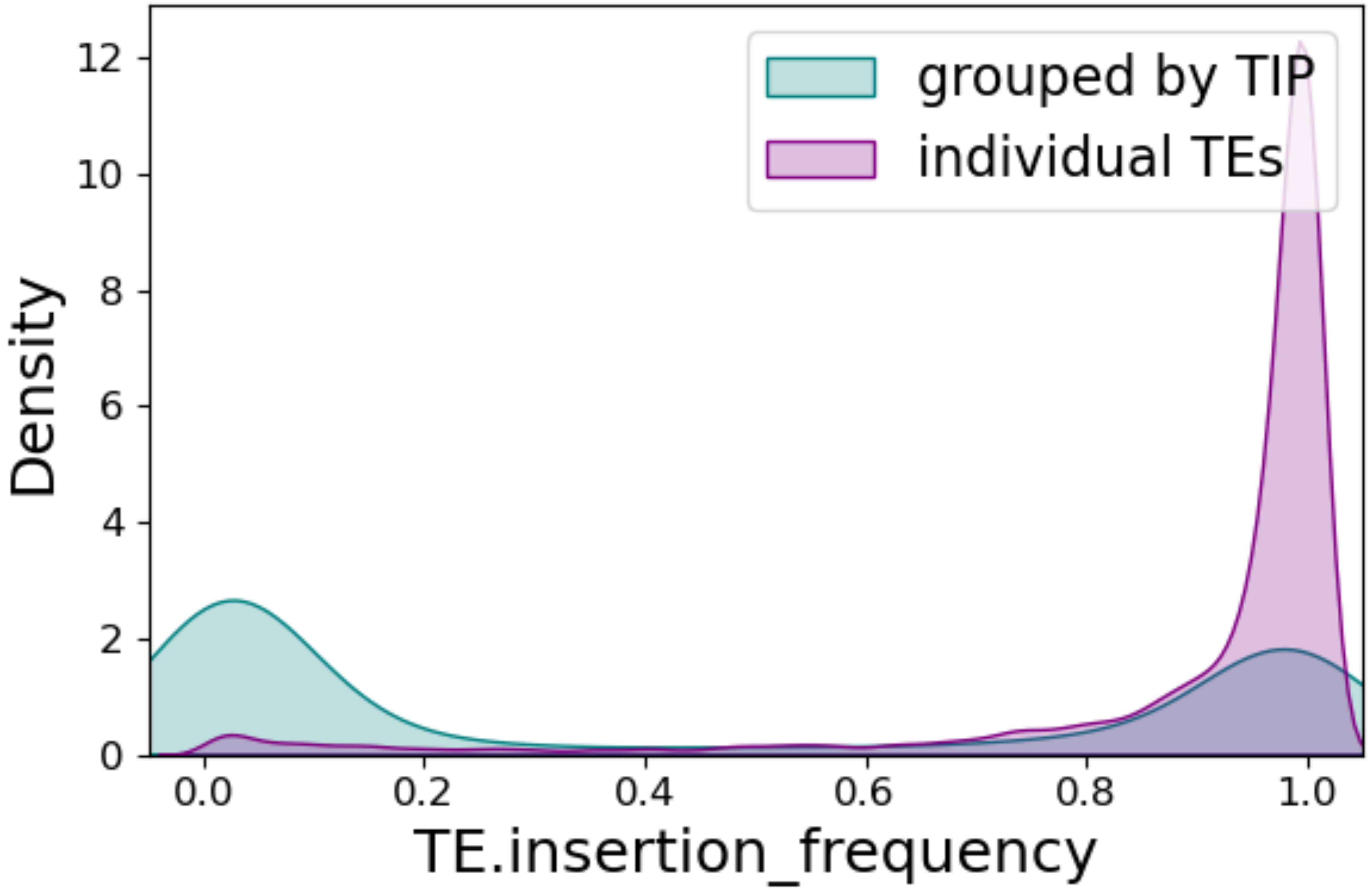
Kawakatsu *et al. Cell 2016*, Alonso-Blanco *et al. Cell 2016*, Quadrana *et al. eLife 2016*

Our data: *Arabidopsis Thaliana*

- 328.043 TEs annotated across N=87 genomes
- 8.795 Transposon Insertion Polymorphisms (TIPs)
- TE age ~ **Insertion frequency (n / N)** (dataset-dependent)
- TE age ~ **% length wrt reference** (<u>dataset-agnostic</u>)

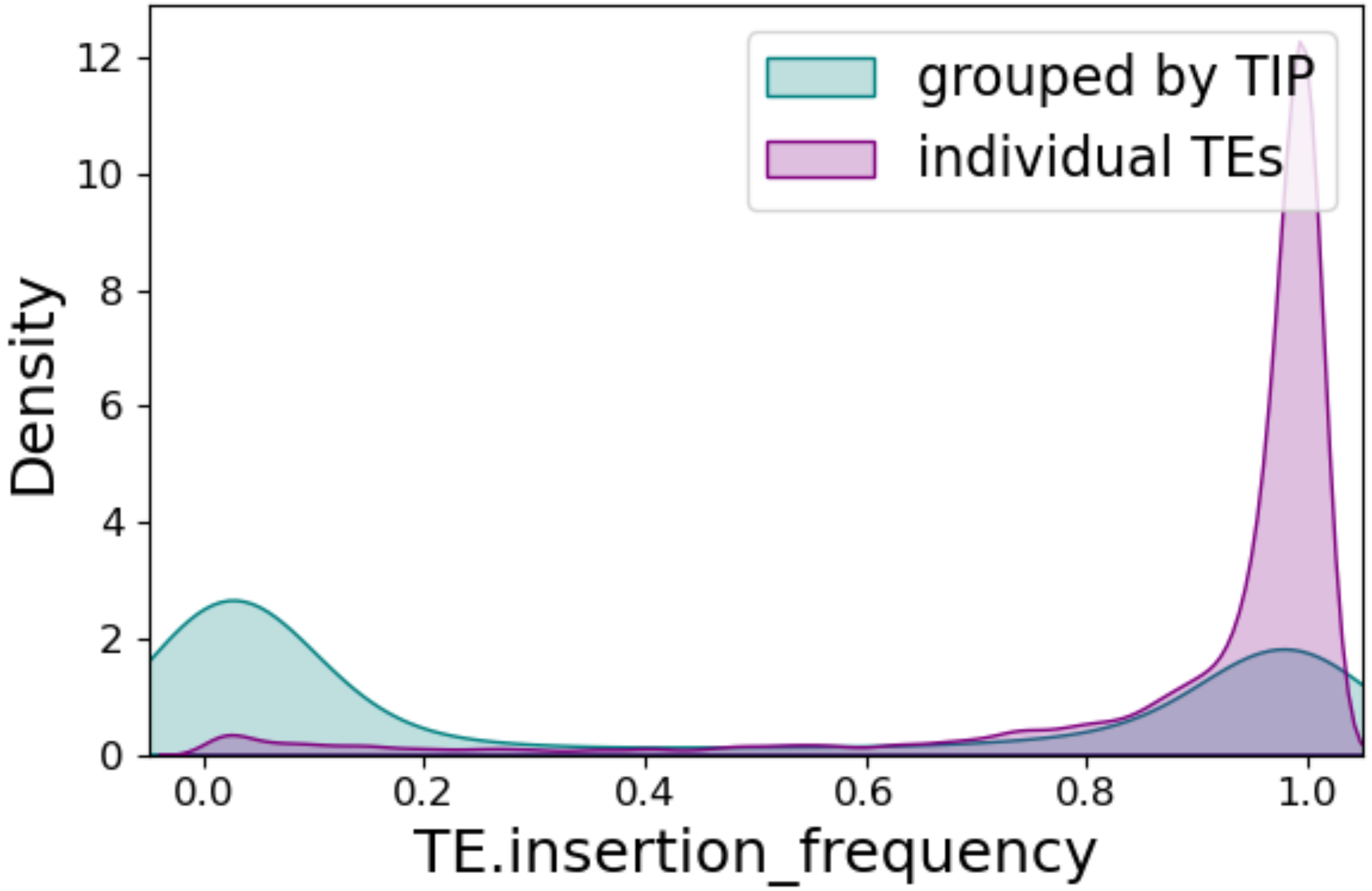
Our data: *Arabidopsis Thaliana*

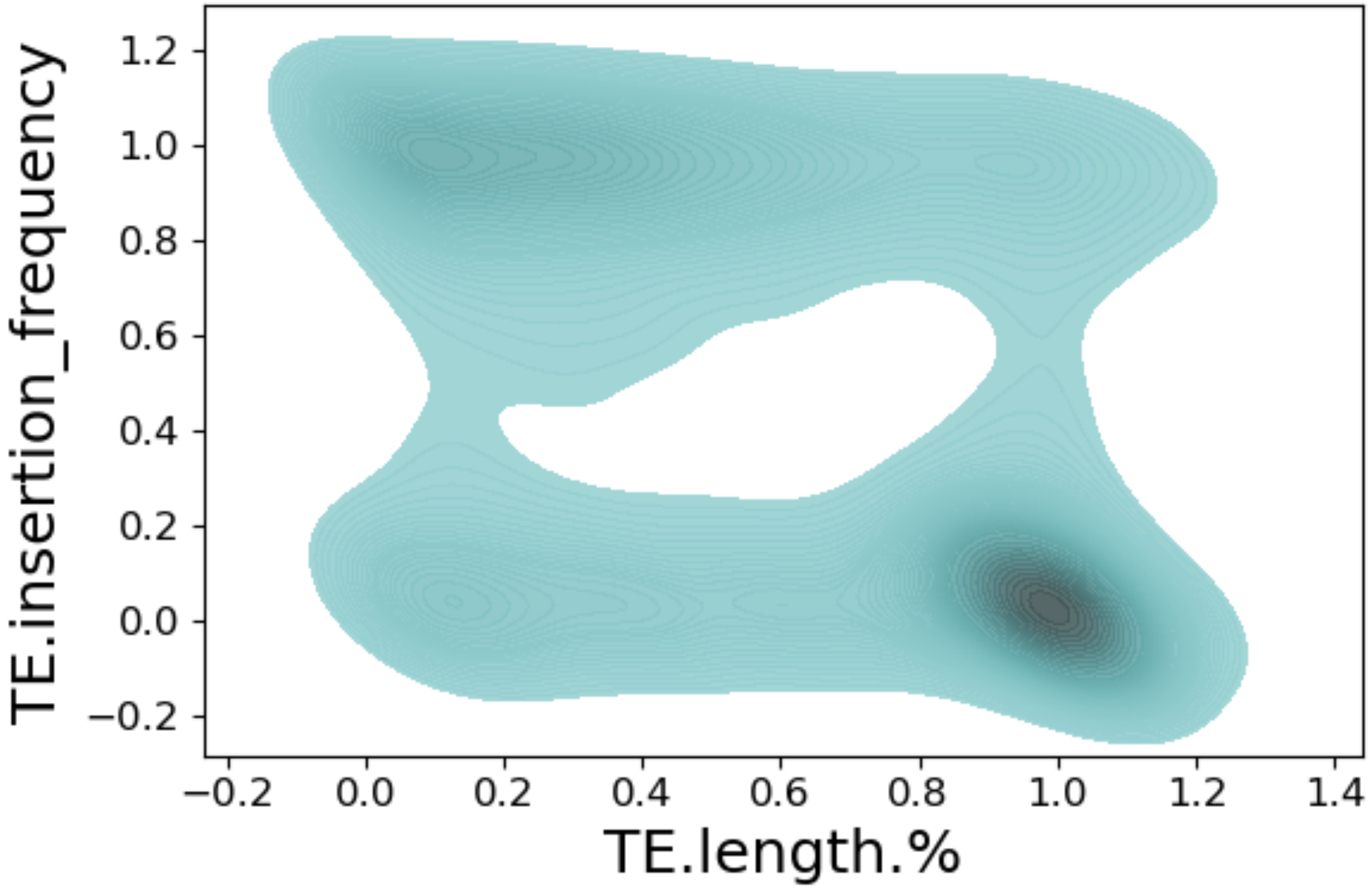
- 328.043 TEs annotated across N=87 genomes
- 8.795 Transposon Insertion Polymorphisms (TIPs)
- TE age ~ **Insertion frequency (n / N)** (dataset-dependent)
- TE age ~ **% length wrt reference** (dataset-agnostic)



Our data: *Arabidopsis Thaliana*

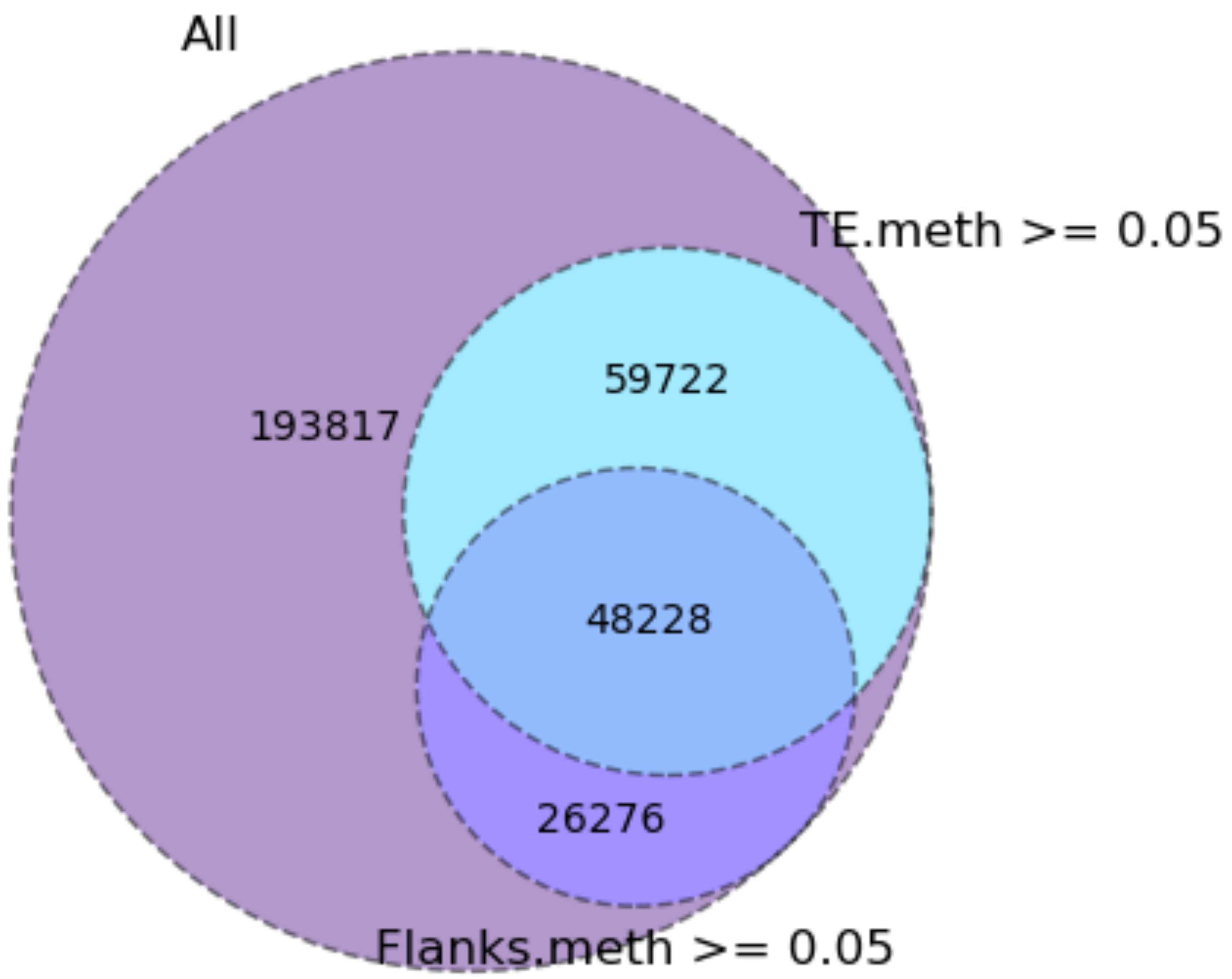
- 328.043 TEs annotated across N=87 genomes
- 8.795 Transposon Insertion Polymorphisms (TIPs)
- TE age ~ **Insertion frequency (n / N)** (dataset-dependent)
- TE age ~ **% length wrt reference** (dataset-agnostic)





Our data: *Arabidopsis Thaliana*

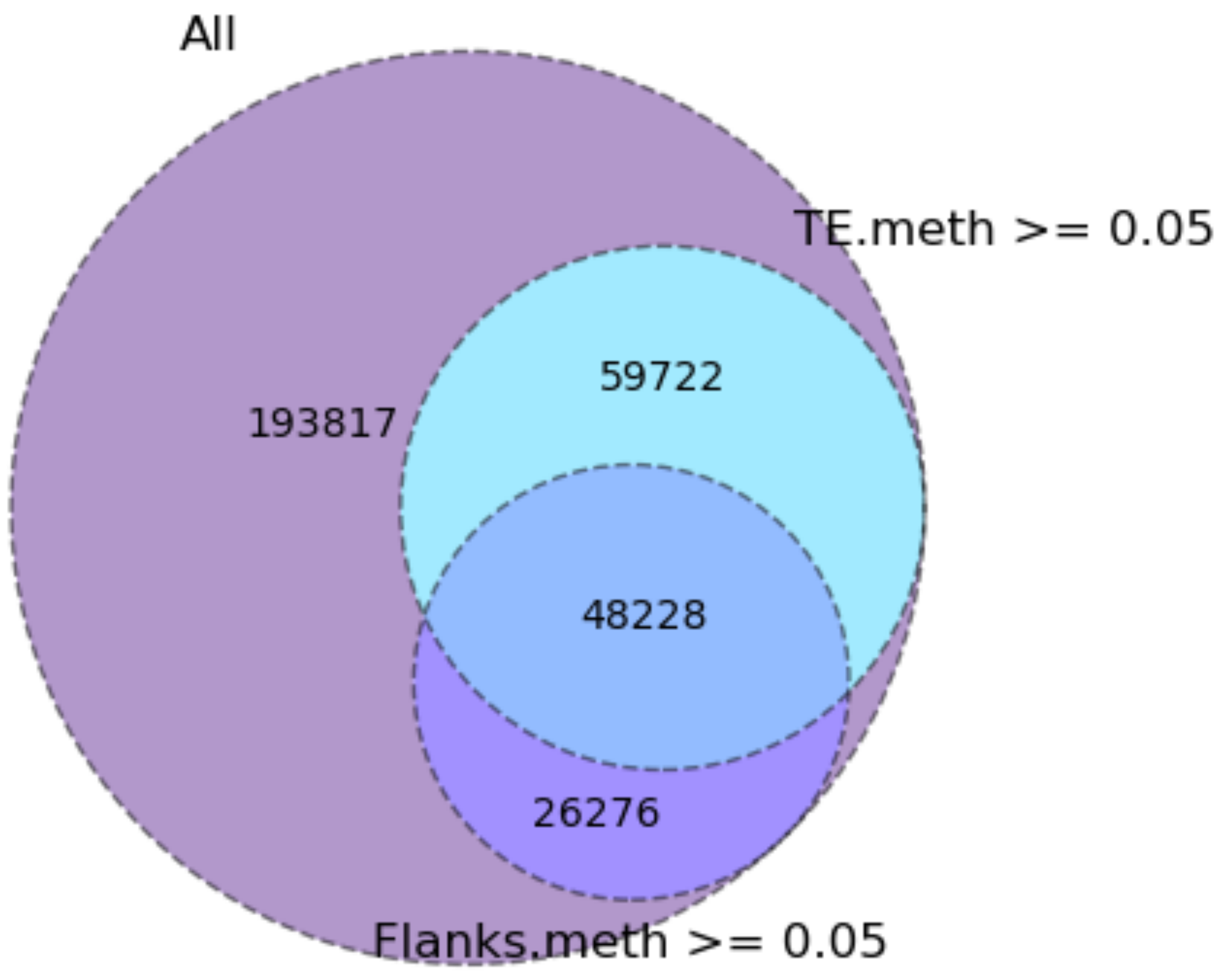
- 328.043 TEs annotated across N=87 genomes
- 8.795 Transposon Insertion Polymorphisms (TIPs)
- TE age ~ **Insertion frequency (n / N)** (dataset-dependent)
- TE age ~ **% length wrt reference** (dataset-agnostic)



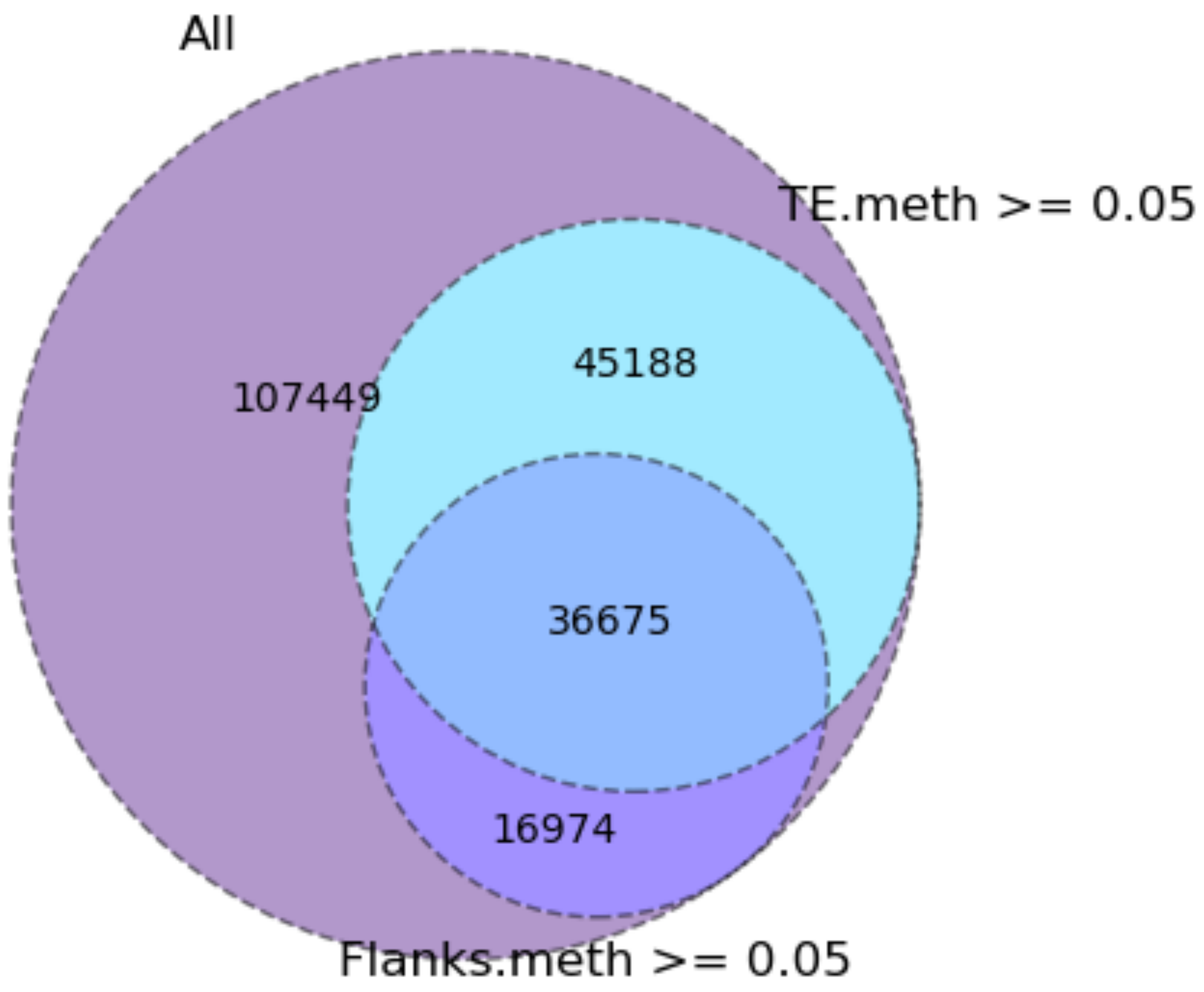
All TEs

Our data: *Arabidopsis Thaliana*

- 328.043 TEs annotated across N=87 genomes
- 8.795 Transposon Insertion Polymorphisms (TIPs)
- TE age ~ **Insertion frequency (n / N)** (dataset-dependent)
- TE age ~ **% length wrt reference** (dataset-agnostic)



