MRC | Biostatistics Unit

# Retroviruses integrate into a shared, non-palindromic motif

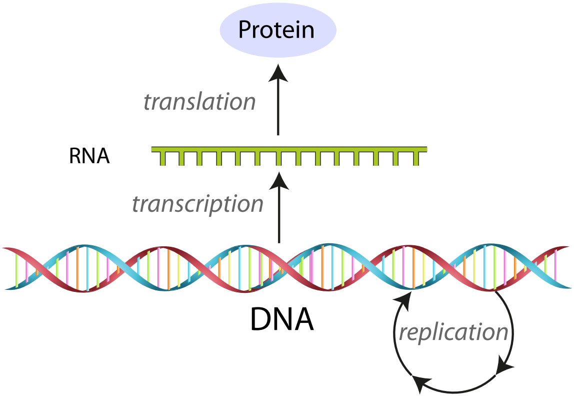
Paul Kirk

MASAMB 2016, Cambridge

October 4, 2016

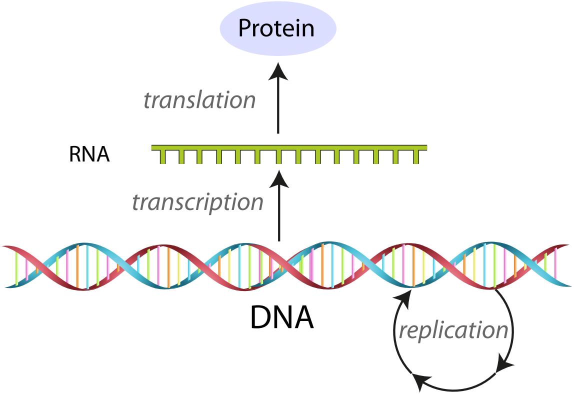
# Central dogma of molecular biology (Crick, 1956)

**General** transfers of biological sequential information:



# Central dogma of molecular biology (Crick, 1956)

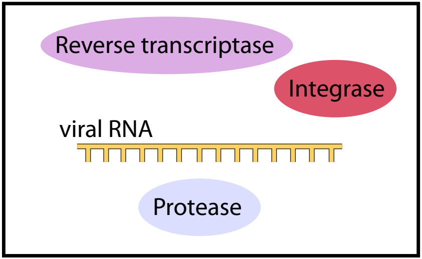
**General** transfers of biological sequential information:



There are also **special** transfers of sequential information.

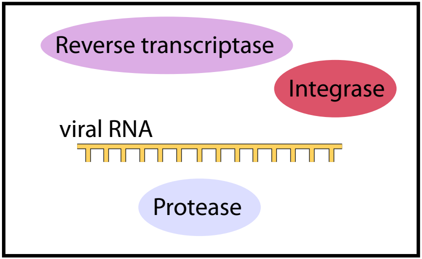
# For example: retroviruses

A retrovirus:



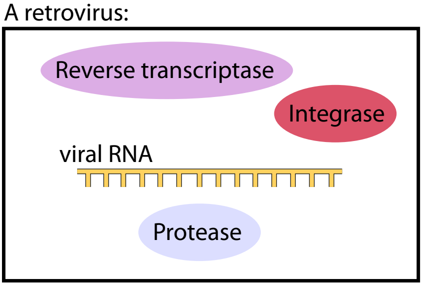
# For example: retroviruses

A retrovirus:



Retroviruses are *obligate parasites*: they require a host cell to complete their “life”-cycle.

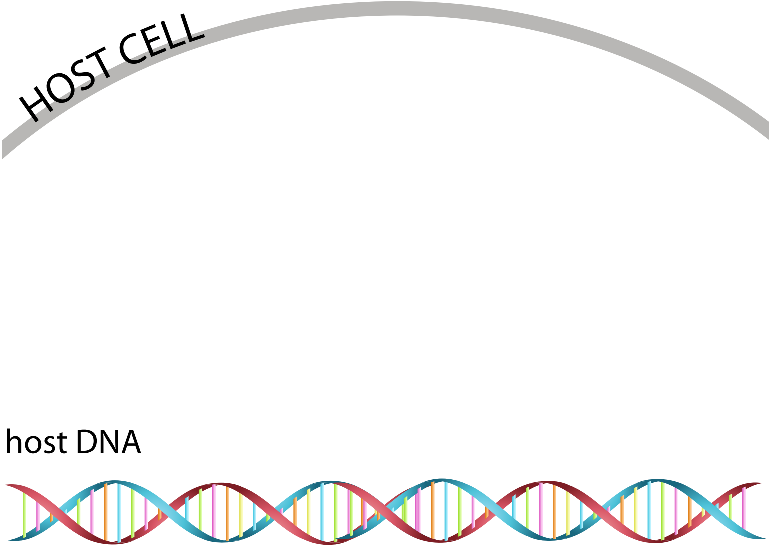
# For example: retroviruses



Retroviruses are *obligate parasites*: they require a host cell to complete their “life”-cycle.

