

From sequences to knowledge,

Improving and learning from sequence alignments

Jury:

Brona	Brejova	Reviewer
Macha	Nikolski	Reviewer
Élodie	Laine	Examinator
Olivier	Gascuel	Examinator
Jean-Philippe	Vert	Examinator
Paul	Medvedev	Invited Member
Rayan	Chikhi	PhD. Supervisor

Luc Blassel – PhD Defense - December 2nd 2022



PhD context

- 2 different projects:
 1. Exploring drug resistance using HIV **MSAs** and ML
 2. Improving long-read **mapping**
- Both linked by **sequence alignment**
- Ignore the chronological **order** for **thematic** coherence

Presentation Outline

1

Introduction

2

Improving
long-read mapping

3

An introduction
to HIV

4

Exploring resistance in
HIV with ML

5

Learning alignment &
Other perspectives

6

Conclusion

Presentation Outline

1

Introduction

2

Improving
long-read mapping

3

An introduction
to HIV

4

Exploring resistance in
HIV with ML

5

Learning alignment &
Other perspectives

6

Conclusion

Introduction

Biological sequences

- Sequences are just a **succession of characters**
- Sequences **encode life**
- **Foundation** of bioinformatics and **modern biology**

ATGGTGCACCTG...

DNA



AUGGUGCACCUG...

RNA



MVHLTPEEKSAV...

Protein

Introduction

Sequencing

- Sequencing \Leftrightarrow technique for **reading** the biological sequence
- A **read** \Leftrightarrow **approximate subsequence** of the **original** sequence
- Many **technological** advances since **Sanger** in **1977**
- Long Reads: PacBio **2011**, ONT **2014**
- Sequencers make **mistakes**:
 - **Substitutions** **ATG** \rightarrow **ACG**
 - **Indels** **ATG** \rightarrow **ATCG** **ATG** \rightarrow **AG**

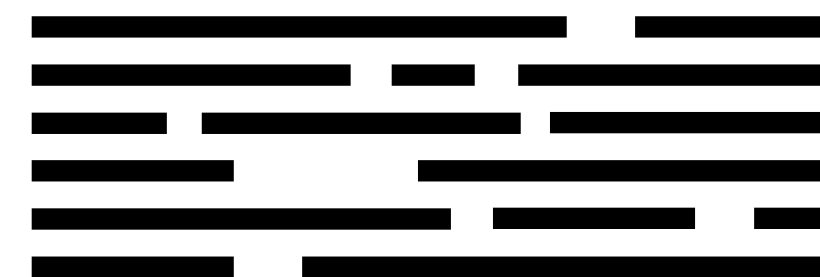
Introduction

Sequence alignment

- Alignment \Leftrightarrow **finding homologies** between sequences
- Alignment \Leftrightarrow **successive operations** to go **from one sequence to another**
- **Hard** problem \rightarrow Often rely on **heuristics**



pairwise alignment



multiple alignment

Presentation Outline

1

Introduction

2

Improving
long-read mapping

3

An introduction
to HIV

4

Exploring resistance in
HIV with ML

5

Learning alignment &
Other perspectives

6

Conclusion

What is read-mapping ?

- Special case of **sequence alignment**
- Finding where a **short subsequence** comes from in a (or several) **long sequence**
- Usually a sequencing **read** is **mapped** to a **reference** genome
- **Fundamental task** in many analyses pipelines

Why long reads ?

A rich information source

Published: 10 November 2014

Resolving the complexity of the human genome using single-molecule sequencing

[Mark J. P. Chaisson](#), [John Huddleston](#), [Megan Y. Dennis](#), [Peter H. Sudmant](#), [Maika Malig](#), [Fereydoun Hormozdiari](#), [Francesca Antonacci](#), [Urvashi Surti](#), [Richard Sandstrom](#), [Matthew Boitano](#), [Jane M. Landolin](#), [John A. Stamatoyannopoulos](#), [Michael W. Hunkapiller](#), [Jonas Korlach](#) & [Evan E. Eichler](#) ✉

[Nature](#) **517**, 608–611 (2015) | [Cite this article](#)

37k Accesses | 465 Citations | 257 Altmetric | [Metrics](#)

Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene *CACNA1C* in human brain

[Michael B. Clark](#), [Tomasz Wrzesinski](#), [Aintzane B. Garcia](#), [Nicola A. L. Hall](#), [Joel E. Kleinman](#), [Thomas Hyde](#), [Daniel R. Weinberger](#), [Paul J. Harrison](#), [Wilfried Haerty](#) ✉ & [Elizabeth M. Tunbridge](#) ✉

[Molecular Psychiatry](#) **25**, 37–47 (2020) | [Cite this article](#)

9230 Accesses | 59 Citations | 57 Altmetric | [Metrics](#)



Mapping and phasing of structural variation in patient genomes using nanopore sequencing

[Mircea Cretu Stancu](#), [Markus J. van Roosmalen](#), [Ivo Renkens](#), [Marleen M. Nieboer](#), [Sjors Middelkamp](#), [Joep de Ligt](#), [Giulia Pregno](#), [Daniela Giachino](#), [Giorgia Mandrile](#), [Jose Espejo Valle-Inclan](#), [Jerome Korzelius](#), [Ewart de Bruijn](#), [Edwin Cuppen](#), [Michael E. Talkowski](#), [Tobias Marschall](#), [Jeroen de Ridder](#) & [Wigard P. Kloosterman](#) ✉

[Nature Communications](#) **8**, Article number: 1326 (2017) | [Cite this article](#)

21k Accesses | 176 Citations | 125 Altmetric | [Metrics](#)

Nanopore sequencing and assembly of a human genome with ultra-long reads

[Miten Jain](#), [Sergey Koren](#), [Karen H Miga](#), [Josh Quick](#), [Arthur C Rand](#), [Thomas A Sasani](#), [John R Tyson](#), [Andrew D Beggs](#), [Alexander T Dilthey](#), [Ian T Fiddes](#), [Sunir Malla](#), [Hannah Marriott](#), [Tom Nieto](#), [Justin O'Grady](#), [Hugh E Olsen](#), [Brent S Pedersen](#), [Arang Rhie](#), [Hollian Richardson](#), [Aaron R Quinlan](#), [Terrance P Snutch](#), [Louise Tee](#), [Benedict Paten](#), [Adam M Phillippy](#), [Jared T Simpson](#), ... [Matthew Loose](#) ✉ + Show authors

[Nature Biotechnology](#) **36**, 338–345 (2018) | [Cite this article](#)

156k Accesses | 853 Citations | 1412 Altmetric | [Metrics](#)

Structural variant calling: the long and the short of it

[Medhat Mahmoud](#), [Nastassia Gobet](#), [Diana Ivette Cruz-Dávalos](#), [Ninon Mounier](#), [Christophe Dessimoz](#) ✉ & [Fritz J. Sedlazeck](#) ✉

[Genome Biology](#) **20**, Article number: 246 (2019) | [Cite this article](#)

50k Accesses | 156 Citations | 99 Altmetric | [Metrics](#)

Why long reads ?

A high error rate

	Illumina	PacBio	ONT
Length	100 - 200	10,000 - 60,000	12,000 - 2.5 10⁶
Accuracy	99.9 %	85 - 92%	87 - 98%

- Errors **complicate** downstream **mapping** (*Gusfield, 1997*)
- Long reads plagued by **errors** (*Dohm et al., 2020*):
 - Short **indels**
 - Particularly in **homopolymers**

AAA → AAAAAAAAAA

What is homopolymer compression ?

- HPC transforms sequences

HPC(**AAATTTGGGCCCAA**) → **ATGCA**

- **Empirically improves** analyses, **no guarantee** it's the **best**
- Can we **find functions** that **improve** long read **mapping** more than **HPC** ?

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$

- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$



- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$



- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$



- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$



- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$



- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$



- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

A formal definition of HPC

Deriving Mapping-friendly Sequence Reductions (MSR)

- Let us define $\Sigma = \{A, C, G, T\}$ and ε the empty character
- $\forall (x_1, x_2) \in \Sigma^2$

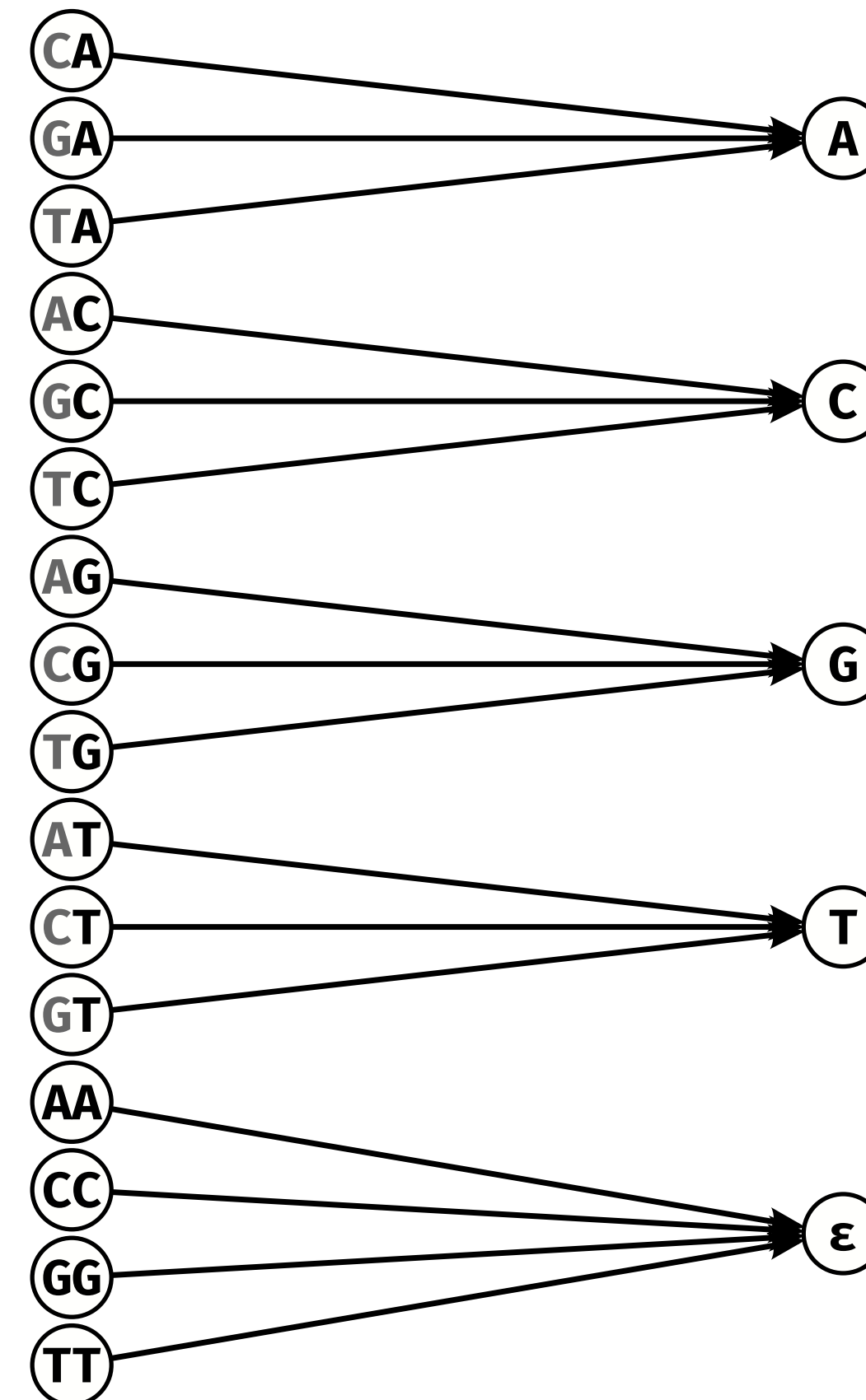
$$g^{HPC}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$



- $HPC(x) \rightarrow$ applying g^{HPC} on a **sliding window** of size 2 **along x** and **concatenating** outputs.
- Different $g = \mathbf{MSR}$

MSRs as directed graphs

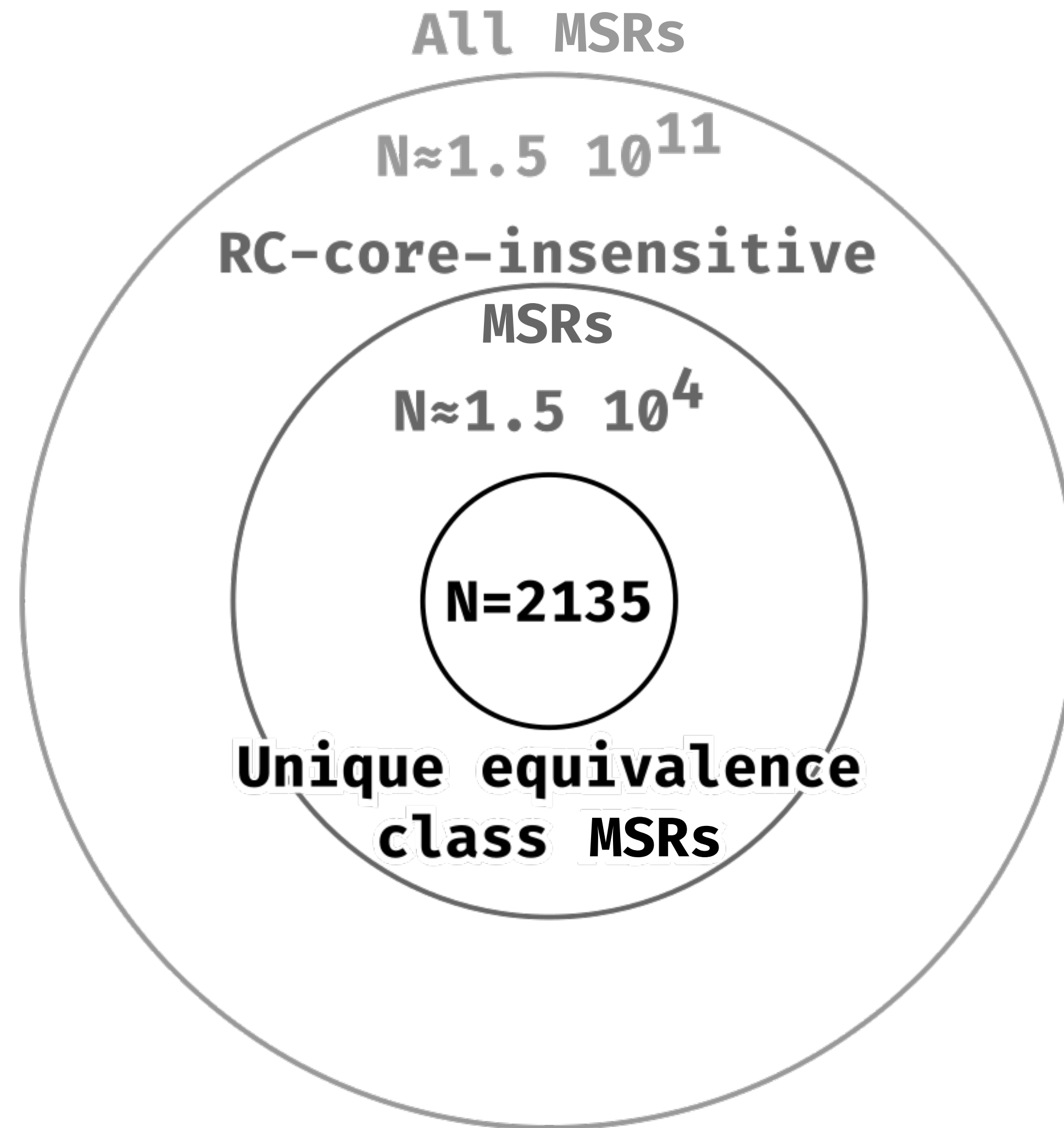
- Each g **function** can be visualised as a **directed graph** defined by a mapping between $|\Sigma^\ell|$ inputs and $|\Sigma| + 1$ outputs
- HPC as a directed graph ($n=16$ inputs $k=5$ outputs)
- There are 5^{16} functions
 $g : \Sigma^2 \rightarrow \Sigma \cup \{\varepsilon\}$
- **Cannot** all be tested



Reducing the search space

- **2 space-reducing strategies:**
 1. **MSRs must commute** with the **reverse complement** operation
 2. We define **equivalence classes**, based on **RC symmetries**

Reducing the search space



Evaluating MSRs

Datasets

- Simulate ONT reads, with **nanosim**, on 4 references:
 - **Whole human genome**, CHM13hTERT human cell line by the T2T
 - **Whole *Drosophila melanogaster* genome**, *Adams et al. (2022)*
 - **Whole *Escherichia coli* genome**, *Blattner et al. (1977)*
 - **Synthetic human centromeric sequence**, *Mikheenko et al. (2020)*
- tandemtools mapper test data

Can MSRs **improve mapping** of simulated reads?

