Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data

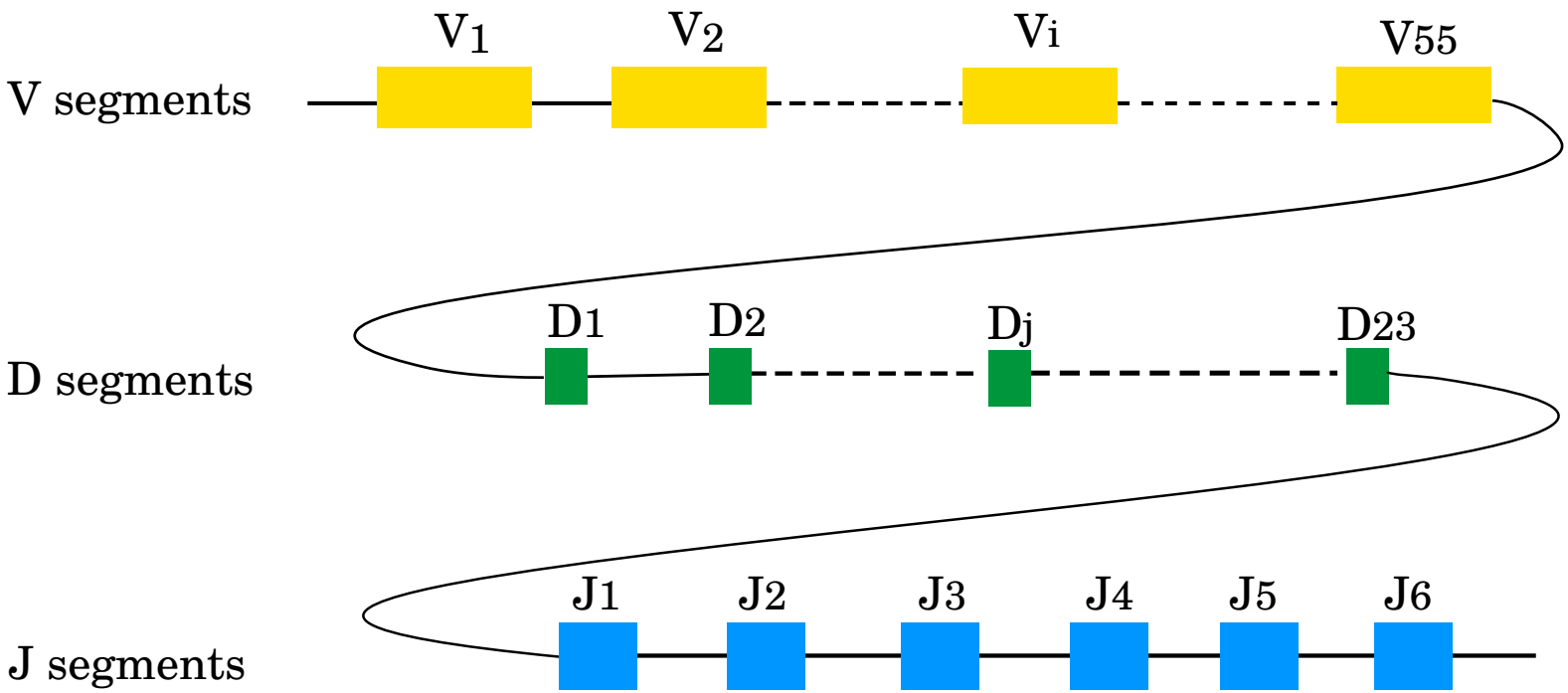
Andrey Bzikadze

February 12, 2016

Outline

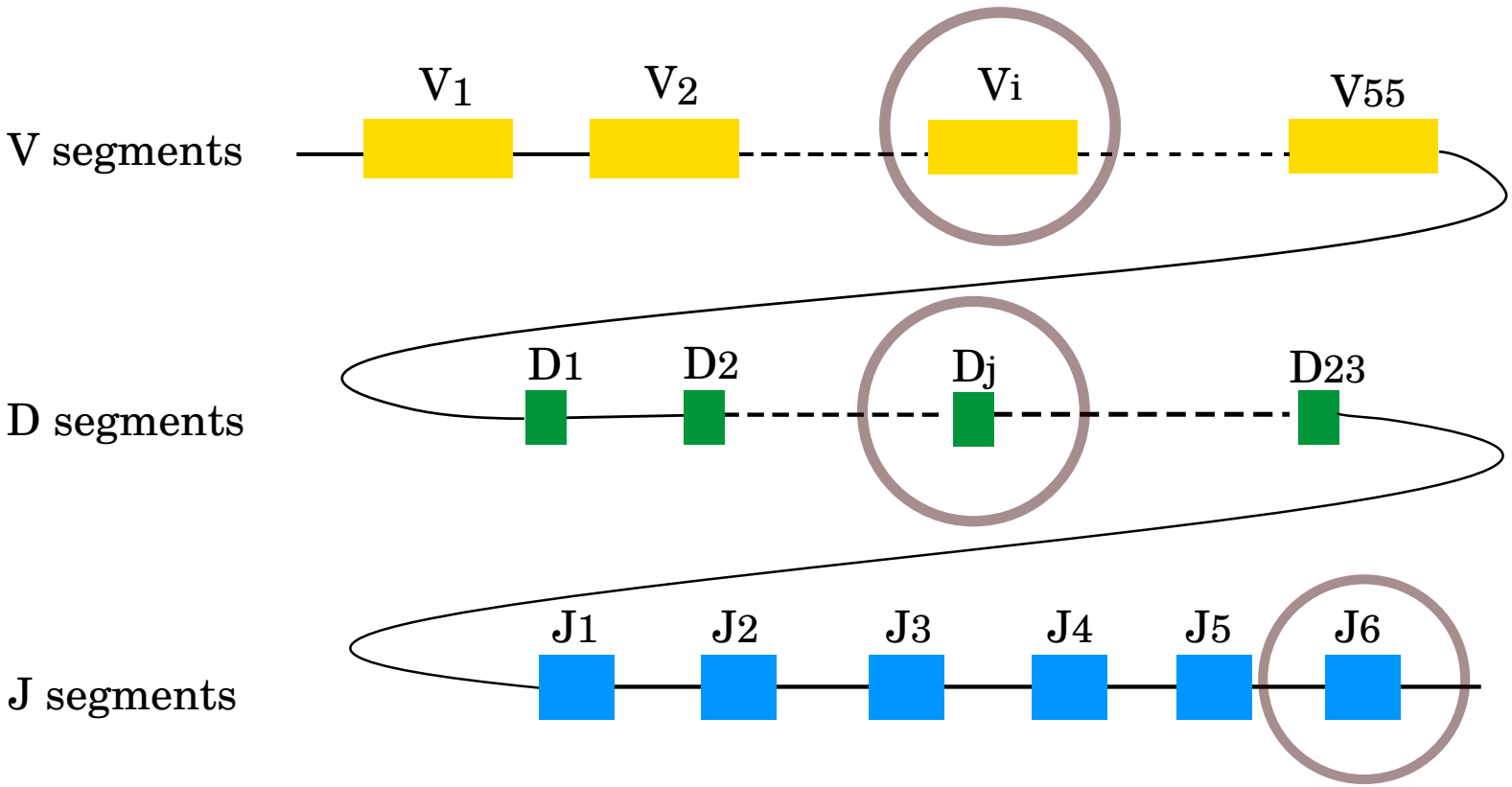