Examples: HIV, HTLV-1, . . . .

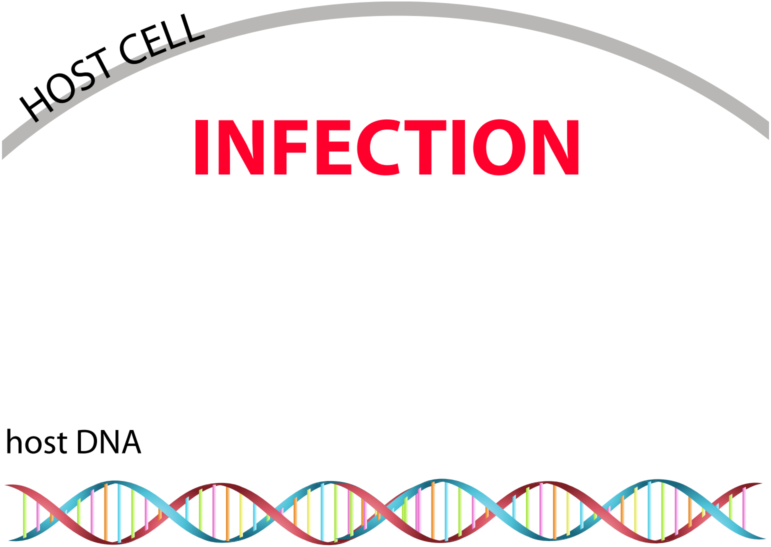
# For example: retroviruses



HOST CELL

host DNA

# For example: retroviruses

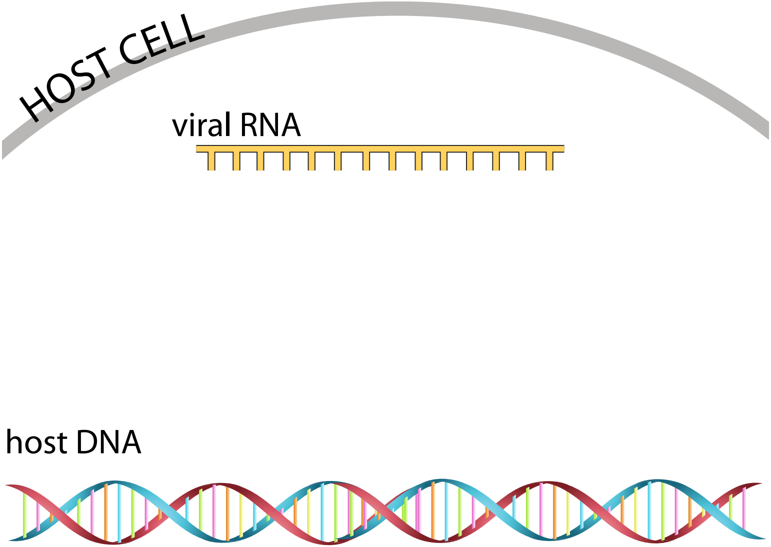


HOST CELL

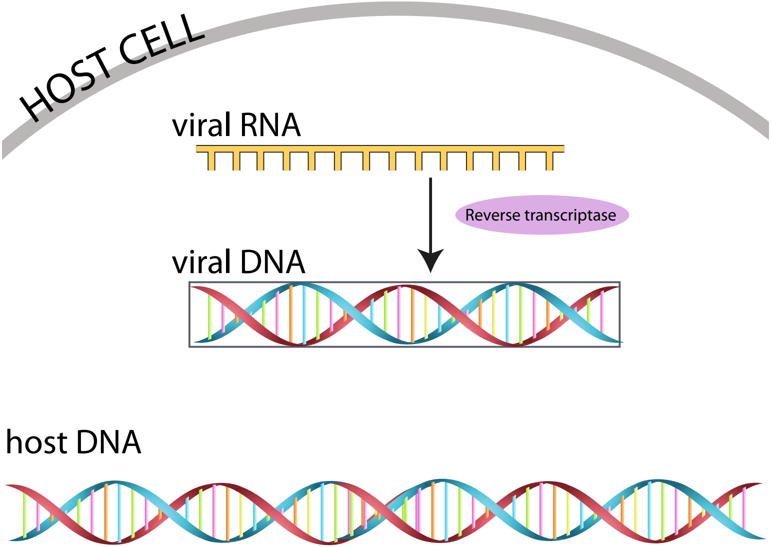
INFECTION

host DNA

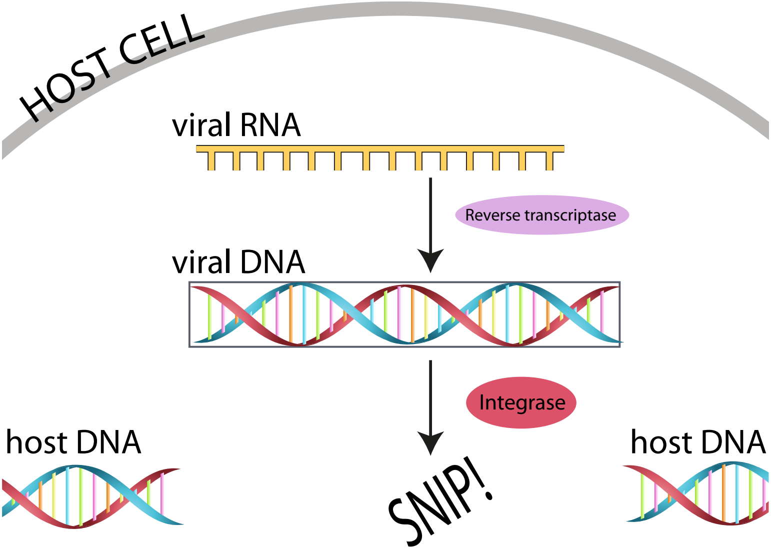