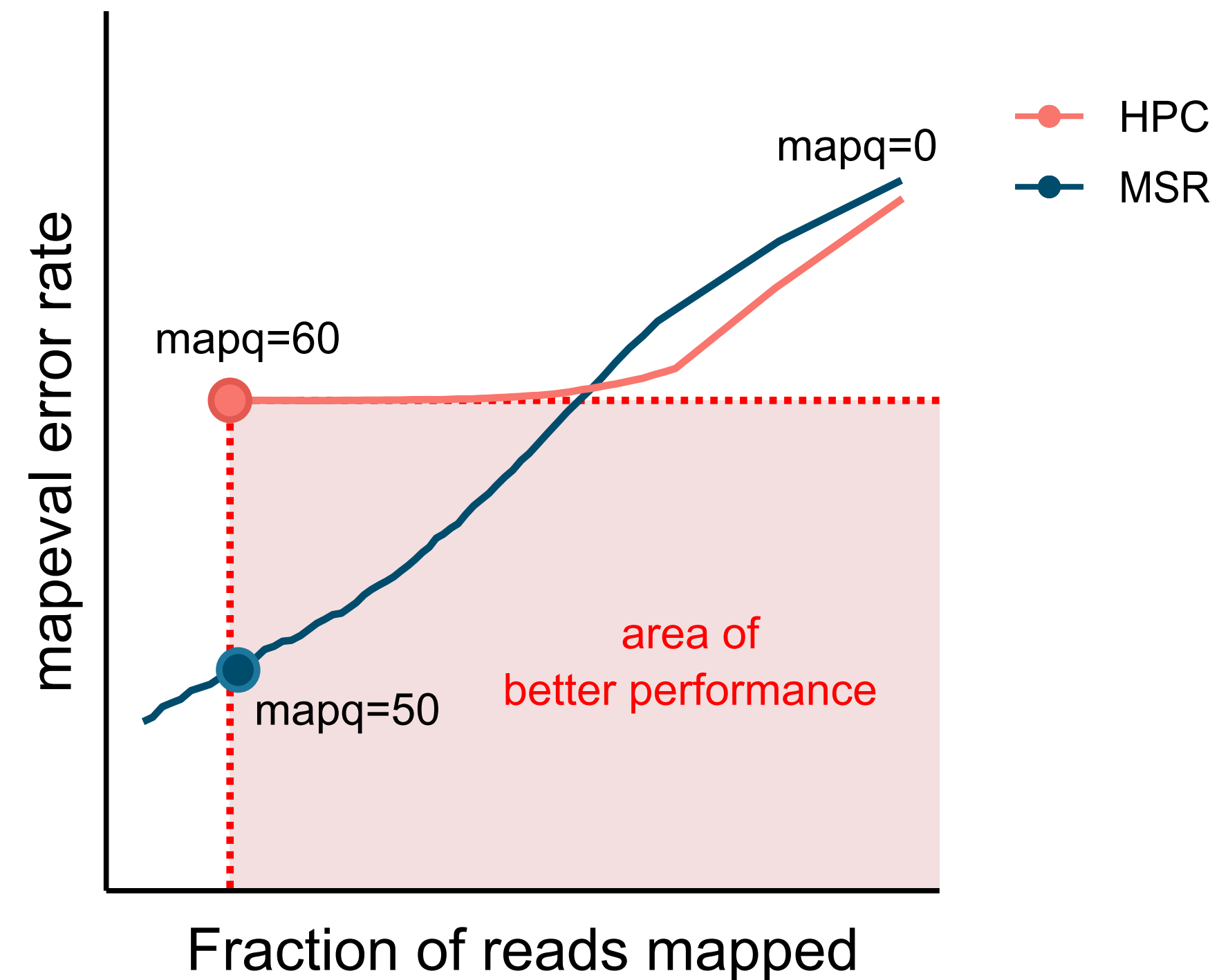
Evaluating MSRs

Evaluation Pipeline

- For **each (MSR, reference) pair** (and no MSR i.e. *raw*):
 1. **Transform** the reference and reads with the MSR
 2. **Map** transformed reads to transformed reference with `minimap2`
 3. **Evaluate** mapping with `paftools mapeval`