Introduction: V(D)J-recombination

Immunoglobulin heavy chain locus:



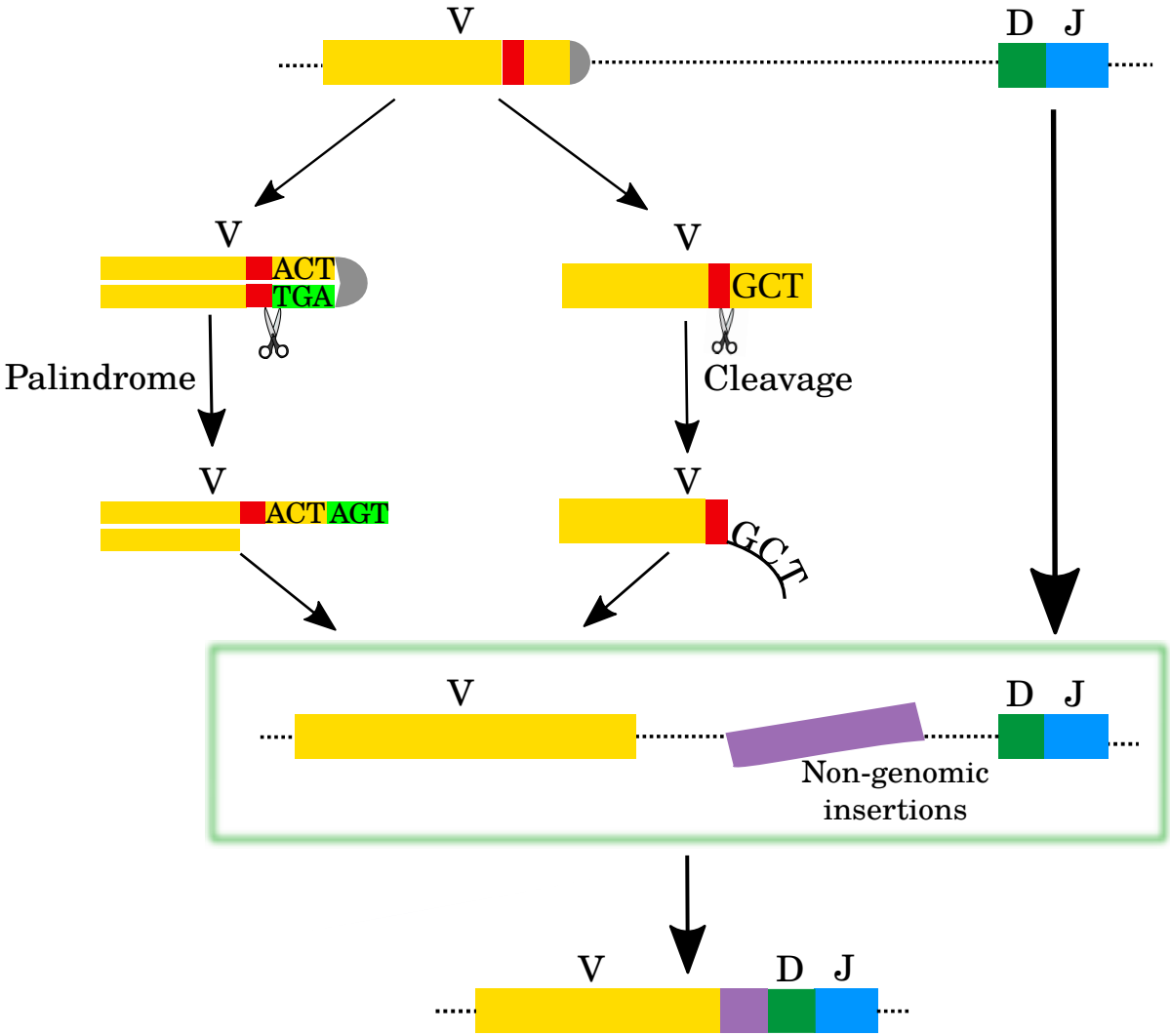
Introduction: V(D)J-recombination

Segment of each type is selected:



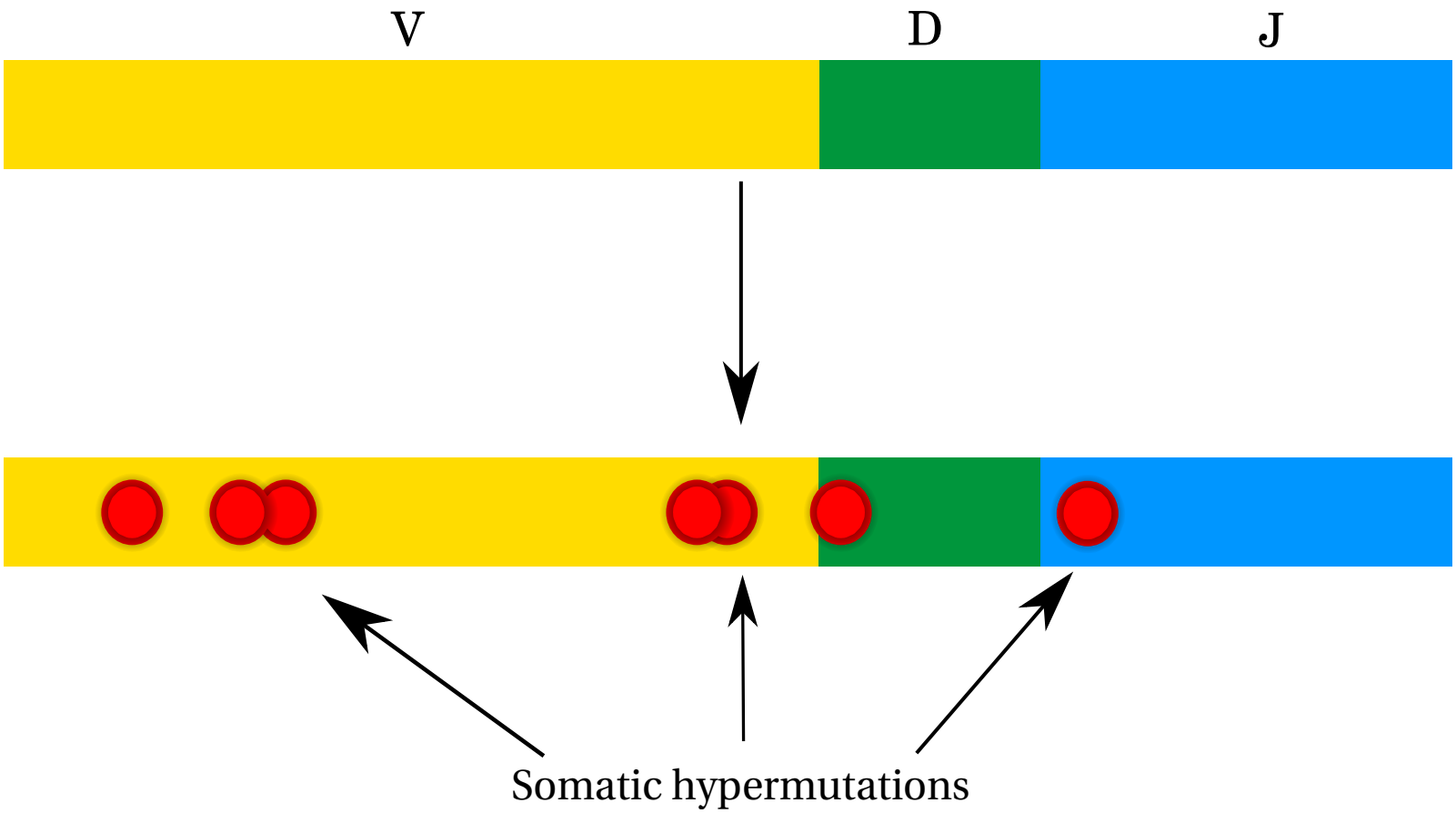
Introduction: V(D)J-recombination

3 types of biochemical events: *palindrome*, *cleavage*, *non-genomic insertion*.



Introduction: V(D)J-recombination

Further optimization of antibody affinity is achieved through extensive mutations referred as *somatic hypermutations*:



Introduction: V(D)J-recombination

Because we do not know the deterministic nature of the V(D)J-recombination and hypermutations, it is reasonable to consider it as a random (stochastic) process.