# For example: retroviruses



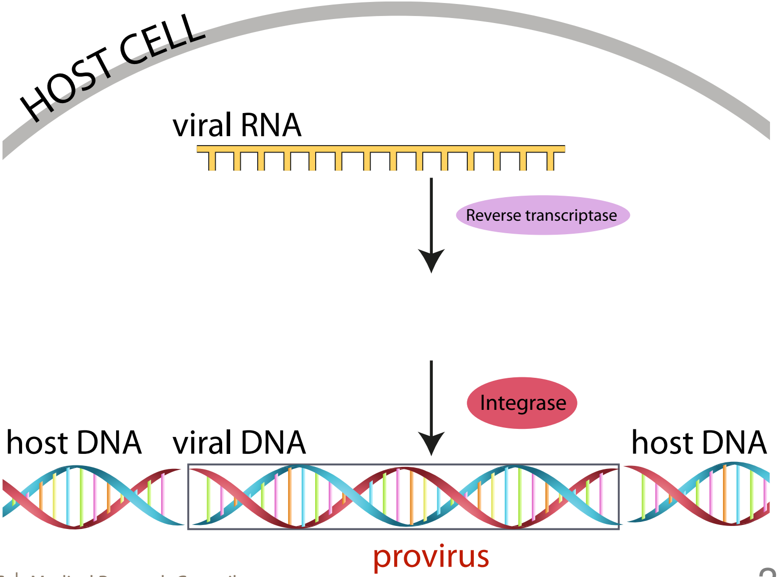
# For example: retroviruses



# For example: retroviruses



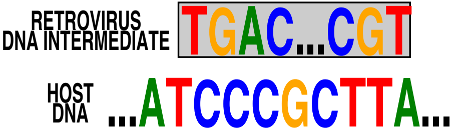
# For example: retroviruses



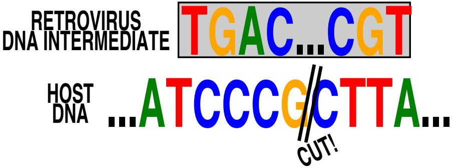
# Characterising retroviral integration sites

HOST
DNA ...ATCCCGCTTA...