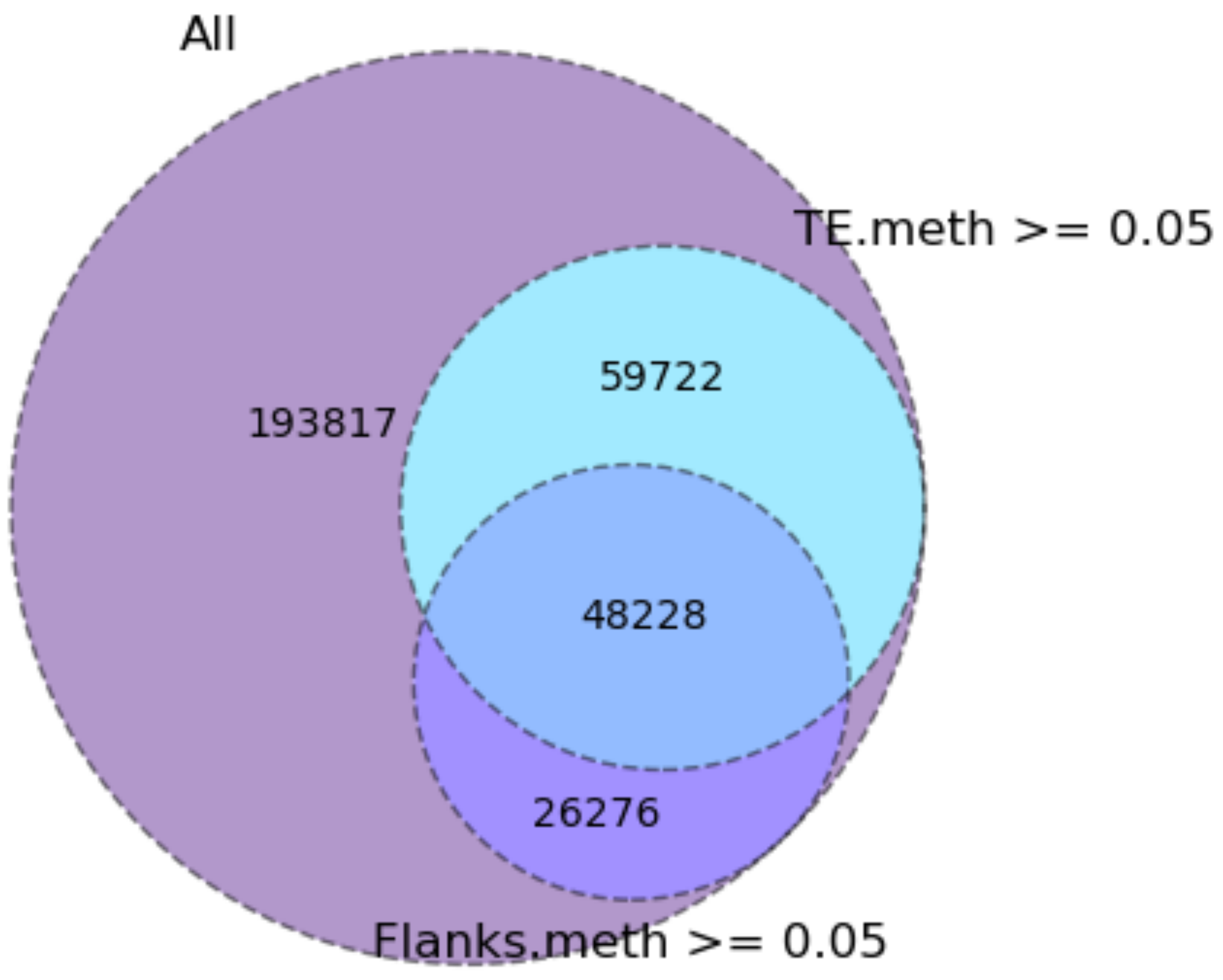
All TEs



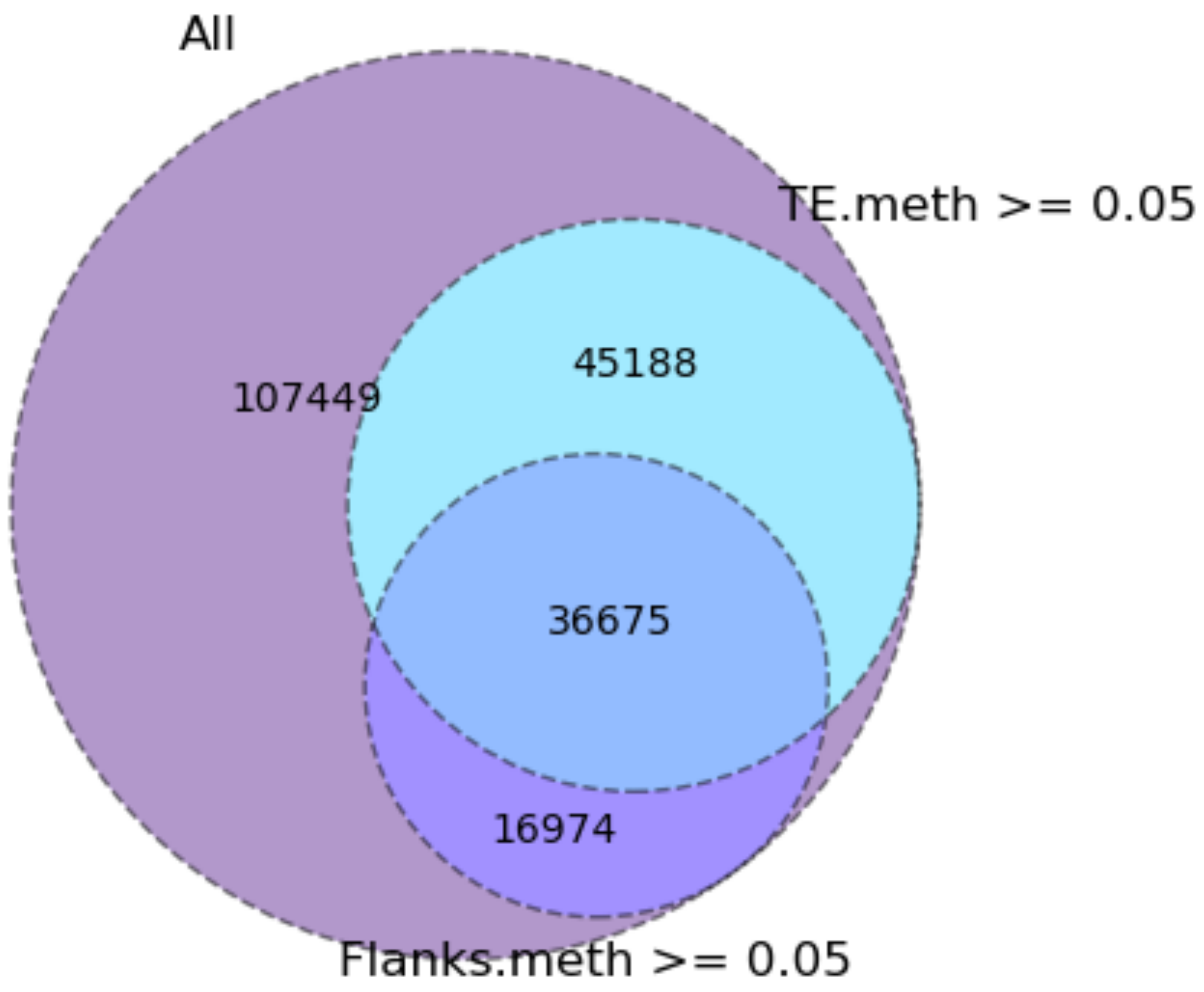
Only polymorphic TEs

Our data: *Arabidopsis Thaliana*

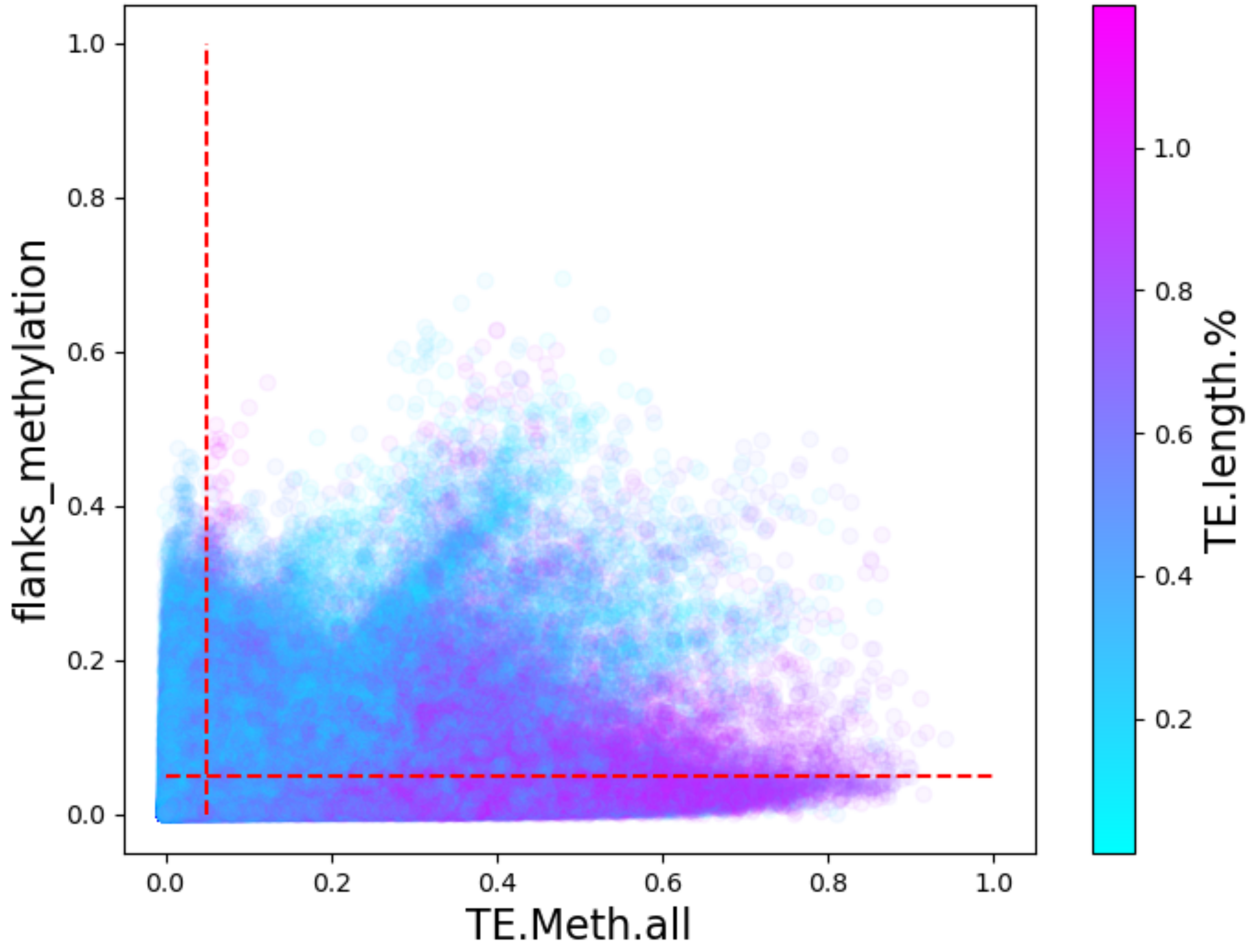
- 328.043 TEs annotated across N=87 genomes
- 8.795 Transposon Insertion Polymorphisms (TIPs)
- TE age ~ **Insertion frequency (n / N)** (dataset-dependent)
- TE age ~ **% length wrt reference** (dataset-agnostic)



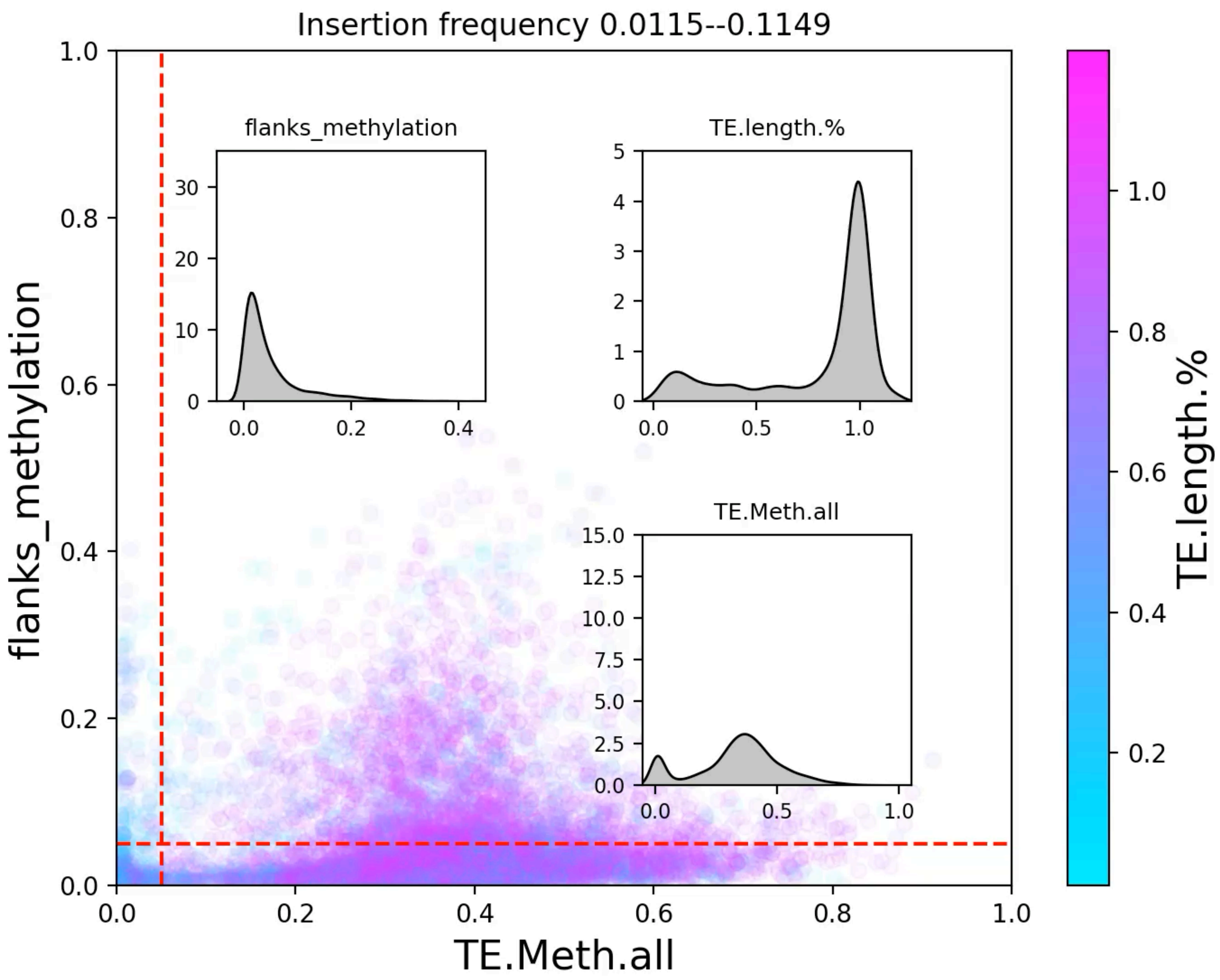
All TEs



Only polymorphic TEs

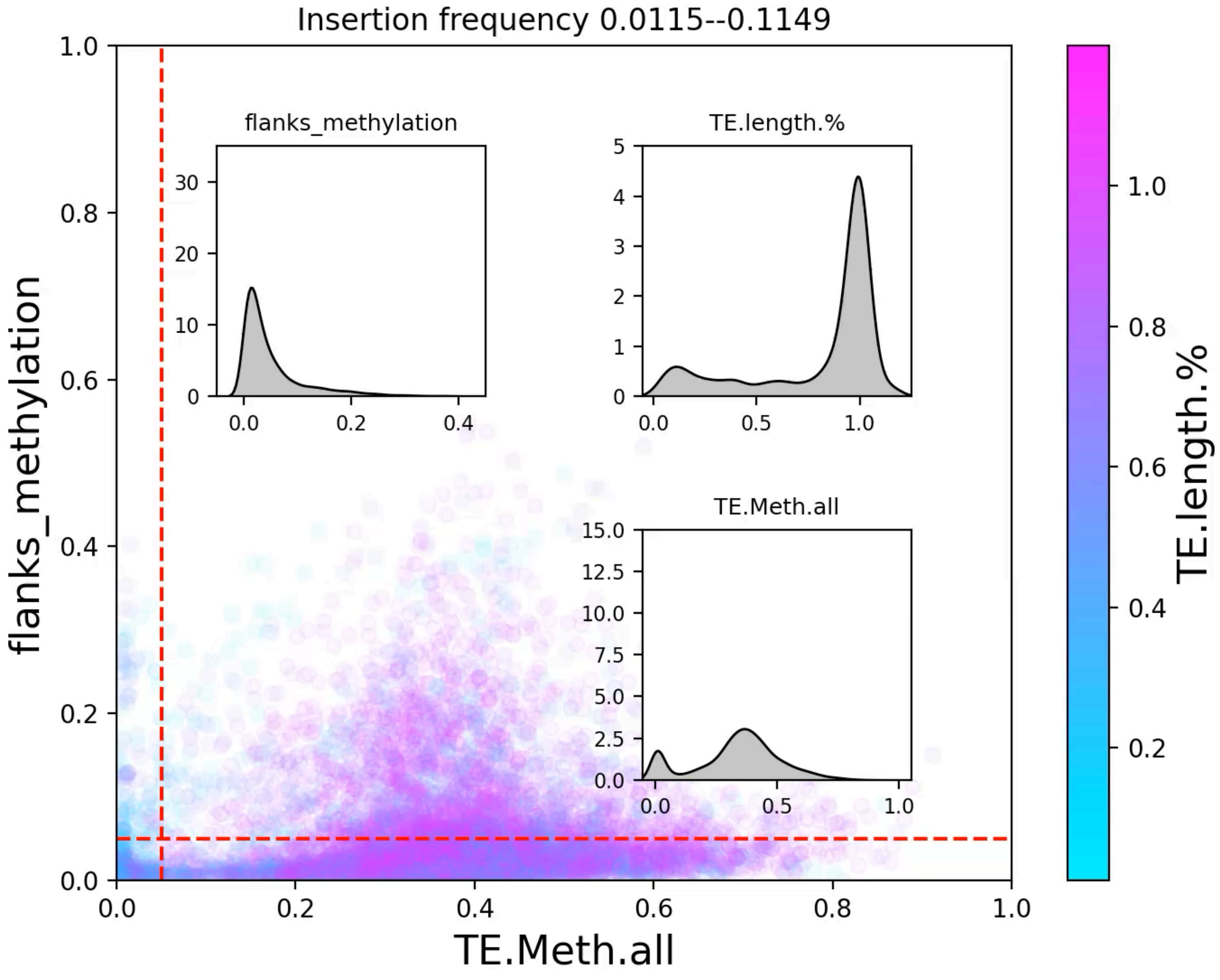


Our data: *Arabidopsis Thaliana*



- **Young TEs** tend to be **methylated** and (sometimes) to spread
- It is not rare to observe **non-methylated but spread** TEs (mostly old)
- **<u>Hypothesis</u>**: the effect is due to secondary de-methylation of decayed TEs (but spreading remains)

Our data: *Arabidopsis Thaliana*



- **Young TEs** tend to be **methylated** and (sometimes) to spread
- It is not rare to observe **non-methylated but spread** TEs (mostly old)
- **<u>Hypothesis</u>**: the effect is due to secondary de-methylation of decayed TEs (but spreading remains)

$\Longrightarrow$ Which **features** (and further, biological mechanisms) **define methylation**?

$\Longrightarrow$ Can we **predict the methylation using genetic features only**?

Part II: understanding methylation

Modeling TE methylation

Model:

- **Random Forest** (hyper parameters tuned via cross-validation stratified by TIPs)

Features:

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

Data: all TEs (328.037)

Modeling TE methylation

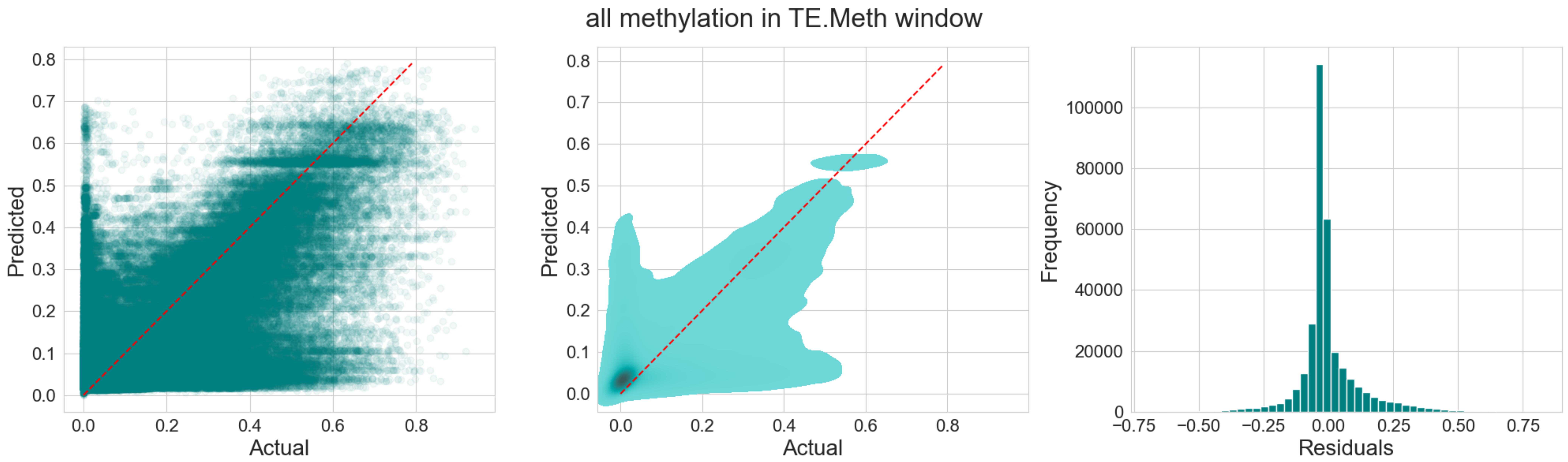
Model:

- **Random Forest** (hyper parameters tuned via cross-validation stratified by TIPs)

Features:

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

Data: all TEs (328.037)



Pearson correlation = 0.688

Modeling TE methylation

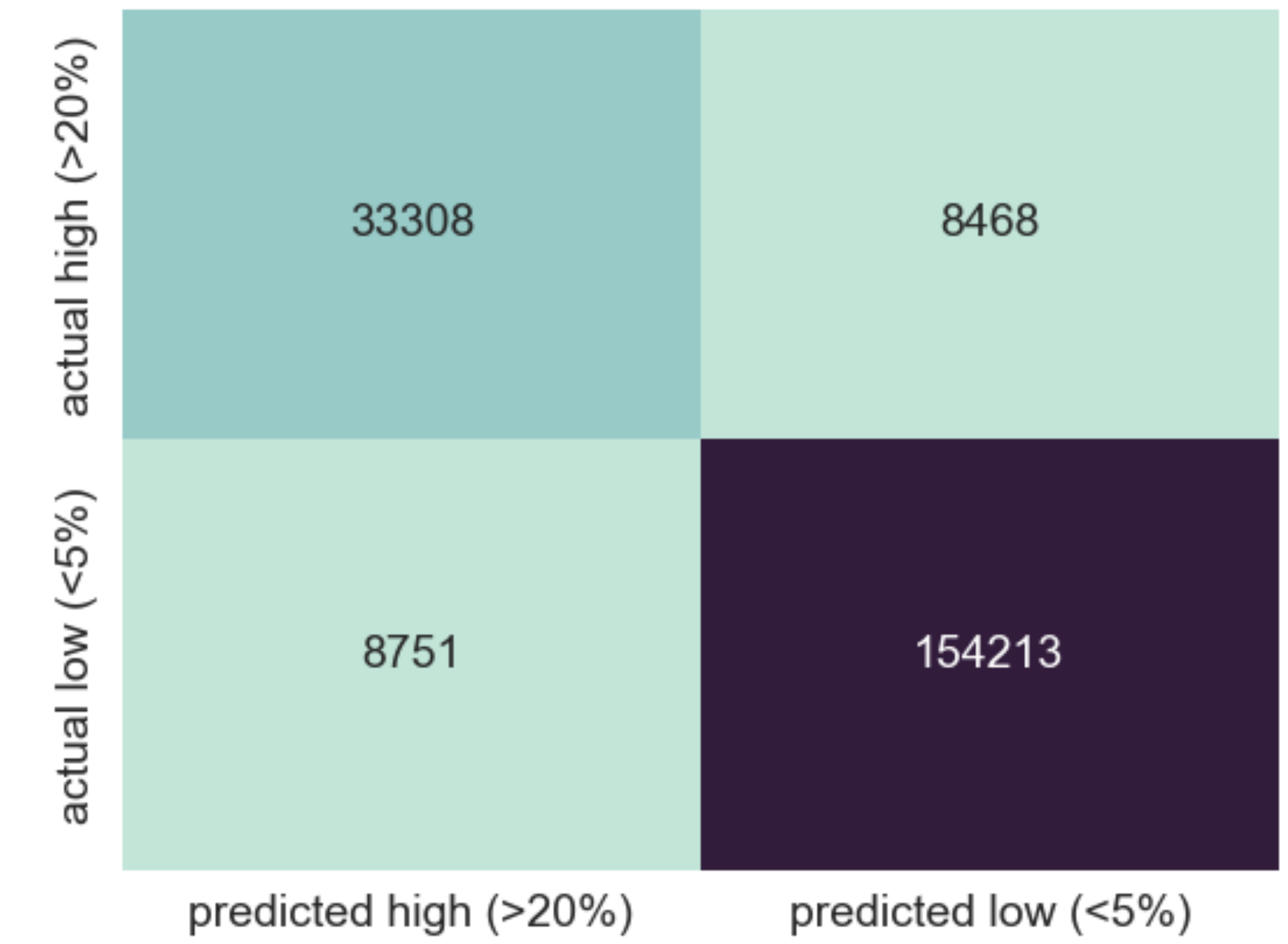
Model:

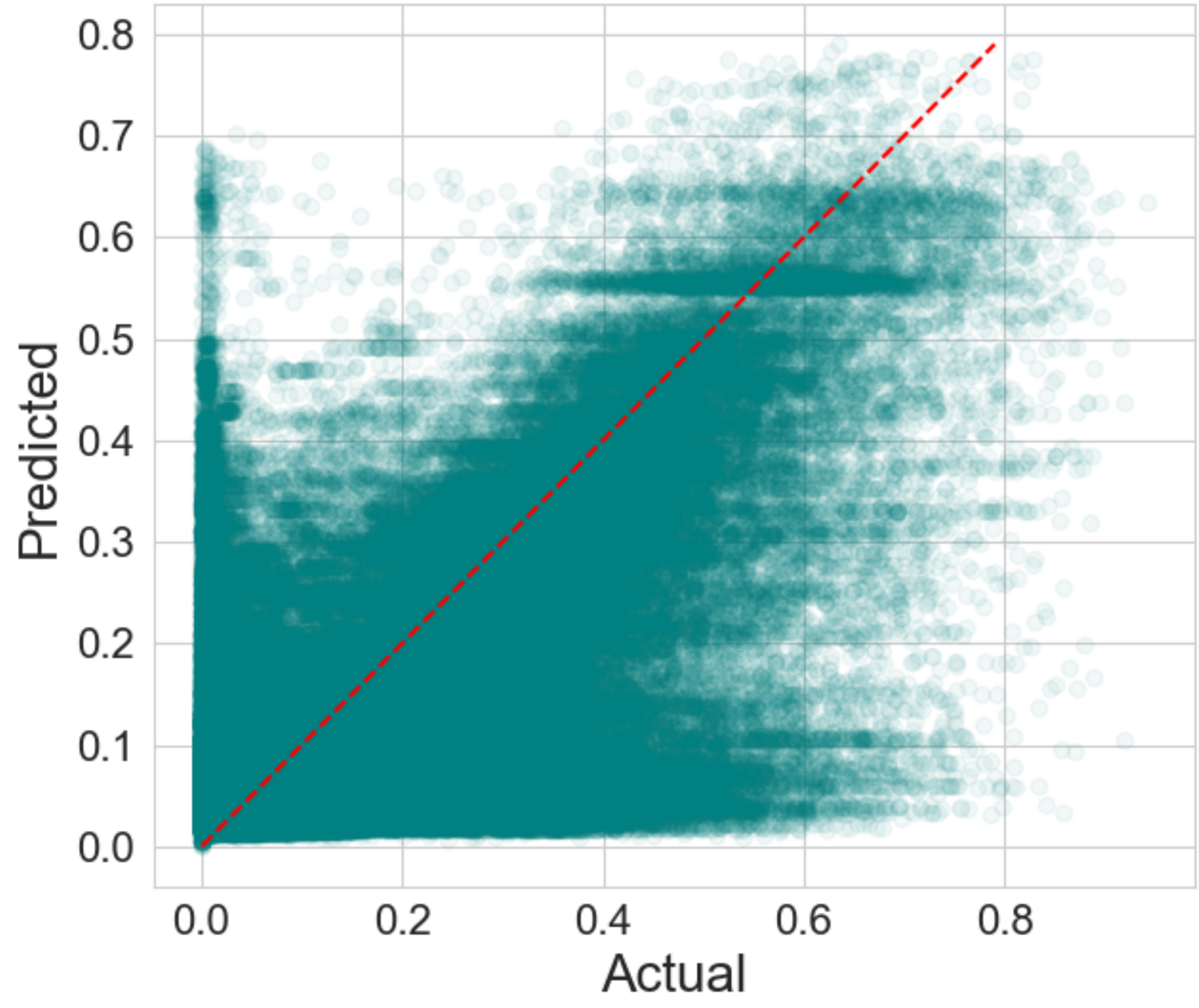
- **Random Forest** (hyper parameters tuned via cross-validation stratified by TIPs)

