DETECTION OF LOW FREQUENCY CXCR4-USING HIV-1 WITH ULTRA-DEEP PYROSEQUENCING

John Archer

Faculty of Life Sciences
University of Manchester

HIV Dynamics and Evolution, 2008, Santa Fe, New Mexico.

Overview

The Roche 454 GS FLX Sequencing platform

HIV co-receptors and HIV phenotypes

Detecting HIV phenotype

Our data

Protocol for managing 454 data

Results

Software

Conclusions

Increase in Sequence Data

Recently, due to the emergence of new sequencing systems, huge amounts of sequence data have become available.

One such system, the Roche 454 GS FLX Sequencer, is a parallel pyrosequencing platform capable of sequencing 25 million bases within a four hour run.

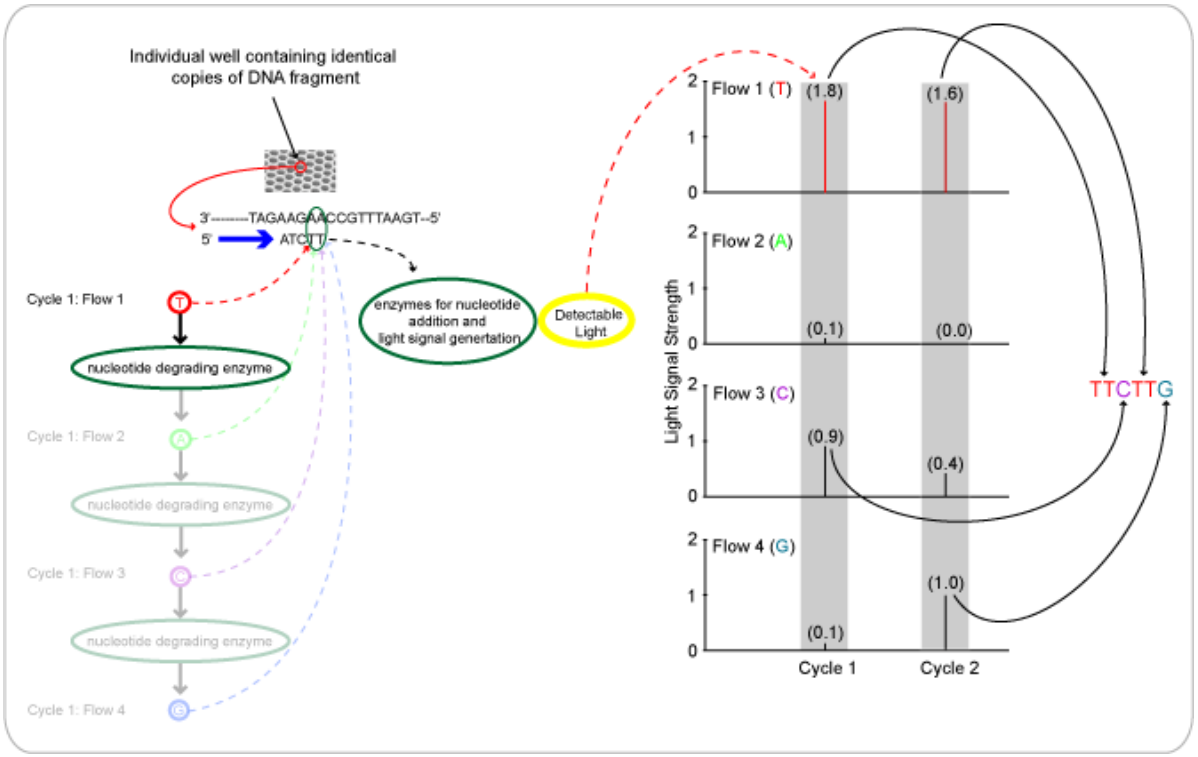
The mean read length of the sequence segments produced is ≈200.

Produces a huge number of reads to process.

For example: NCBI's Trace Archive contains >182 million reads representing just 123 species.