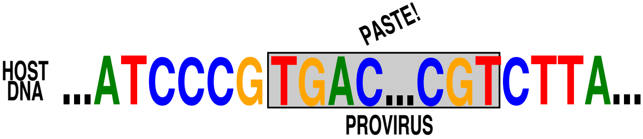
# Characterising retroviral integration sites



# Characterising retroviral integration sites



# Characterising retroviral integration sites

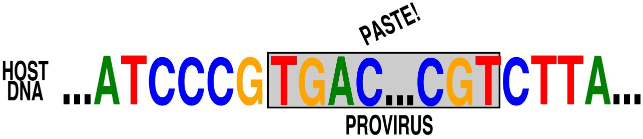


PASTE!

HOST DNA ...ATCCCGTGAC...CGTCTTA...

PROVIRUS

# Characterising retroviral integration sites

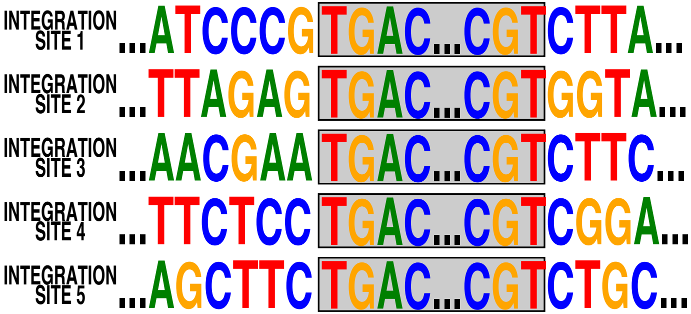


We would like to characterise the target integration site

- i.e. the regions flanking the provirus
- Is there a motif?

# Aligning integration sites

Given a collection of integration sites, we can align them according to the position of the provirus. . .



INTEGRATION SITE 1 ...ATCCCGTGAC...CGTCTTA...

INTEGRATION SITE 2 ...TTAGAGTGAC...CGTGGTA...

INTEGRATION SITE 3 ...AACGAATGAC...CGTCTTC...

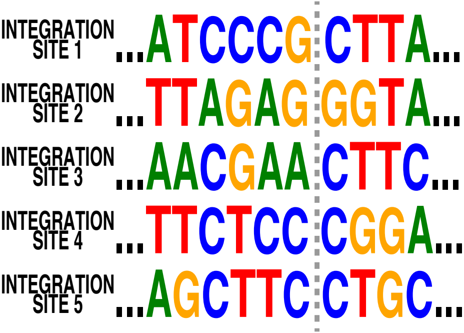
INTEGRATION SITE 4 ...TTCTCCTGAC...CGTCGGA...

INTEGRATION SITE 5 ...AGCTTCTGAC...CGTCTGC...

# Aligning integration sites

Given a collection of integration sites, we can align them according to the position of the provirus. . .

. . . and then ignore/remove/mask the provirus sequence, so that we just look at the target sites:



INTEGRATION SITE 1 ...ATCCCG CTTA...

INTEGRATION SITE 2 ...TTAGAG GGTA...

INTEGRATION SITE 3 ...AACGAA CTTC...

INTEGRATION SITE 4 ...TTCTCC CGGA...

INTEGRATION SITE 5 ...AGCTTC CTGC...

# Summarising a collection of target sites

Sequences

**Example (5 sequences)**

```
...ATC...
...TTA...
...AAC...
...TTC...
...AGC...
```

## Consensus sequence

Just take the most frequent letter at each position: `...ATC...`

## Position probability matrix (PPM), $P$

Estimate the probability of each letter at each position:

$$P = \begin{array}{c} A \\ T \\ C \\ G \end{array} \begin{pmatrix} \dots & 3/5 & 1/5 & 1/5 & \dots \\ \dots & 2/5 & 3/5 & 0 & \dots \\ \dots & 0 & 0 & 4/5 & \dots \\ \dots & 0 & 1/5 & 0 & \dots \end{pmatrix}$$

# Summarising a collection of target sites

| | Sequences | Complements | Reverse complements |
|---|---|---|---|
| **Example (5 sequences)** | ...ATC... | ...TAG... | ...GAT... |
| | ...TTA... | ...AAT... | ...TAA... |
| | ...AAC... | ...TTG... | ...GTT... |
| | ...TTC... | ...AAG... | ...GAA... |
| | ...AGC... | ...TCG... | ...GCT... |