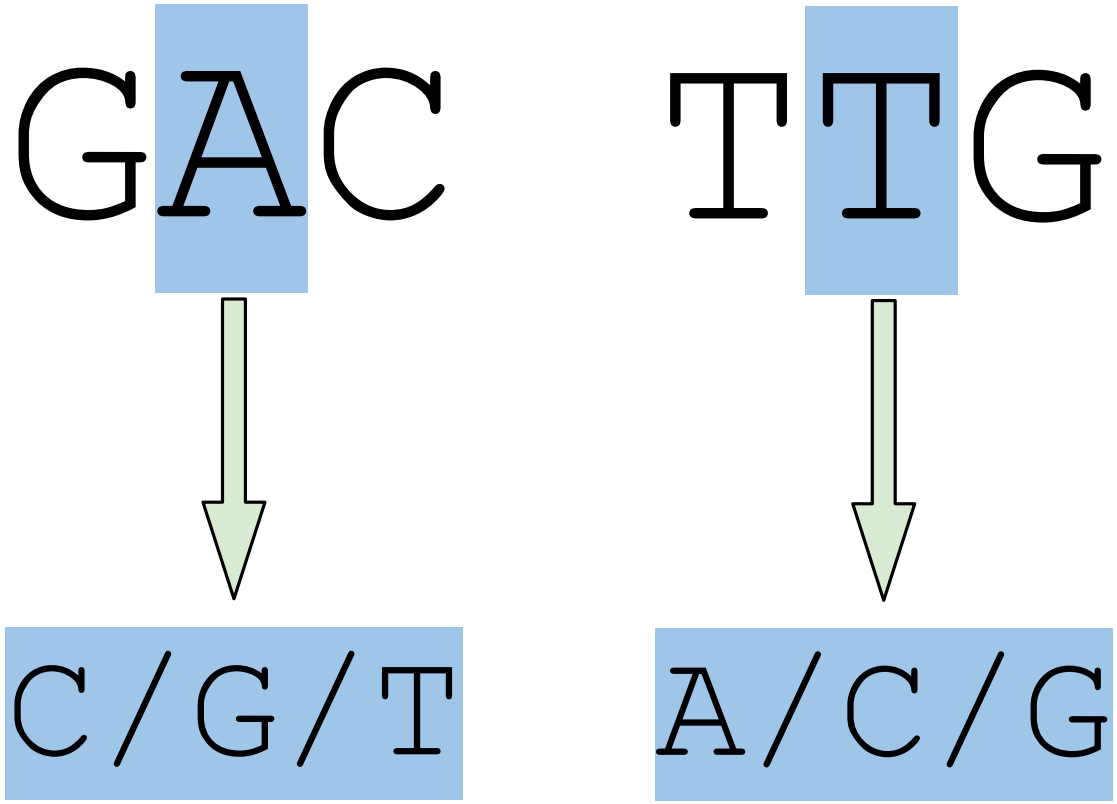
Hence the analysis of somatic recombination and hypermutations can be done in statistical and simulation terms.

Introduction: “hot” and “cold” spots

“hot”	WRCY/RGYW	W={A,T} Y={C,T} R={G,A}
	WA/TW	
	WRCH/DGYW	H={A,C,T} D={A,G,T}
“cold”	SYC/GRS	S={C,G}

Introduction: surrounding basis

Reuma Magori Cohen, Steven H. Kleinstein, Yoram Louzoun, Somatic hypermutation targeting is influenced by location within the immunoglobulin V region, Molecular Immunology, 2011.



Outline

SHM models based on 5-mers:

Gur Yaari *et. al.*, Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data — Front. Immunol., 2013.

SHM models based on 5-mers:

Gur Yaari *et. al.*, Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data — Front. Immunol., 2013.

A **targeting** model.

5-mer	Mutability
...	...
GCCTC	0.12
GCGAC	0.16
ACACT	0.48
AGCTA	3.17
...	...

SHM models based on 5-mers:

Gur Yaari *et. al.*, Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data — Front. Immunol., 2013.

A **targeting** model.

5-mer	Mutability
...	...
GCCTC	0.12
GCGAC	0.16
ACACT	0.48
AGCTA	3.17
...	...

A nucleotide **substitution** model.

5-mer	A	C	G	T
...	...	...	...	...
ACAAC	0	.24	.48	.29
GGCGT	.22	0	.12	.65
CCGTC	.35	.52	0	.13
TCTAC	.31	.54	.14	0
...	...	...	...	...

Bias of the selection

The standard solution: use **non-productively** rearranged Ig genes. (for example, Anand Murugana, Thierry Morab, Aleksandra M. Walczak and Curtis G. Callan — 2012).

Bias of the selection

The standard solution: use **non-productively** rearranged Ig genes. (for example, Anand Murugana, Thierry Morab, Aleksandra M. Walczak and Curtis G. Callan — 2012).

Authors point: "non-productively rearranged Ig genes may still be influenced by selection".

Authors solution: "developed a new methodology for constructing models from **synonymous** mutations only, thus avoiding the need to limit analysis to non-productive Ig sequences".

Outline

Used datasets

Subj.	Tech.	Raw	Processed
1	MiSeq	3,641,633	79,777
2	MiSeq	3,714,152	106,006
3	MiSeq	10,917,517	231,387
4	MiSeq	7,691,509	99,519
5	MiSeq	3,851,658	55,606
5	MiSeq	3,946,514	59,611
5	MiSeq	4,543,353	48,971
5	MiSeq	3,121,884	52,844
5	454	117,188	71,043
6	454	178,584	92,055
7	454	398,517	248,363
Total	—	42,122,509	1,145,182

Each sample was uniquely barcoded.

Pre-processing (pRESTO): quality control

- Removal of low-quality reads (mean Phred quality sc. < 20).
- For the MiSeq data, sets of sequences with identical molecular IDs were identified. Sets were collapsed into one consensus sequence per set.

Pre-processing (pRESTO): quality control

- Removal of low-quality reads (mean Phred quality sc. < 20).
- For the MiSeq data, sets of sequences with identical molecular IDs were identified. Sets were collapsed into one consensus sequence per set.
- Removal of sequences that do not appear in a single sample at least **twice**.

Pre-processing (pRESTO)

- Alignment — IMGT/HighV-QUEST.
- Non-mutated sequences: choice of V gene if $> 0.1\%$ of the sequences; choice of V gene alleles if $> 10\%$ of the assignments to this V gene.
- Mutated sequences: closest V segment due Hamming distance.

Pre-processing (pRESTO): clonally related sequences

Clones — the sequences related by a common ancestor.

Identification of clonally related sequences:

1. Clusterization based on V, J alignment and junction.
2. Clones, if junctions differ by ≤ 3 mutations.

> *"The threshold of three was determined after manual inspection of the mutation patterns in resulting clones identified through building lineage trees."*

Outline

Substitution model

Consider only 5-mers where **any** mutation at central position is synonymous.