Species	No. Of Bases
Listeria monocytogenes	6620471
Drosophila mauritiana	2569374
Francisella tularensis	1218271
Pseudomonas putida	1047806
Fossil metagenome	590264
Pseudomonas aeruginosa	588691
Staphylococcus aureus	575197
Clostridium difficile	417941

Overview of 454 System



During each cycle bases are added to an individual well in a predetermined order.

If a particular base is incorporated to the growing read a detectable light signal is emitted.

Noise in the System

The light signal produced can become ambiguous potentially resulting in an error.

The estimated overall error rate is approximately 1% of the total number of bases sequenced.

Insertions and deletions are the most common errors as a result of misinterpreting the lengths of the homopolymeric runs due to the light signal being disrupted.

Light signal disruption can be caused by:

- Multiple different templates on a bead.
- Signal contamination for nearby wells.
- Loss of synchrony between the identical templates within a single well.

Noise in the System

In Wang *et al*. 2007 (*Genome Research* **17**:1195-1201) a mean error rate of 0.98% was observed within a dataset consisting of 6,827 reads from HIV-1's RT gene. This was subdivided into:

Insertions - 0.73%
Deletions - 0.16%
Mismatches - 0.12%

The error rate within homopolymeric regions of size three or greater was 6.2 times higher than outside of these region.

8% of the dataset was unusable as it had no sequence identity with the template sequence.

Characterizing low frequency variants

Because of the depth of coverage provided, pryosequencing permits the "ultra-deep" sequencing of a viral population.

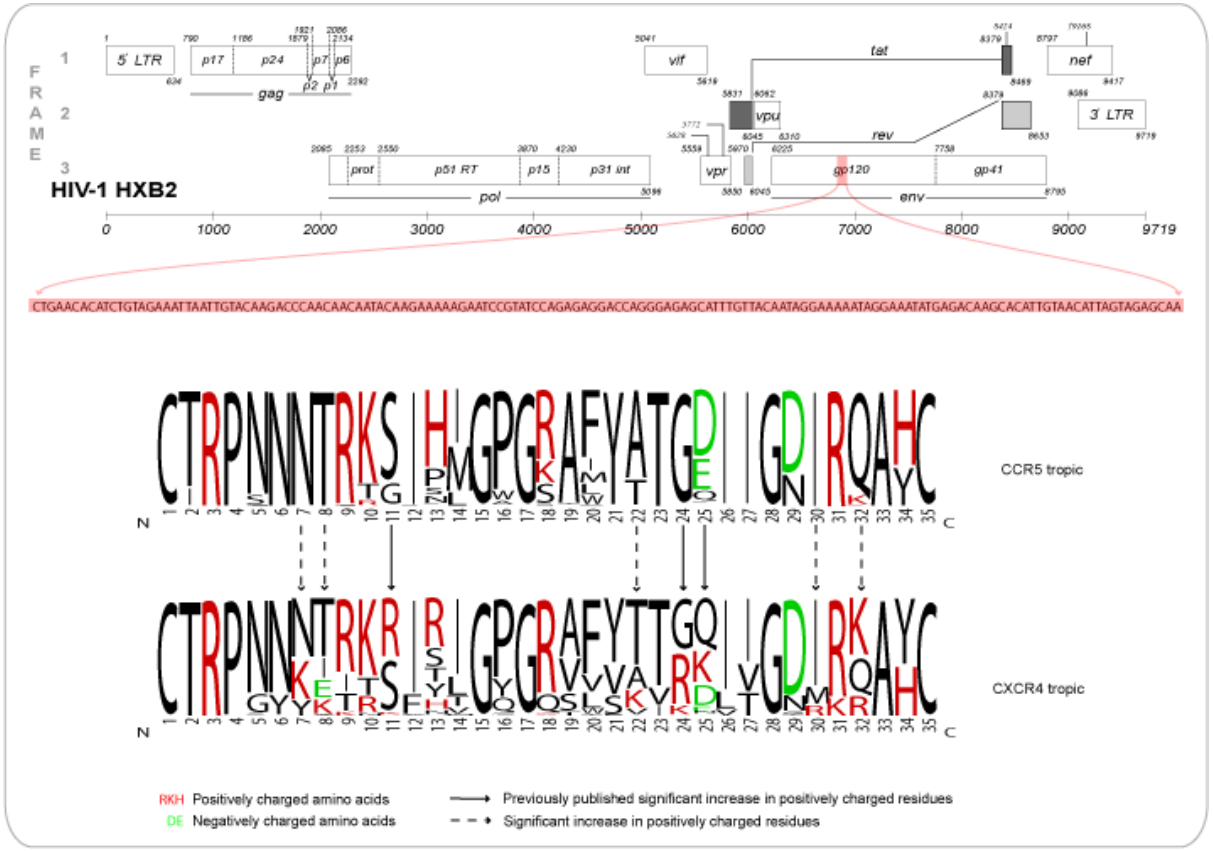
This has the potential to permit the quantification of the full range of sequence variants present within an HIV-1 sample.