Features:

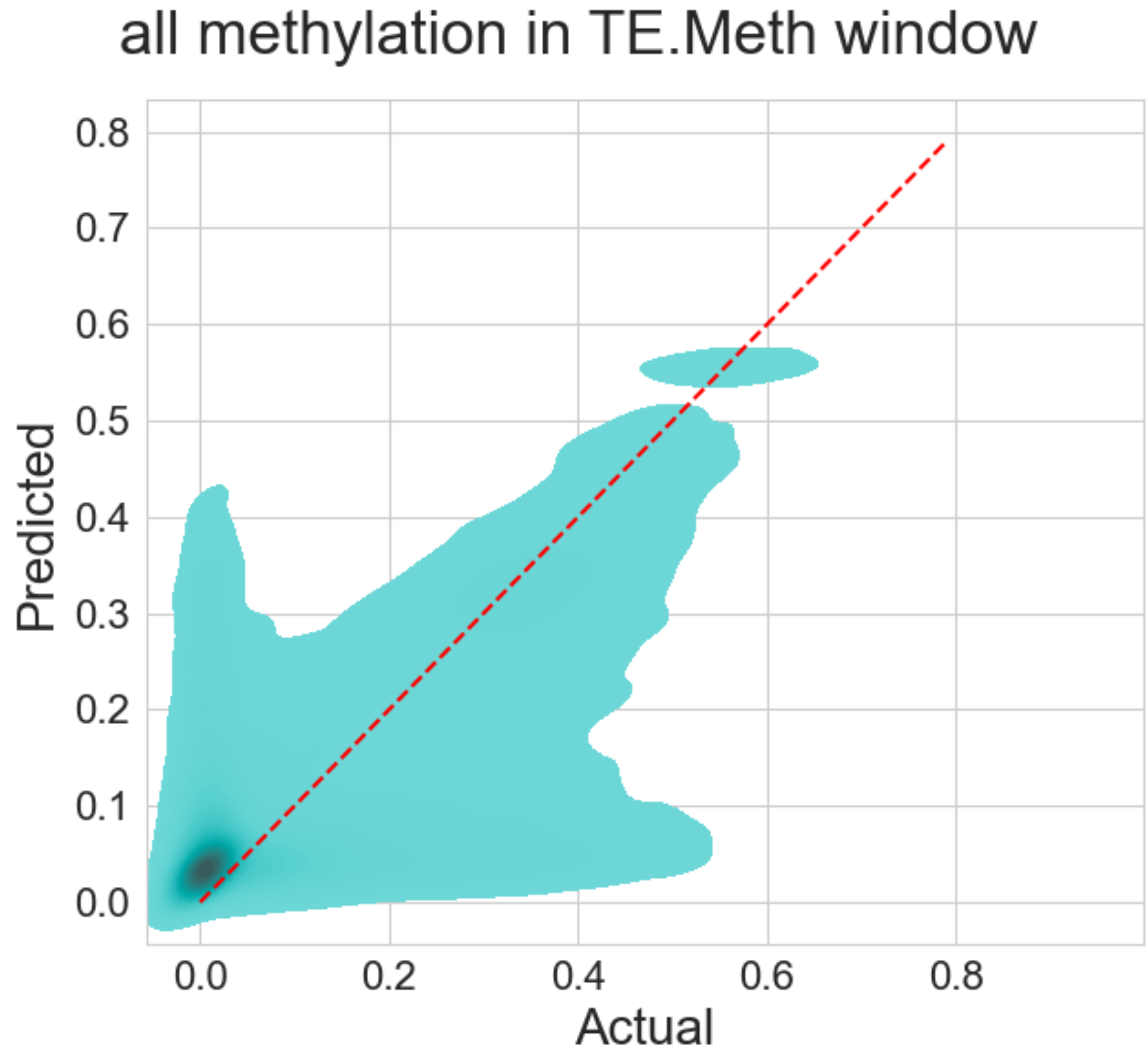
- **TE** (length, distance to pericentromere, superfamily, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Average genome-wide methylation** in CG, CHG, CHH contexts
- **Densities** of CG, CHG, CHH contexts

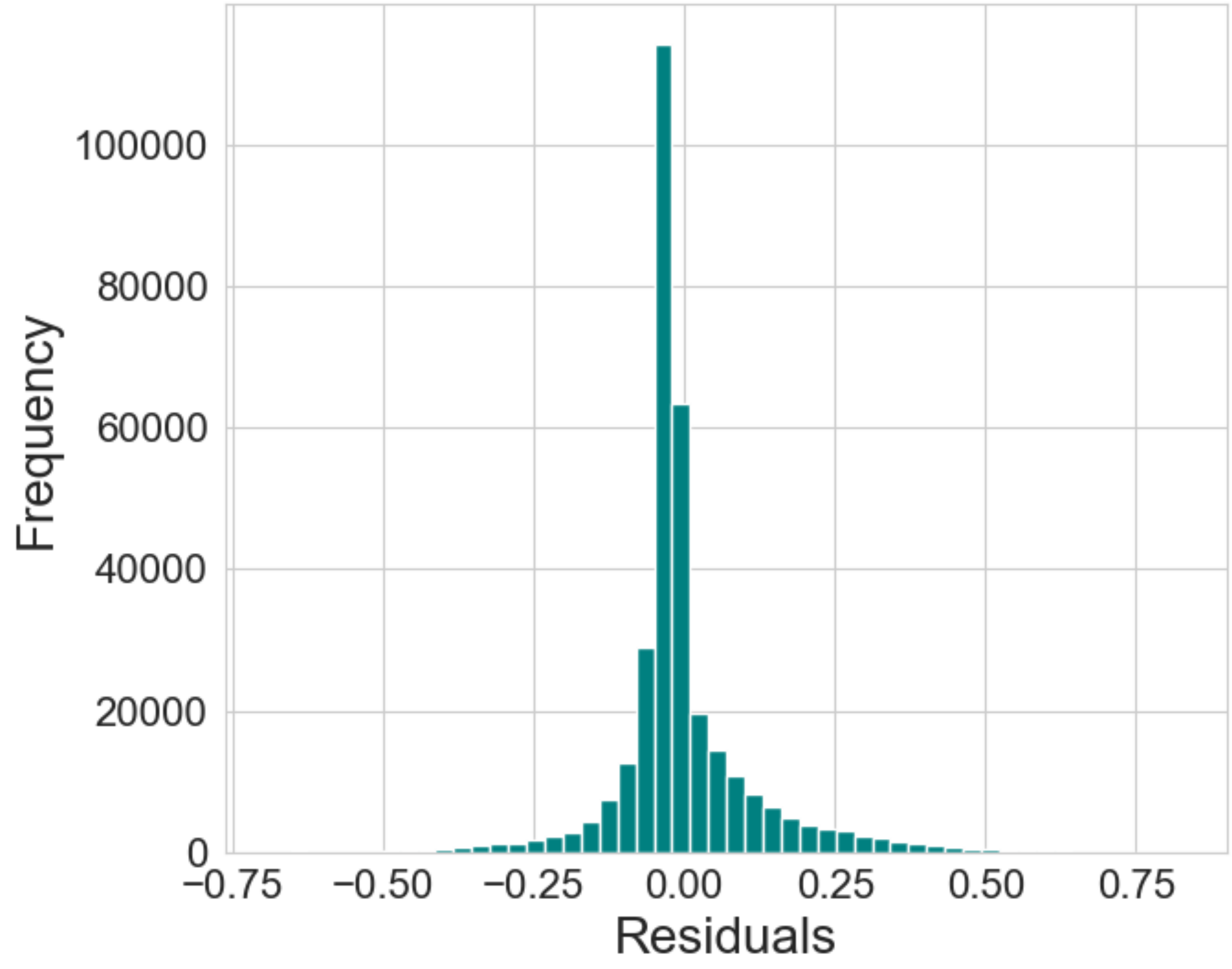
Data: all TEs (328.037)





Pearson correlation = 0.688





Modeling TE methylation

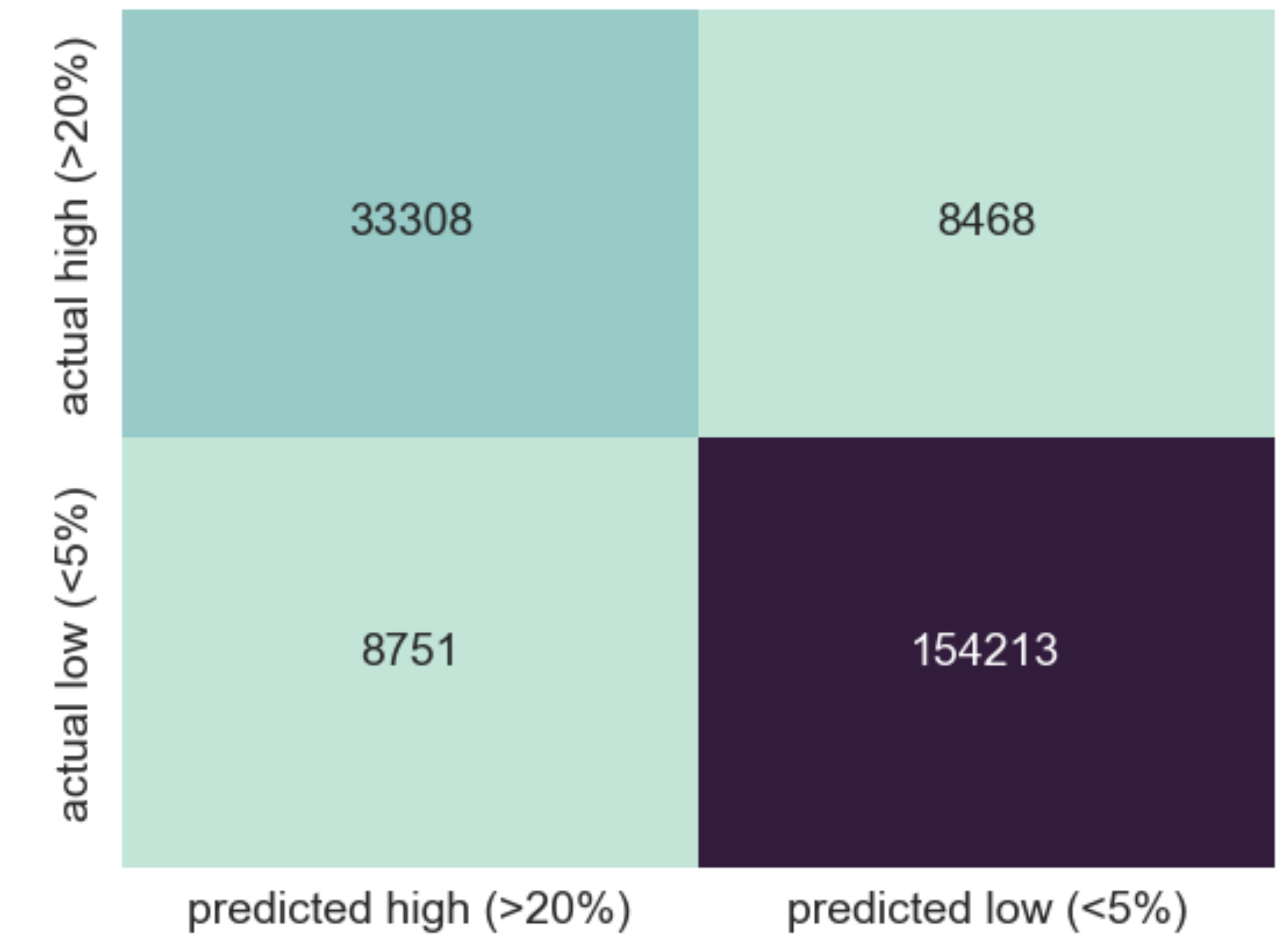
Model:

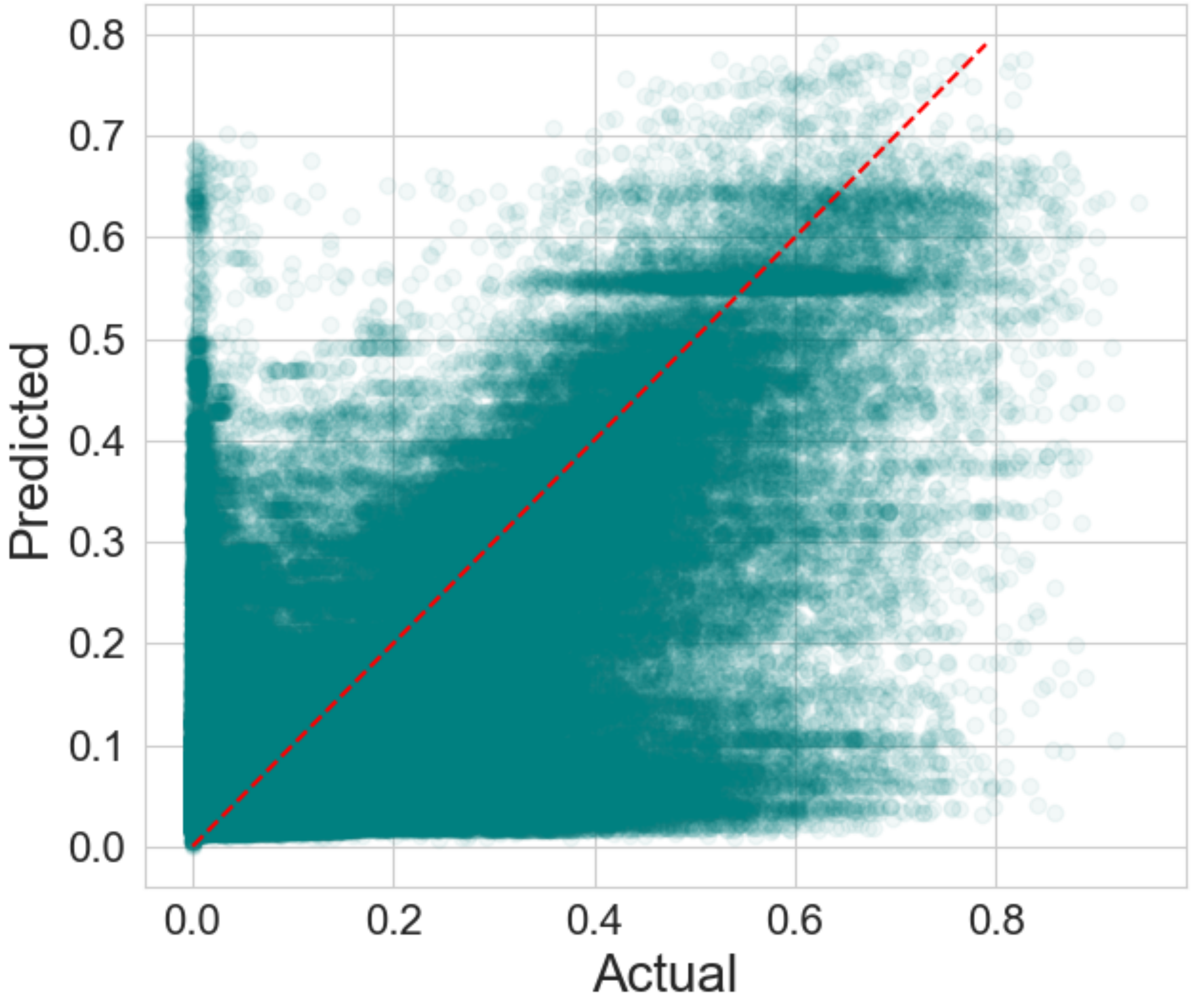
- **Random Forest** (hyper parameters tuned via cross-validation stratified by TIPs)

Features:

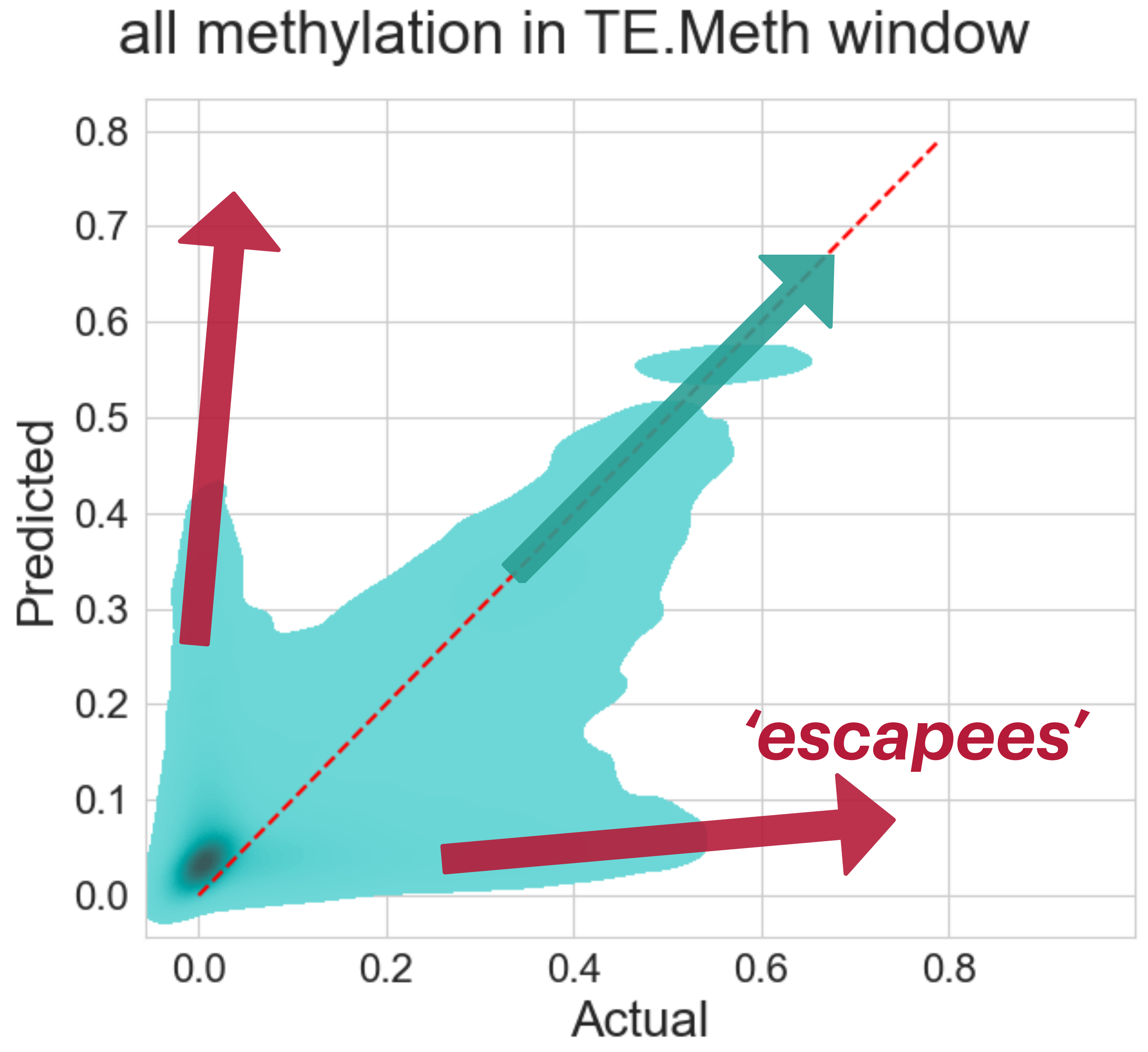
- **TE** (length, distance to pericentromere, superfamily, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Average genome-wide methylation** in CG, CHG, CHH contexts
- **Densities** of CG, CHG, CHH contexts

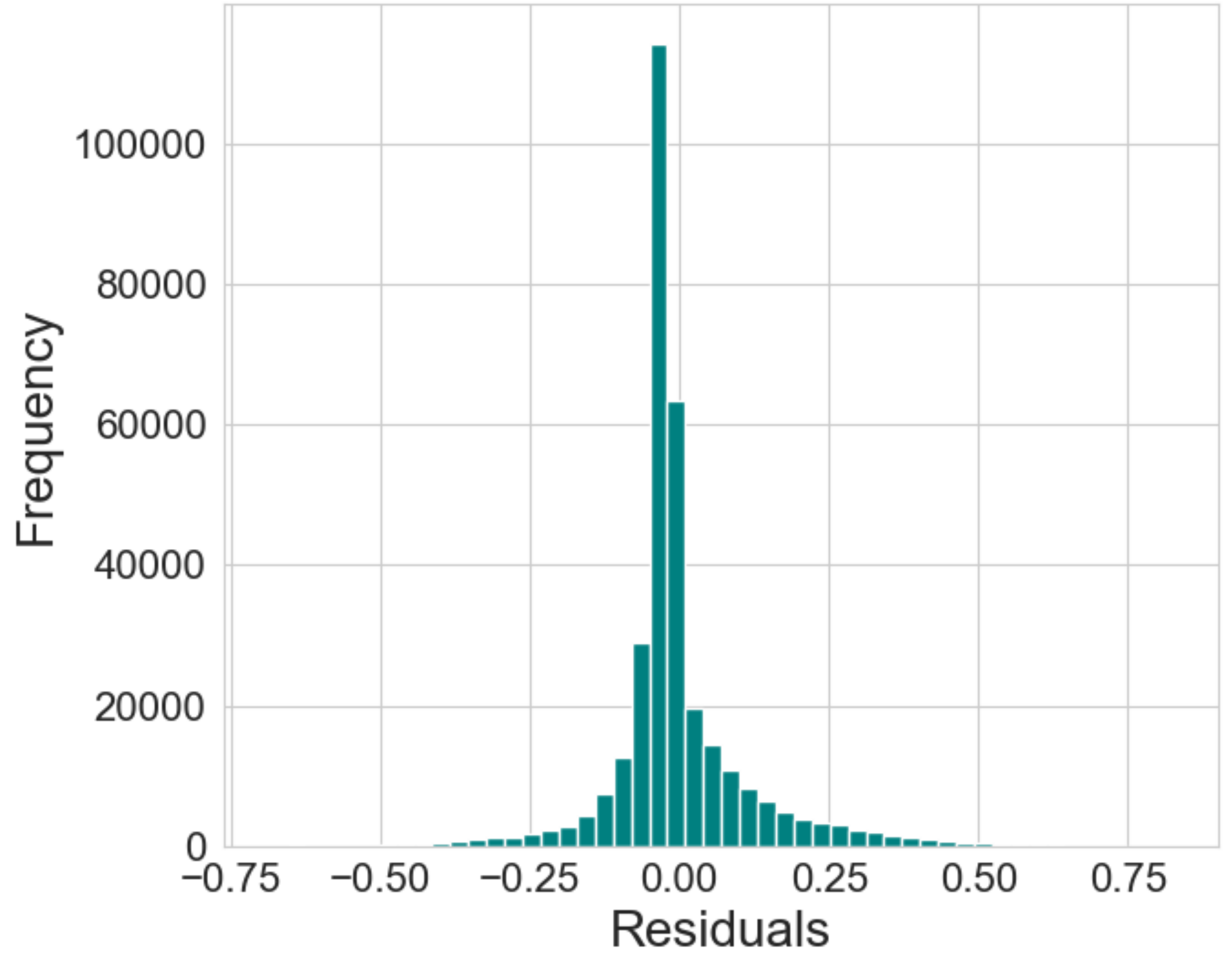
Data: all TEs (328.037)