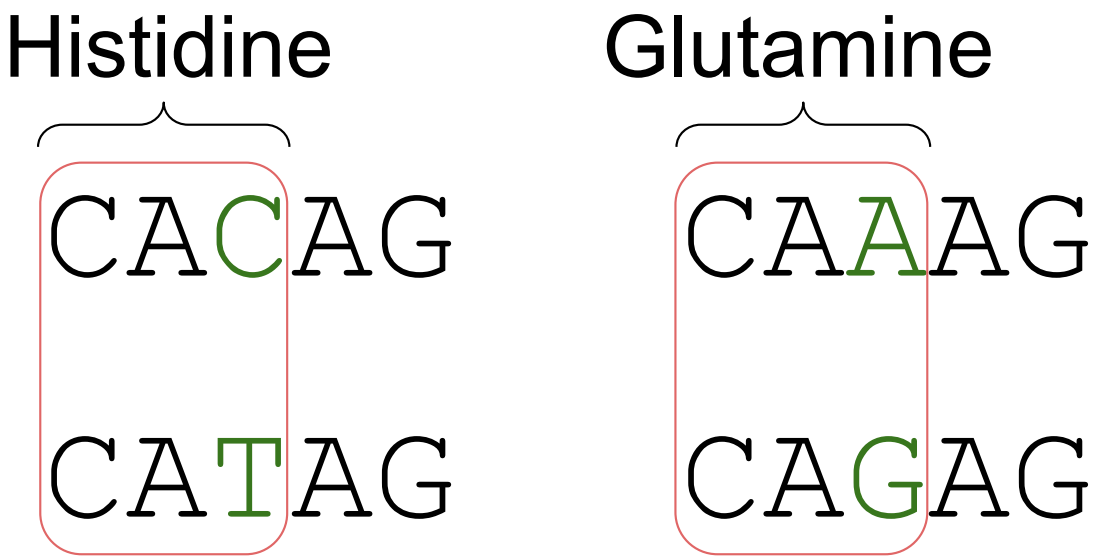
Subj.	**Tech.**	**Processed**	**# Subst. mut.**
1	MiSeq	79,777	25,307
2	MiSeq	106,006	57,215
3	MiSeq	231,387	108,591
4	MiSeq	99,519	68,051
5	MiSeq	55,606	23,939
5	MiSeq	59,611	24,971
5	MiSeq	48,971	20,865
5	MiSeq	52,844	23,243
5	454	71,043	8,209
6	454	92,055	23,260
7	454	248,363	24,771
Total	—	1,145,182	408,422

Substitution model: definition

For each 5-mer M let central base mutate to $B \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.
The model is the set of probabilities for mutation of M to B.
Estimations are frequencies.

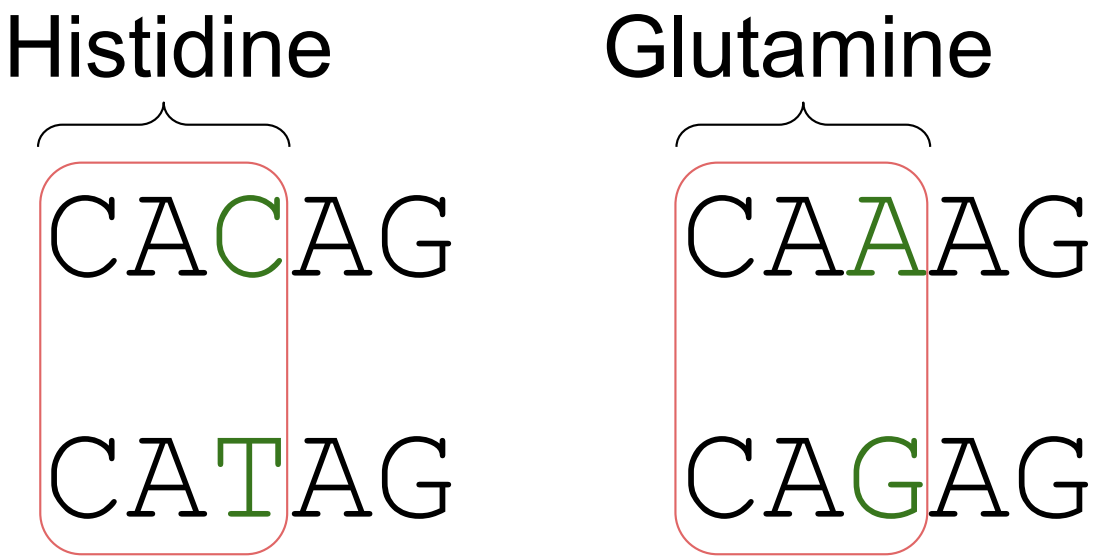
Substitution model: inferred estimations

- Not all 5-mers appear in datasets.
- Some 5-mers can never appear: not all substitutions are synonymous.



Substitution model: inferred estimations

- Not all 5-mers appear in datasets.
- Some 5-mers can never appear: not all substitutions are synonymous.



Substitutions for 717 (of 1024) 5-mers were not estimated!