Evaluating MSRs

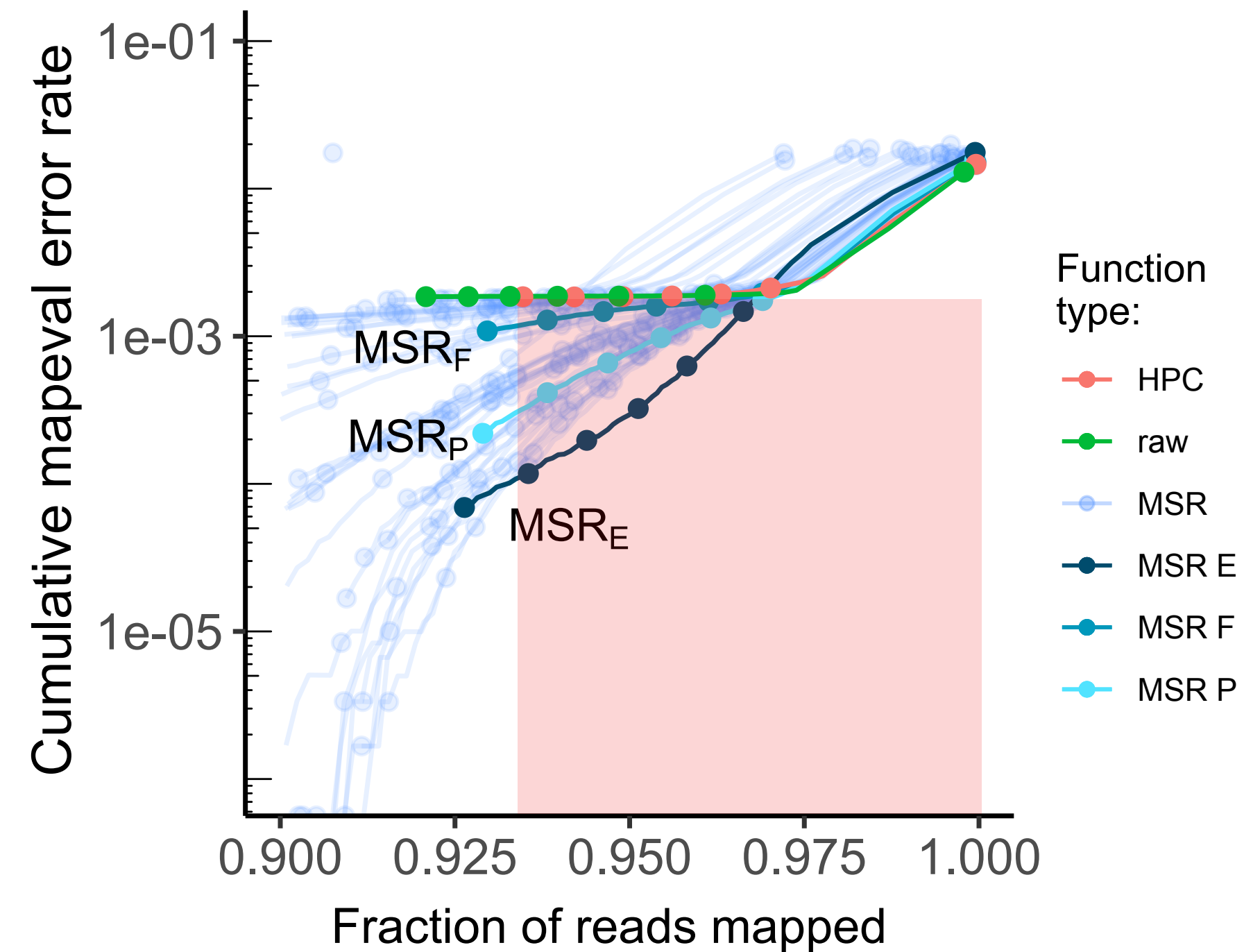
Comparing to HPC



We compare **MSRs** to **HPC** at **mapq 60**

Results

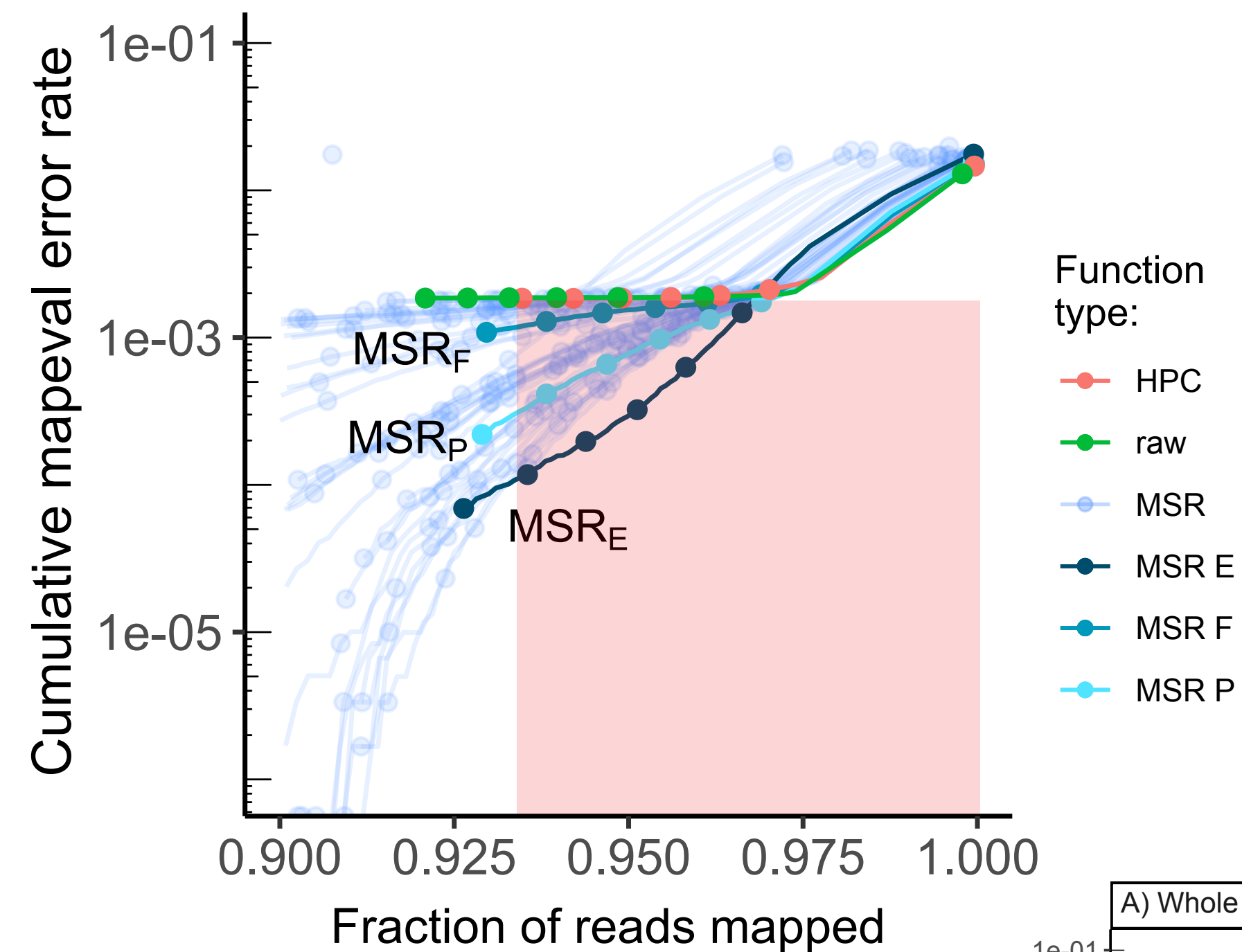
Across whole human genome



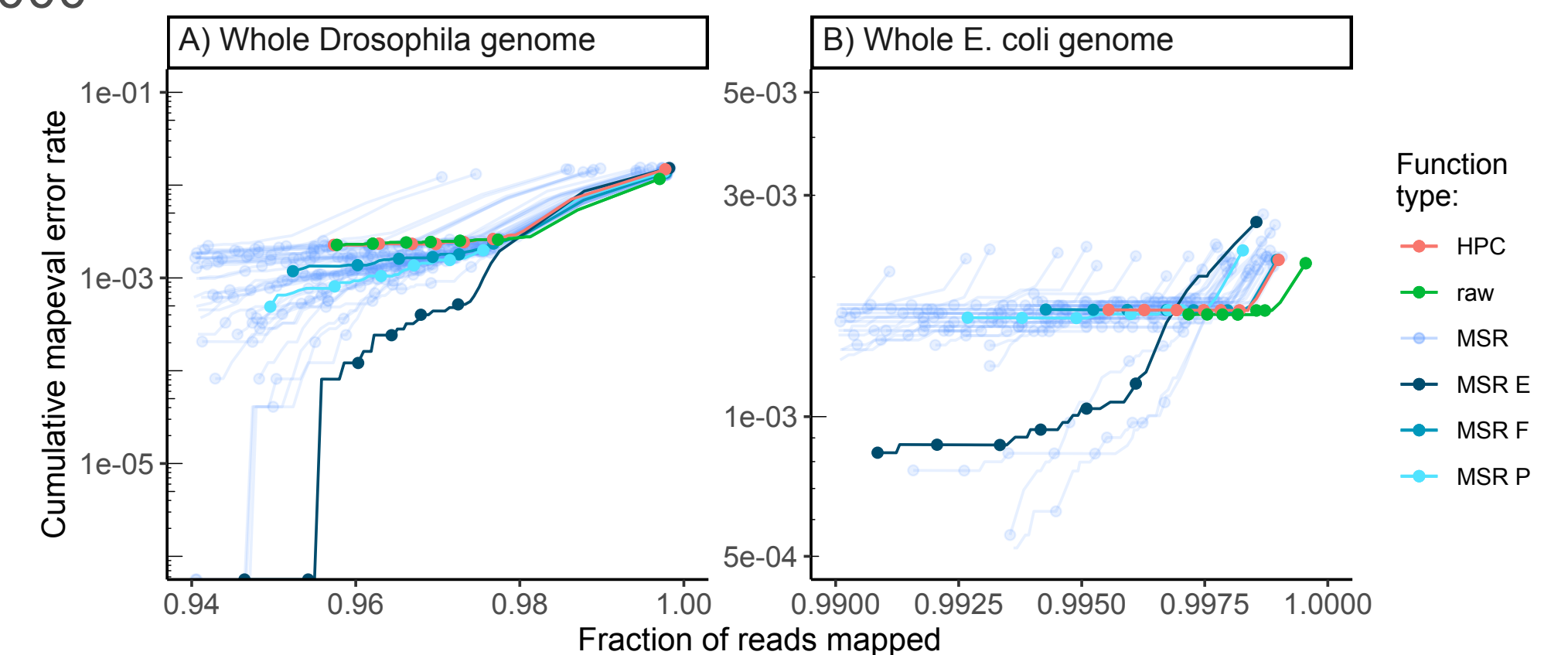
Many MSRs are **better** than HPC60

Results

Across whole human genome

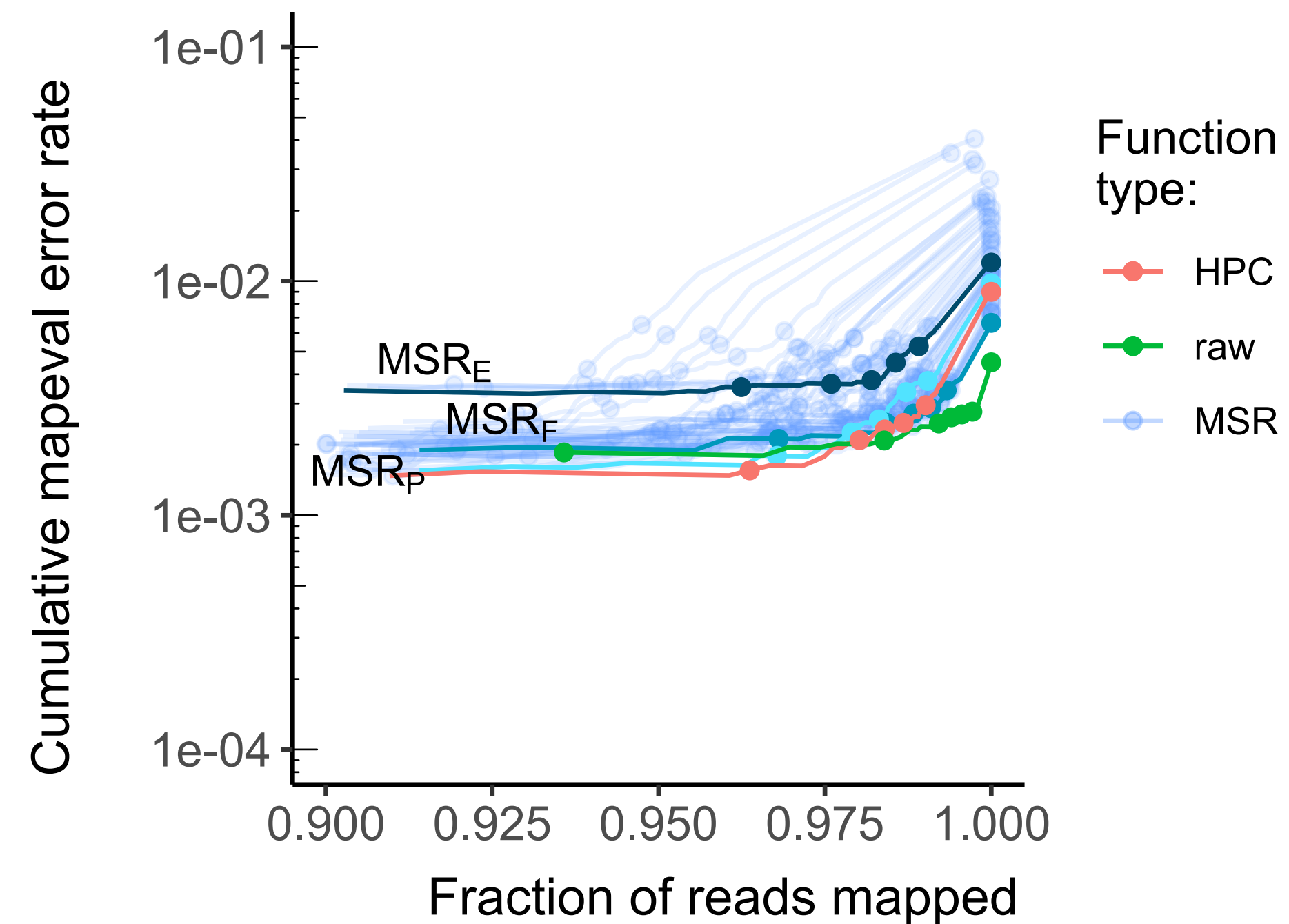


Many MSRs are **better** than HPC60



Results

Centromeric sequence



Mapping to **centromeres** is hard,
best **not to apply any function**

Take home message

- Some **MSRs** are **better** than **HPC**
- In some cases, the **mapping error rate** goes from 10^{-3} to 10^{-6}
- **MSRs** are **easy to implement** in existing aligners,
i.e. **cheap performance gains**

iScience



Volume 25, Issue 11, 18 November 2022, 105305

Article

Mapping-friendly sequence reductions: Going beyond homopolymer compression

Luc Blassel ^{1, 2} , Paul Medvedev ^{3, 4, 5}, Rayan Chikhi ^{1, 6} 


Show more 

+ Add to Mendeley  Cite

<https://doi.org/10.1016/j.isci.2022.105305>

Under a Creative Commons [license](#)

[Get rights and content](#)

 Open access

Perspectives

- MSRs work on simulated data → How do we evaluate on **real datasets** ?
(*fraction of mapped reads, mismatch rate, ...*)
- Explore **higher-order MSRs** ($N(3) \approx 3 \cdot 10^{21}$ and $N(4) \approx 10^{85}$):
 - **Reduce** the search space
 - **Explore** search space **better**:
 - Define objective function and **optimise**
 - **“Learn”** MSRs

Presentation Outline

1

Introduction

2

Improving
long-read mapping

3

An introduction
to HIV

4

Exploring resistance in
HIV with ML

5

Learning alignment &
Other perspectives

6

Conclusion

What is HIV ?

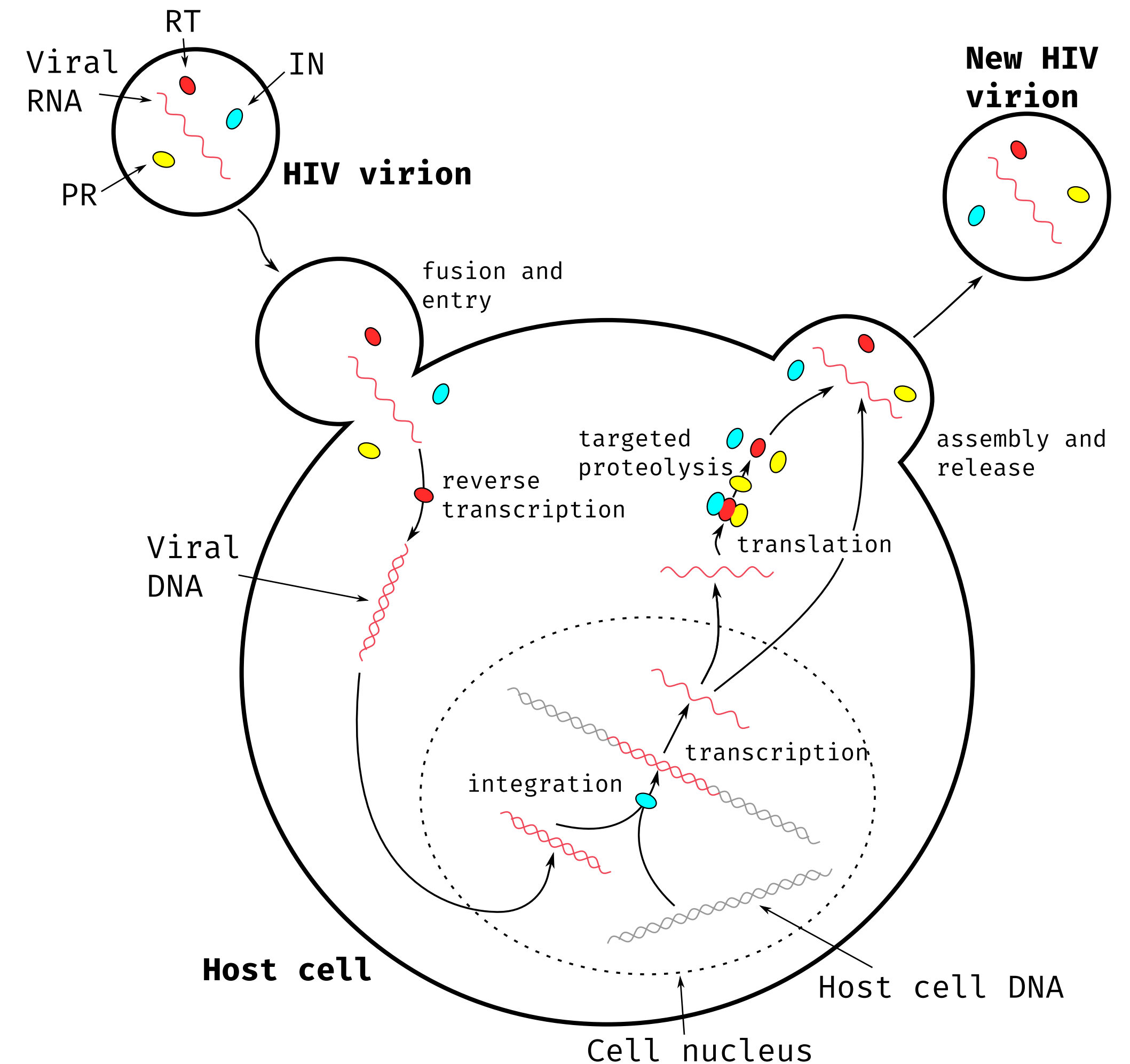
- Human **Immunodeficiency** Virus, discovered in **1983**
- Transmission: **sexual** contact, **blood**
- **40M** total deaths, **650k** in 2021
- **40M** living with HIV in 2021
- **Global** health **problem**



UNAIDS Global AIDS Update 2022 report cover

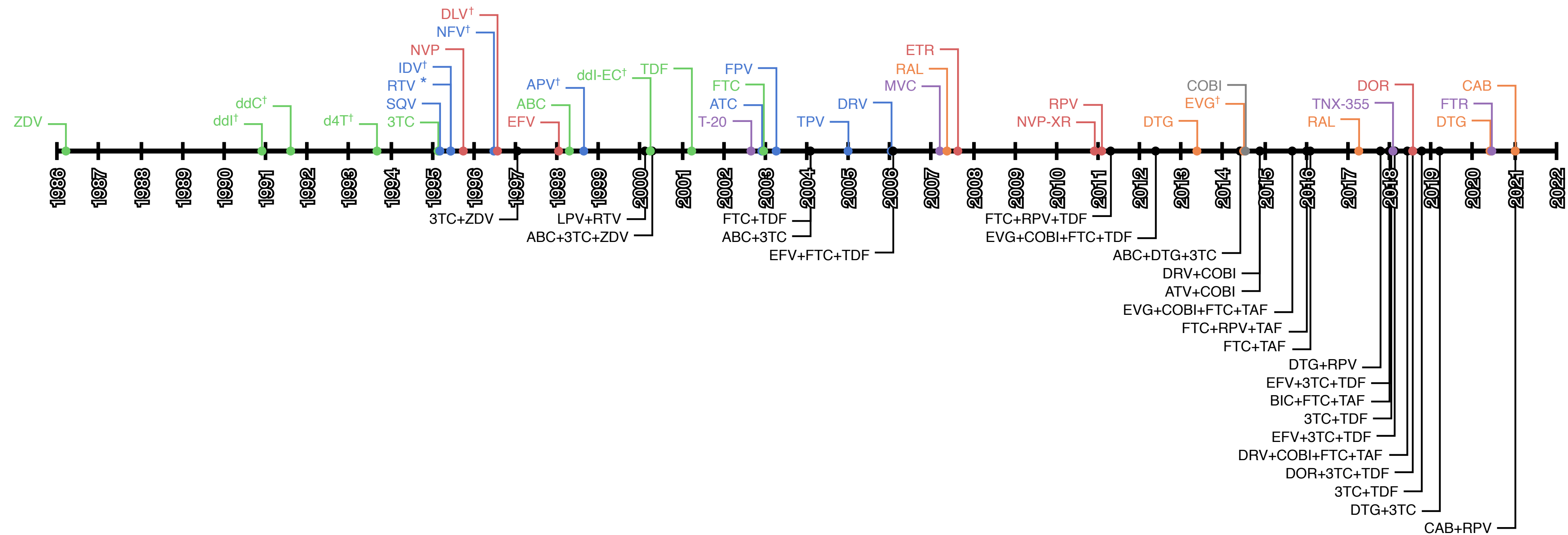
How does HIV work ?

- HIV is a **Retrovirus**
- Genetic **information** contained in **RNA**
- Key **proteins**:
 - Reverse Transcriptase **RT**
 - Integrase **IN**
 - Protease **PR**



How do we treat HIV ?