This is especially useful when it comes to detecting low frequency variants that may be responsible for drug resistance or drug failure.

HIV Phenotypes

Two HIV phenotypes exist that are dependent on coreceptor usage:

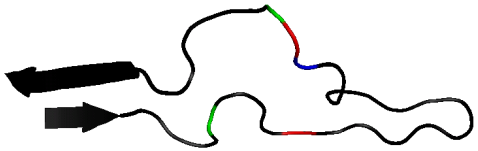
CCR5 using strains -> infect macrophages.
CXCR4 using strains -> infect T-Cells.



Phenotypes Detection

Amino acid sequence variations giving rise to a more positive charge within the V3 loop at sites 11, 24 and 25 have been strongly linked with the more virulent X4 phenotype (Cardozo *et al*. 2007 and Rosen *et al*. 2006).

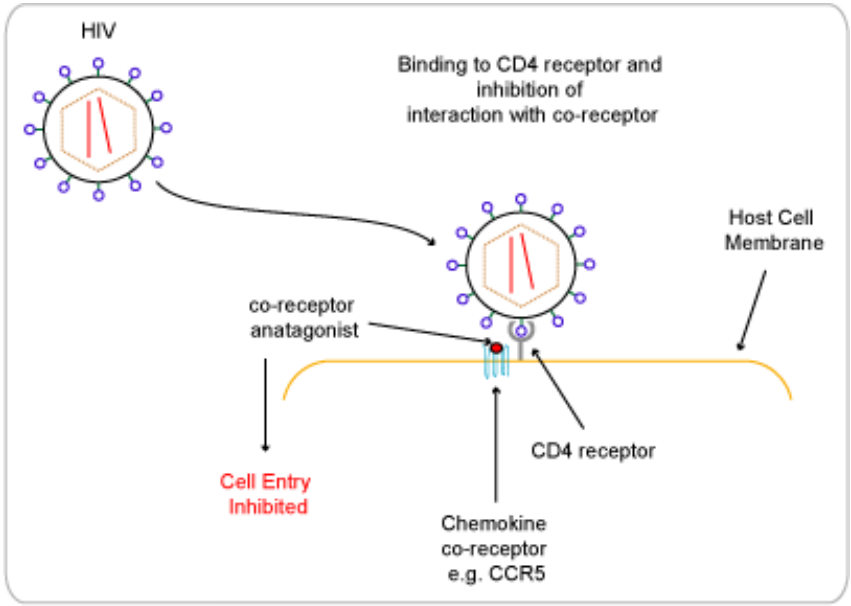
A single positive charge at any of these sites results in the usage of the CXCR4 coreceptor. This well characterized observation is referred to as the "charge rule".



Other less well understood interactions within the V3 loop may also be involved in determining phenotype.

Web PSSM and geno2pheno are genotyping algorithms that attempt to account for these.

Inhibition of Cell Entry



Co-receptor blocking is a novel way of attempting to control the progression of HIV within the host.

Targeting the CCR5 co-receptor is seen to be a viable option within humans as a natural polymorphism exists (CCR5Δ32) with few deleterious effects.

Individuals that are heterozygous for this mutation were found to have a slower disease progression to AIDS .

Individuals that are homozygotic for the mutation show strong resistance to HIV-1.