## Reverse complement PPM, $P^{(RC)}$

The PPM for the reverse complement sequences:

$$P^{(RC)} = \begin{array}{c} A \\ T \\ C \\ G \end{array} \left( \begin{array}{cccccc} \dots & 0 & 3/5 & 2/5 & \dots & \\ \dots & 1/5 & 1/5 & 3/5 & \dots & \\ \dots & 0 & 1/5 & 0 & \dots & \\ \dots & 4/5 & 0 & 0 & \dots & \end{array} \right)$$

Note: we can get $P^{(RC)}$ from $P$ (and vice versa) by swapping the rows A $\leftrightarrow$ T and C $\leftrightarrow$ G, and reversing the order of the columns.

# Palindromic consensus sequences for HTLV-1 and HIV-1 target integration sites

From 4,521 HTLV-1 target integration sites, we find the consensus:

**AAGTGGATATCCACTT**

From 13,442 HIV-1 target integration sites, we find the consensus:

**TTTGGTAACCAAA**

# Palindromic consensus sequences for HTLV-1 and HIV-1 target integration sites

From 4,521 HTLV-1 target integration sites, we find the consensus:

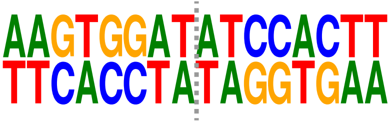
AAGTGGATATCCACTT

From 13,442 HIV-1 target integration sites, we find the consensus:

TTTGGTAACCAAA

# Palindromic consensus sequences for HTLV-1 and HIV-1 target integration sites

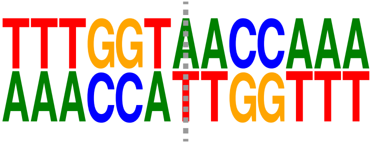
From 4,521 HTLV-1 target integration sites, we find the consensus:



AAGTGGATATCCACTT

TTCACCTATAGGTGAA

From 13,442 HIV-1 target integration sites, we find the consensus:



TTTGGTAACCAAA

AAACCATTGGTTT

# Palindromic consensus sequences for HTLV-1 and HIV-1 target integration sites

From 4,521 HTLV-1 target integration sites, we find the consensus:

AAGTGGATATCCACTT
TTCACCTATAGGTGAA

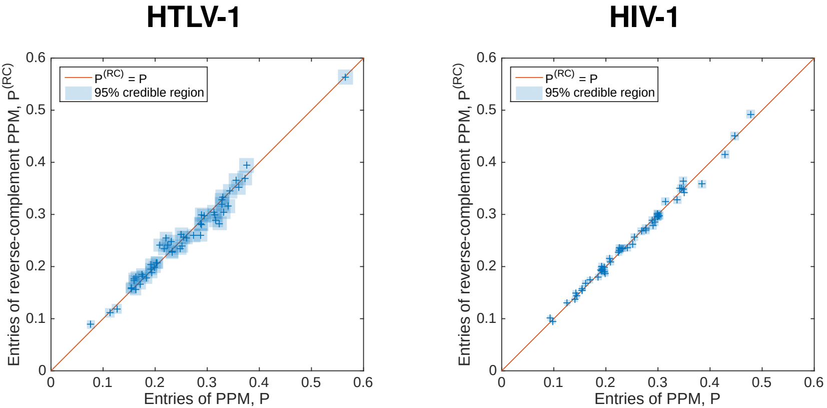
From 13,442 HIV-1 target integration sites, we find the consensus:

TTTGGTAACCAAA
AAACCATTGGTTT

The target integration sites are palindromic (as already known!)

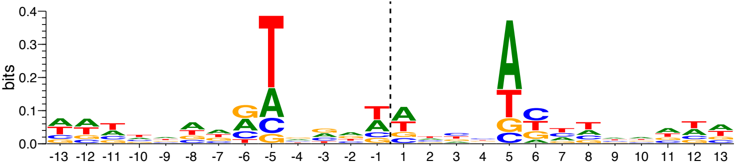
# Palindromic PPMs for HTLV-1 and HIV-1 target integration sites

For both HTLV-1 and HIV-1, we have $P^{(RC)} \approx P$

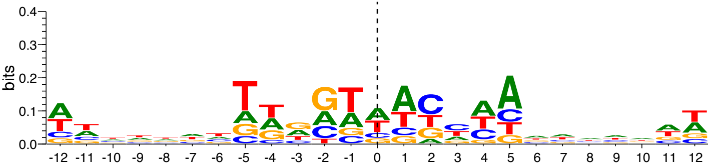


# Palindromic sequence logos

**HTLV-1:**



**HIV-1:**



# An attack of aibohphobia

- There is an almost unbelievable amount of symmetry (!)