- Antiretroviral Therapy (**ART**)
- **Most** drugs target **RT, IN** or **PR** → **RTI, INI** or **PI**



What are DRMs ?

- **Resistance** arises in **response** to treatment **pressure**
- Drug resistance mutations (**DRMs**) have been found for **every drug**
- To **mitigate DRM** effects:
 - Treatment **switching**
 - **Combination** therapy

Presentation Outline

1

Introduction

2

Improving
long-read mapping

3

An introduction
to HIV

4

Exploring resistance in
HIV with ML

5

Learning alignment &
Other perspectives

6

Conclusion

Why study DRMs ?

- In **lower-income** countries, **access** to treatment is **not easy**
- In **higher-income** countries, **transmission to and within** treatment-**naïve** populations
- DRMs **limit** treatment **options** at **population level**

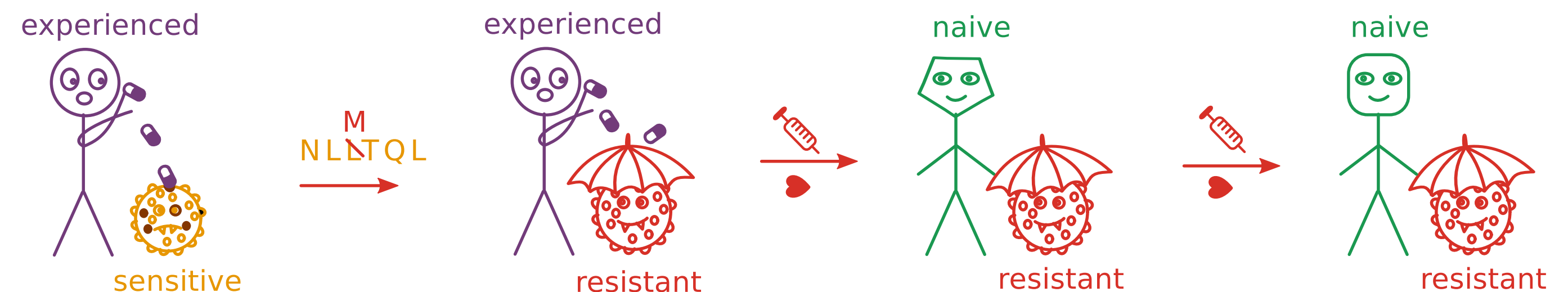
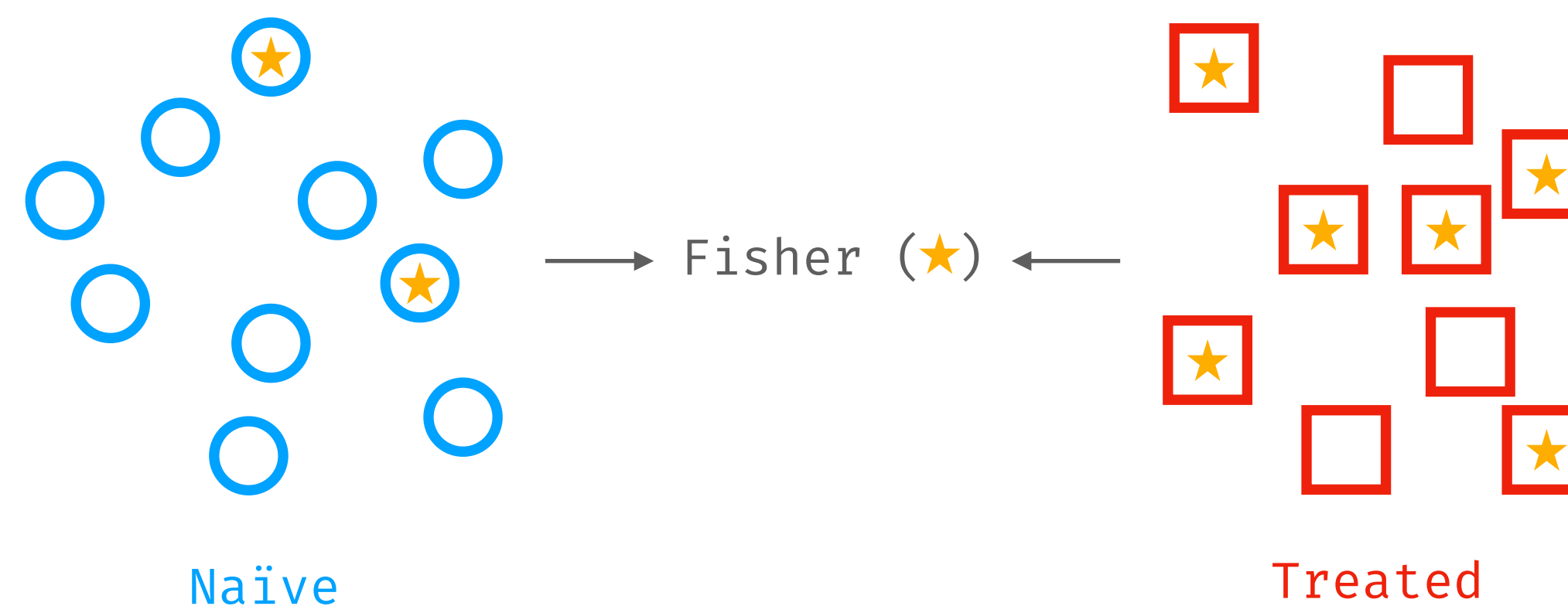


Fig A. Zhukova

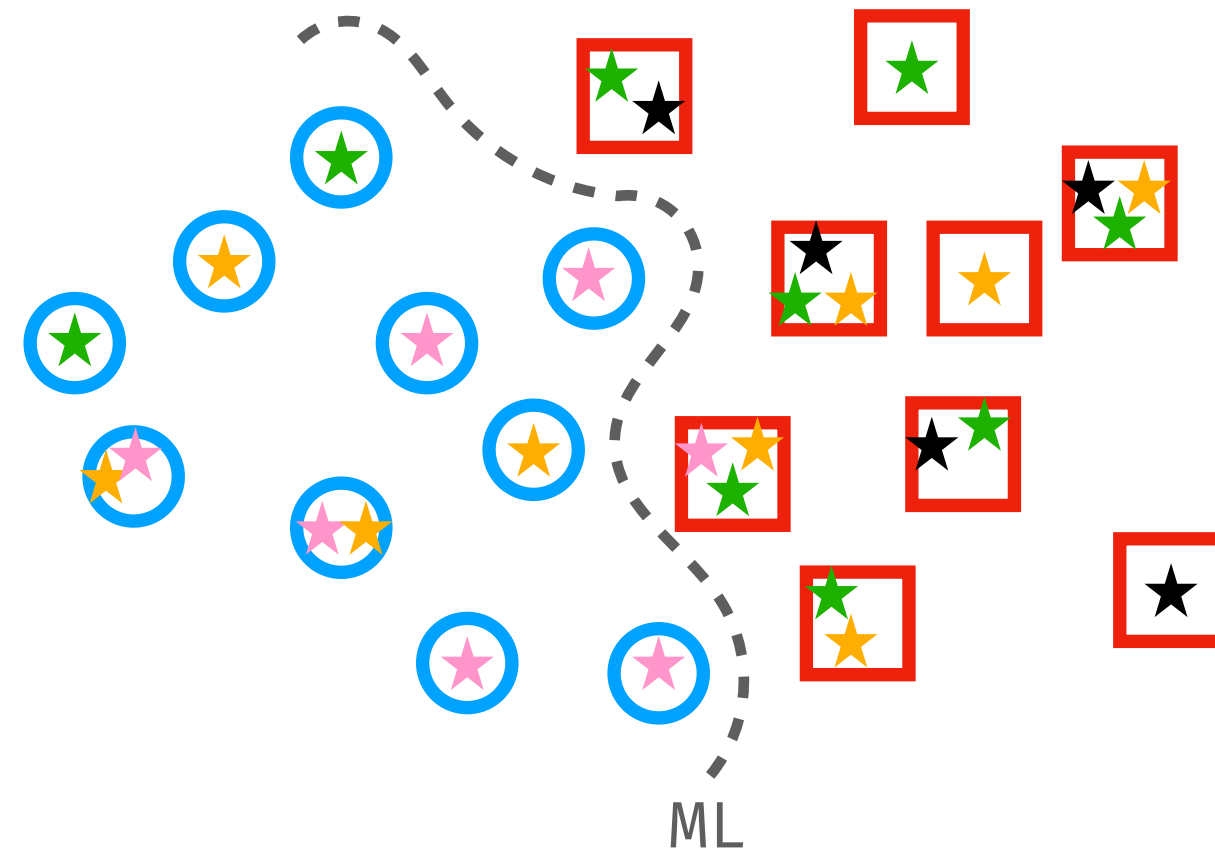
How do we find DRMs ?

- **Test statistical association** to treatment for **each mutation**
- **Multiple testing** correction → usually **decrease** in statistical **power**
- **Epistasis** and **groups** of mutations **worsen problem**



Using machine learning to find DRMs

- **Encode** each **mutation** as **binary feature**
- Train model to **discriminate experienced** from **naive** sequences
- **Important** model **features** might be **DRMs**
- **Treatment** status is a **proxy**



What models do we use ?

- **Random Forest**, can capture complex interaction between features
- **LASSO Logistic Regression**, class-specific weights & feature selection
- **Naive Bayes**, simple and statistical interpretation
- All classifiers are **easy to train** and **easy to interpret**

What do we learn from ?

- **UK Drug Resistance Database:**
 - **55,000 RT sequences**
 - Subtypes **B 68%** & **C 32%**
 - Naive **75%** & Experienced **25%**
- **African dataset:**
 - **4,000 RT sequences**
 - **24 subtypes**
 - Naive **58%** & Experienced **42%**

Confounding factors

- **Unbalanced** classes in **training** data
 - Use **adapted** performance **metrics**
- **Sequences** are evolutionarily **related** (*i.e. not independent*)
 - **Separate subtypes** during **training** & **testing**
- **Known DRMs** have very **strong signal**
 - **Remove** known DRM **signal**

Removing known signal

DRM features

	181V	181K	182D	182F	184V	184E	187K
Seq 1	1	0	0	0	0	0	0
Seq 2	0	0	1	0	0	0	0
Seq 3	0	0	0	0	1	0	0
Seq 4	0	1	0	0	0	1	1
Seq 5	0	0	0	1	0	0	0
	Known				Known		

Removing known signal

DRM features

	181V	181K	182D	182F	184V	184E	187K
Seq 1	1	0	0	0	0	0	0
Seq 2	0	0	1	0	0	0	0
Seq 3	0	0	0	0	1	0	0
Seq 4	0	1	0	0	0	1	1
Seq 5	0	0	0	1	0	0	0
Known				Known			

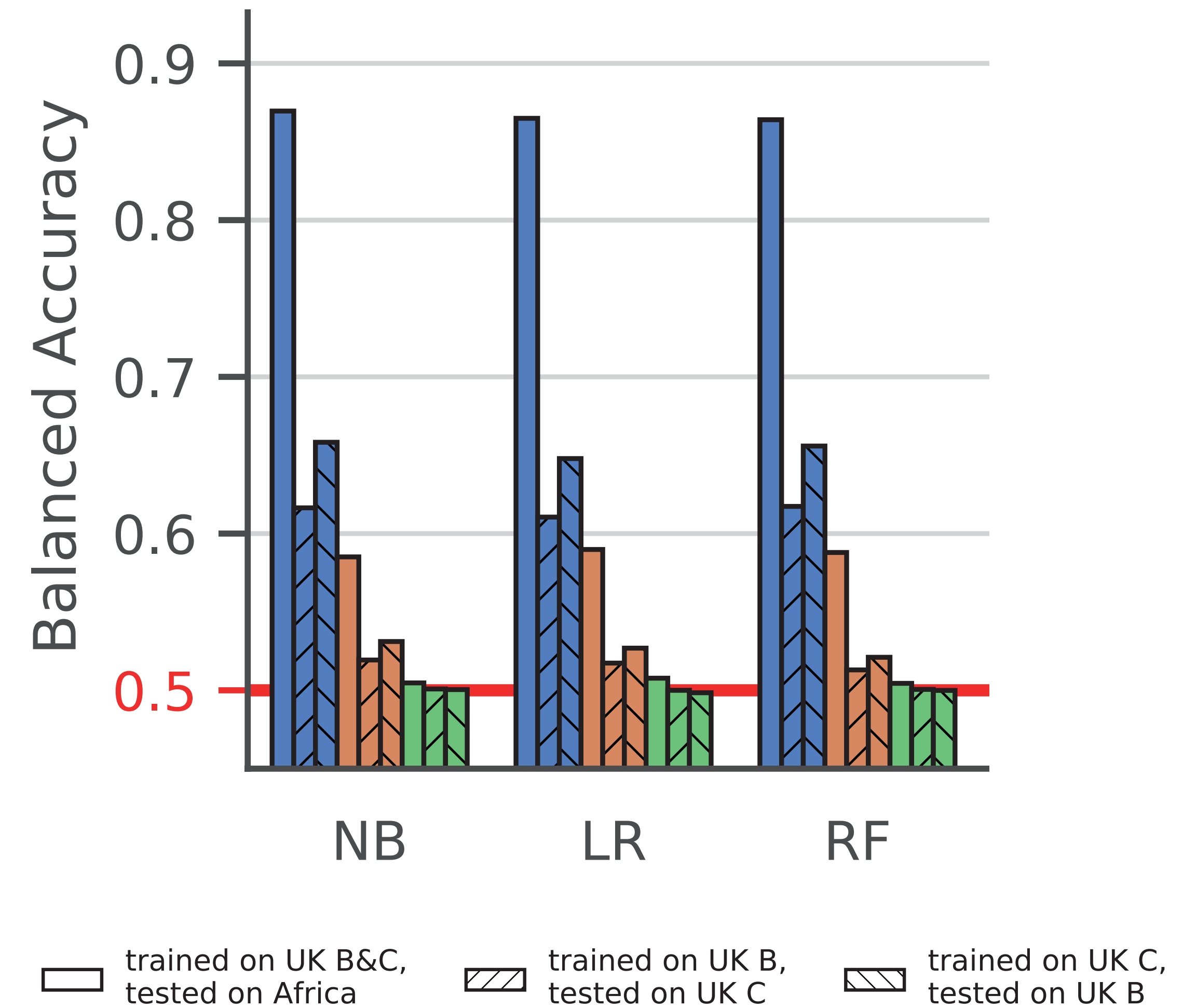
Removing known signal

DRM features & DRM sequences

	181V	181K	182D	182F	184V	184E	187K
Seq 1	1	0	0	0	0	0	0
Seq 2	0	0	1	0	0	0	0
Seq 3	0	0	0	0	1	0	0
Seq 4	0	1	0	0	0	1	1
Seq 5	0	0	0	1	0	0	0
Known				Known			