Inhibition of Cell Entry

Recently Pfizer has developed a small-molecule drug, Maraviroc, that binds to the CCR5 receptor making it unavailable for HIV-1 cell entry.

Treatment with Maraviroc effectively prevents use of the CCR5 co-receptor.

It is the only CCR5 antagonist that is in phase 3 trials at the moment???

However:
In drug-failure patients the presence of CXCR4-using virus prior to treatment is predictive of failure (Westby *et al*., 2007; *Journal of Virology* **80**:4909-4920).

Aims

To develop a protocol to allow for the rapid analysis of pyrosequenced data.

To detect the presence of low frequency CXCR4-using virus within a patient that had previously been screened for the absence of this phenotype using pyrosequencing data.

To incorporate the protocol into a user friendly software.

Data

Two pyrosequencing datasets from patient A (Westby *et al*., 2007; *Journal of Virology* **80**:4909-4920) were generated using a Roche 454 GS FLX Sequencer.

The first dataset, Day 1, corresponded to a sample taken pre-treatment with Maraviroc.

The second dataset, Day 11, corresponded to a sample taken 11 days post Maraviroc treatment (after failure due to emergence of CXCR4-using virus).

The datasets were amplified form the *gp160* region of the HIV-1 genome.

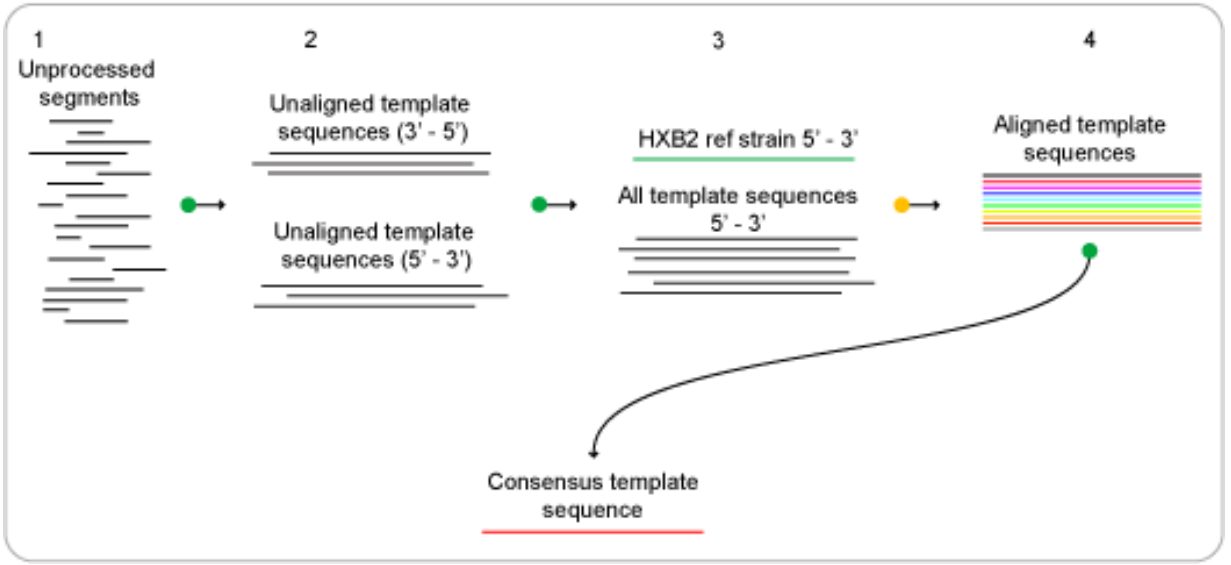
For day 1 and day 11 104,628 and 191,637 nucleotide sequence segments were generated, respectively.

11 clones from day 1 and 12 from day 11 were also available (Westby *et al*., 2007).