Substitution model: inferred estimations

To estimate, for example, $\mathbb{P}(\mathrm{CALAG} \to \mathrm{CAMAG})$:

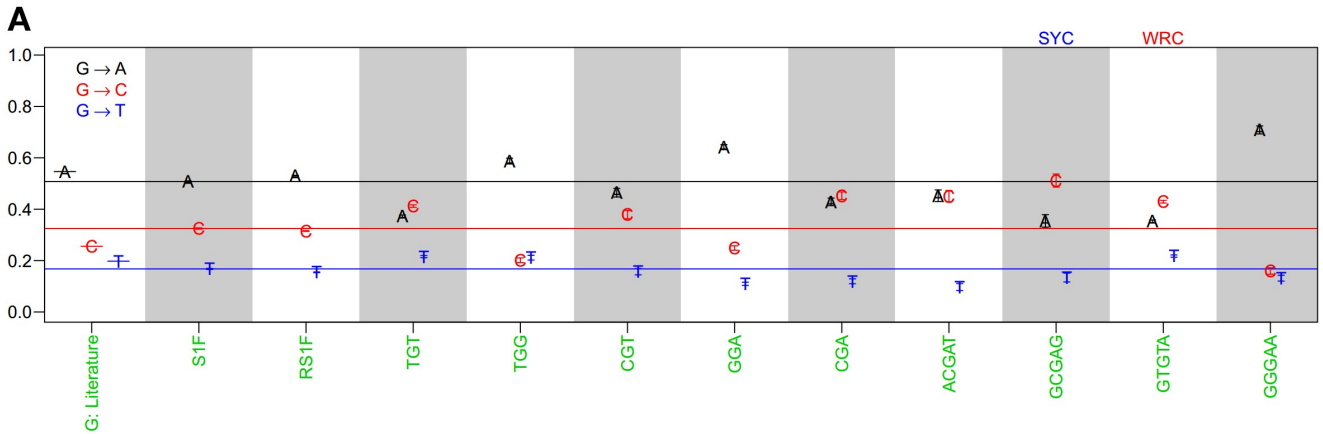
- "inner 3-mer": $\mathbb{P}(*\,\mathrm{ALA}\,* \to\ *\mathrm{A}\,\mathrm{M}\,\mathrm{A}*)$;
- 2 upstream nucleotides: $\mathbb{P}(\mathrm{CAL}\,**\to \mathrm{CAM}\,**)$;
- 2 downstream nucleotides: $\mathbb{P}(**\,\mathrm{LAG} \to **\,\mathrm{MAG})$;
- "hot-spot" method:

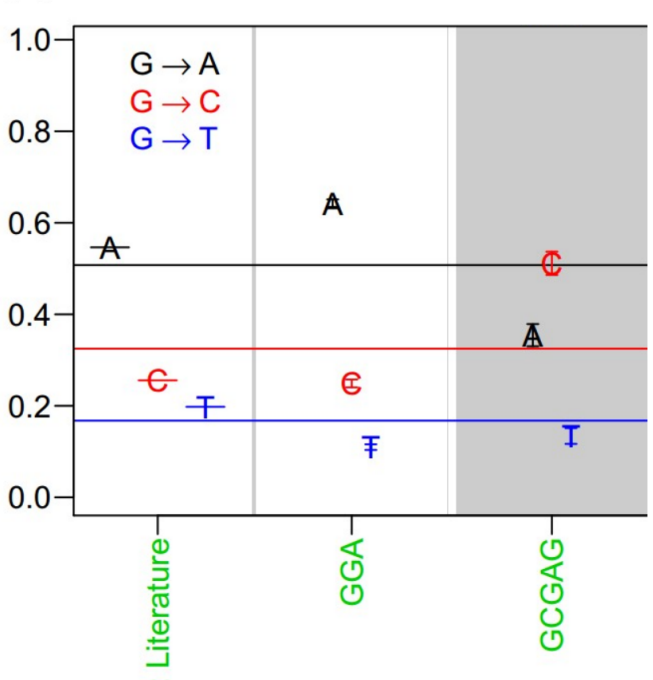
$$\begin{cases} \mathbb{P}(\mathrm{CAL}\,**\to \mathrm{CAM}\,**), & \mathrm{L},\mathrm{M} \in \{\mathrm{A},\mathrm{C}\}; \\ \mathbb{P}(**\,\mathrm{LAG} \to **\,\mathrm{MAG}), & \mathrm{L},\mathrm{M} \in \{\mathrm{G},\mathrm{T}\}. \end{cases}$$

Table : Corr. between true estimations and inferred for synonymous mutations.

Correlation	Inner	Upstream	Downstream	Hot spots
Pearson	.4	.37	.15	.04
Spearman	.2	.24	.23	.09

Substitutions are affected by adjacent nucleotides





Substitution model: comments

Consistency on different datasets: in my opinion, the arguments are not convincing

- Bootstrap 95% CIs often do not overlap.
- Correlation between mutations estimated on some of the datasets (esp. 454) is quite low.

Statistical research: absence of information about

- Significance of the results.
- Hypothesis testing about non uniformity of the 5-mer mutations distributions.
- Hypothesis testing about non identical distribution of the 5-mer mutations distributions.

Outline

Mutability model

Only mutations that are synonymous were considered.

Subj.	Tech.	Processed	# Targ. mut.
1	MiSeq	79,777	53,840
2	MiSeq	106,006	106,265
3	MiSeq	231,387	208,338
4	MiSeq	99,519	132,795
5	MiSeq	55,606	48,558
5	MiSeq	59,611	50,117
5	MiSeq	48,971	42,737
5	MiSeq	52,844	47,049
5	454	71,043	48,838
6	454	92,055	50,899
7	454	248,363	17,424
Total	—	1,145,182	806,860

Mutability model: definition

"The mutability of a motif is defined here as the (non-normalized) probability of the central base in the motif being targeted for SHM relative to all other motifs."

2 steps:

- Calculating the **background frequency** of the different 5-mers based on the germline (unmutated) version of the sequence.
- Creating a table of the 5-mers that were mutated in the sequence.