Pearson correlation = 0.688

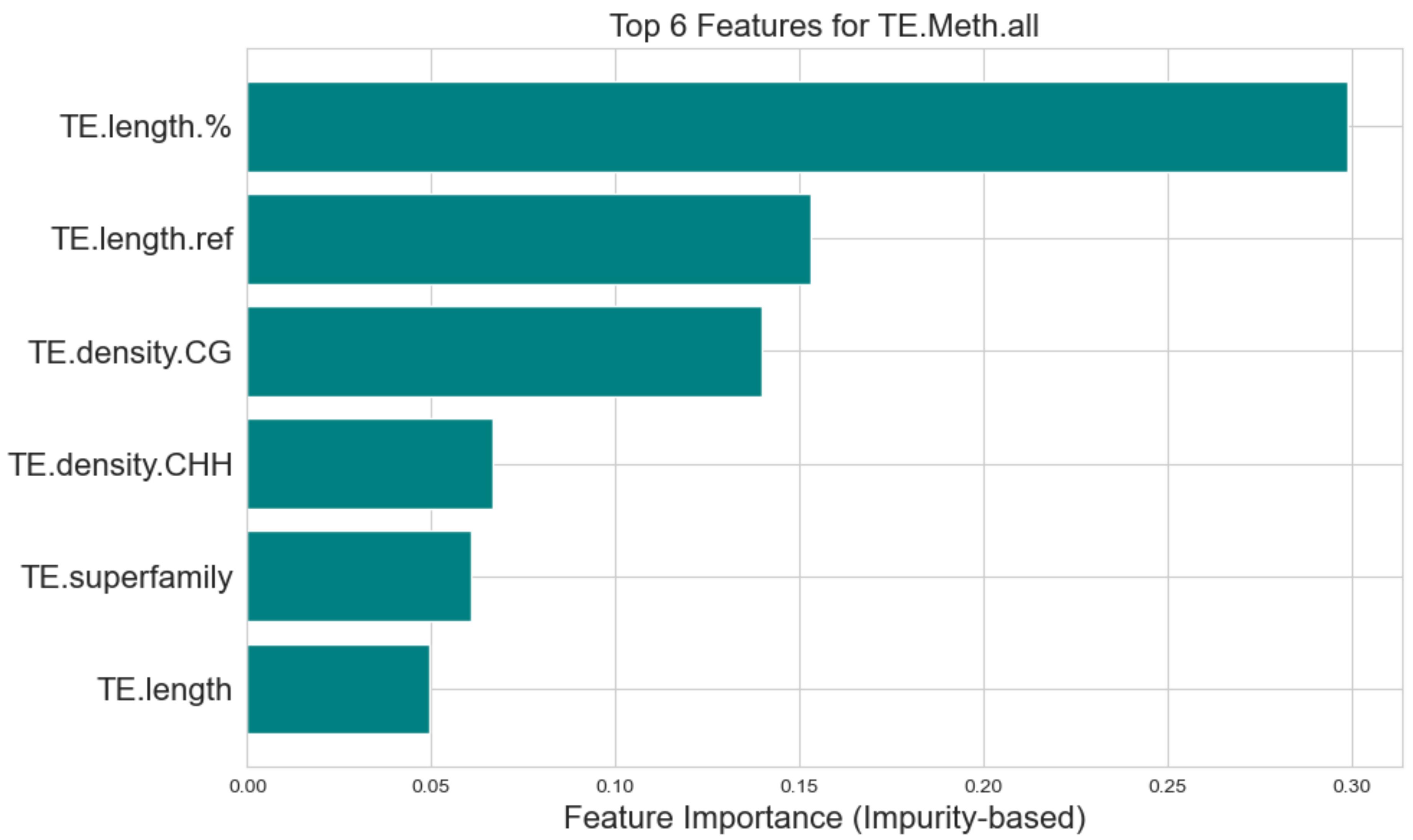




Modeling TE methylation

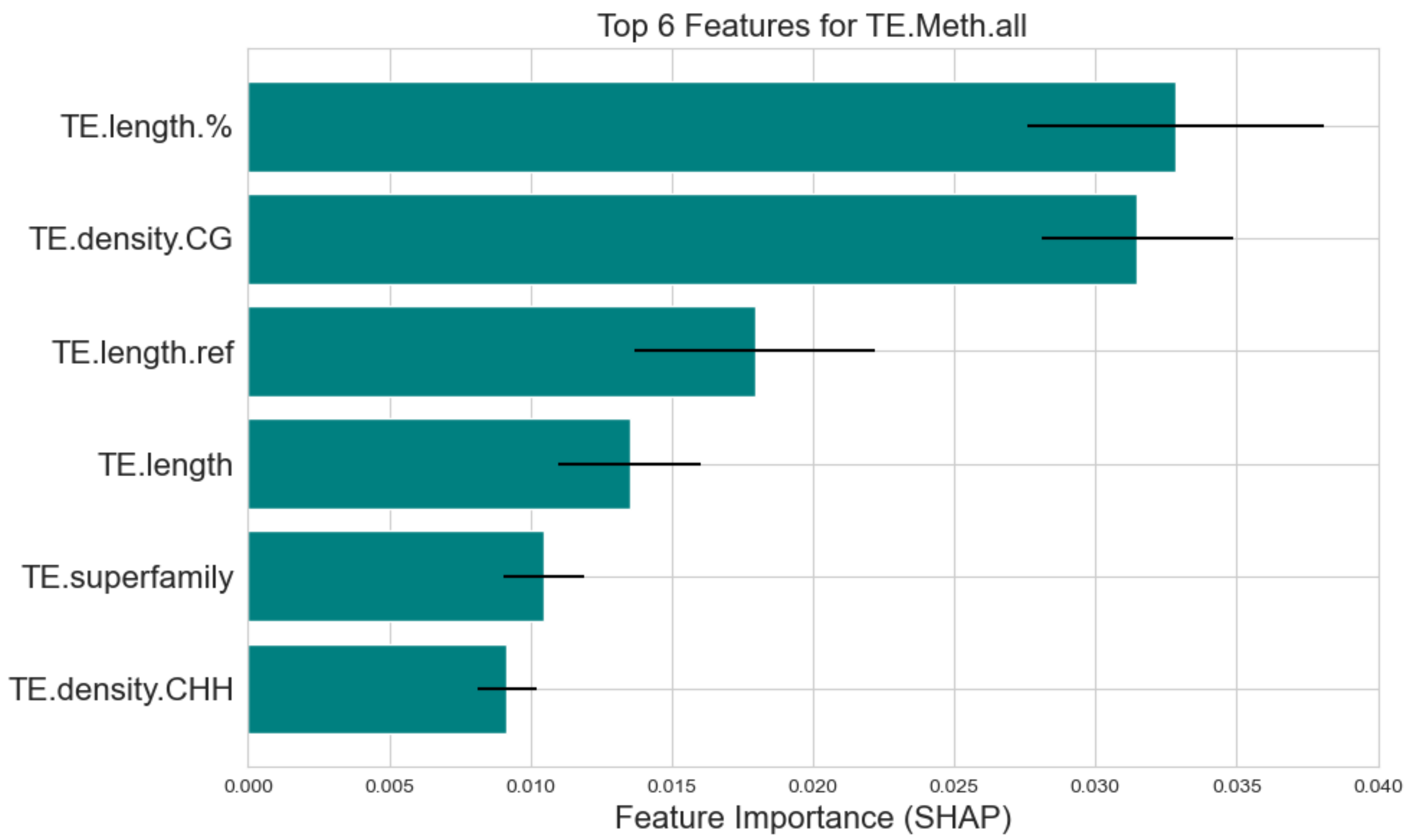
Impurity-based feature importances

* 10 independent runs with different random seeds



SHAP values

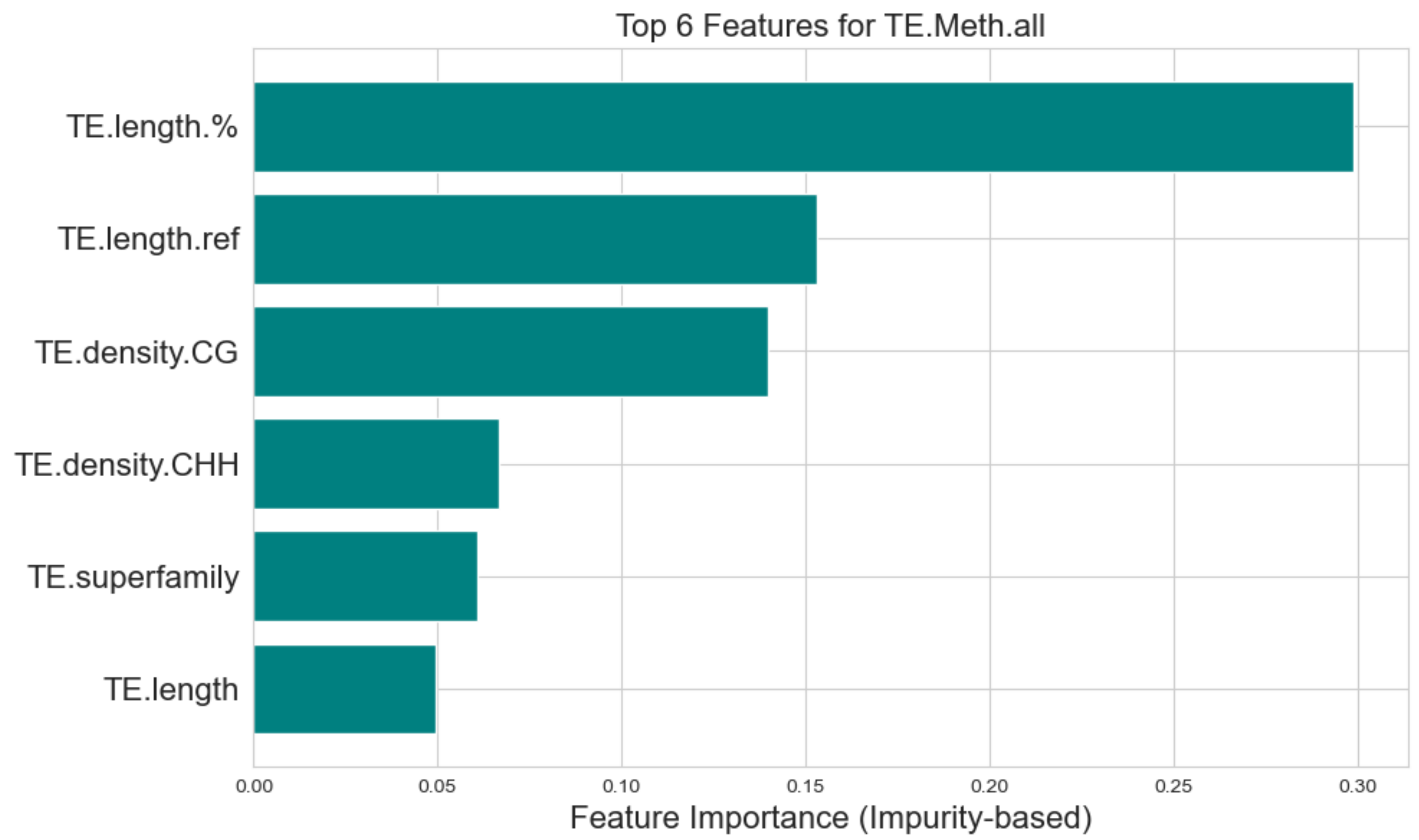
* 10 folds in a cross-validation manner



Modeling TE methylation

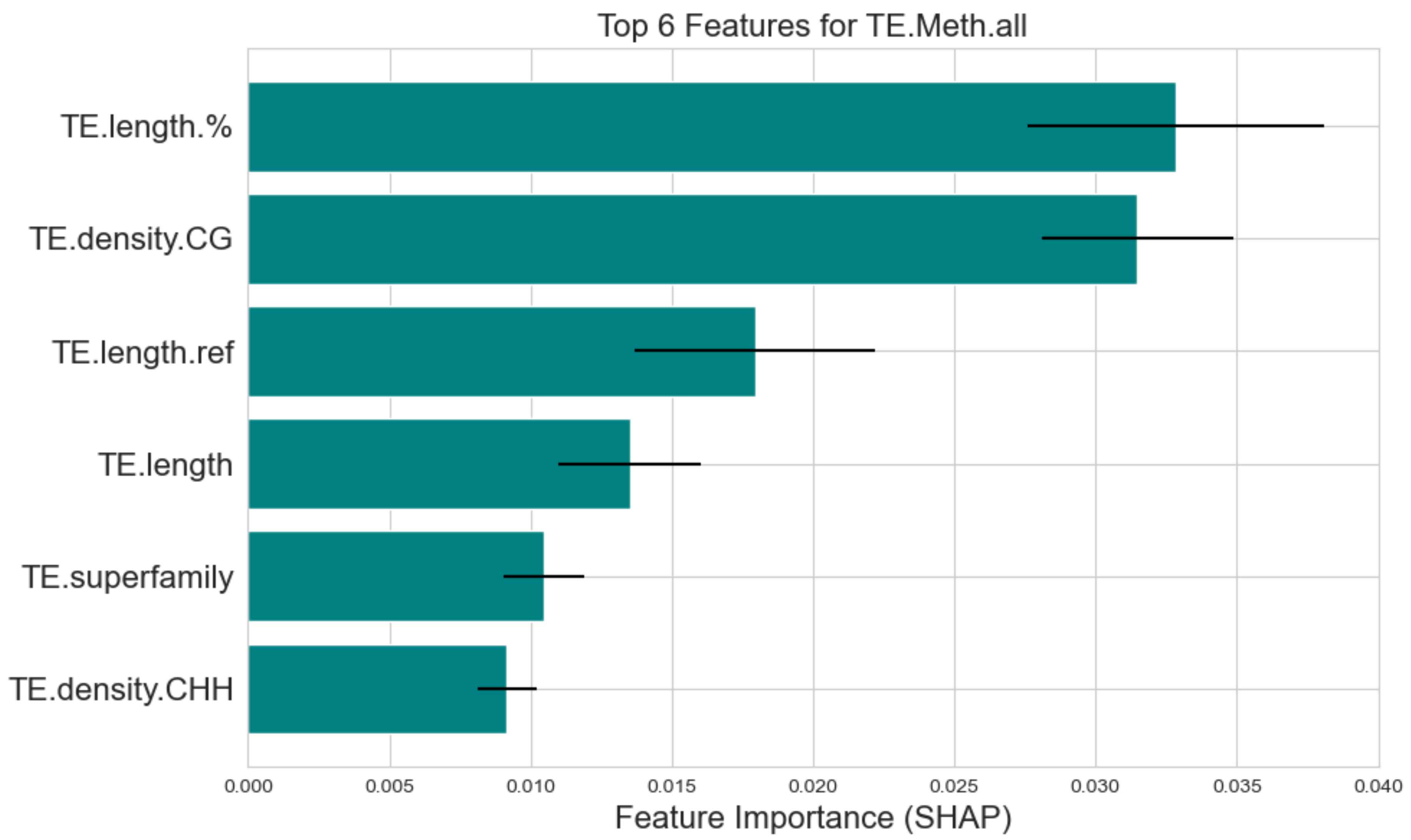
Impurity-based feature importances

* 10 independent runs with different random seeds

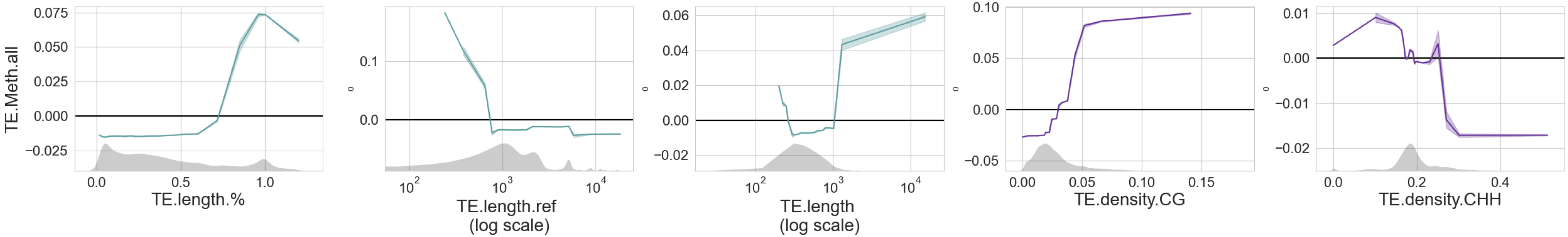


SHAP values

* 10 folds in a cross-validation manner



Accumulated Local Effects (ALE) * 10 folds in a cross-validation manner



Modeling TE methylation

Who are the escapees?

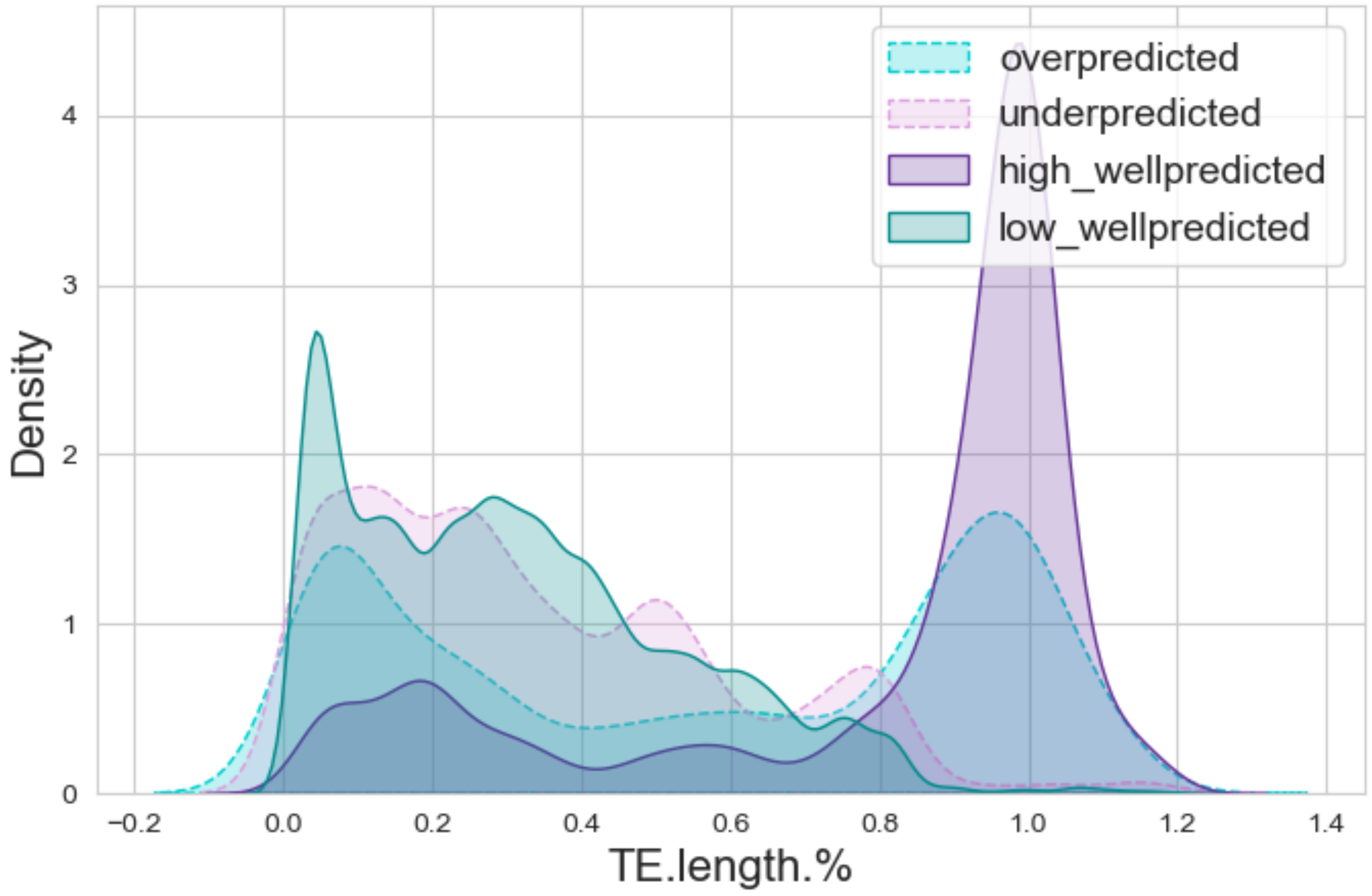
- Can we separate {*underpredicted* vs. *low_wellpredicted*} and {*overpredicted* vs. *high_wellpredicted*}?

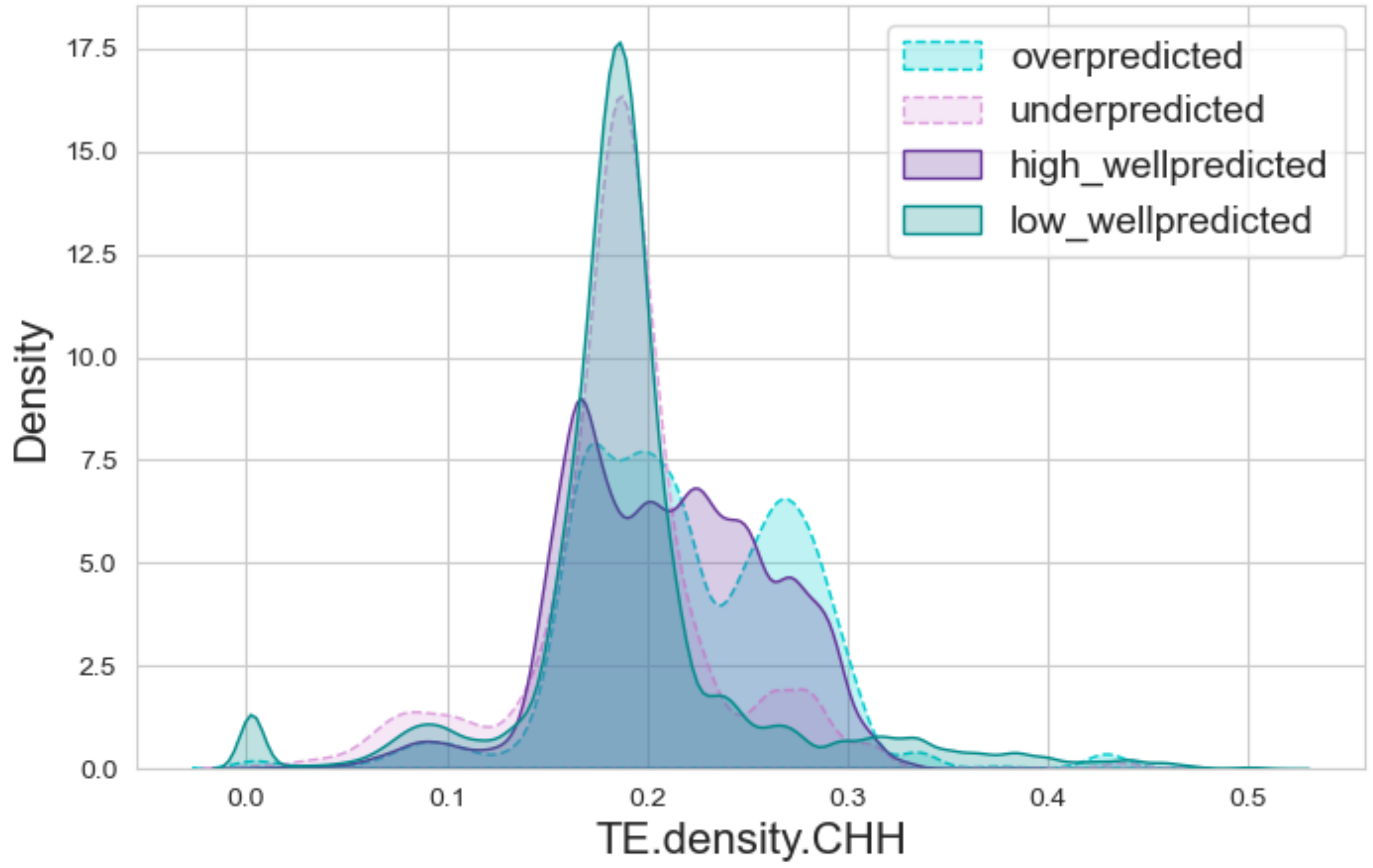
Modeling TE methylation

Who are the escapees?

- Can we separate {*underpredicted* vs. *low_wellpredicted*} and {*overpredicted* vs. *high_wellpredicted*}?

- For the most of the features, the groups have similar distributions
  (so, the model is reasonable wrt the data)



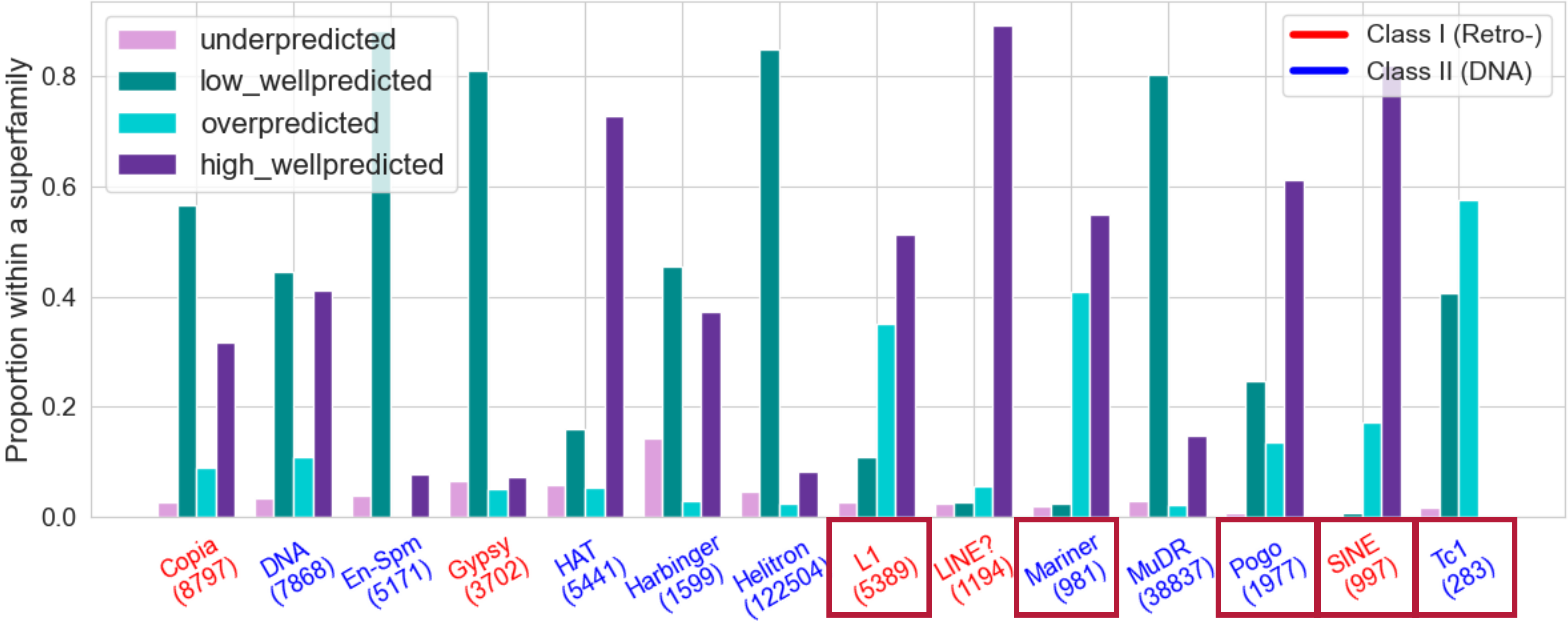


Modeling TE methylation

Who are the escapees?

- Can we separate {*underpredicted* vs. *low_wellpredicted*} and {*overpredicted* vs. *high_wellpredicted*}?

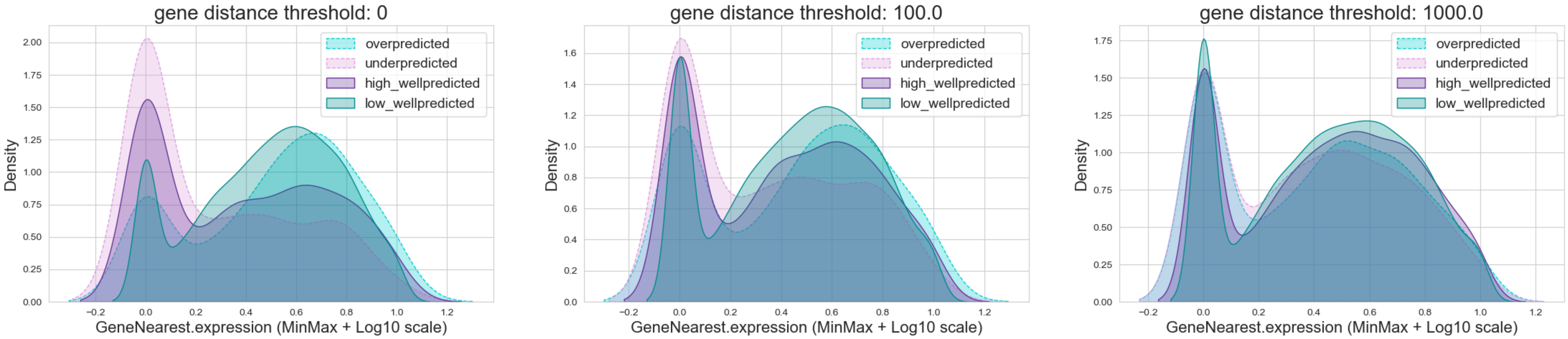
- For the most of the features, the groups have similar distributions
  (so, the model is reasonable wrt the data)



Modeling TE methylation

Who are the escapees?