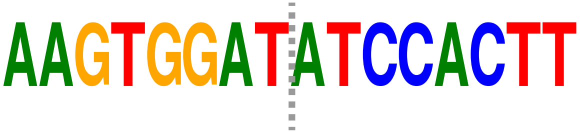
# An attack of aibohphobia

- There is an almost unbelievable amount of symmetry (!)
- Is this “real”? Do we see evidence of the symmetry within individual sequences, or just at the level of these summaries?

# An attack of aibohphobia

- There is an almost unbelievable amount of symmetry (!)
- Is this "real"? Do we see evidence of the symmetry within individual sequences, or just at the level of these summaries?
- **We introduce a palindrome index to quantify "how palindromic" each sequence is**

# The palindrome index



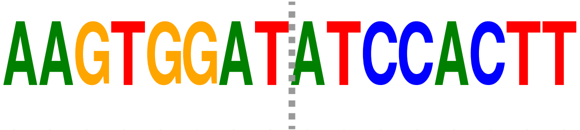
AAGTGGATATCCACTT

# The palindrome index

AAGTGGAT ATCCACTT

$$\mathbf{S} = s_{-8}\, s_{-7}\, s_{-6}\, s_{-5}\, s_{-4}\, s_{-3}\, s_{-2}\, s_{-1}\, s_{1}\, s_{2}\, s_{3}\, s_{4}\, s_{5}\, s_{6}\, s_{7}\, s_{8}$$

# The palindrome index



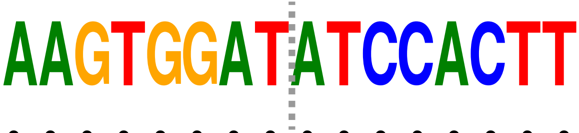
$$\mathbf{S} = s_{-8}\, s_{-7}\, s_{-6}\, s_{-5}\, s_{-4}\, s_{-3}\, s_{-2}\, s_{-1}\, s_{1}\, s_{2}\, s_{3}\, s_{4}\, s_{5}\, s_{6}\, s_{7}\, s_{8}$$

Define

$$\rho(\mathbf{S}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(s_i = c(s_{-i})),$$

where $2n$ is the sequence length, $\mathbb{I}$ is the indicator function, and $c(x)$ is the complement of $x$ (e.g. $c(T) = A$).

# The palindrome index


AAGTGGATATCCACTT

$$\mathbf{S} = s_{-8}\ s_{-7}\ s_{-6}\ s_{-5}\ s_{-4}\ s_{-3}\ s_{-2}\ s_{-1}\ s_{1}\ s_{2}\ s_{3}\ s_{4}\ s_{5}\ s_{6}\ s_{7}\ s_{8}$$

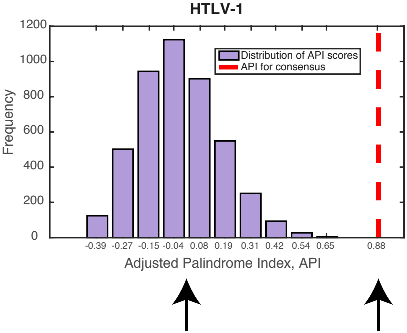
Define

$$\rho(\mathbf{S}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(s_i = c(s_{-i})),$$

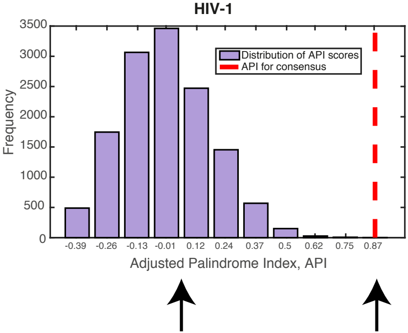
where $2n$ is the sequence length, $\mathbb{I}$ is the indicator function, and $c(x)$ is the complement of $x$ (e.g. $c(T) = A$).

(In practice, we use an “adjusted for chance” version, which is maximally 1, and is 0 if **S** is no more palindromic than expected by chance.)