Mutability model: definition

For each string S let denote GL — germline string for string S. Then for each 5-mer M background frequency is

$$B_M^S = \sum_{i=1}^{\mathrm{Len}(S)} \sum_{b \in ACGT} \mathrm{PrSubst}(M, b)\mathrm{IsSynonymous}(\mathrm{GL}[i], b|M).$$

$$C_M^S = \sum_{i=1}^{\mathrm{Len}(S)} \mathrm{IsSynonymous}(\mathrm{GL}[i], \mathrm{OS}[i]|M).$$

Mutability model: definition

Mutability score μ and normalized mutability score are defined as follows

$$\mu_M^S = C_M^S / B_M^S;$$
$$\overline{\mu}_M^S = \mu_M^S / \sum_m \mu_m^S.$$

Final mutability score for the 5-mer M is defined as

$$\mathrm{Mut}_M = \frac{1}{\#S} \sum_S \overline{\mu}_M^S \left(\sum_m C_m^S \right).$$

And what if $B_M^S = 0$?

Mutability model: inferred estimations

Same 4 methods were proposed.
To estimate, for example, $\mathbb{P}(\mathrm{CALAG} \to \mathrm{CAMAG})$:

- "inner 3-mer": $\mathbb{P}(*\,\mathrm{ALA}\,* \to\ *\mathrm{A}\,\mathrm{M}\,\mathrm{A}*)$;
- 2 upstream nucleotides: $\mathbb{P}(\mathrm{CAL}\,**\to \mathrm{CAM}\,**)$;
- 2 downstream nucleotides: $\mathbb{P}(**\,\mathrm{LAG} \to **\,\mathrm{MAG})$;
- "hot-spot" method:

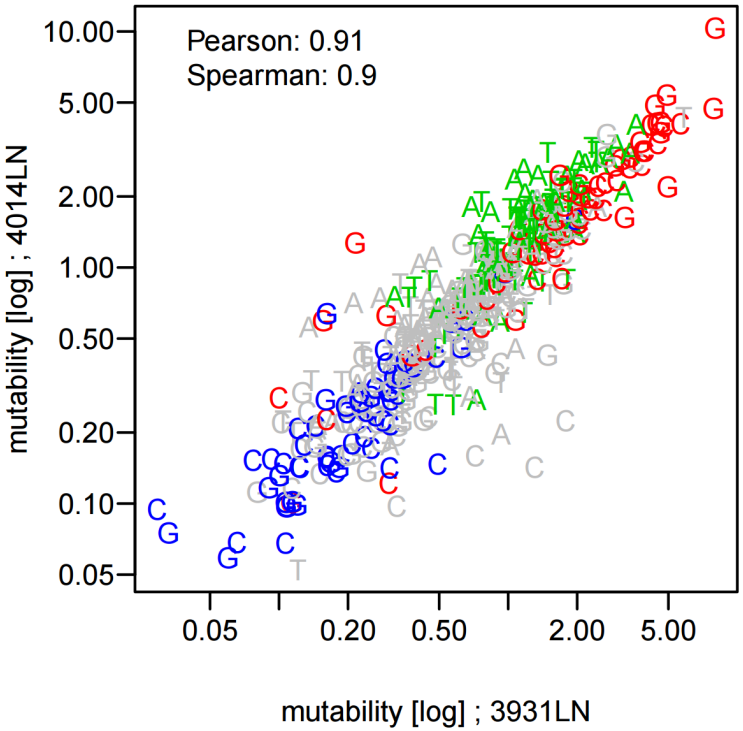
$$\begin{cases} \mathbb{P}(\mathrm{CAL}\,** \to \mathrm{CAM}\,**), & \mathrm{L},\mathrm{M} \in \{\mathrm{A},\mathrm{C}\}; \\ \mathbb{P}(**\,\mathrm{LAG} \to **\,\mathrm{MAG}), & \mathrm{L},\mathrm{M} \in \{\mathrm{G},\mathrm{T}\}. \end{cases}$$

Table : Corr. between true estimations and inferred for targeting model.

Correlation	Inner	Upstream	Downstream	Hot spots
Pearson	.58	.57	.61	.73
Spearman	.61	.58	.64	.79

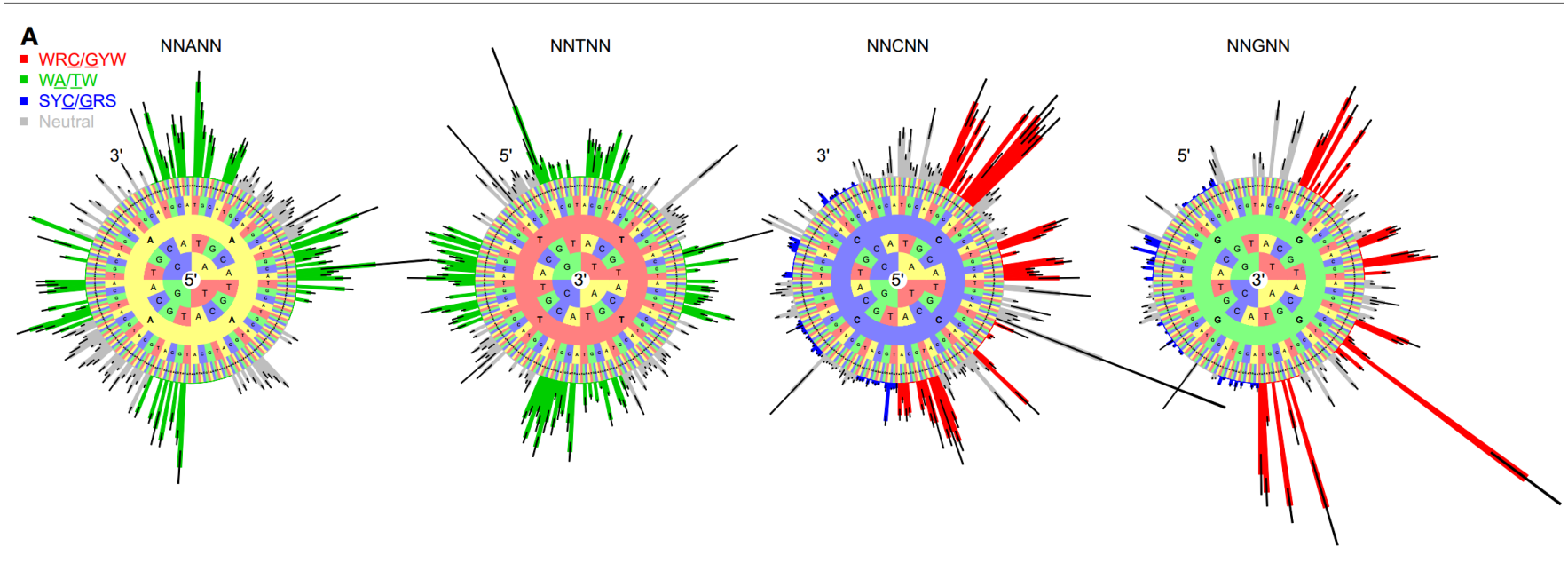
Targeting model: consistency on different datasets