Methods

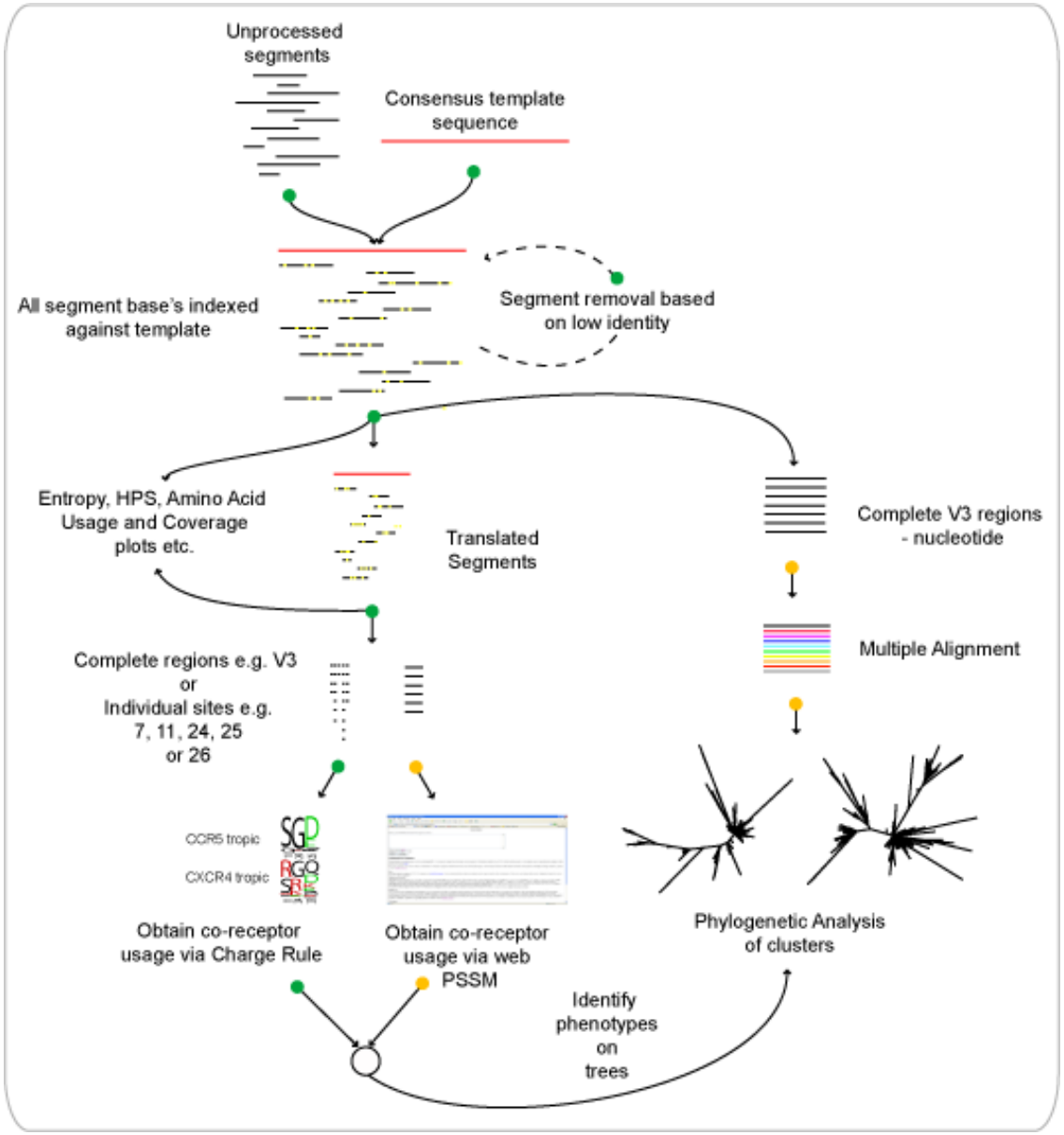
A template sequence was required in order to obtain the locations of individual segments within the *gp160* gene.

The Roche 454 GS FLX Sequencer does not provide such a template across the region being sequenced.