Results

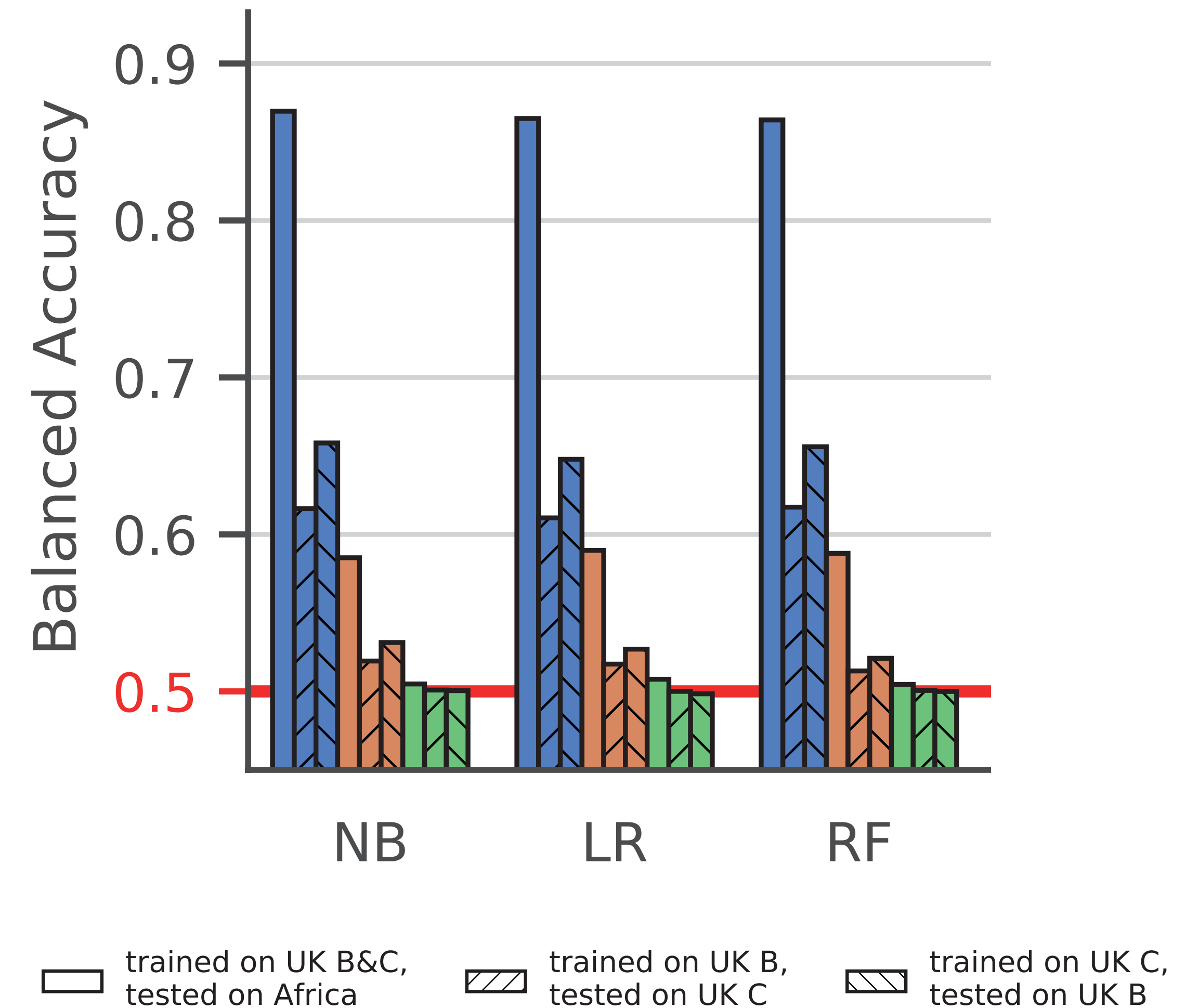
Classifier performance



Results

Classifier performance

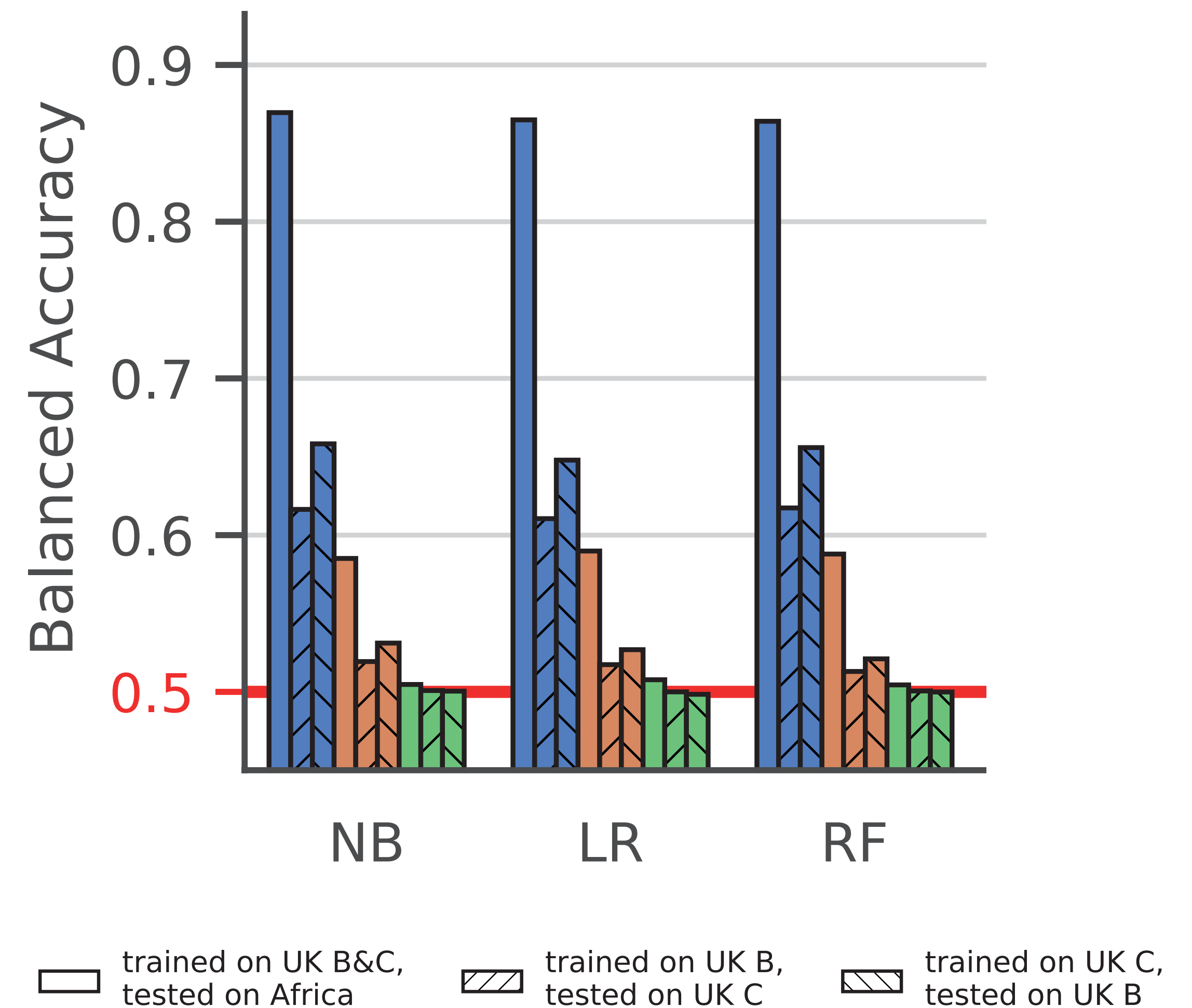
- **High accuracy** with all signal



Results

Classifier performance

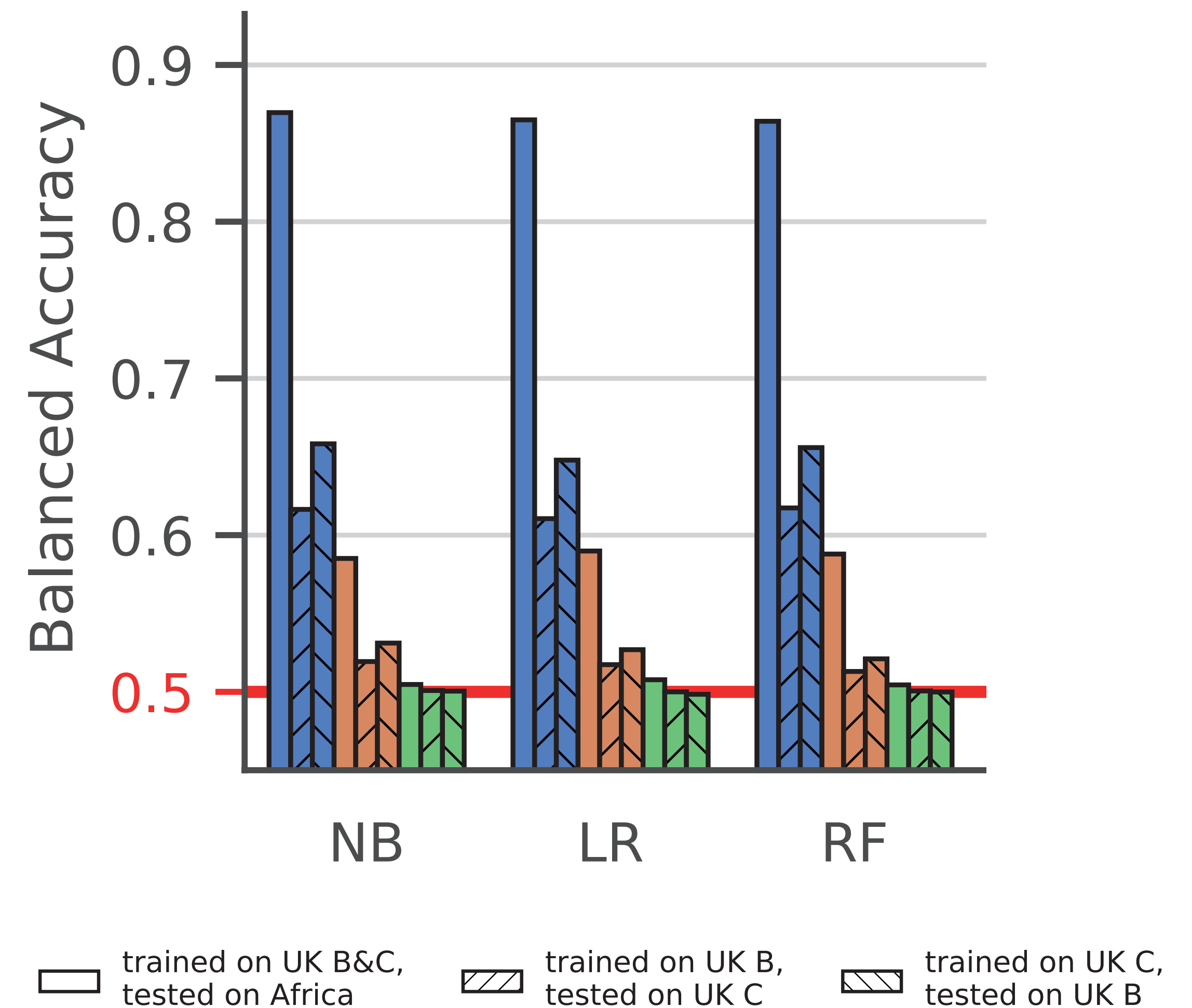
- **High accuracy** with all signal
- **Significantly better** than random when **removing DRM** features



Results

Classifier performance

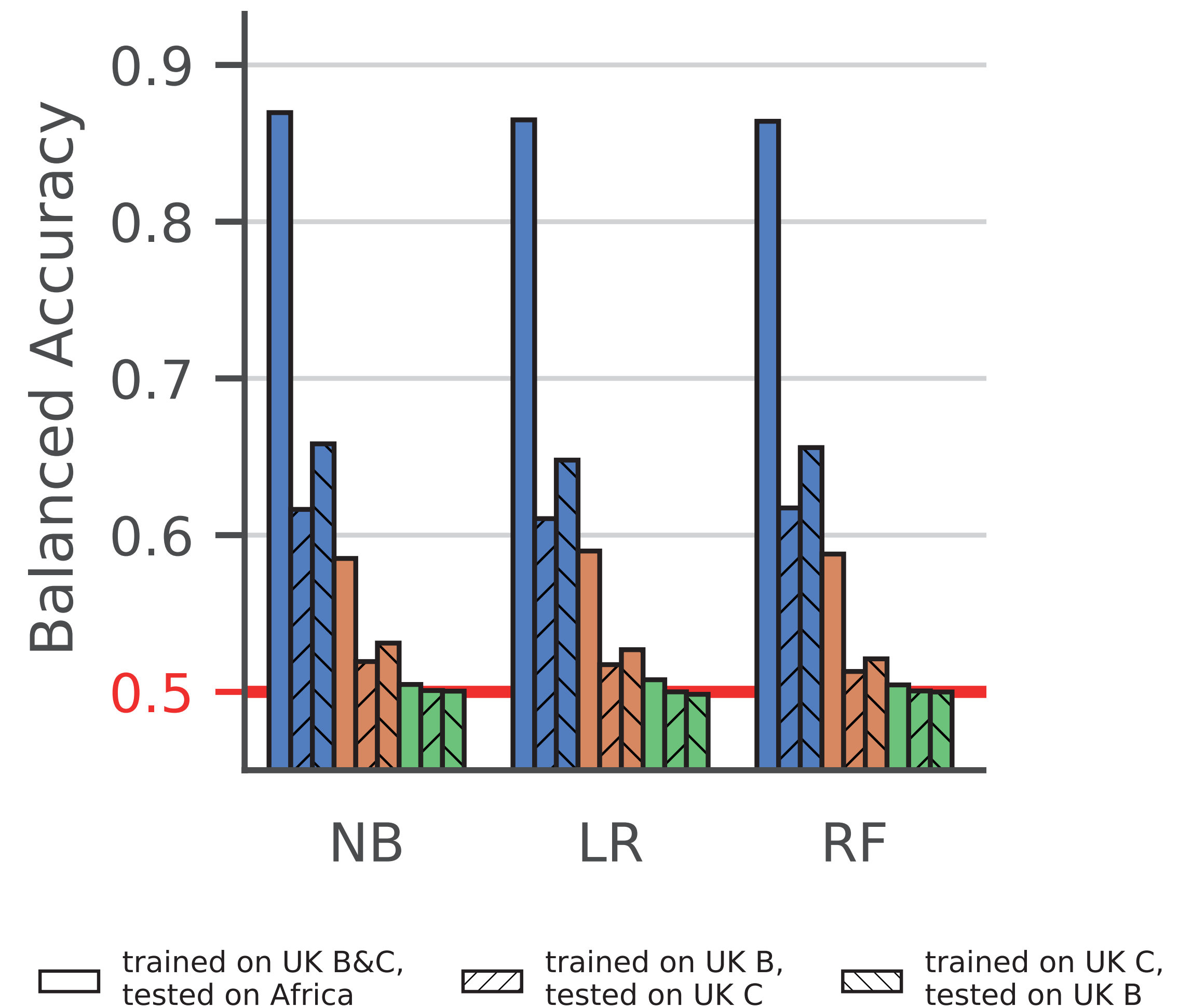
- **High accuracy** with all signal
- **Significantly better** than random when **removing DRM** features
- **No signal left** when also **removing DRM sequences**



Results

Classifier performance

- **High accuracy** with all signal
- **Significantly better** than random when **removing DRM** features
- **No signal left** when also **removing DRM sequences**
- **Probably** means that **all** primary **DRMs** are **known**



Results

Finding new DRMs

- We studied the **most important features**:
 - Across different **training settings**
 - Across different **classifiers**
- We identified **6 potential** DRMs:

L228R E203K D218E L228H I135L H208Y

- These **potential** DRMs are most likely **accessory mutations**

Results

New ?