- <u>**Hypothesis**</u>: ***selective pressure*** may be one of the factors
  (some genes need to be expressed, some need to be silenced)

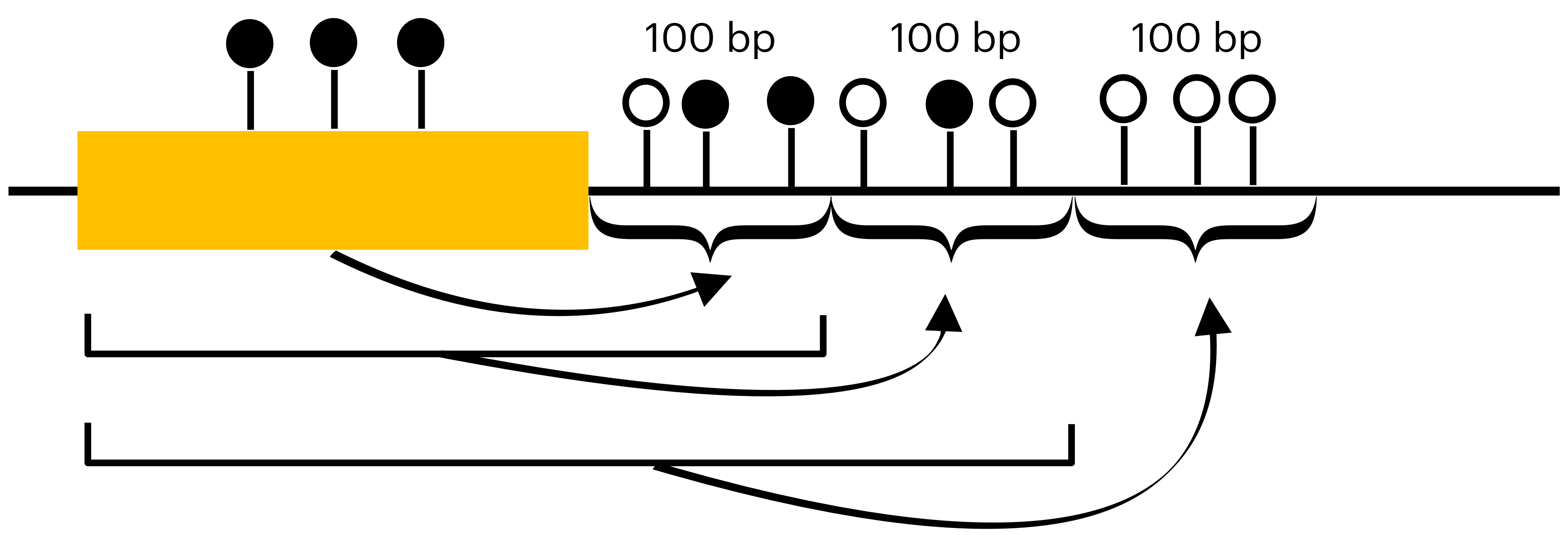


Distance from the gene

Modeling TE methylation

- The **predictive model is accurate** within appropriate range
- The **length in % wrt to the reference** (proxy for age) is the most informative feature
  (= **young TEs tend to be methylated**)
- **Context densities** play an important role, as well as superfamilies
- There are **escapees** in both directions (therefore, some missed factors)
- An example of possible factor: **selective pressure for gene expression**

Modeling methylation spreading

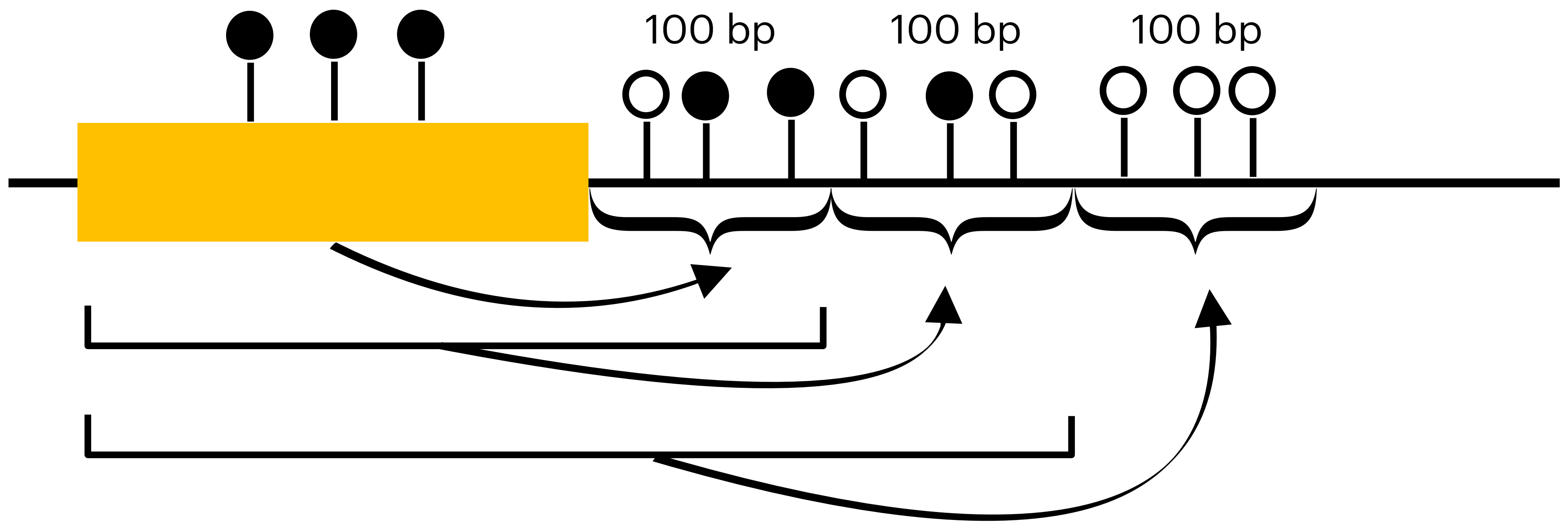


Model:
Random Forest
(hyper parameters tuned via cross-validation stratified by TIPs)

Features:
- **TE** (length, distance to pericentromere, superfamily, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Methylation** in CG, CHG, CHH contexts (average genome-wide, TE, edges of TE, previous windows)
- **Densities** of CG, CHG, CHH contexts

Data:
Only methylated TEs (107.950)

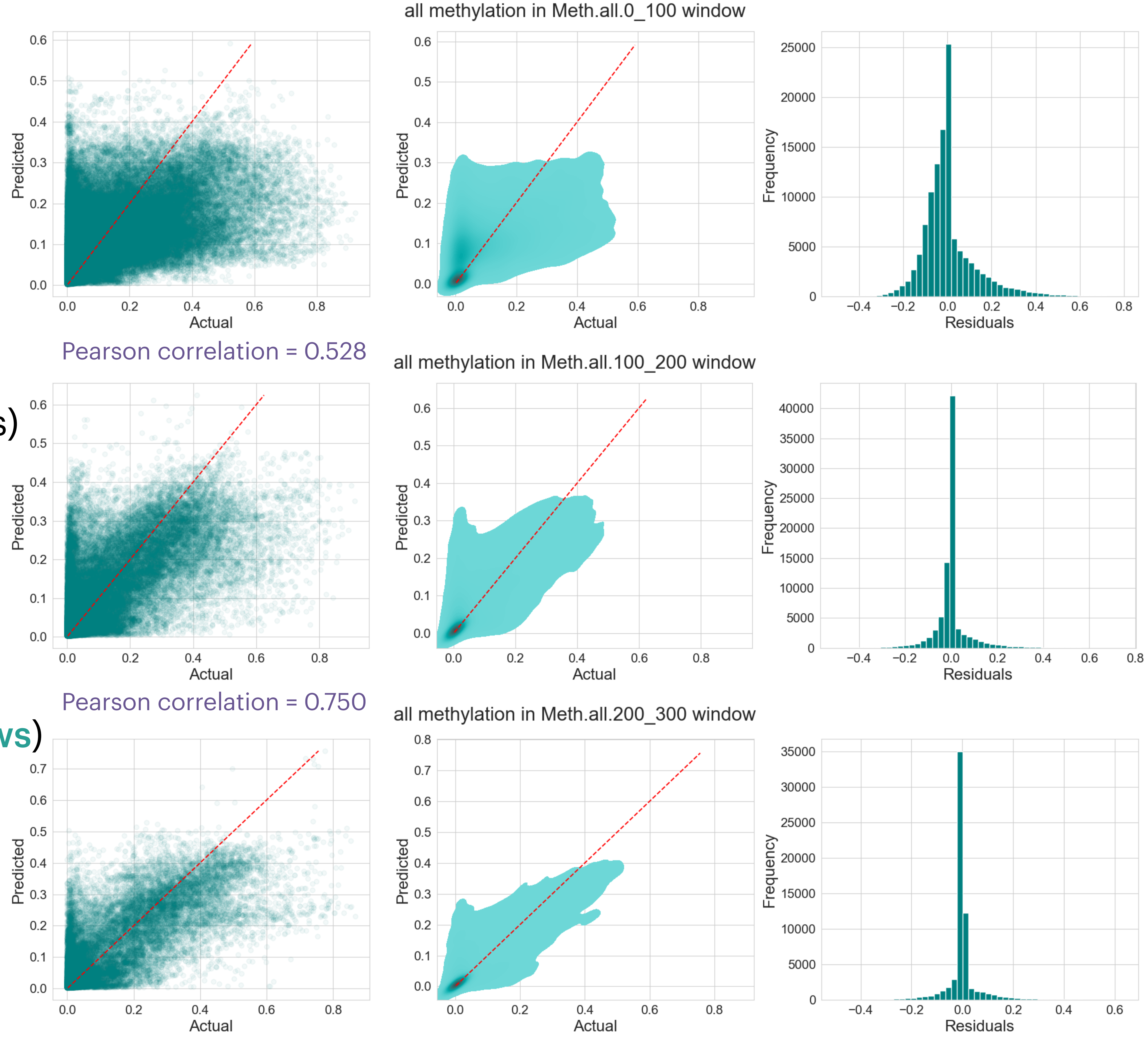
Modeling methylation spreading



Model:
Random Forest
(hyper parameters tuned via cross-validation stratified by TIPs)

Features:
- **TE** (length, distance to pericentromere, superfamily, if inside a gene)
- **Nearest 2 genes** (length, distance, relative direction)
- **Methylation** in CG, CHG, CHH contexts (average genome-wide, TE, edges of TE, previous windows)
- **Densities** of CG, CHG, CHH contexts

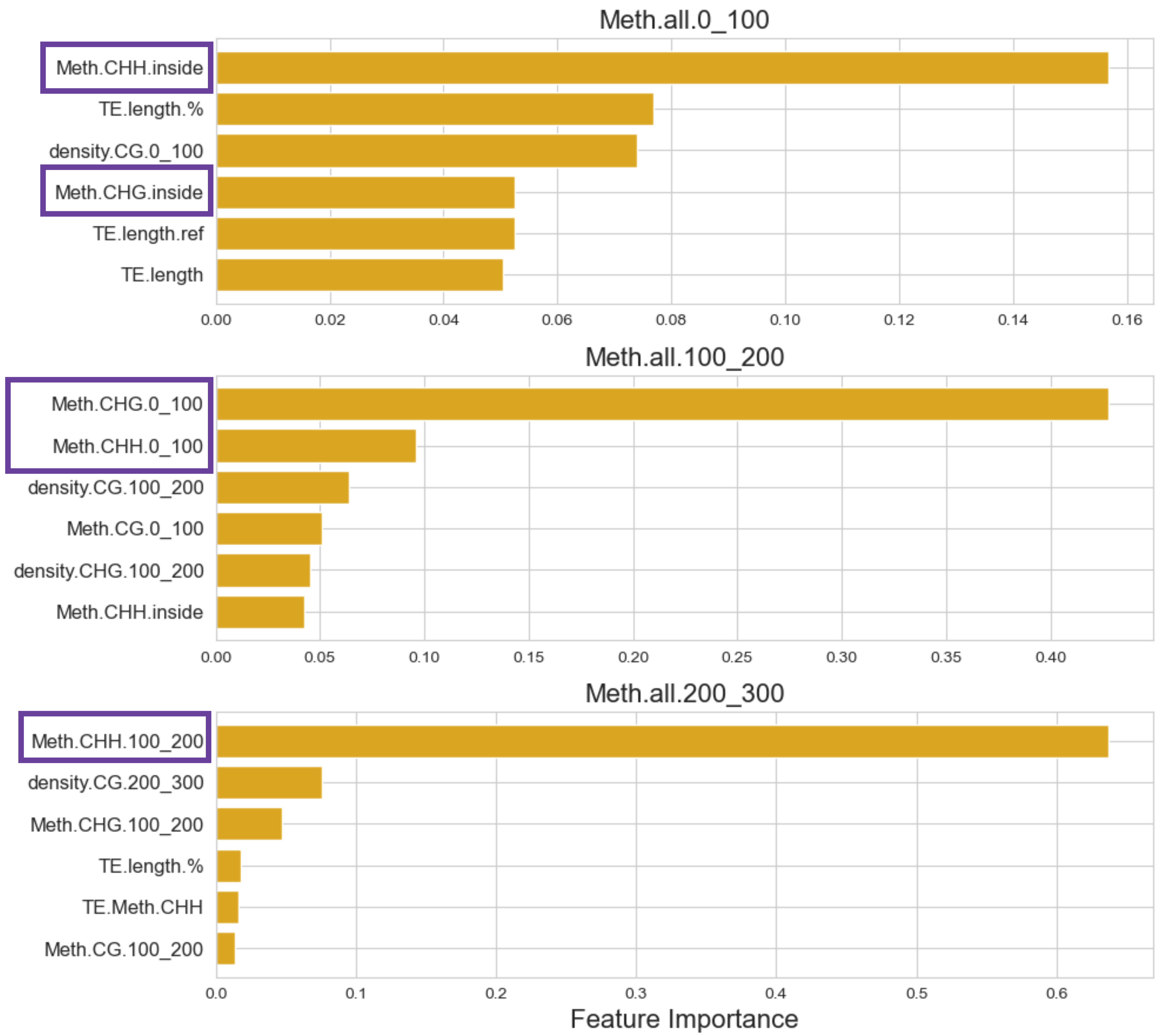
Data:
Only methylated TEs (107.950)



Modeling methylation spreading

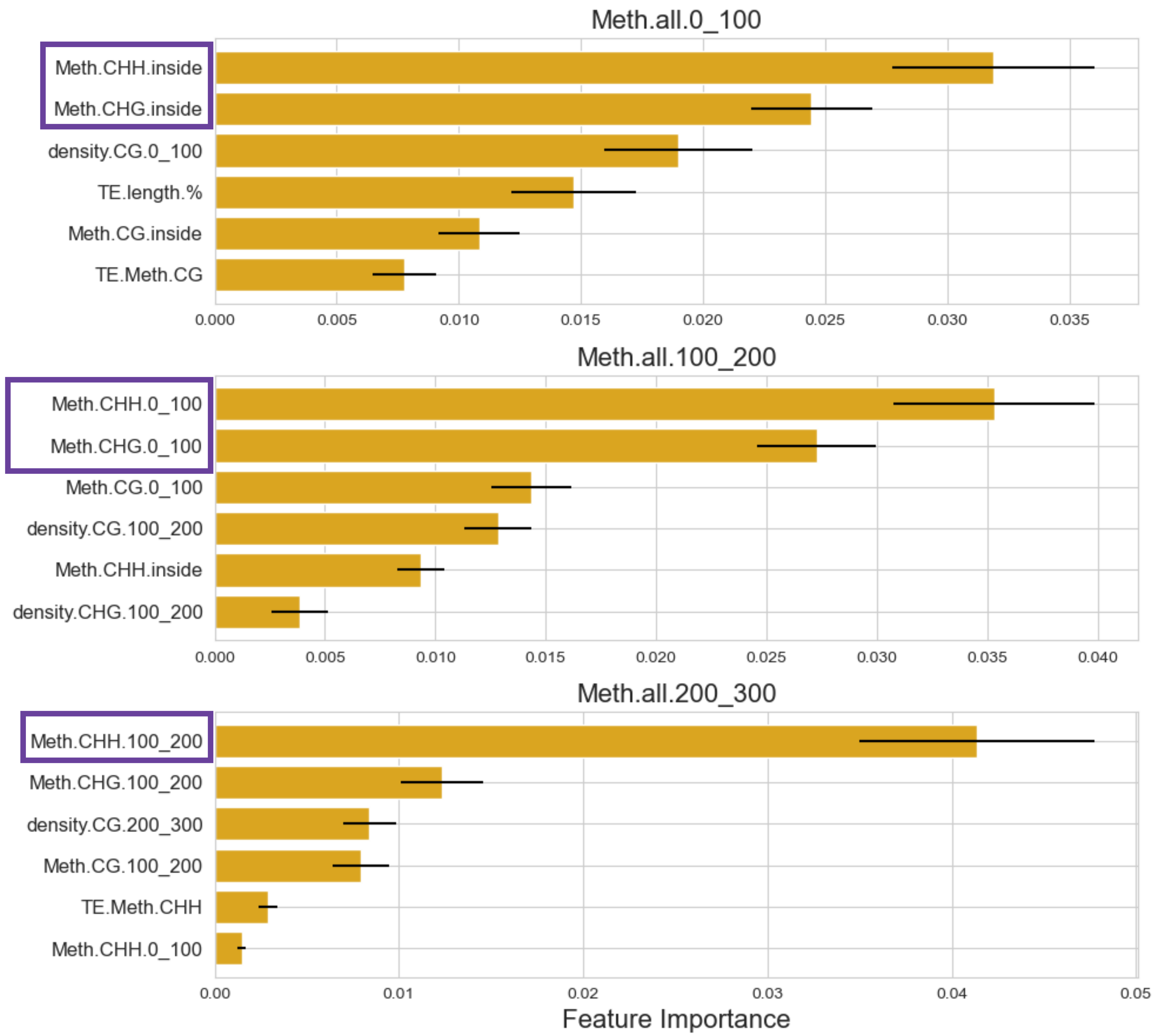
Impurity-based feature importances

* 10 independent runs with different random seeds



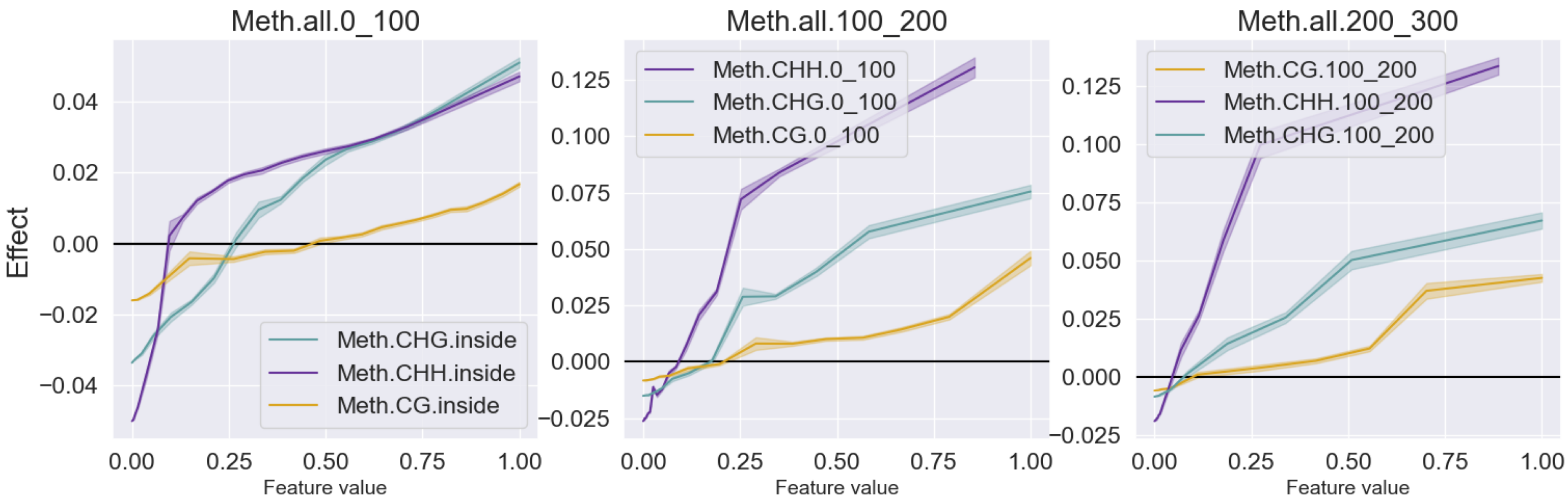
SHAP values

* 10 folds in a cross-validation manner

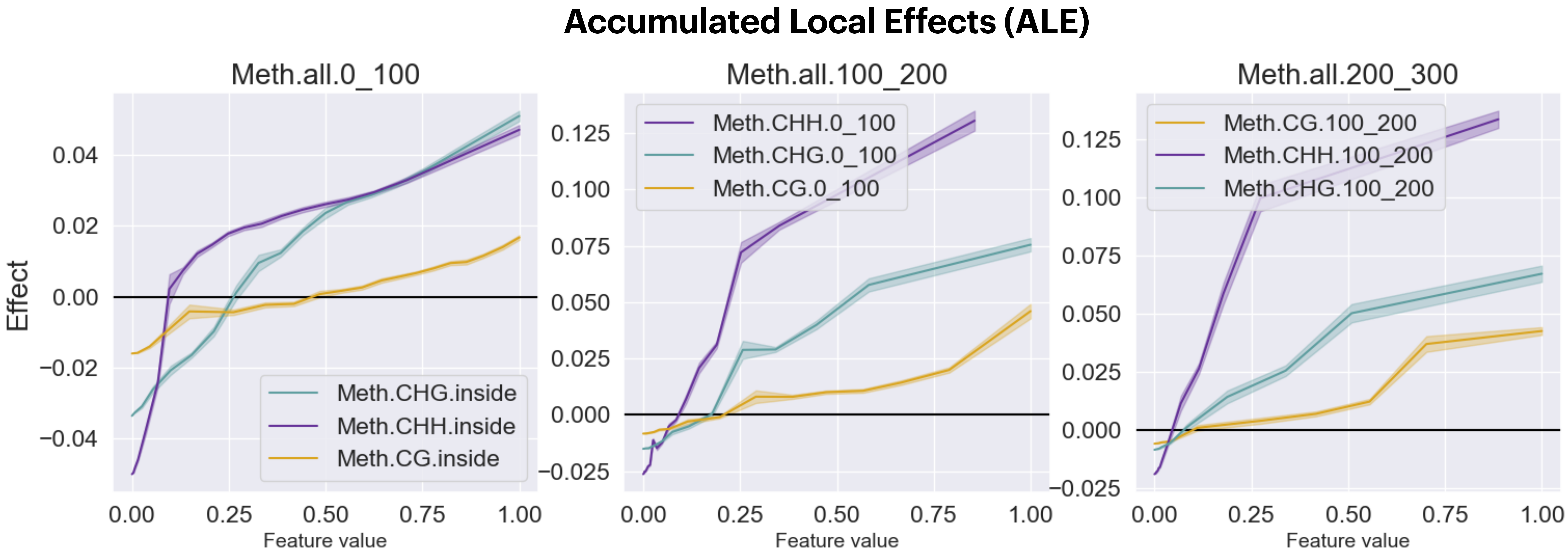


Modeling methylation spreading

Accumulated Local Effects (ALE)



Modeling methylation spreading

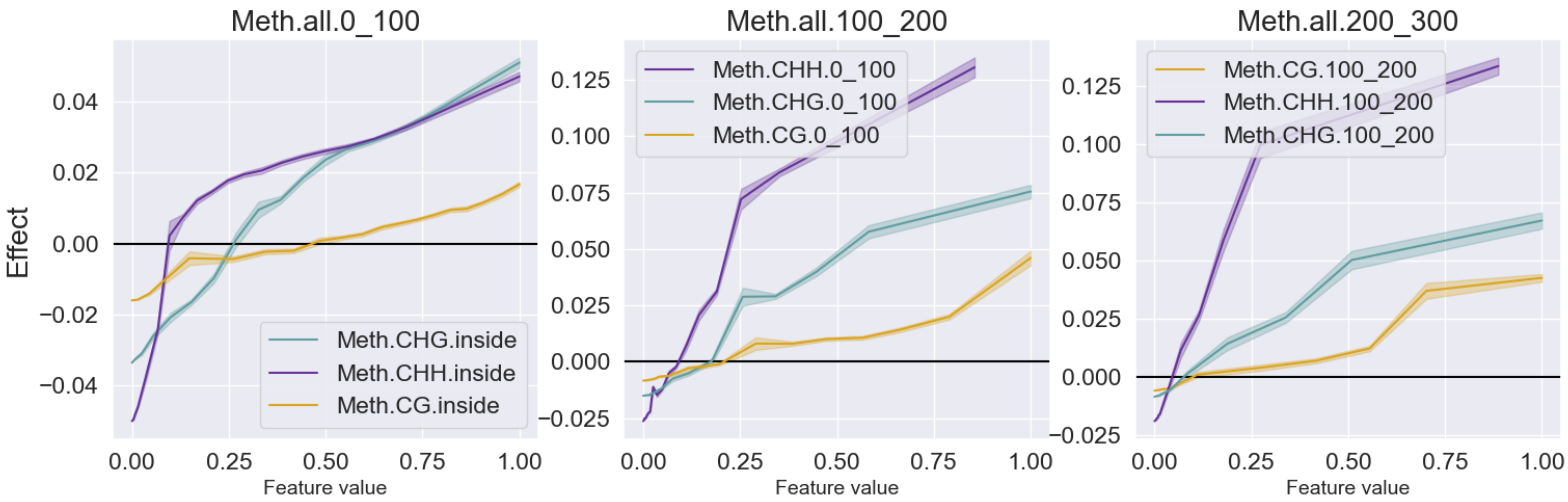


Conclusion:

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

Modeling methylation spreading

Accumulated Local Effects (ALE)