"Comparison of the motif mutabilities between pairs of samples showed that the models were highly consistent, with Pearson correlation ≈ 0.9*."*

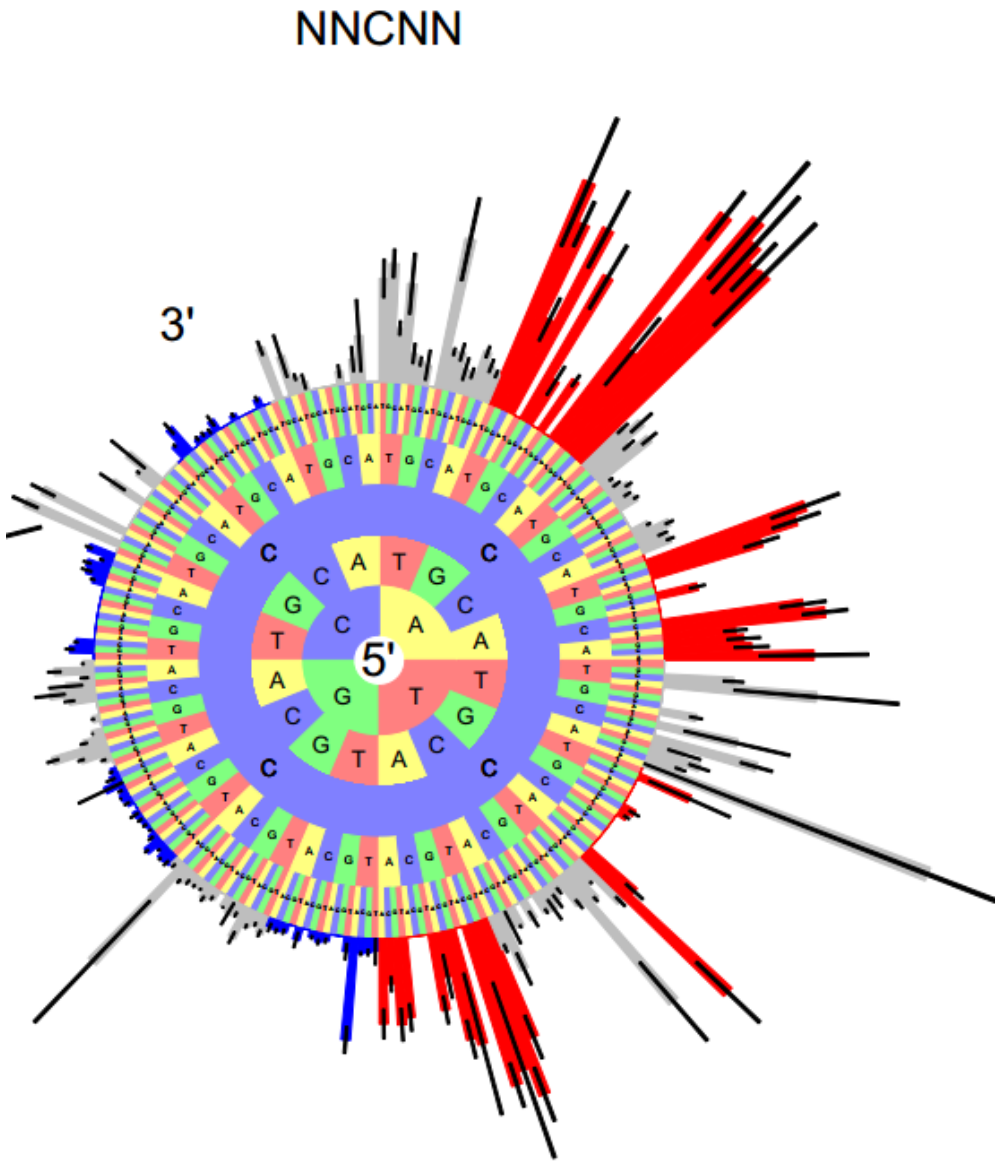


Still... I do not find it really convincing because it is based on substitution model.

Targeting model: Hedgehog plots



Targeting model: Hedgehog plots



Outline

The **R** codebase is opened and 2 models computed on **all** datasets are uploaded.

Unfortunately, we find it impossible to use opened codebase, because of the script errors.