PLOS COMPUTATIONAL BIOLOGY

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation

Soo-Yon Rhee, Tommy F Liu, Susan P Holmes, Robert W Shafer 

Published: May 11, 2007 • <https://doi.org/10.1371/journal.pcbi.0030087>

JOURNAL OF MEDICAL VIROLOGY

Research Article |  Free Access

Impact of unreported HIV-1 reverse transcriptase mutations on phenotypic resistance to nucleoside and non-nucleoside inhibitors

A. Saracino, L. Monno, L. Scudeller, D.C. Cibelli, A. Tartaglia, G. Punzi, C. Torti, S. Lo Caputo, F. Mazzotta, G. Scotto, G. Carosi, G. Angarano 

First published: 18 November 2005 | <https://doi.org/10.1002/jmv.20500> | Citations: 25

American Society for Microbiology
Journal of Virology
Volume 74, Issue 22, 15 November 2000, Pages 10269–10273
<https://doi.org/10.1128/JVI.74.22.10269-10273.2000>


Vaccines and Antiviral Agents

Reduced Susceptibility of Human Immunodeficiency Virus Type 1 (HIV-1) from Patients with Primary HIV Infection to Nonnucleoside Reverse Transcriptase Inhibitors Is Associated with Variation at Novel Amino Acid Sites

Andrew J. Leigh Brown^{1,*}, Heather M. Precious¹, Jeannette M. Whitcomb², Joseph K. Wong³, Marlynne Quigg¹, Wei Huang², Eric S. Daar⁴, Richard T. D'Aquila⁵, Philip H. Keiser⁶, Elizabeth Connick⁷, Nicholas S. Hellmann², Christos J. Petropoulos², Douglas D. Richman³, and Susan J. Little³

Emergence of the H208Y mutation in the reverse transcriptase (RT) of HIV-1 in association with nucleoside RT inhibitor therapy

G. Nebbia, Caroline A. Sabin, D. T. Dunn,

Anna Maria Geretti  on behalf of the UK Collaborative Group on HIV Drug Resistance and the UK Collaborative HIV Cohort (CHIC) Study Group

Journal of Antimicrobial Chemotherapy, Volume 59, Issue 5, May 2007, Pages 1013–1016, <https://doi.org/10.1093/jac/dkm067>

Published: 13 March 2007 Article history ▾

Improved Interpretation of Genotypic Changes in the HIV-1 Reverse Transcriptase Coding Region That Determine the Virological Response to Didanosine

Andrea De Luca , Simona Di Giambenedetto, Maria Paola Trotta, Manuela Colafigli, Mattia Prosperi, Lidia Ruiz, John Baxter, Philippe Clevenbergh, Roberto Cauda, Carlo-Federico Perno ... Show more

The Journal of Infectious Diseases, Volume 196, Issue 11, 1 December 2007, Pages 1645–1653, <https://doi.org/10.1086/522231>

Published: 01 December 2007 Article history ▾

Antiviral Therapy 11:693–699

Impact of HIV-1 reverse transcriptase polymorphism at codons 211 and 228 on virological response to didanosine

Anne-Genevieve Marcelin^{1*}, Philippe Flandre², Andre Furco³, Marc Wirden¹, Jean-Michel Molina² and Vincent Calvez¹ on behalf of the AI454-176 Jaguar Study Team[†]

Reverse transcriptase mutations 118I, 208Y, and 215Y cause HIV-1 hypersusceptibility to non-nucleoside reverse transcriptase inhibitors

Clark, Shauna A^a; Shulman, Nancy S^b; Bosch, Ronald J^c; Mellors, John W^a

Author Information 

AIDS: April 24, 2006 - Volume 20 - Issue 7 - p 981-984
doi: 10.1097/01.aids.00002222069.14878.44

Take home message

- We **found 6 new** potential **DRMs**
- Most **likely accessory** mutations
- **All primary resistance** mutations are **probably known**

PLOS COMPUTATIONAL BIOLOGY

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Using machine learning and big data to explore the drug resistance landscape in HIV

Luc Blassel , Anna Tostevin, Christian Julian Villabona-Arenas, Martine Peeters, Stéphane Hué, Olivier Gascuel ,

On behalf of the UK HIV Drug Resistance Database 

Version 2



Published: August 26, 2021 • <https://doi.org/10.1371/journal.pcbi.1008873>

Perspectives

- **Experimental** confirmation of **DRMs**
- **Search** for complex **epistasis** with more **refined models**
 - **Deep Learning** → **black box**
 - Neural Network **interpretation** is an **active field**
- **Fine-grained** knowledge with **more metadata**

Presentation Outline

1

Introduction

2

Improving
long-read mapping

3

An introduction
to HIV

4

Exploring resistance in
HIV with ML

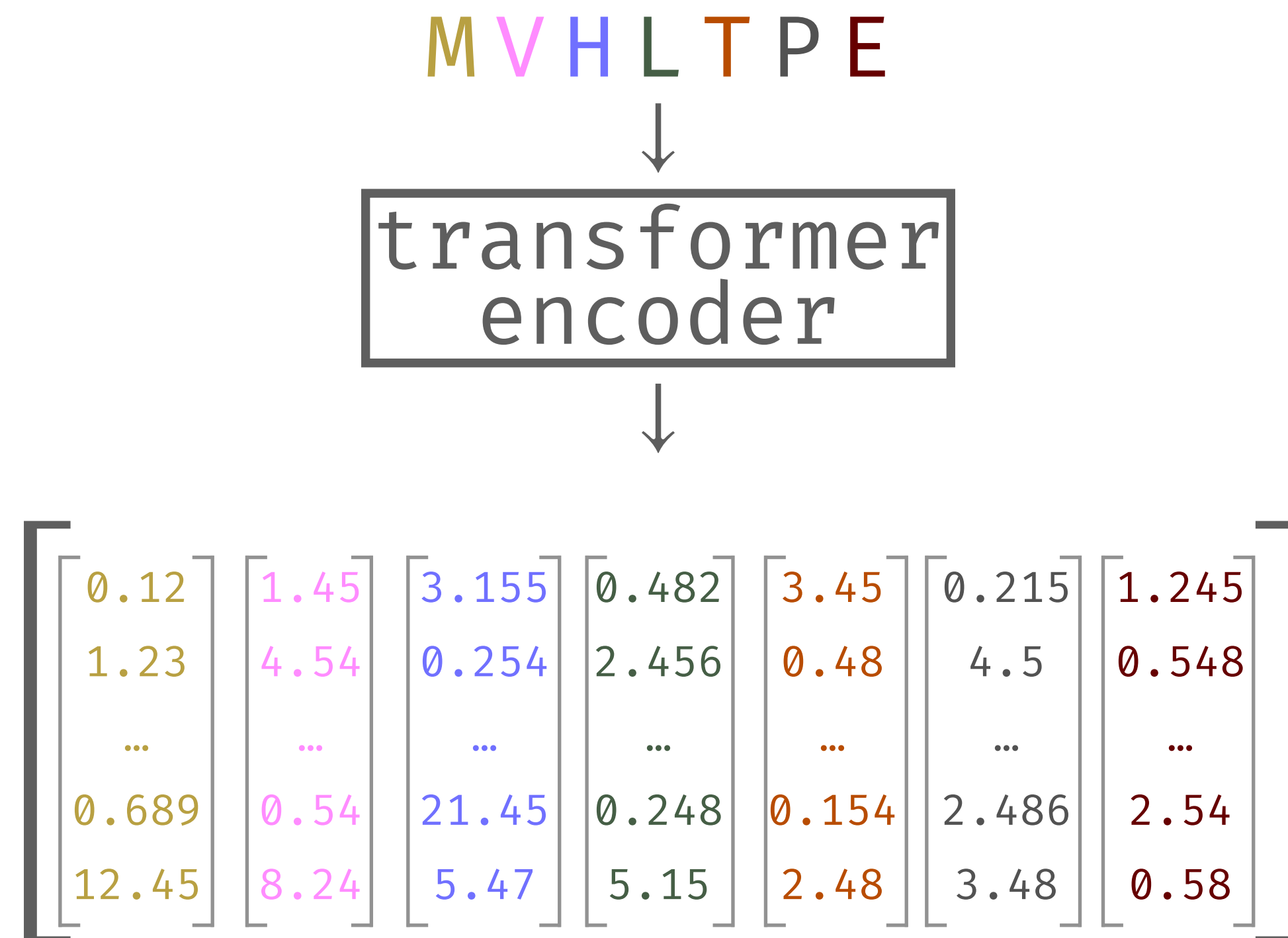
5

Learning alignment &
Other perspectives

6

Conclusion

Transformers and their embeddings

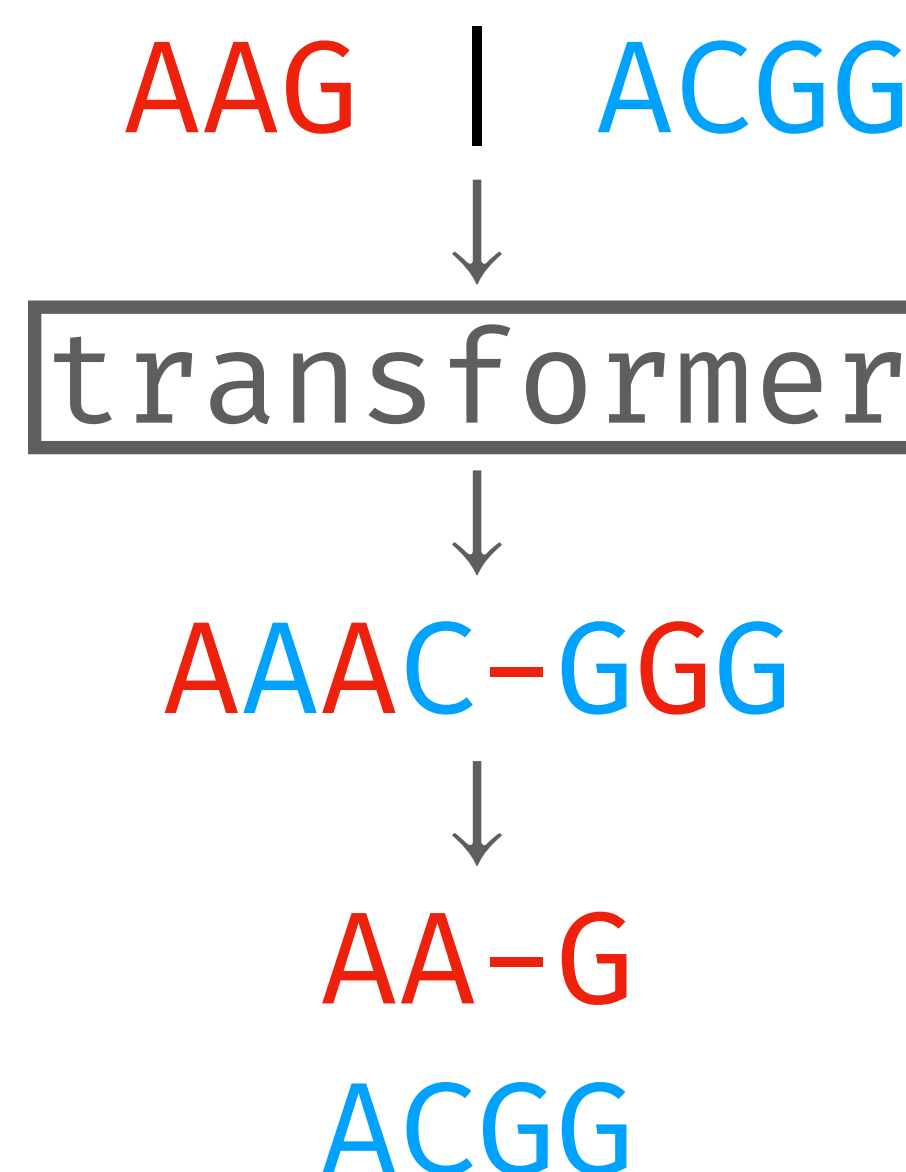


Using embeddings for alignments

- With embedding we can **learn** custom **parameters** for sequence **alignment**
- **DEDAL** **learns** a custom **substitution** matrix, for aligning 2 sequences
- **Better** pairwise **alignments** on **remote homologs** than standard methods

Alignment as a translation task

- Translate “Unaligned language” to “**aligned** language”



Caveats

- Main **problem** is **scaling** up
- **Self-attention** mechanism is very **memory hungry**
 - Approximations
 - Other mechanisms
- **Inference** time can be long

Presentation Outline

1

Introduction

2

Improving
long-read mapping

3

An introduction
to HIV

4

Exploring resistance in
HIV with ML

5

Learning alignment &
Other perspectives

6

Conclusion

Conclusion

- **Improving** sequence **alignments**:
 - We **improve** long-read **mapping**
 - **Better** than Homopolymer compression
 - Centromeres are **hard**
- **Learn** from sequence **alignments**:
 - Searching for **resistance** in HIV
⇒ sequence **classification**
 - Found potential **new resistance mutations**
 - Primary **resistance mutations** are **known**
- **Learning** sequence alignment is an exciting **perspective**

First Author

Co-first Author

Middle Author

50

Thank you all!



Gascuel O.



Lemoine F.



Morel M.



Voznica J.



Zhukova A.



Bernardini-Ridel M.



Andreace F.



Chikhi R.



Denti L.



Duitama-Gonzales C.



Dufresne Y.



Lemane T.



Vicedomini R.



Medvedev P.



Carcano A.



Holtz A.



Cadet-Diaby F.