# Observed palindrome indices



0 0.88



0 0.87

# Where do the palindromes come from?

- The individual sequences are not palindromic

# Where do the palindromes come from?

- The individual sequences are not palindromic

- So why do we see palindromes when we average over a large number of sequences?

# Where do the palindromes come from?

- One possible explanation is that we have a mix of “forward” and “reverse complement” sequence orientations,

# Where do the palindromes come from?

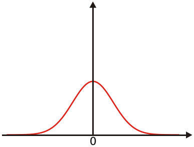
- One possible explanation is that we have a mix of “forward” and “reverse complement” sequence orientations, e.g. in the noiseless case

```
Sequence 1:  AATTTAAGTGGAT (Forward)
Sequence 2:  ATCCACTTAAATT (Reverse complement)
Sequence 3:  ATCCACTTAAATT (Reverse complement)
Sequence 4:  AATTTAAGTGGAT (Forward)
Sequence 5:  ATCCACTTAAATT (Forward)
Sequence 6:  AATTTAAGTGGAT (Reverse complement)
```

$$P = \begin{array}{c} \mathtt{A} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{G} \end{array} \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0.5 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0.5 & 0.5 & 0.5 & 0 & 0 & 0.5 & 1 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0.5 & 0 & 0 \end{pmatrix} = P^{(RC)}$$

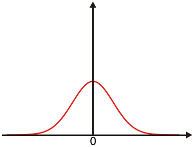
# Analogy

If we have a sample of many real numbers, and we take their mean and find it to be **exactly zero**, one possibility is that this mean is representative of the sample:

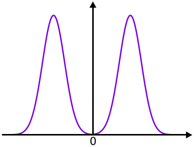


# Analogy

If we have a sample of many real numbers, and we take their mean and find it to be **exactly zero**, one possibility is that this mean is representative of the sample:



Another possibility is that we have 2 symmetric components, one positive and one negative:



# Mixture modelling

- We model the sequences as coming from two populations
  - one with PPM $P$; and
  - one with reverse complement PPM $P^{(RC)}$.

$$\pi(S) = \omega\pi(S|P) + (1-\omega)\pi(S|P^{(RC)}).$$

# Mixture modelling

- We model the sequences as coming from two populations
  - one with PPM $P$; and
  - one with reverse complement PPM $P^{(RC)}$.

$$\pi(S) = \omega\pi(S|P) + (1-\omega)\pi(S|P^{(RC)}).$$

- Here, $\omega$ is the proportion of sequences coming from the population with PPM $P$.

# Mixture modelling

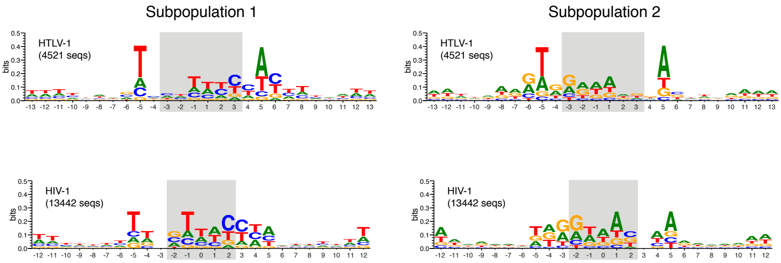
- We model the sequences as coming from two populations
  - one with PPM $P$; and
  - one with reverse complement PPM $P^{(RC)}$.

$$\pi(S) = \omega\pi(S|P) + (1-\omega)\pi(S|P^{(RC)}).$$

- Here, $\omega$ is the proportion of sequences coming from the population with PPM $P$.

- The parameters, $\omega$ and $P$, can be estimated/inferred in numerous ways. I will show results from using an EM-algorithm, but identical results are obtained by: (i) maximum profile likelihood; (ii) Gibbs sampling; (iii) greedy Gibbs.