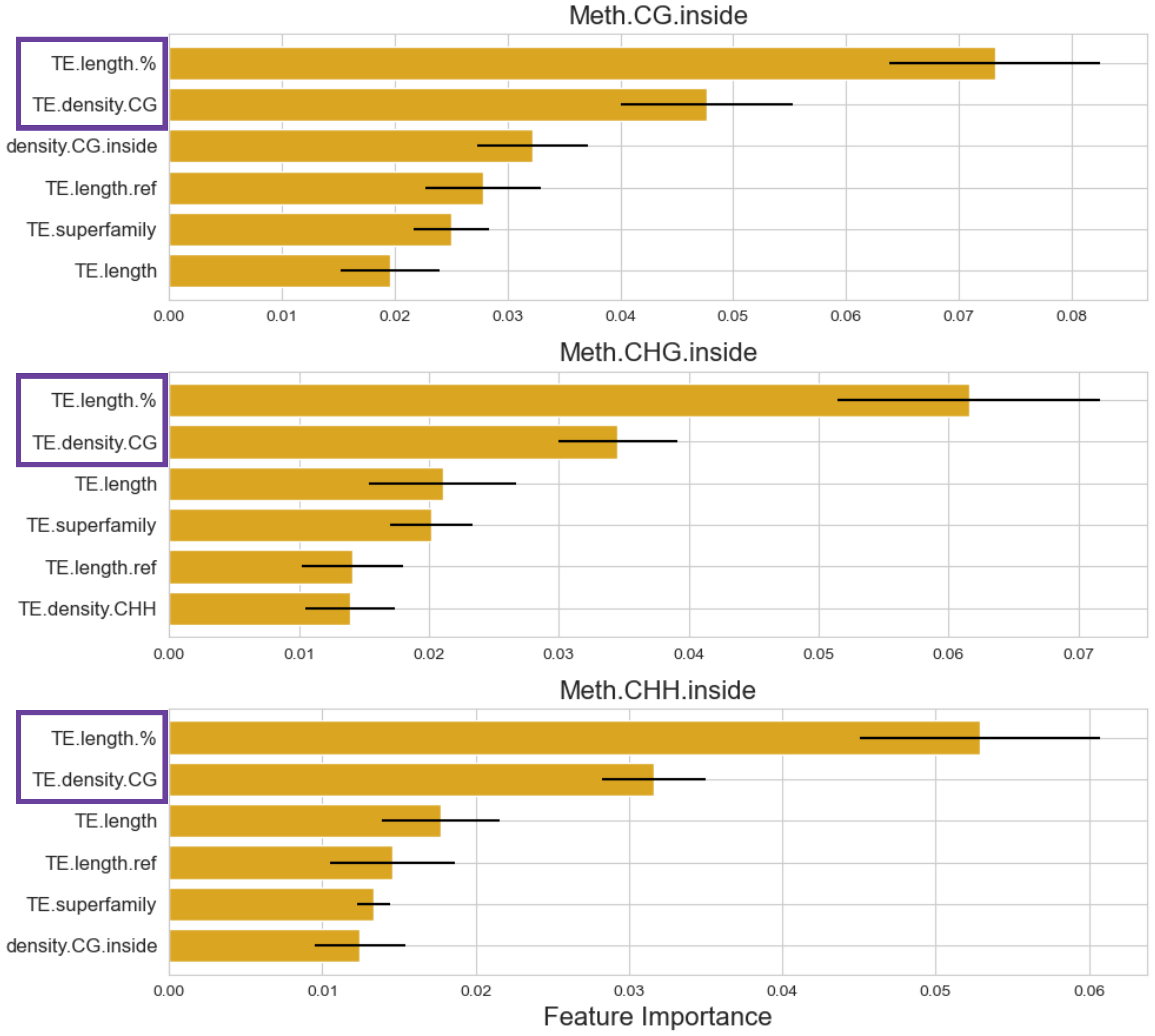
Conclusion:

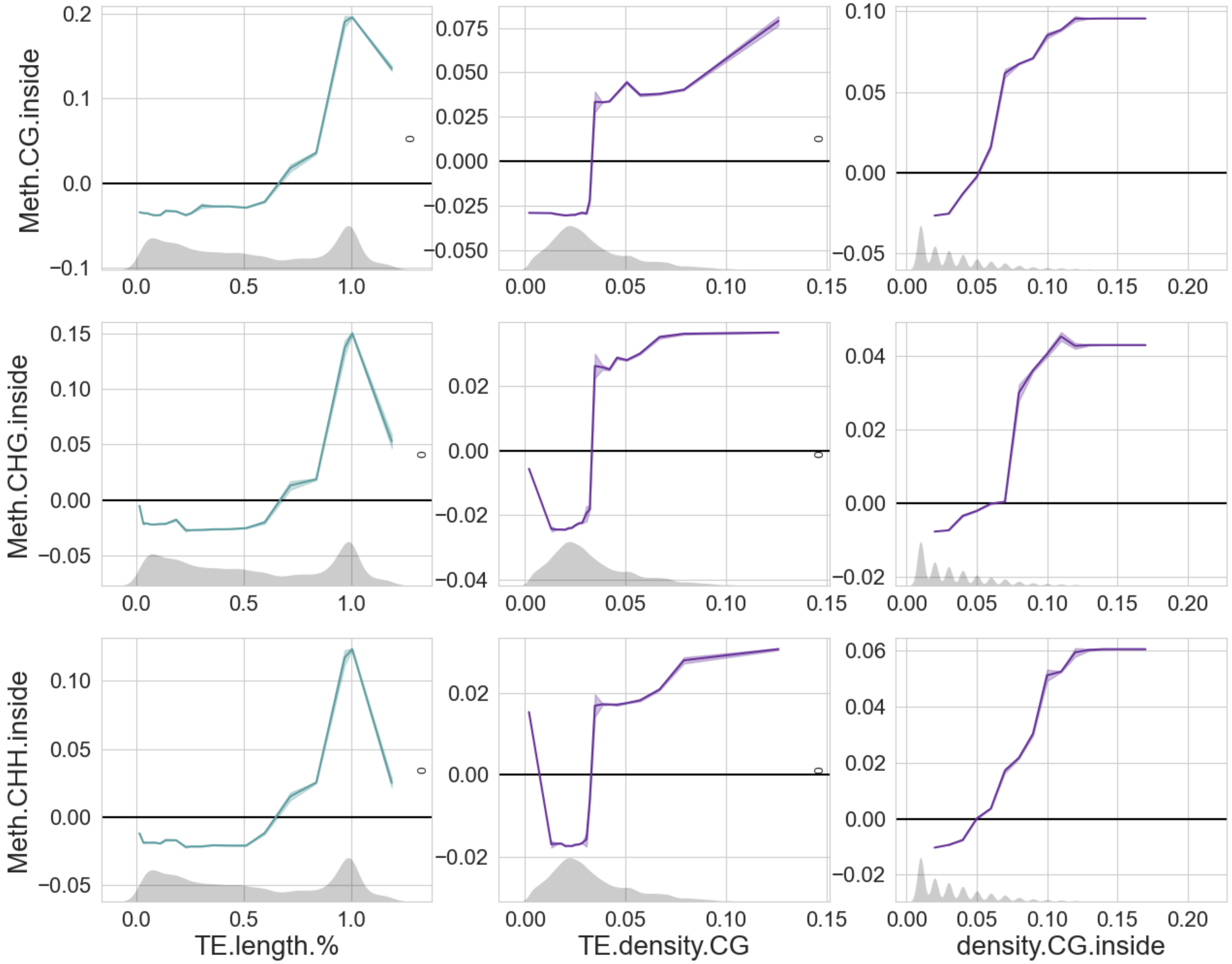
- **Methylation of the TE edges** consistently comes as the most important feature with monotonous effect increase
- TE is **methylated on the edges** $\Longrightarrow$ more **likely to spread**

Question:

- What defines the **methylation of the TE edges**?

Modeling edges methylation





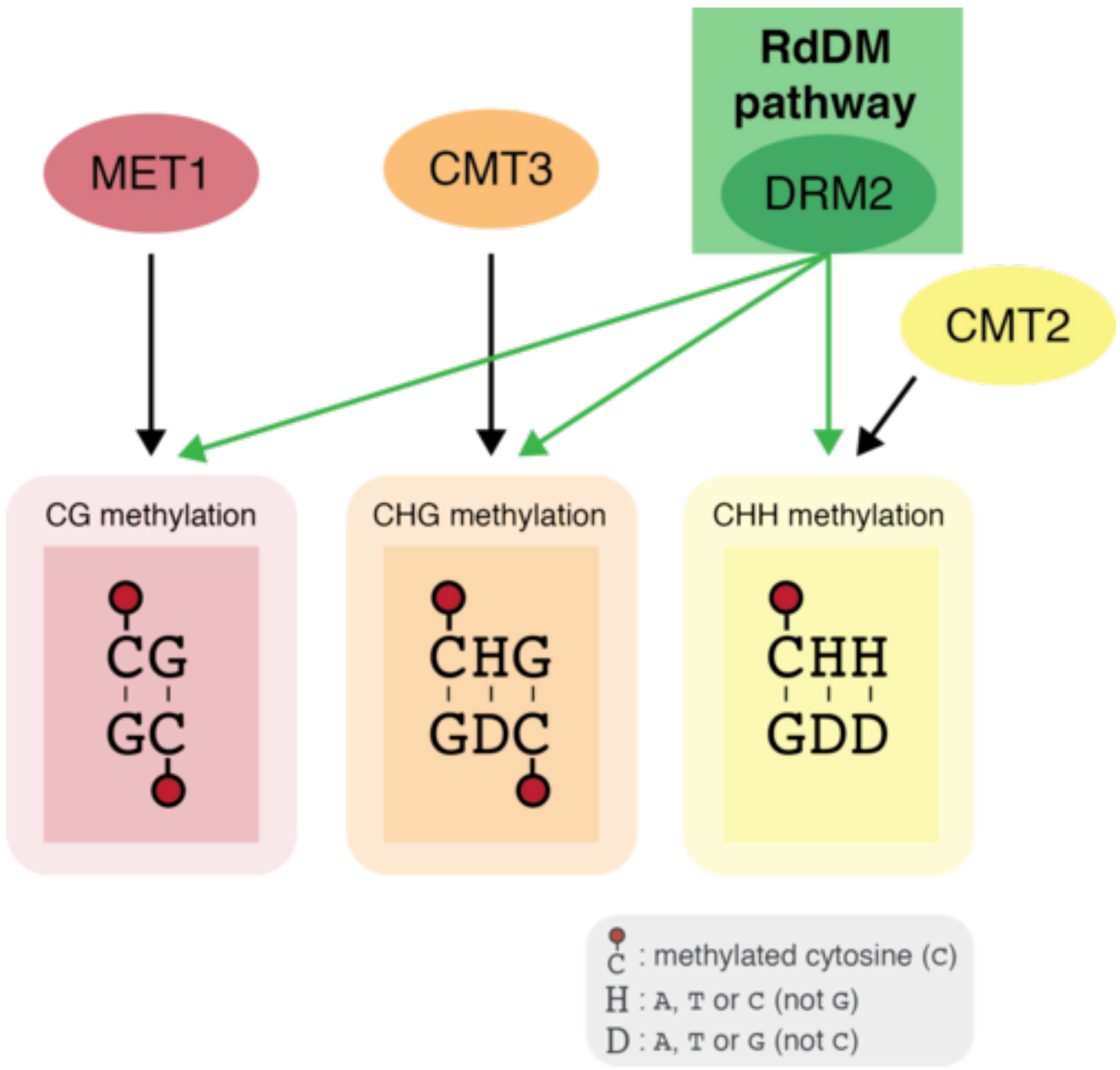
Back to biology of methylation

- **The most important factors for spreading:**
  - methylation of the TE edges in the CHG and CHH contexts
  - % of full length (proxy for the TE age)
  - density of CG contexts

Back to biology of methylation

- **The most important factors for spreading:**
  - methylation of the TE edges in the CHG and CHH contexts
  - % of full length (proxy for the TE age)
  - density of CG contexts

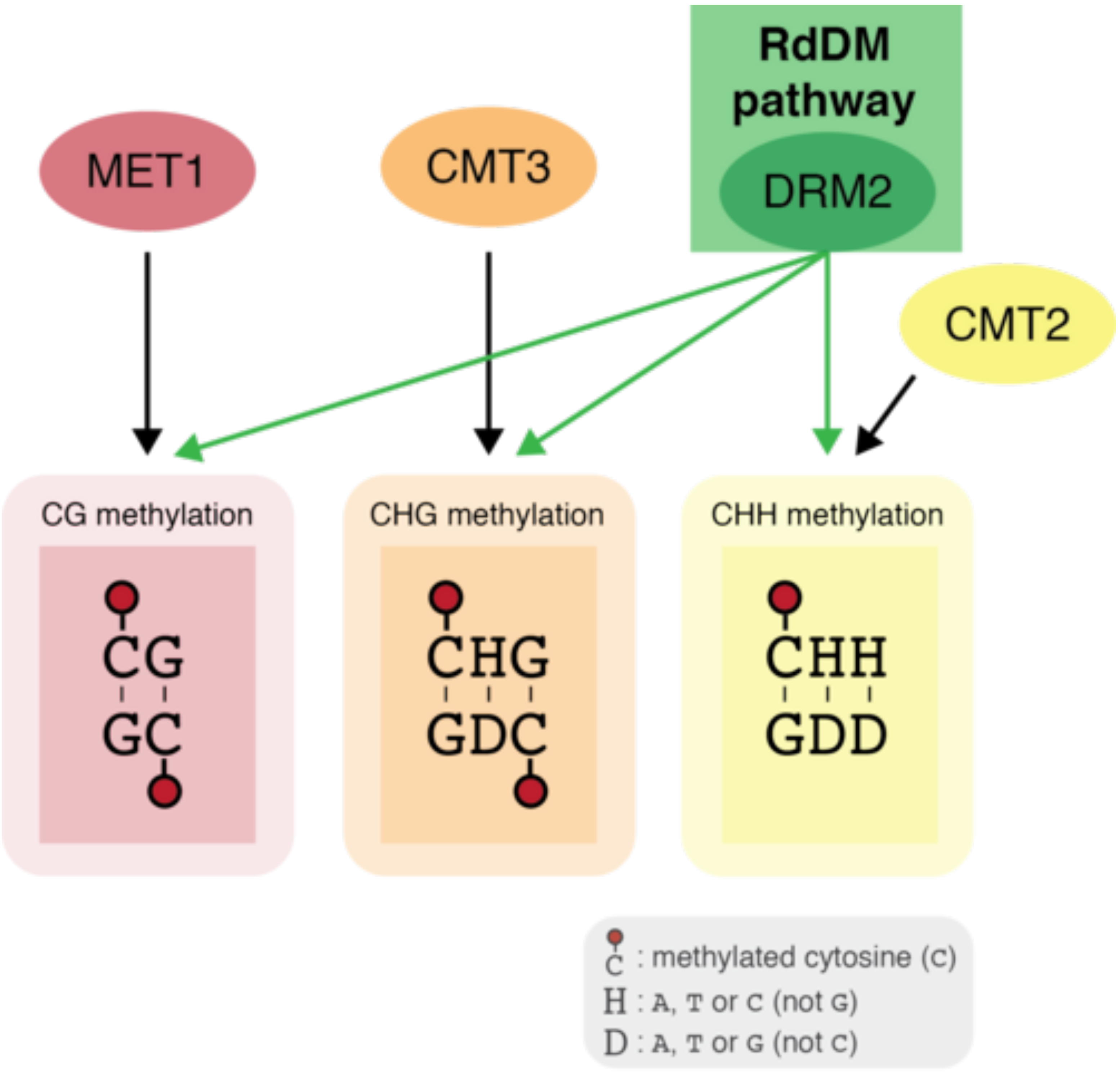
- **Hypothesis:** the **non-canonical RdDM** machinery is responsible for spreading
  - targets all contexts (CG, CHG, CHH)
  - the <u>only</u> pathway capable of adding DNA methylation *de novo*



Erdmann & Picard, *PLOS Genetics 2020*

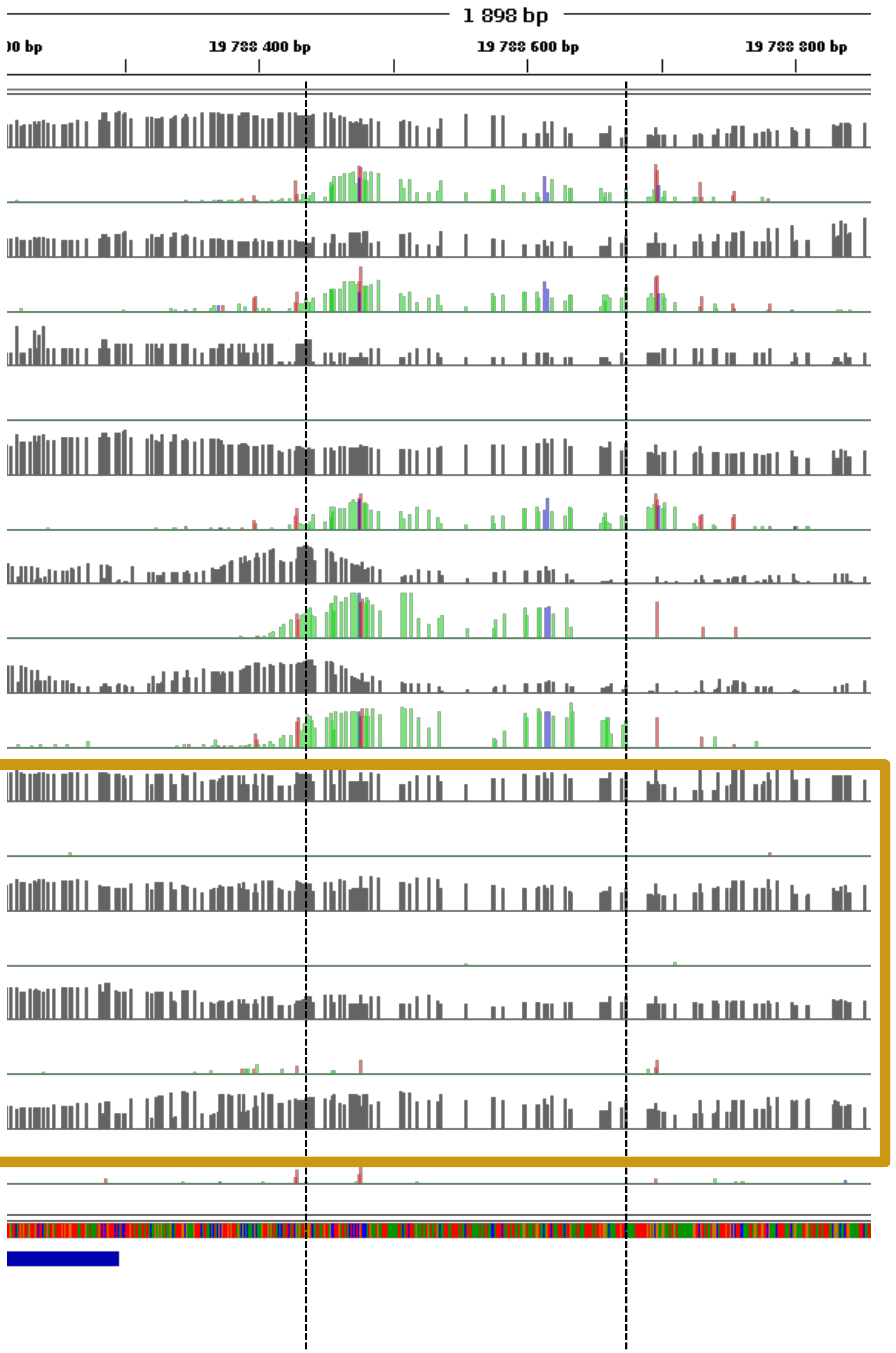
Back to biology of methylation

- **The most important factors for spreading:**
  - methylation of the TE edges in the CHG and CHH contexts
  - % of full length (proxy for the TE age)
  - density of CG contexts
- **Hypothesis:** the **non-canonical RdDM** machinery is responsible for spreading
  - targets all contexts (CG, CHG, CHH)
  - the <u>only</u> pathway capable of adding DNA methylation *de novo*
- **Test:** mutants of Col-0 strain of *A. Thaliana* where different methylation pathways are knocked out