Pasteur:

- Didier Mazel
- Jerome Bourret

Friends:

- Balzac
- EGS
- AgroParisTech

Family (of course)

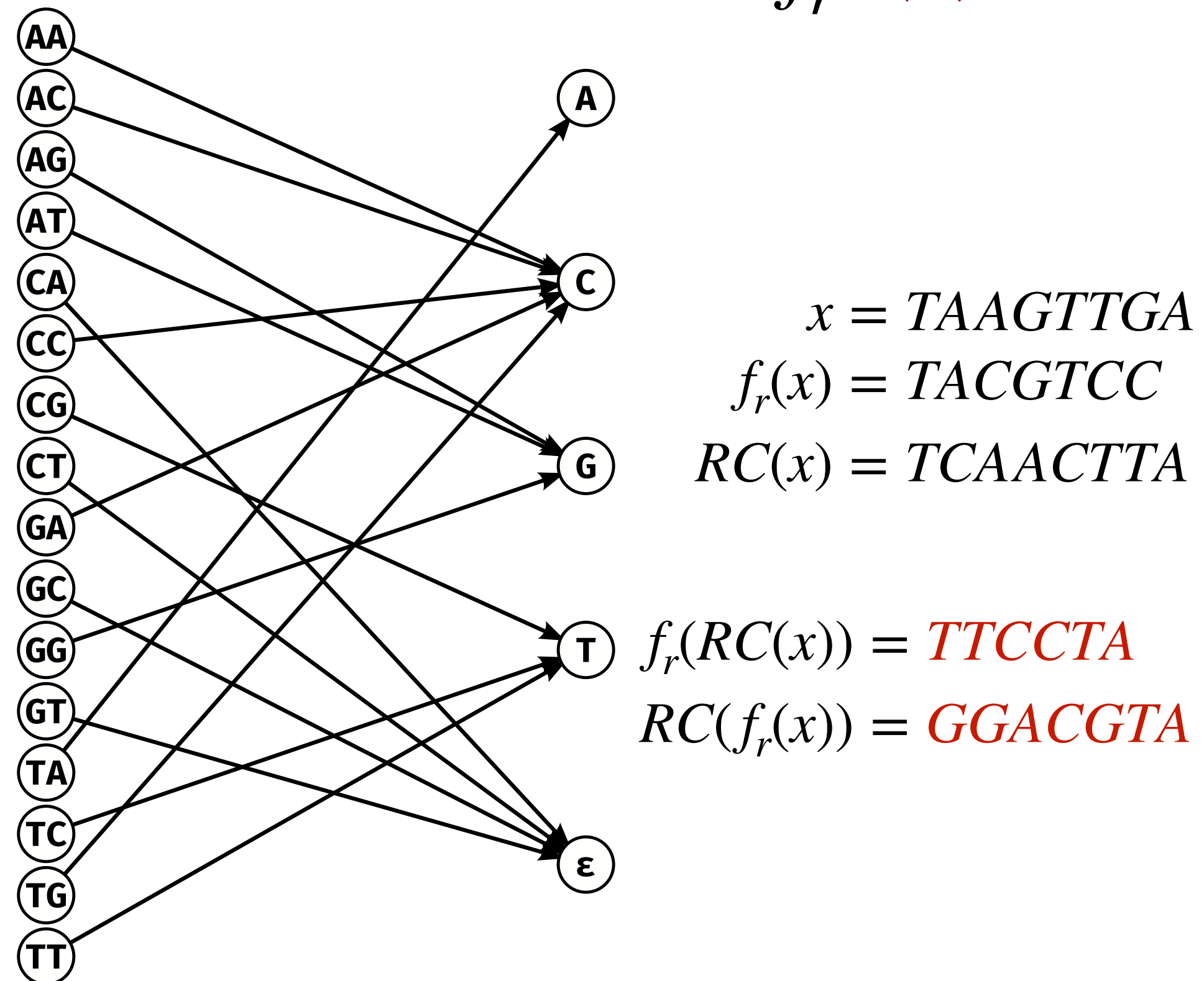
In memory of:

- Pierre Blassel
- John F. Murray

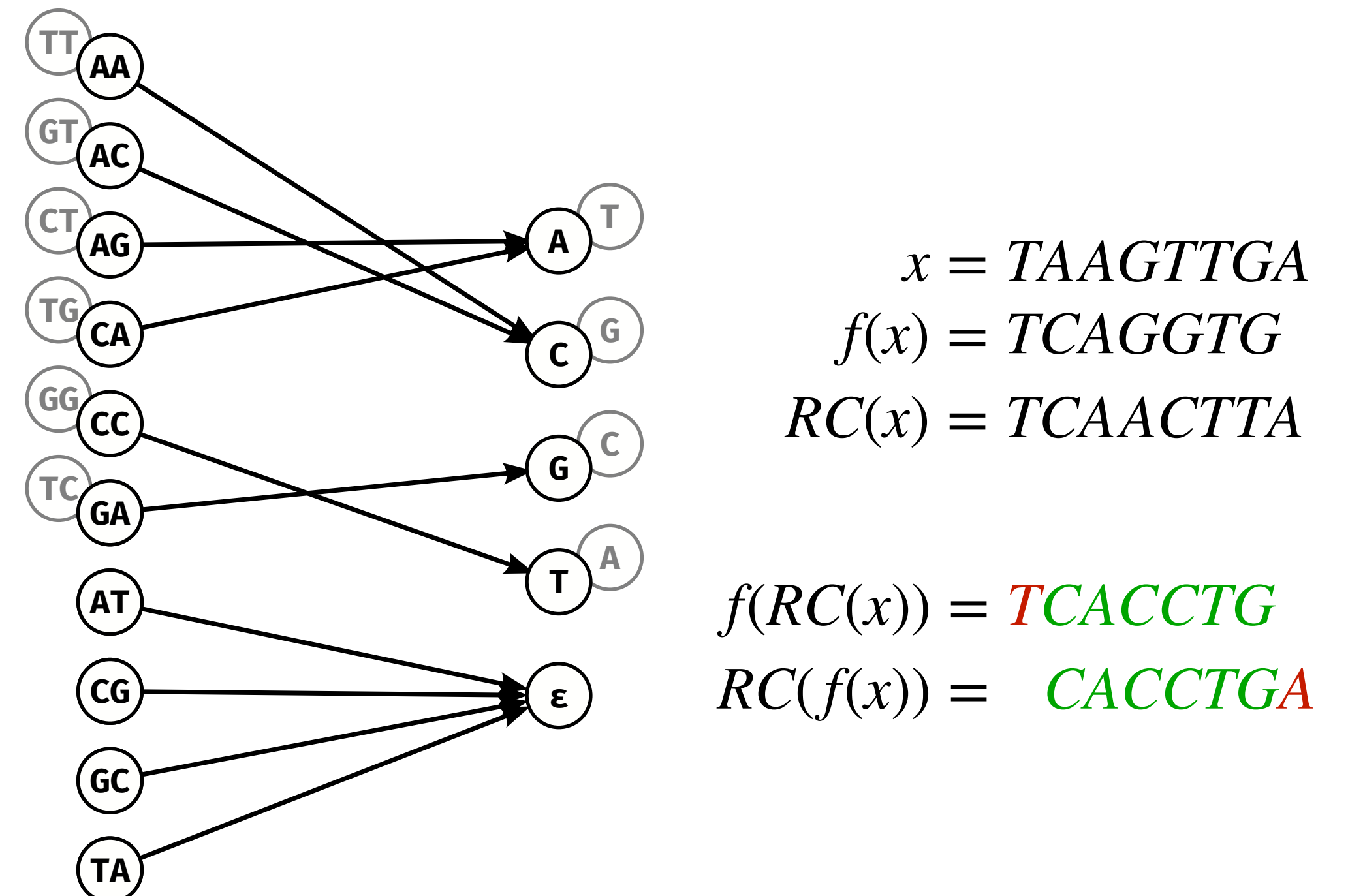
Reducing the search space

Reverse complements

Random SSR f_r ❌



RC-core-insensitive SSR f ✅



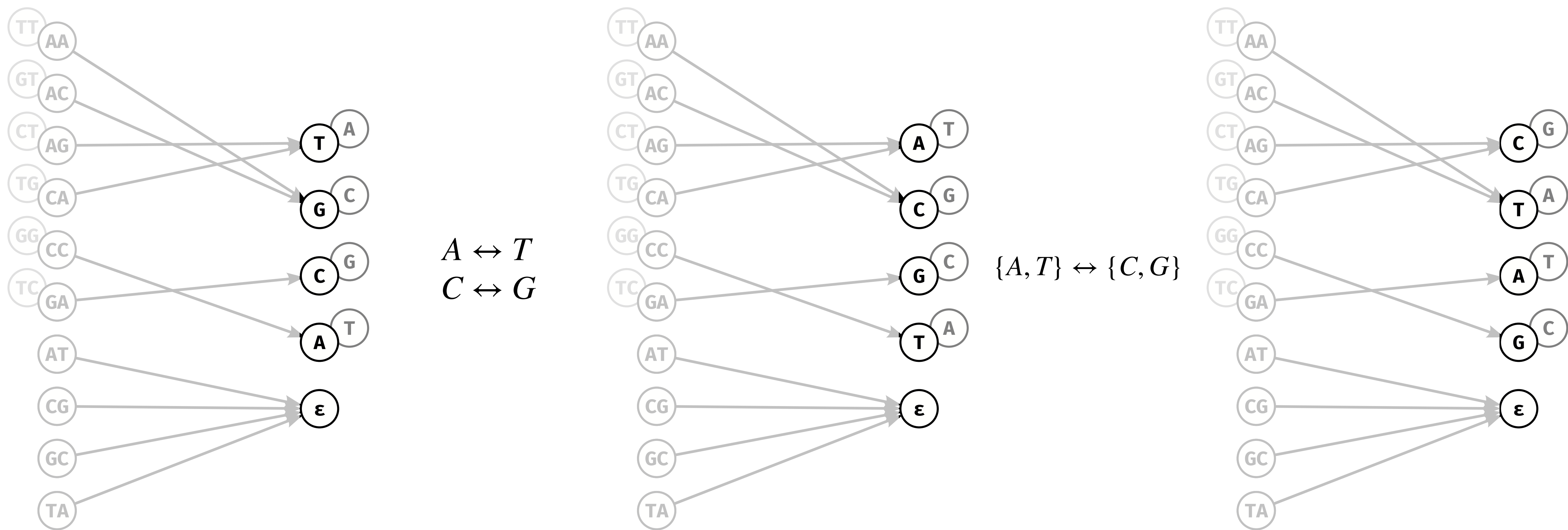
Reducing the search space

Equivalence classes

- Reverse complement **symmetries**:
 - $A \Leftrightarrow T$ and $G \Leftrightarrow C$
 - $\{A, T\}_{pair} \Leftrightarrow \{G, C\}_{pair}$
- We can define **equivalence classes** from them

Reducing the search space

Equivalence classes



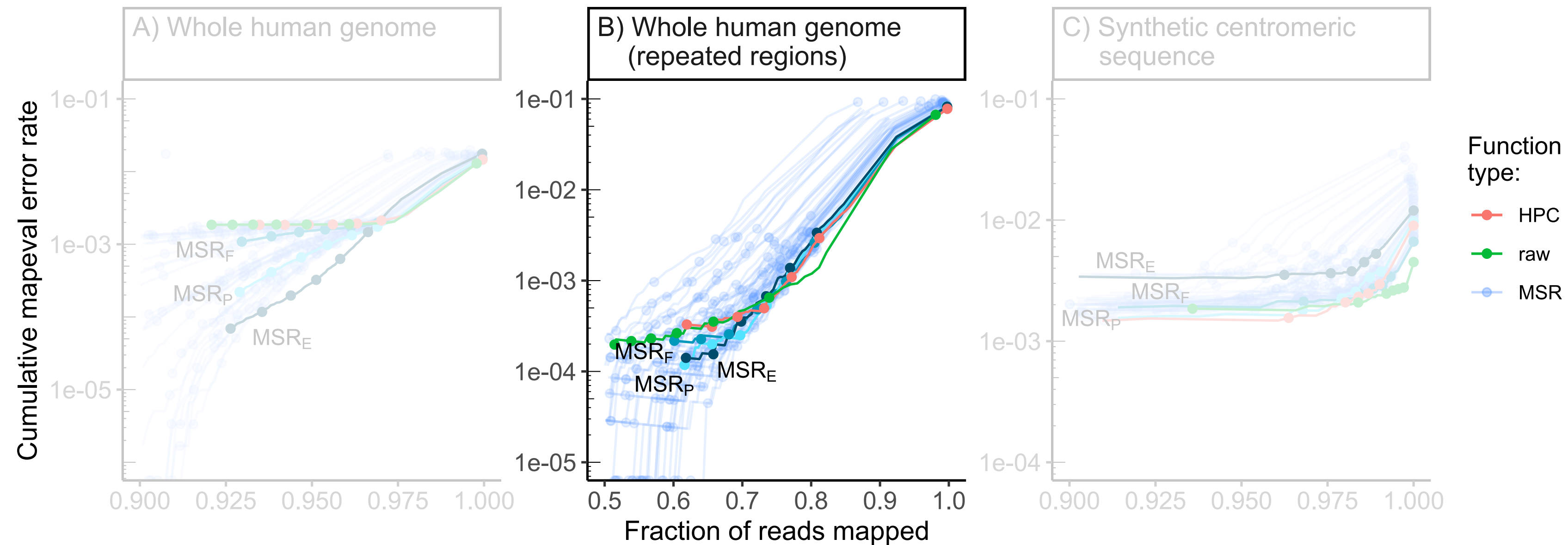
Evaluating MSRs

Evaluation Pipeline

- Mapping quality (**mapq**) is a measure of how confident the aligner is in its read placement. 0 (*worse*) \leq mapq \leq 60 (*best*)
- **mapeval** gives results for **mapq thresholds**
i.e. sets of mapped reads with mapq \geq than a given value
- **mapeval** reports for each threshold:
 - **Number of reads** mapped
 - Mapping **error rate**