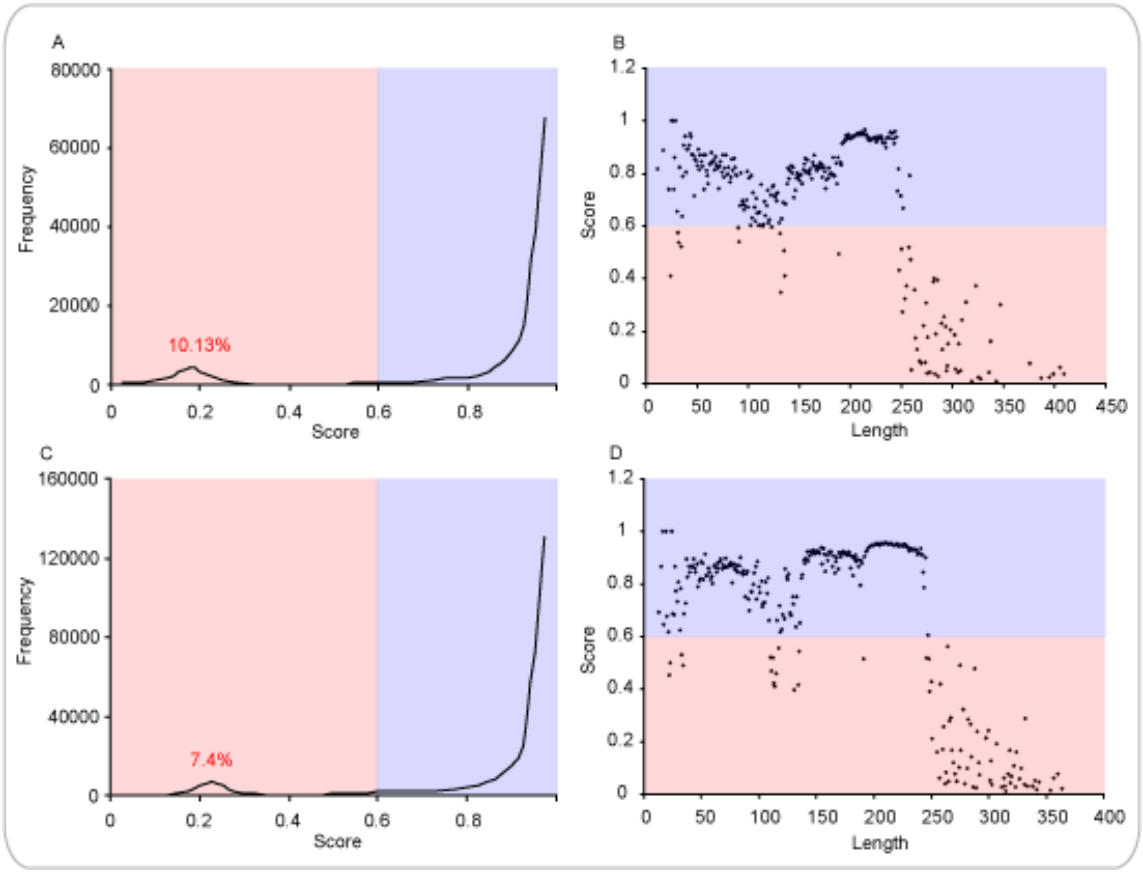


Constructing the template sequence from the read data minimizes the phylogenetic divergence from the dataset.

Methods


Unprocessed segments
Consensus template sequence
All segment base's indexed against template
Segment removal based on low identity
Entropy, HPS, Amino Acid Usage and Coverage plots etc.
Translated Segments
Complete V3 regions - nucleotide
Multiple Alignment
Complete regions e.g. V3 or Individual sites e.g. 7, 11, 24, 25 or 26
CCR5 tropic
CXCR4 tropic
Obtain co-receptor usage via Charge Rule
Obtain co-receptor usage via web PSSM
Phylogenetic Analysis of clusters
Identify phenotypes on trees