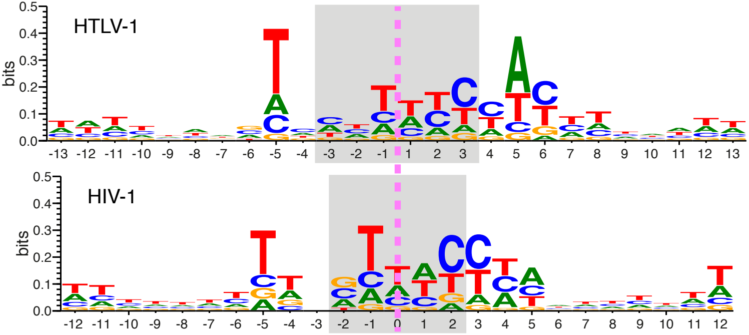
# Unmixing the forward and reverse sequences



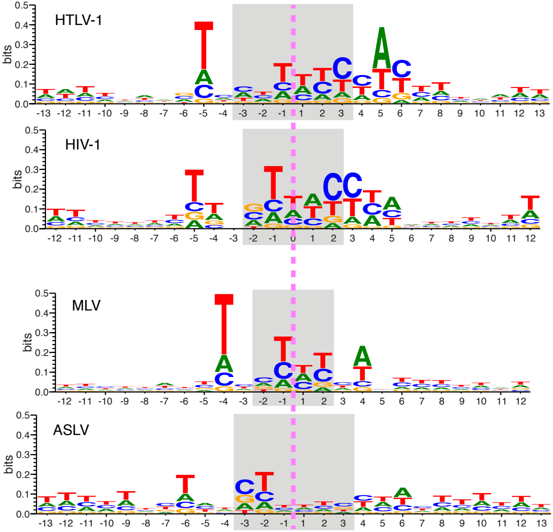
Subpopulation 1

Subpopulation 2

# Unmixing the forward and reverse sequences



# Unmixing the forward and reverse sequences



# Summary

- The palindrome is not observed within individual sequences.

# Summary

- The palindrome is not observed within individual sequences.
- Hypothesis: the palindrome results from a mixture of sequences that contain a non-palindromic motif in approximately equal proportions in “forward” and “reverse complement” orientations

# Summary

- The palindrome is not observed within individual sequences.
- Hypothesis: the palindrome results from a mixture of sequences that contain a non-palindromic motif in approximately equal proportions in “forward” and “reverse complement” orientations
- Modelling this hypothesis revealed a common nucleotide motif across 4 retroviruses:

5’-T(N1/2)[C(N0/1)T|(W1/2)C]CW-3’

# Summary

- The palindrome is not observed within individual sequences.
- Hypothesis: the palindrome results from a mixture of sequences that contain a non-palindromic motif in approximately equal proportions in “forward” and “reverse complement” orientations
- Modelling this hypothesis revealed a common nucleotide motif across 4 retroviruses:

  5’-T(N1/2)[C(N0/1)T|(W1/2)C]CW-3’

- Potential implications for understanding retroviral integration.

# Summary

- The palindrome is not observed within individual sequences.
- Hypothesis: the palindrome results from a mixture of sequences that contain a non-palindromic motif in approximately equal proportions in “forward” and “reverse complement” orientations
- Modelling this hypothesis revealed a common nucleotide motif across 4 retroviruses:

  5’-T(N1/2)[C(N0/1)T|(W1/2)C]CW-3’

- Potential implications for understanding retroviral integration.
- True validation requires further structural information about retroviral intasomes.

# Availability

- Accepted for publication in Nature Microbiology.
- Preprint:
  - Kirk, Huvet, Melamed, Maertens & Bangham (2015). Retroviruses integrate into a shared, non-palindromic motif. bioRxiv.

Matlab code (and the HTLV-1 dataset) are available online:

http://www.mrc-bsu.cam.ac.uk/software/bioinformatics-and-statistical-genomics/

Just click on **retroCode** to download!

# Acknowledgements

Charles Bangham

Maxime Huvet
Anat Melamed
Goedele Maertens

---

Sylvia Richardson
MRC Biostatistics Unit

---

Michael Stumpf
Imperial College Theoretical Systems Biology group

# Thanks for listening!

MRC | Biostatistics Unit

@pauldwkirk

http://www.mrc-bsu.cam.ac.uk/people/paul-kirk/

# SCIENCE SHOWOFF

## THE CHAOTIC SCIENCE COMEDY CABARET

with

**STEVE CROSS**

**SARAH BENNETTO**

and loads of Cambridge
science talent

MONDAY 10th OCTOBER
Portland Arms, Doors 6.30
Tickets £5 from
scienceshowoff.org
or £7 on the door
All ticket money
will go to Parkinson's UK

biology
week
2016

UNIVERSITY OF
CAMBRIDGE

Wellcome Trust - Medical Research Council
Cambridge Stem Cell Institute