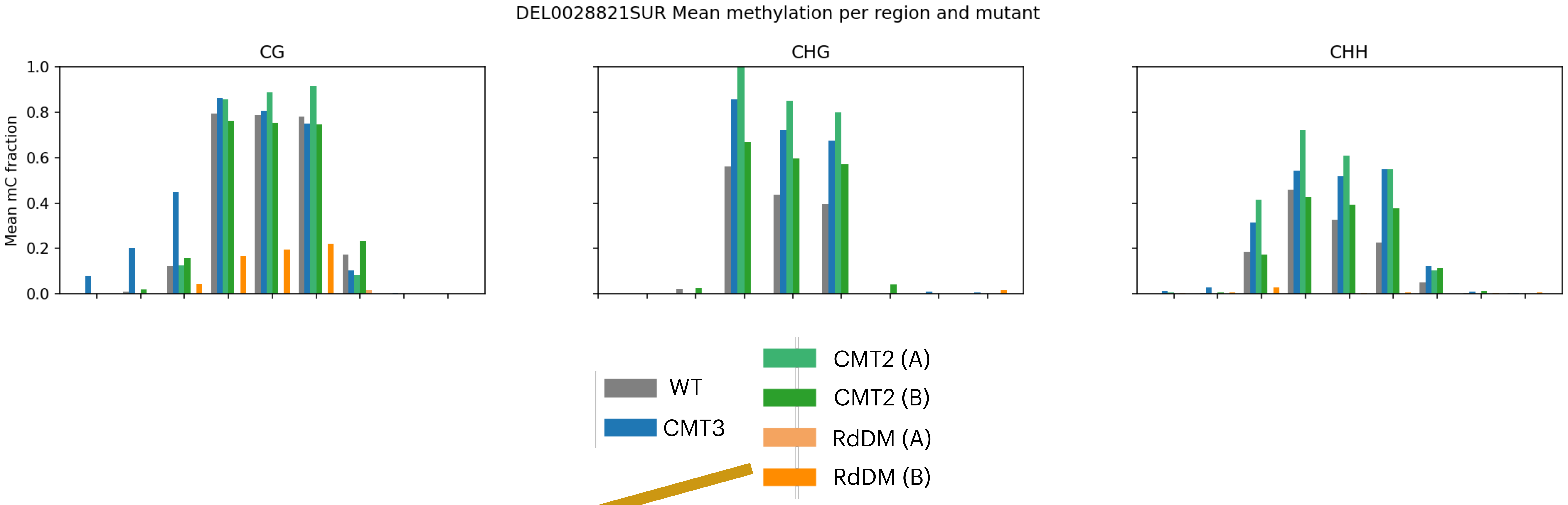


Erdmann & Picard, *PLOS Genetics 2020*

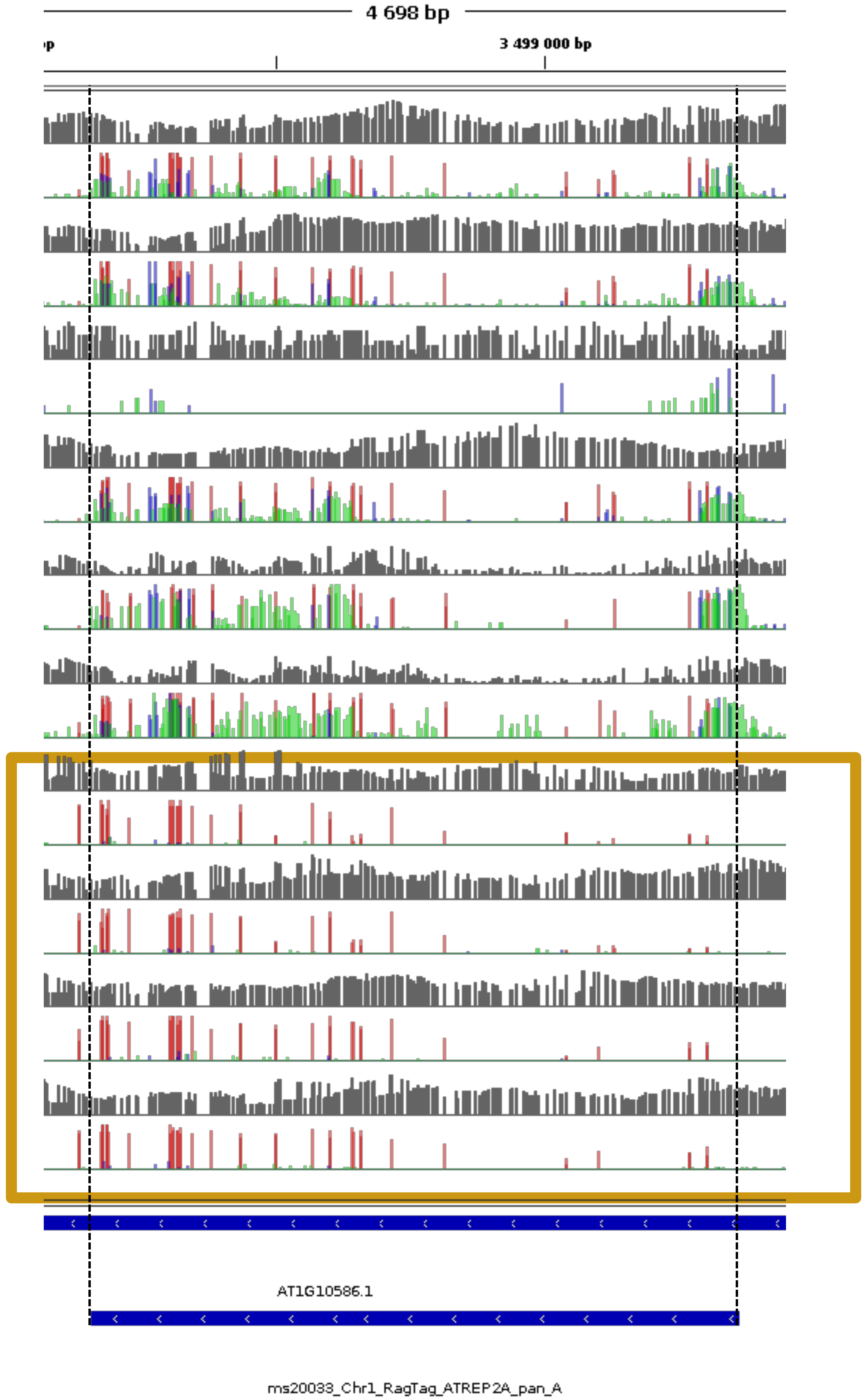
Back to biology of methylation



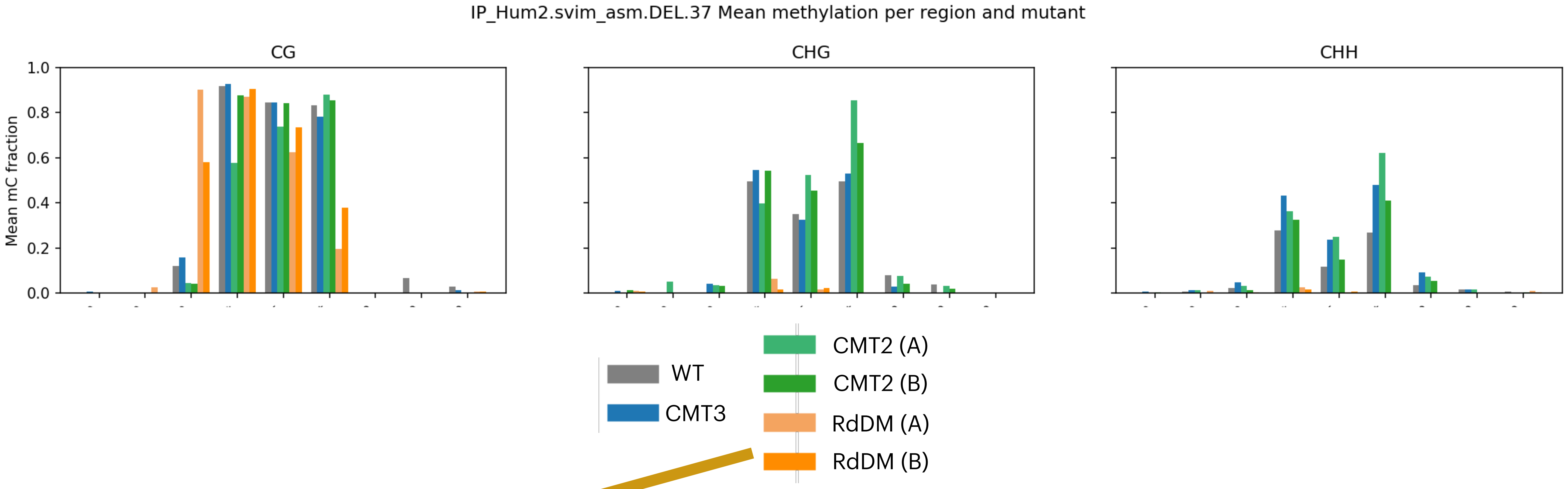
DEL0028821SUR



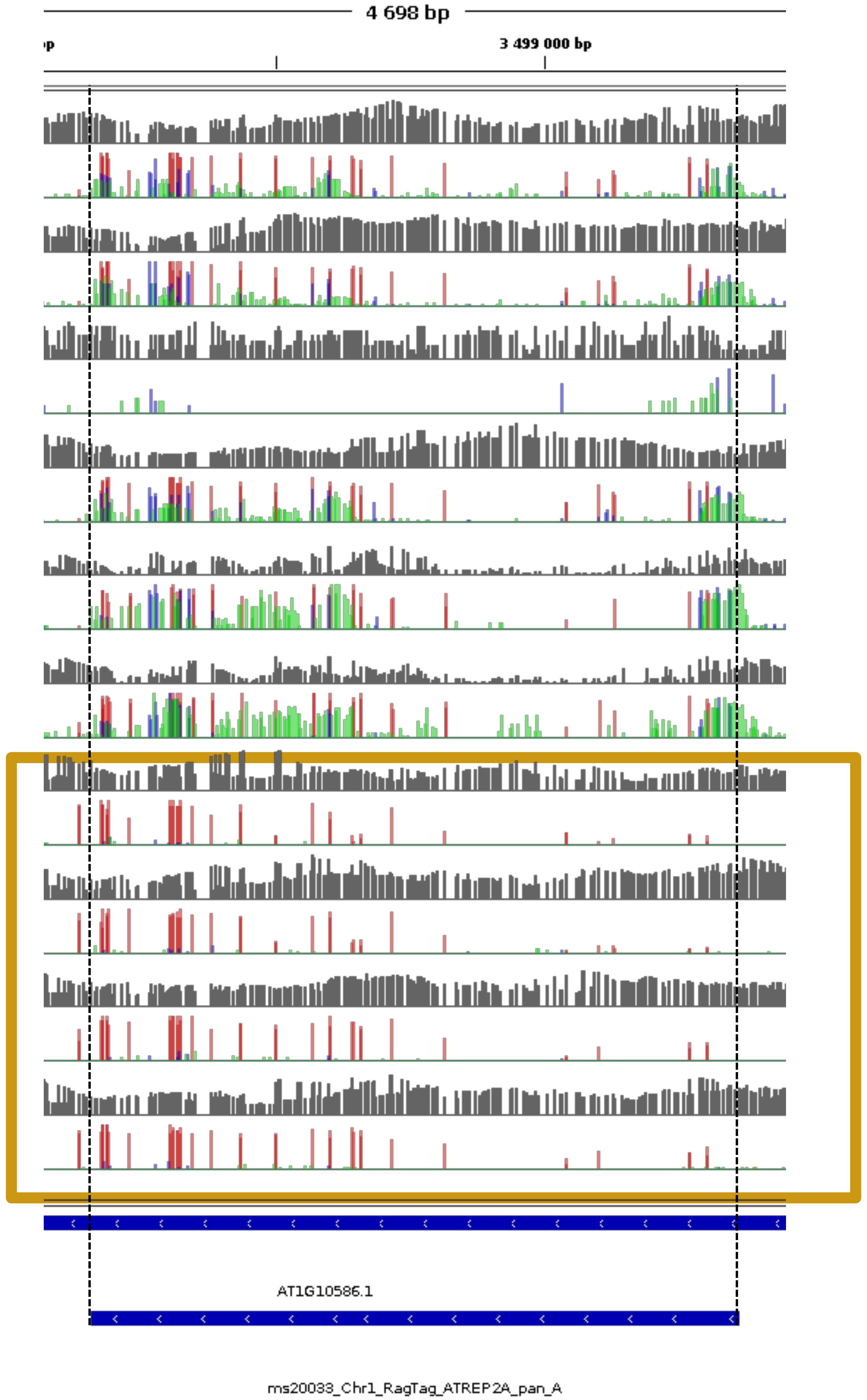
Back to biology of methylation



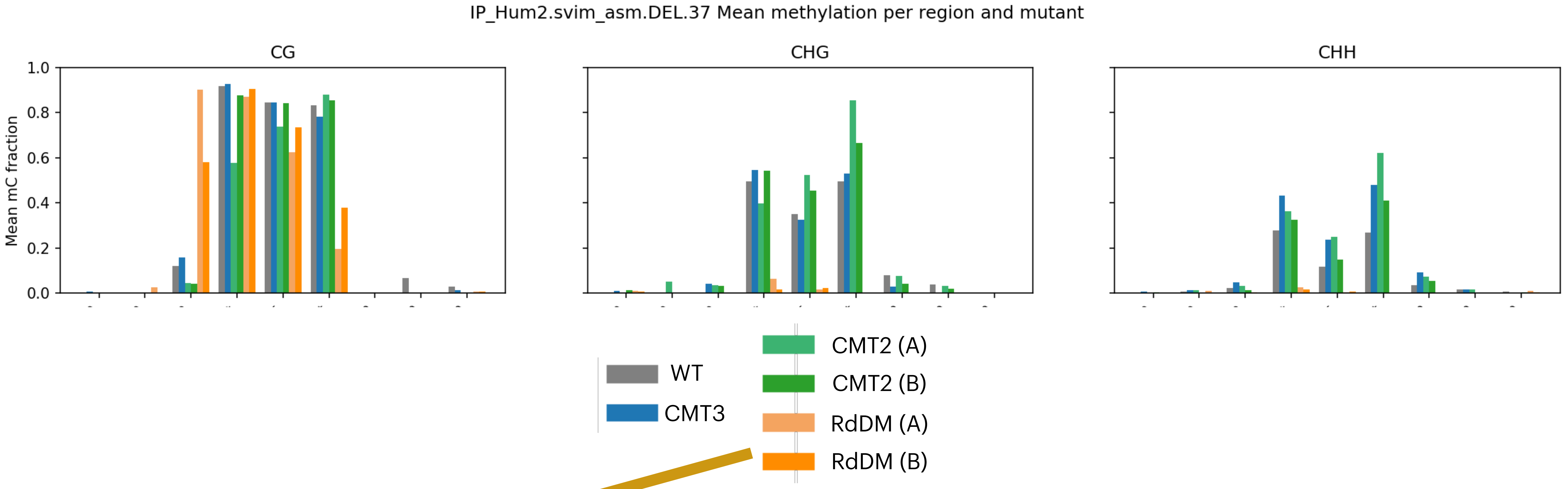
IP_Hum2.svim_asm.DEL.37



Back to biology of methylation



IP_Hum2.svim_asm.DEL.37



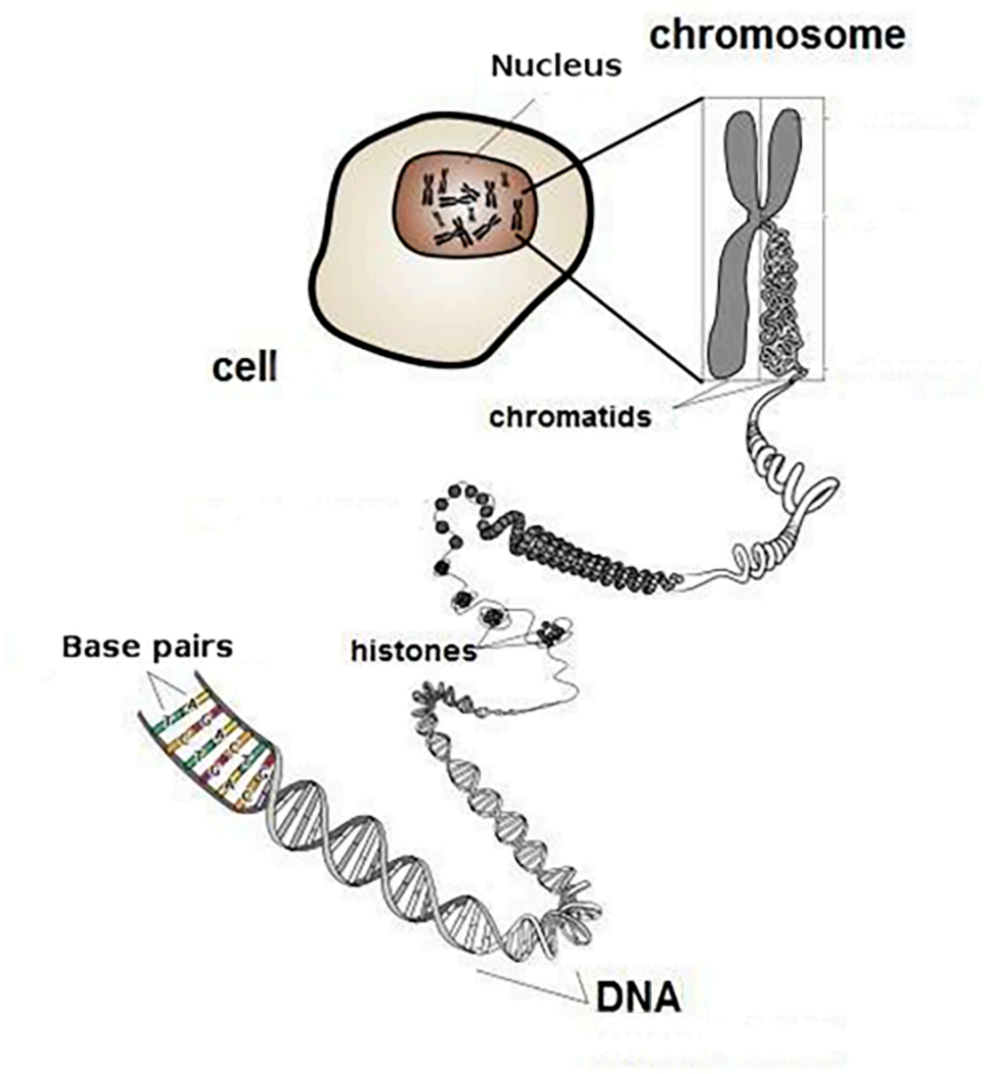
CHG and CHH methylation (and spreading!) disappear in RdDM mutants

Back to biology of methylation

- The **predictive model is accurate** within appropriate range
- Different **explainability tools** have been explored, and they provide **consistent conclusions**
- For spreading, a potential actor (**RdDM**) is identified

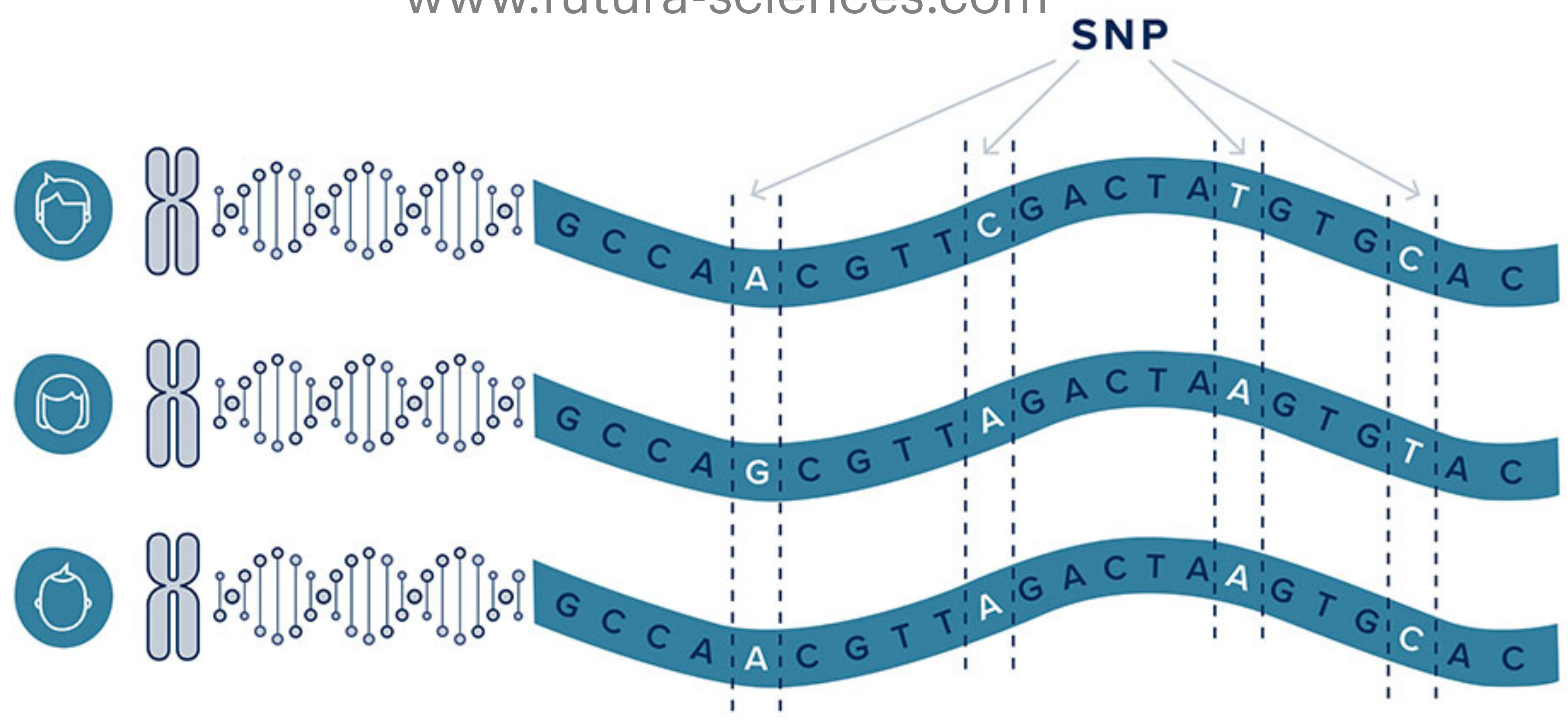
Part III: associations with gene expression

From genotype to phenotype

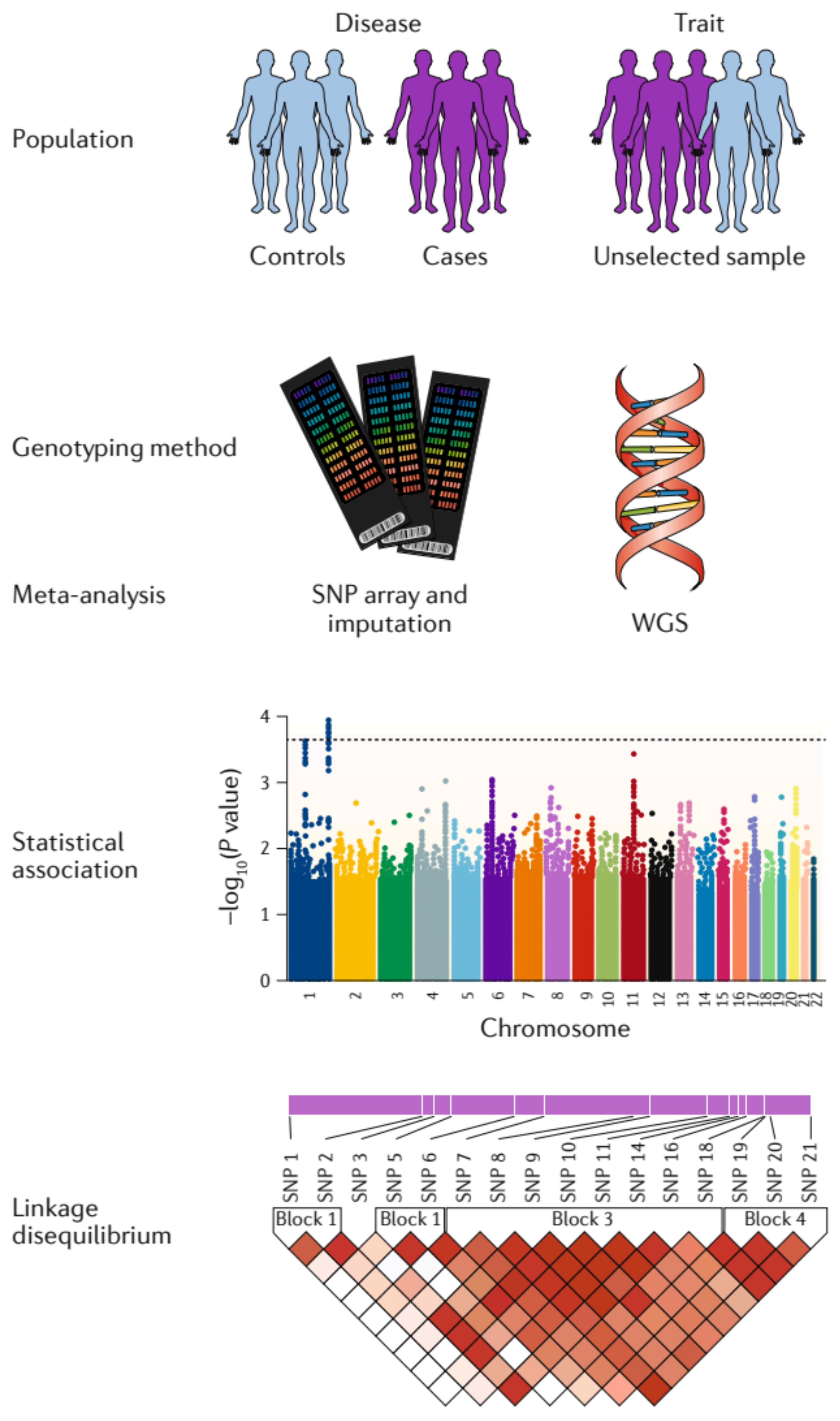


www.futura-sciences.com

Genome-Wide Association Study

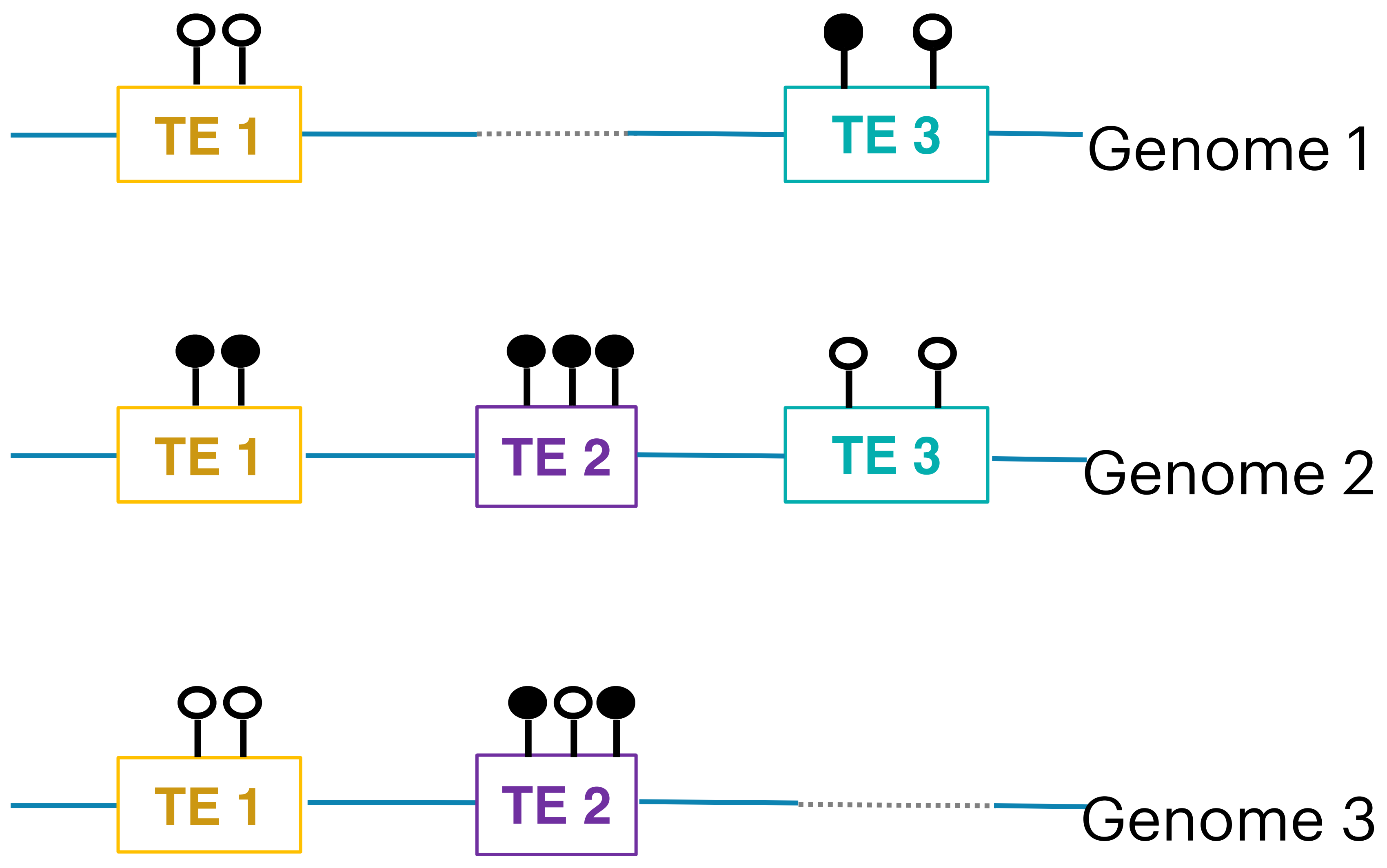


Scientific DX GmbH, 2020



Tam et al., *Nature Reviews Genetics 2019*

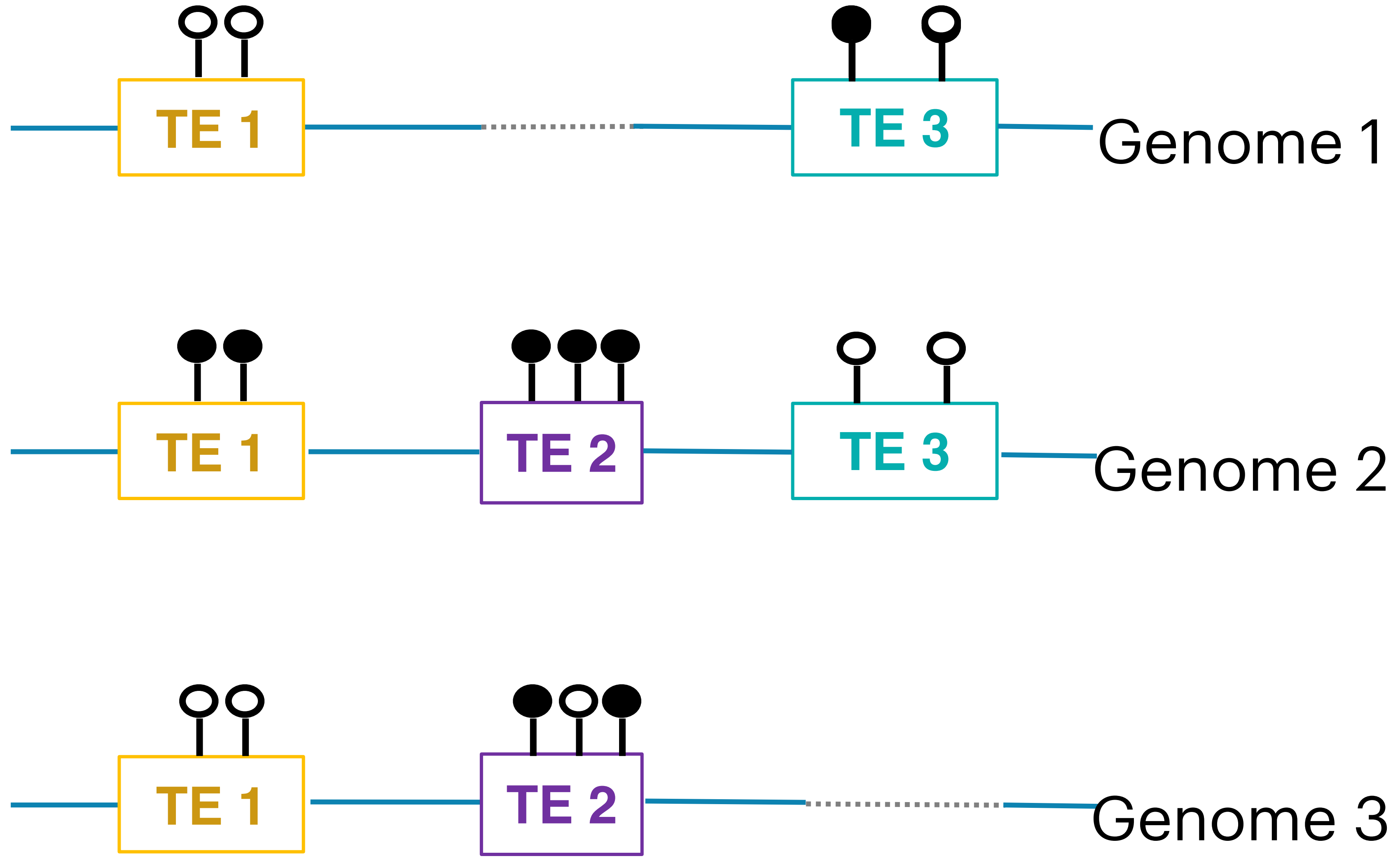
From epi-genotype to phenotype



Genome-Wide Association Study

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

From epi-genotype to phenotype



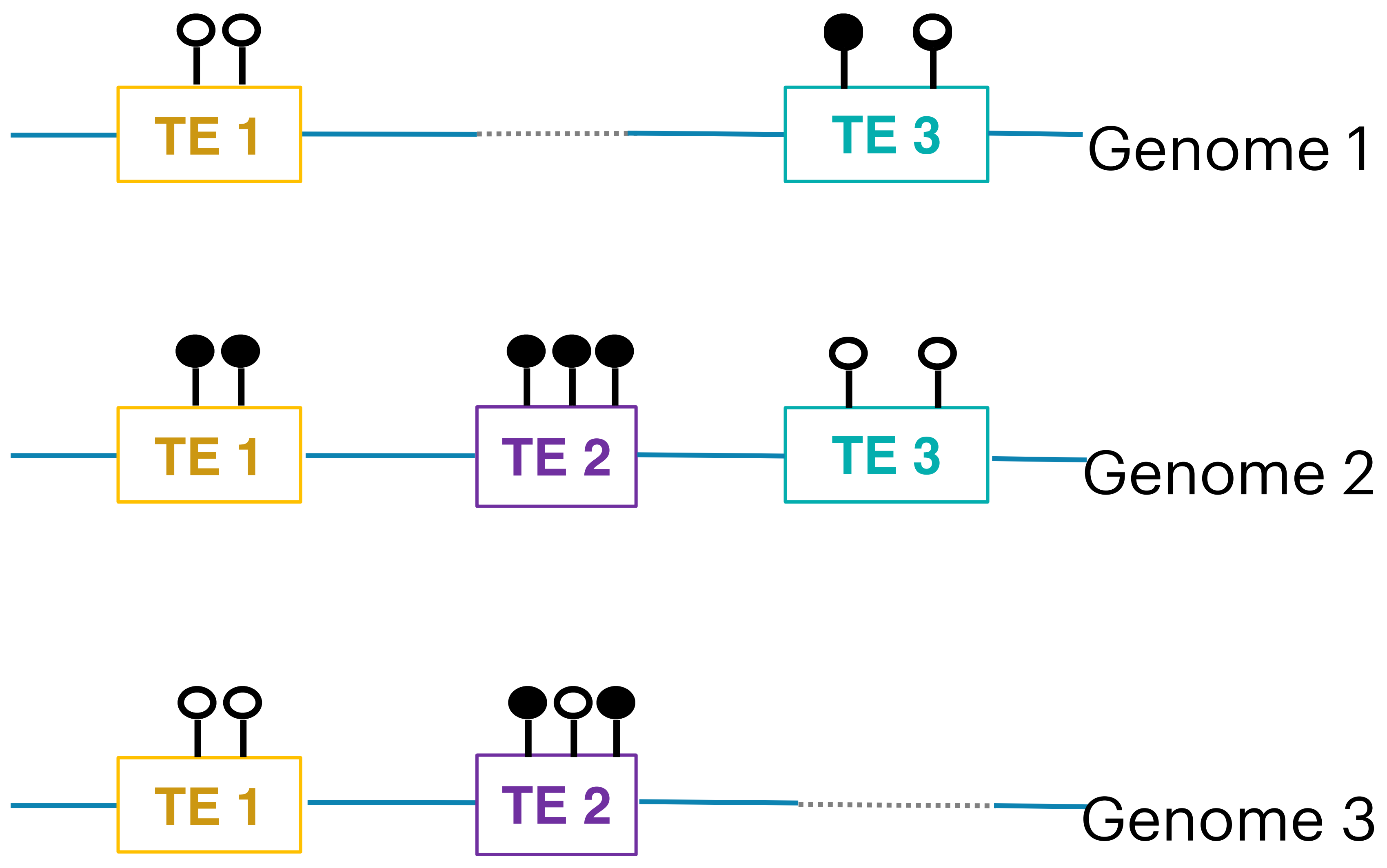
Genome-Wide Association Study

3 groups:

00 = absent
10 = present and not methylated (< 5%)
11 = present and methylated (> 5%)
Kruskal-Wallis test (instead of t-test)

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

From epi-genotype to phenotype



Genome-Wide Association Study

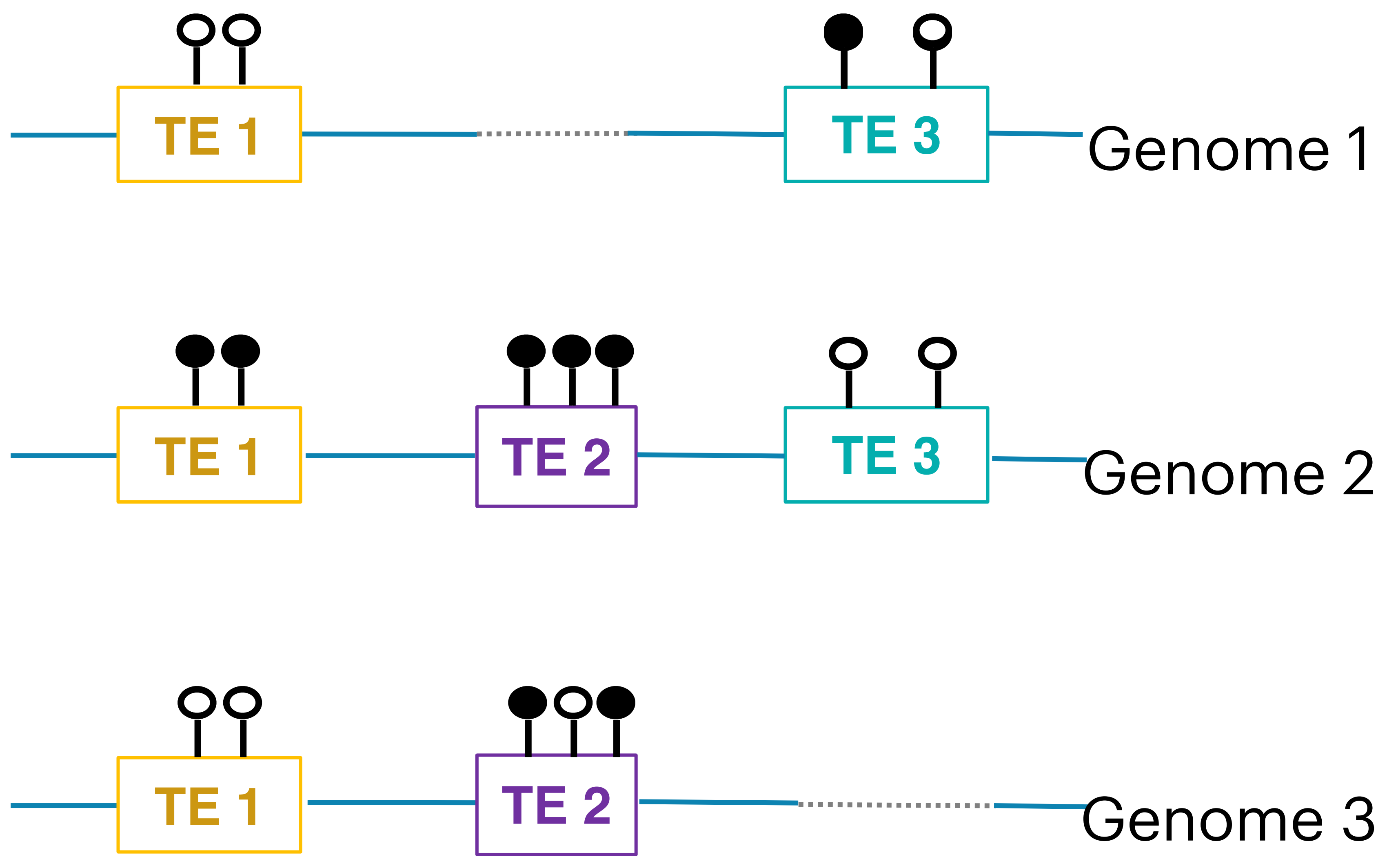
3 groups:

00 = absent
10 = present and not methylated (< 5%)
11 = present and methylated (> 5%)
<u>**Kruskal-Wallis test**</u> (instead of t-test)

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

<u>Setting</u>:
Genotypes: 50 genomes * 9.557 mTIPs
Phenotypes: 37k genes (including alternatively spliced)
Standard GWAS pipeline (quality controls, *statistical testing, Bonferroni corrections)

From epi-genotype to phenotype



Genome-Wide Association Study

3 groups:

00 = absent
10 = present and not methylated (< 5%)
11 = present and methylated (> 5%)
<u>Kruskal-Wallis test</u> (instead of t-test)

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

<u>Setting</u>:
Genotypes: 50 genomes * 9.557 mTIPs
Phenotypes: 37k genes (including alternatively spliced)
Standard GWAS pipeline (quality controls, *statistical testing, Bonferroni corrections)

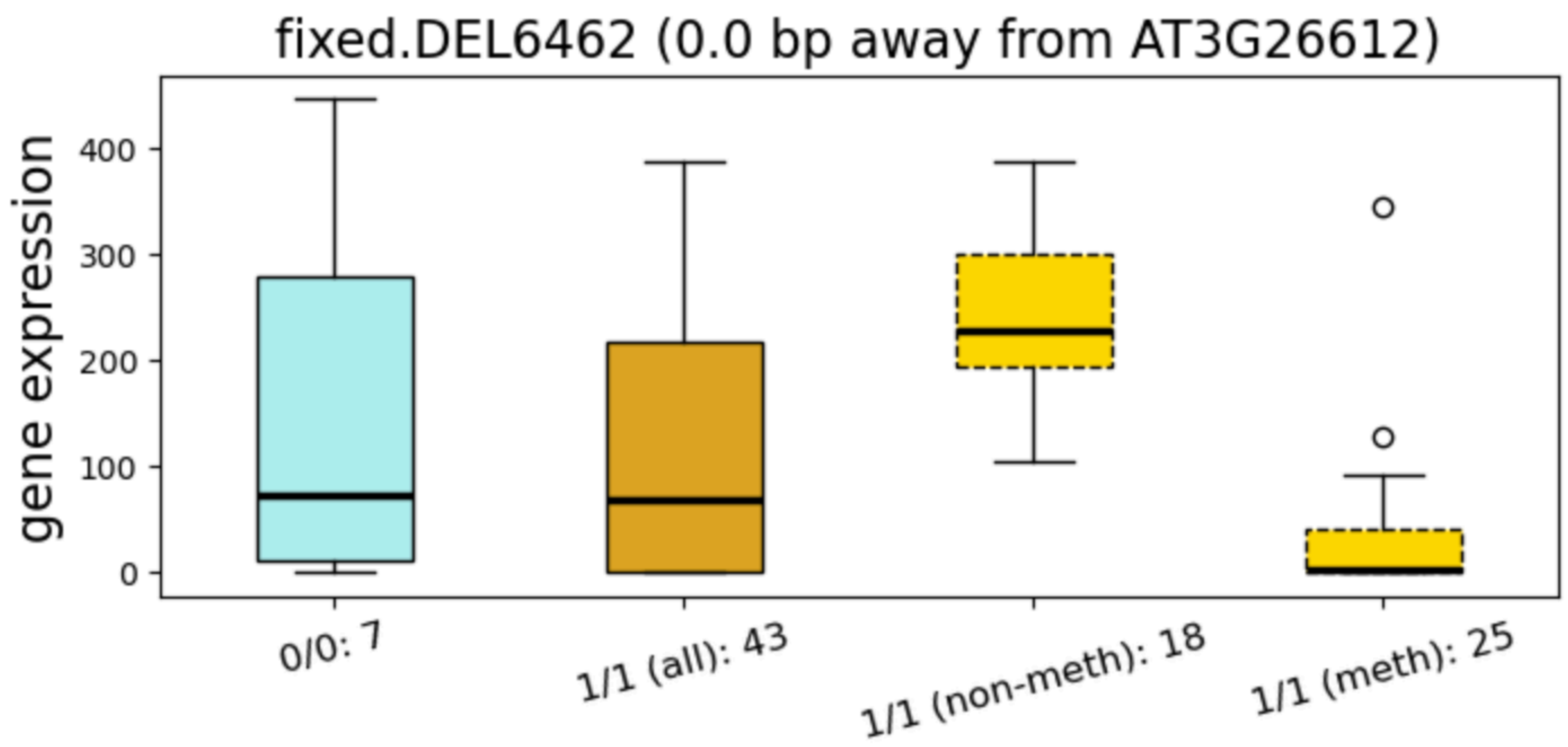
<u>Findings</u>:
All (cis + trans) associations: 1.054 mTIPs for 1.091 genes (corrected by N_{tips}*N_{genes})
Cis-associations (<1.500 bp distance): 457 mTIPs for 633 genes [most are not found with SNPs]

From epi-genotype to phenotype

Examples of cis- effects:

Inside gene

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

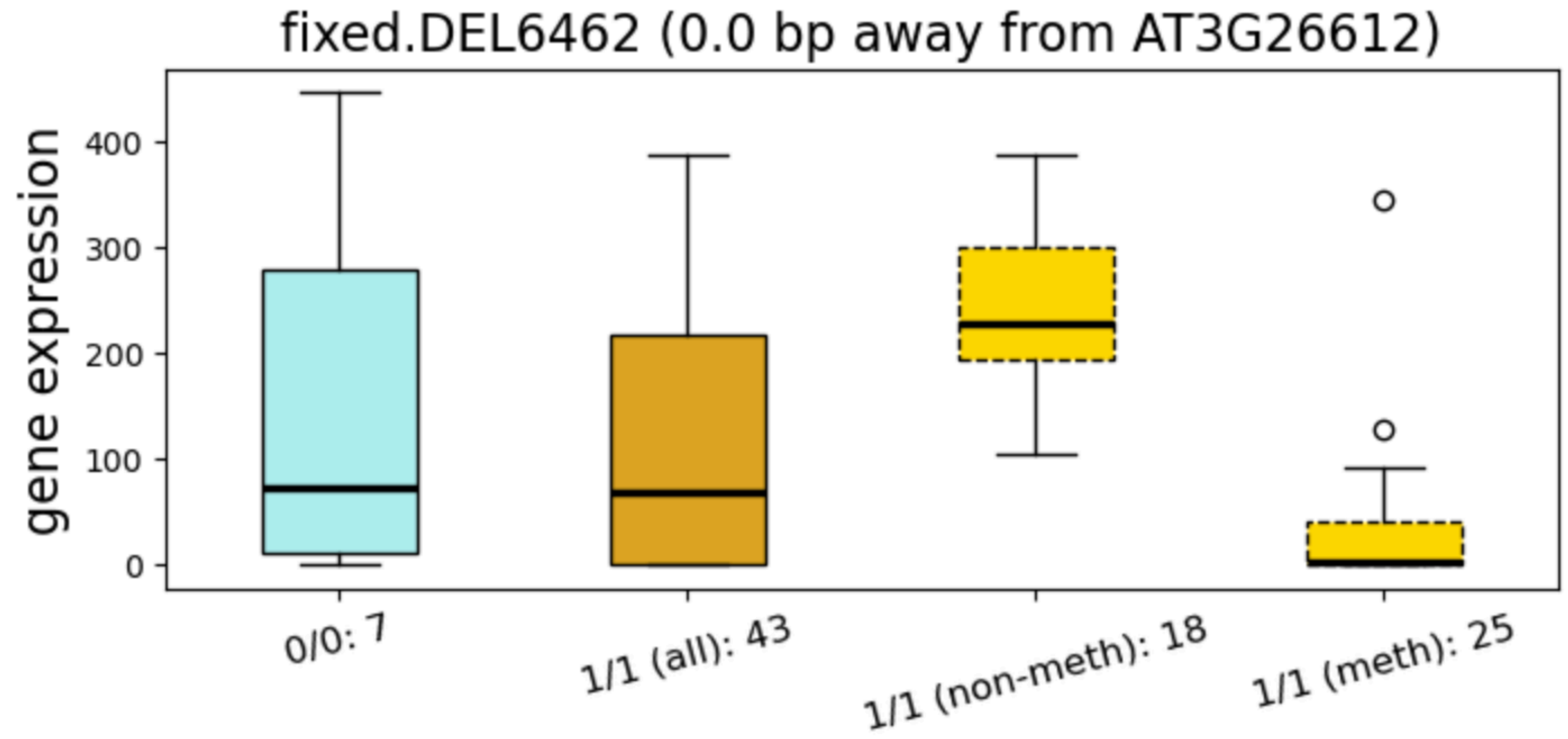


From epi-genotype to phenotype

Examples of cis- effects:

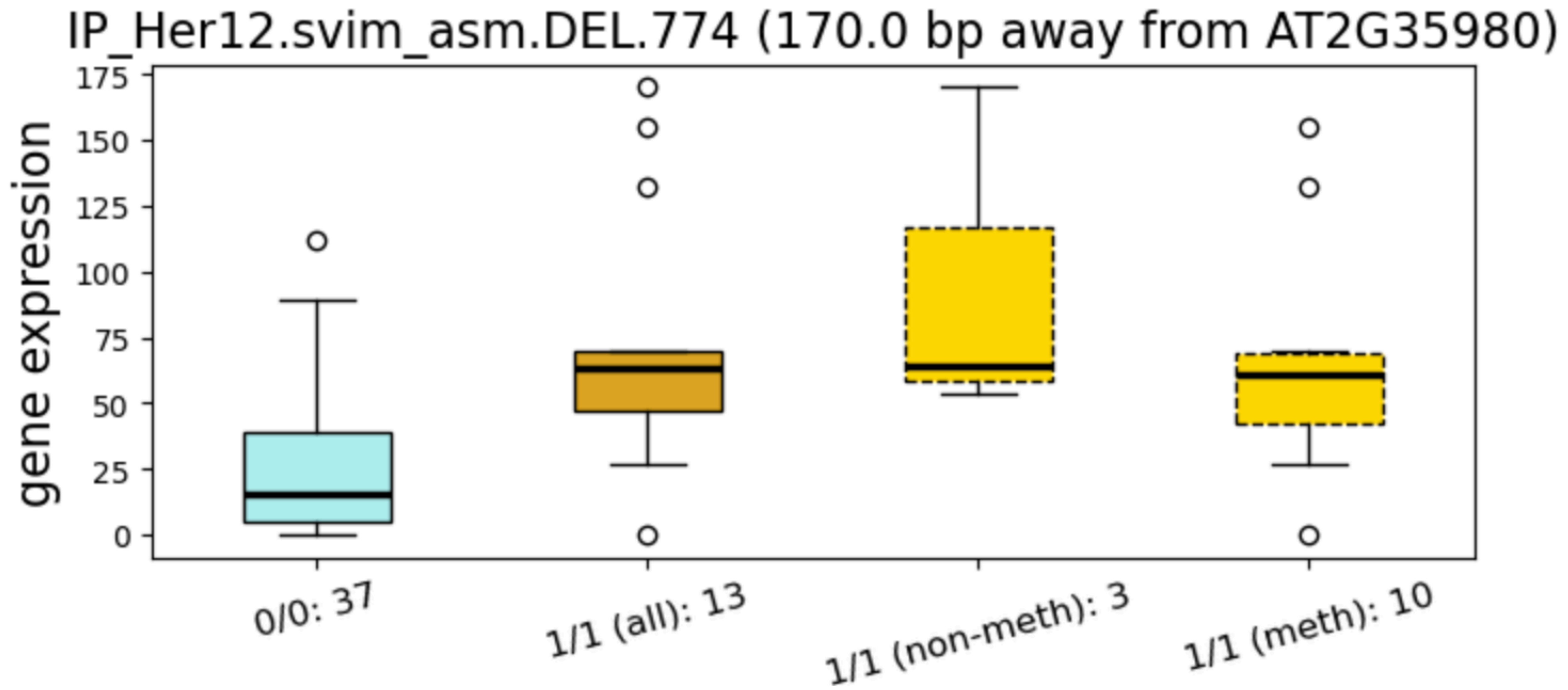
Inside gene

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



Confirmed spreader

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



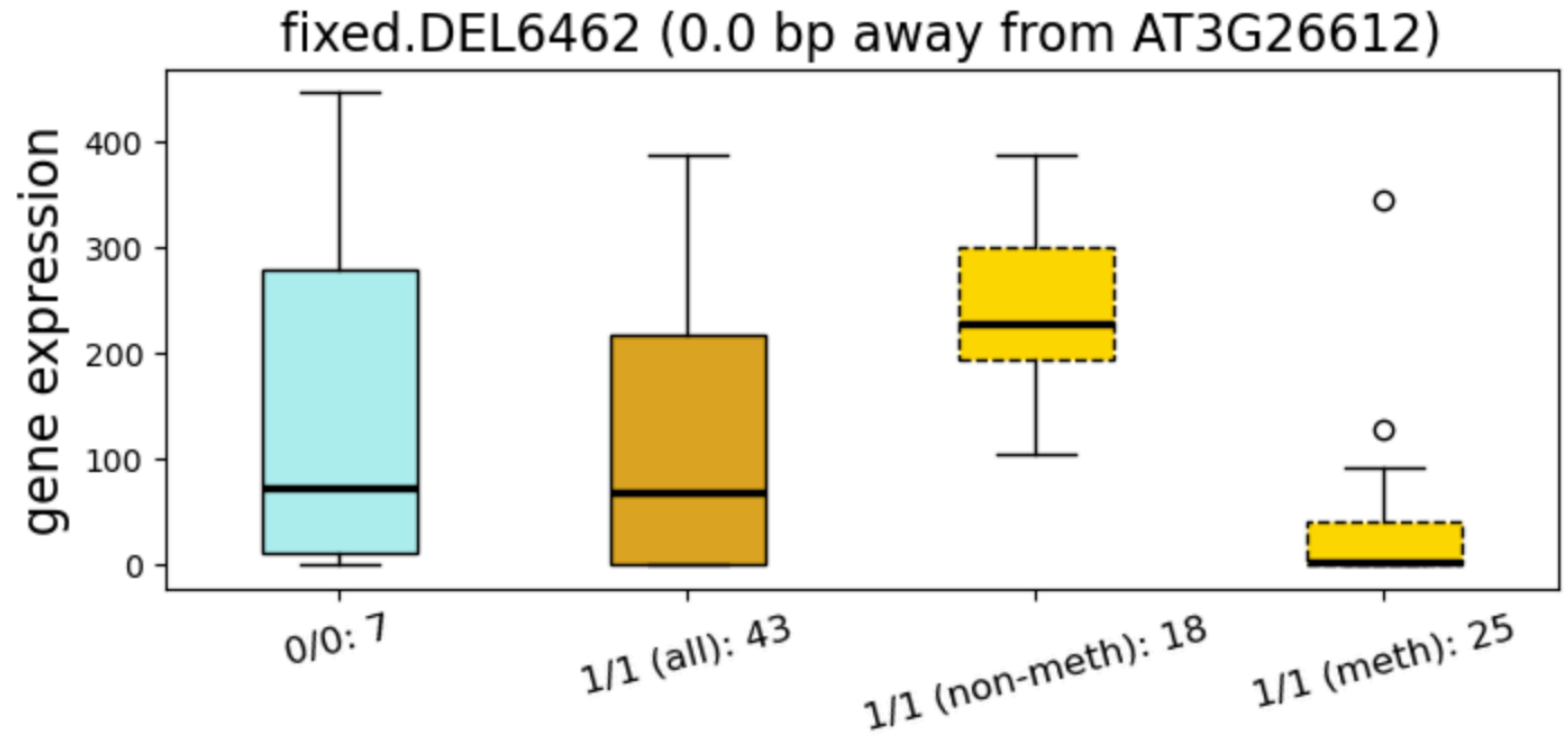
*Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family

From epi-genotype to phenotype

Examples of cis- effects:

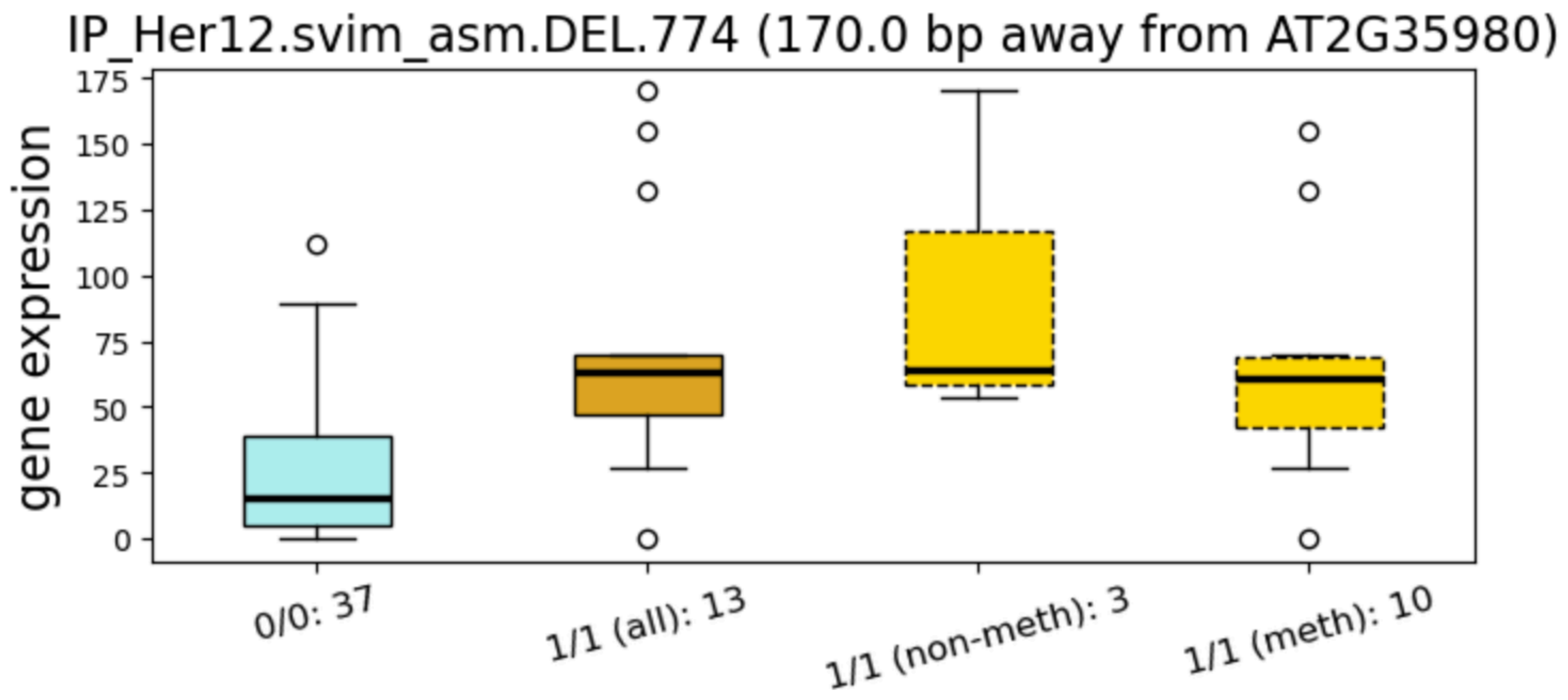
Inside gene

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



Confirmed spreader

	P_tip	P_meth	TIP	Chr	start	end	Distan[illegible]rom gene
535	0.000068	0.002327	IP_Her12.svim_asm.DEL.774	Chr2	15110051	15110322.0	170.0



*Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family

For trans- effects:

Future work: extending and fine-tuning GWAS signals with gene networks

Conclusions

Conclusions

- **Pipeline for precise TIP annotation** (genotyping + positions)
- **Unique dataset:** fully annotated for TIPs, genes, and methylation
- Potential evidence of (a) **secondary demethylation**, and (b) **remaining spreading** in old decayed TEs
- Significant part of methylation may be **explained from a TE sequence**
- But, there are **exceptions in both directions**
- An example of the workflow:

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

- **Genome-wide association studies may be improved** by including TIP and methylation data

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)

institut Curie

Inserm
La science pour la santé
From science to health

MINES PARIS | PSL | CBIO

anr

IBENS Institut de Biologie de l'Ecole Normale Supérieure

ENS | PSL

[✉️] *ekaterina.antonenko@minesparis.psl.eu*

Thank you!