Results

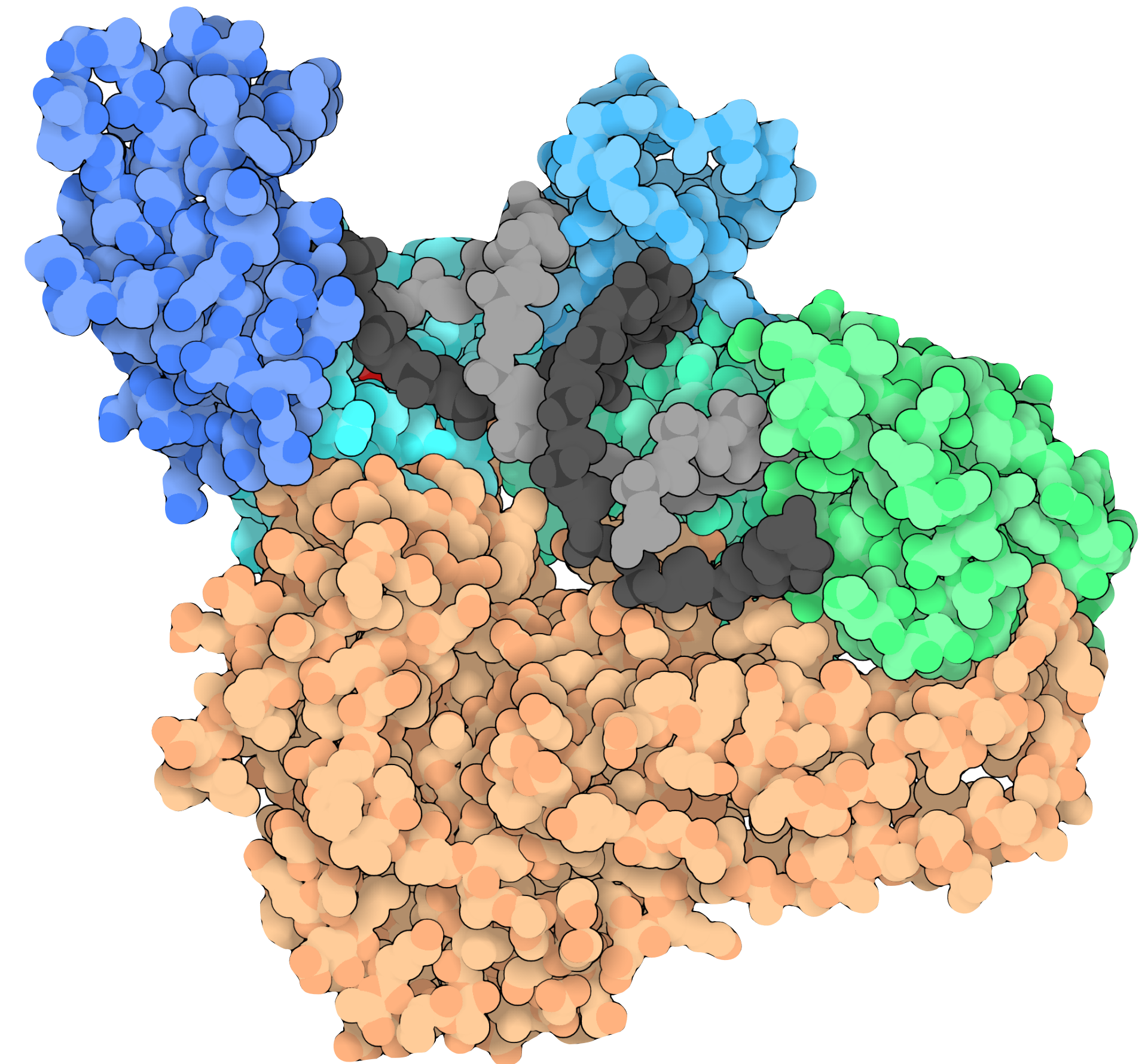
Repeated regions of the genome



MSRs are still **better** than HPC60
but performance **gap is smaller**

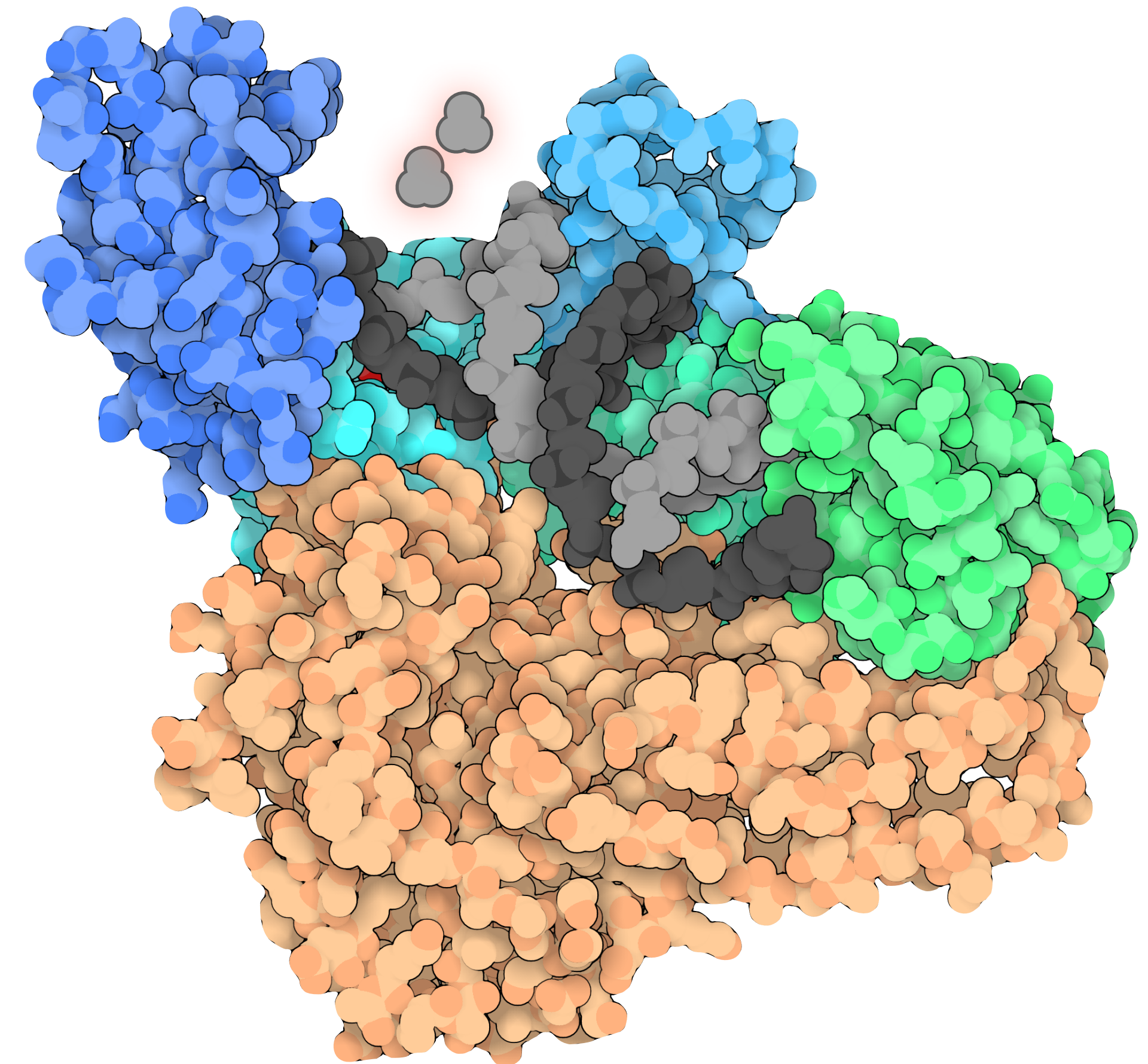
What are DRMs ?

- **Resistance** arises in **response** to treatment **pressure**
- Drug resistance mutations (**DRMs**) have been found for **every drug**
- **DRMs** often incur a **fitness cost**
- To **mitigate DRM** effects:
 - Treatment **switching**
 - **Combination** therapy



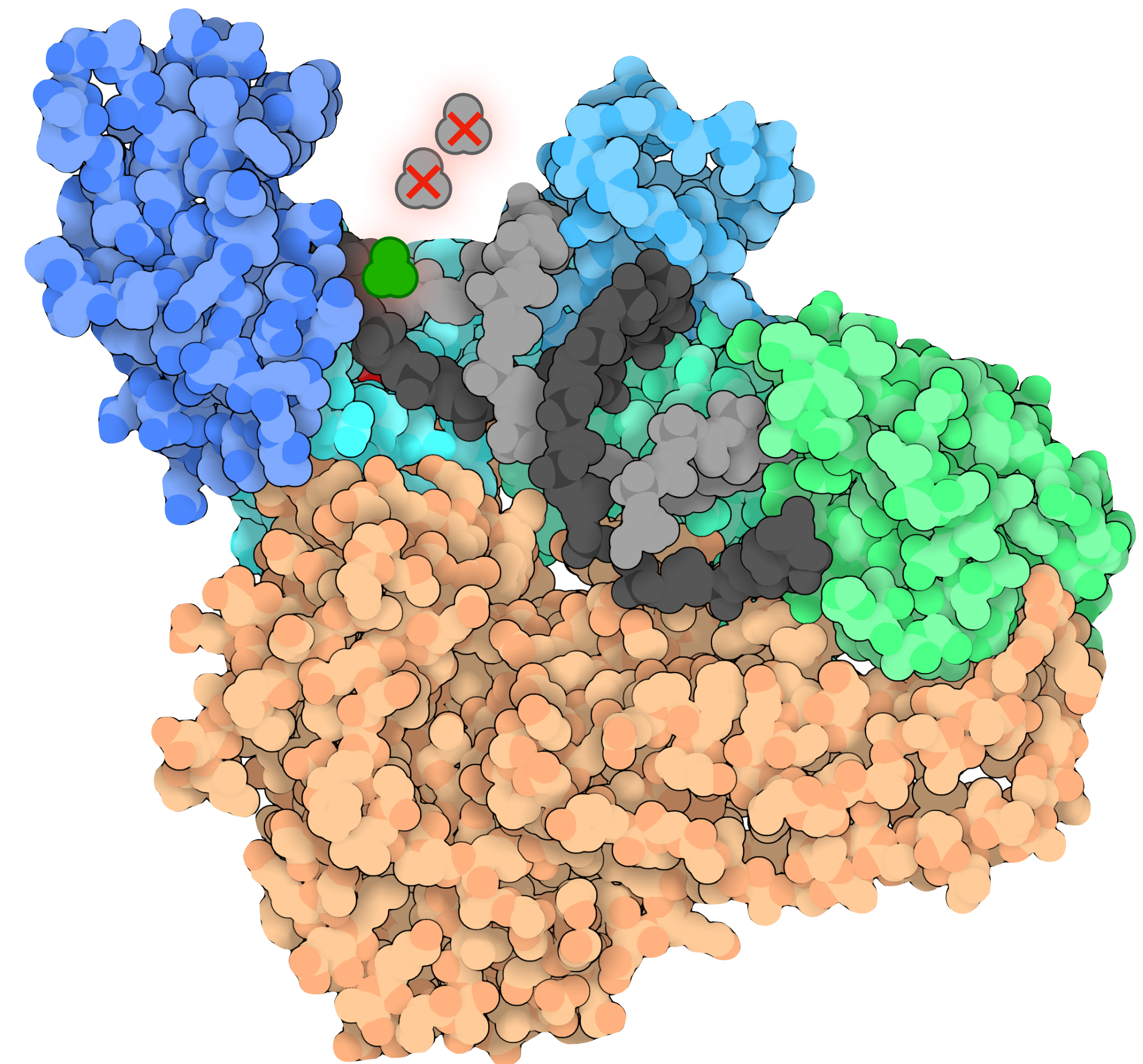
What are DRMs ?

- **Resistance** arises in **response** to treatment **pressure**
- Drug resistance mutations (**DRMs**) have been found for **every drug**
- **DRMs** often incur a **fitness cost**
- To **mitigate DRM** effects:
 - Treatment **switching**
 - **Combination** therapy



What are DRMs ?

- **Resistance** arises in **response** to treatment **pressure**
- Drug resistance mutations (**DRMs**) have been found for **every drug**
- **DRMs** often incur a **fitness cost**
- To **mitigate DRM** effects:
 - Treatment **switching**
 - **Combination** therapy



Preparing our data

Encoding scheme

Seq 1 ...I¹⁸⁰**V**¹⁸⁵QYMDDDL...

Seq 2 ...IY**D**YMDDDL...

Seq 3 ...IYQY**V**DDL...

Seq 4 ...I**K**QY**E**DD**K**...

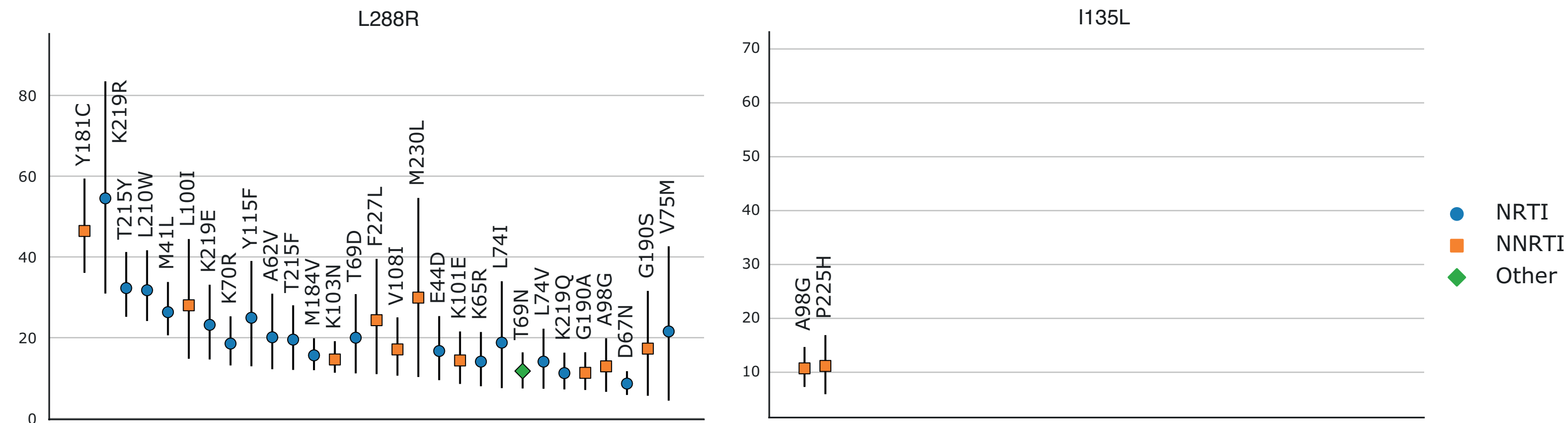
Seq 5 ...IY**F**YMDDDL...

	181V	181K	182D	182F	184V	184E	187K
Seq 1	1	0	0	0	0	0	0
Seq 2	0	0	1	0	0	0	0
Seq 3	0	0	0	0	1	0	0
Seq 4	0	1	0	0	0	1	1
Seq 5	0	0	0	1	0	0	0

Results

Did we find accessory RAMs ?

- **Relative risk** between **new RAMs** and **known DRMs**
→ **Overrepresentation** of RAMs in sequences with DRMs



Results

Structural argument

- **L228R** close to **active site** and **NNIBP**
- **I135L** close to **NNIBP** entrance
- NNIBP → NNRTI
- Active site → NRTI