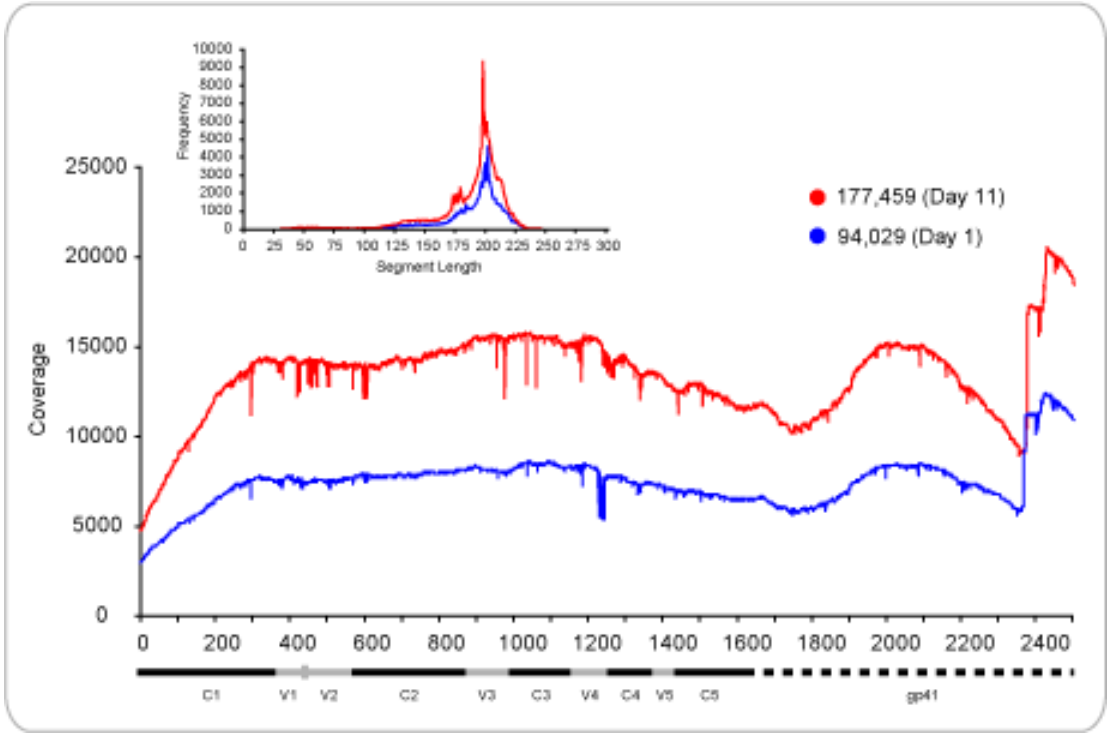
Results



Low identity reads, <0.6, were excluded from further analysis leaving 94,029 and 177,459 reads.

The majority of reads within these lower peaks are greater than 250 bases in length.

Results



A good depth of coverage was provided across the gp160 gene within both datasets.

The primary area that we were interested in was V3.

Results

V3 data summary:

Day 1:
3,384 complete V3 reads - 621 unique
5,986 reads with complete sites 11, 24 and 25 present
5,005 reads with completed sites 7, 8, 11, 24 and 25

Day 11:
6,687 complete V3 reads - 1155 unique
12,086 reads with complete sites 11, 24 and 25 present
9,941 reads with completed sites 7, 8, 11, 24 and 25

Results

A - Day 1

	CCR5		CXCR4	
	Total	% of data	Total	% of data
Charge rule	5968	99.69	18	0.31
Charge rule (plus extra sites)	4985	99.60	20	0.4
Geno2pheno (Complete V3's)				
PSSM (Complete V3's)	3377	99.79	7	0.21
Charge rule (Complete V3's)	3373	99.67	11	0.32

B - Day 11

	CCR5		CXCR4	
	Total	% of data	Total	% of data
Charge rule	2306	19.08	9780	80.92
Charge rule (plus extra sites)	1886	18.97	8055	81.03
Geno2pheno (Complete V3's)				
PSSM (Complete V3's)	1304	19.5	5383	80.5
Charge rule (Complete V3's)	1264	18.56	5423	81.09

For the charge rule less than 4 reads within the complete day 1 V3 sequences (3,384) tested using the charge rule could be expected to be falsely predicted as CXCR4-using.

For the complete and partial sequences (5,986) <7 reads expected to be falsely predicted.

Results

A - Day 1

	CXCR4	
	Total	% of data
Charge rule	25	0.28
Charge rule (plus extra sites)	33	0.32

B - Day 11

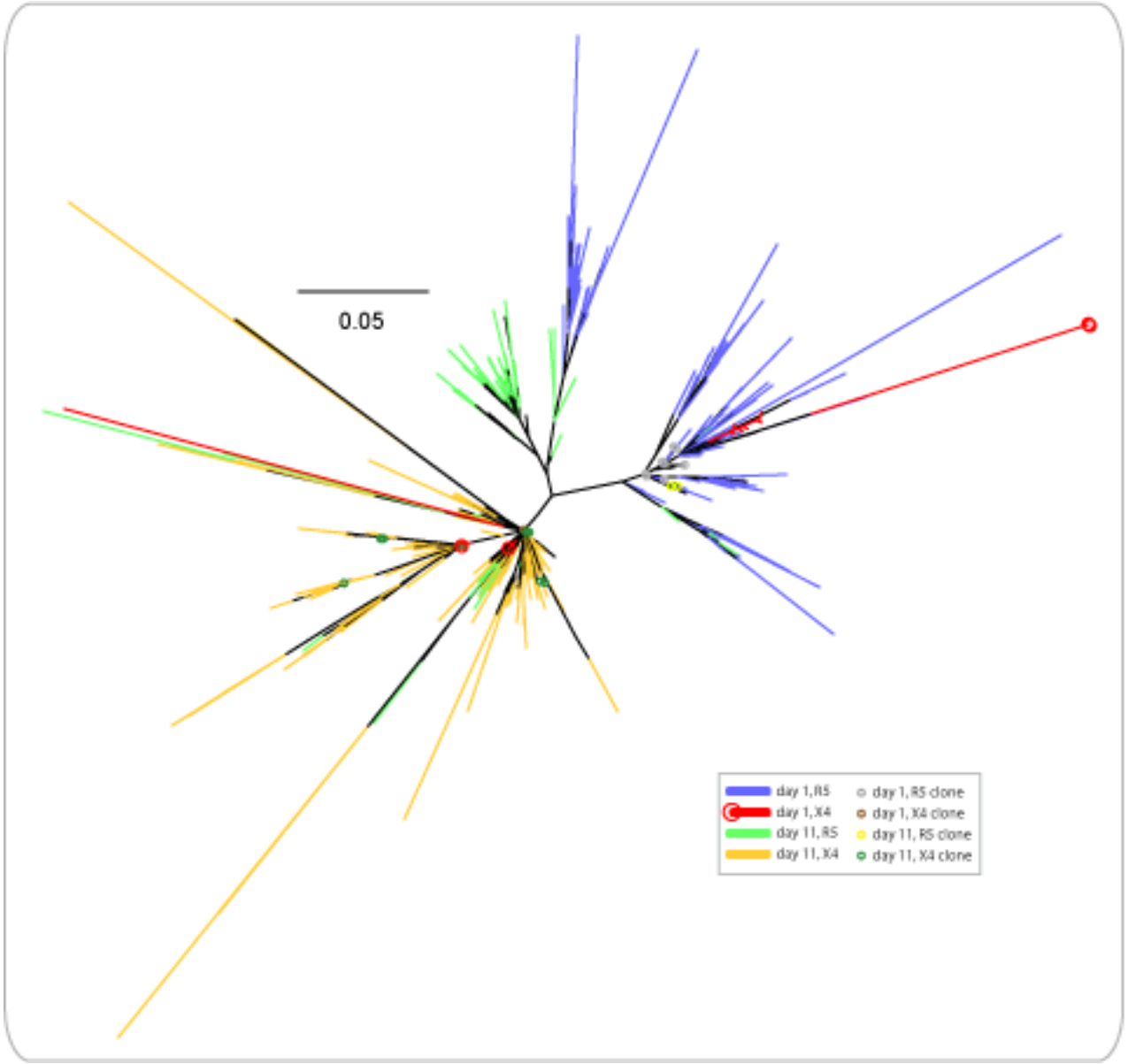
	CXCR4	
	Total	% of data
Charge rule	12147	67.85
Charge rule (plus extra sites)	12159	60.76

However not all 3 sites are required for the detection of X4 so more are possible.

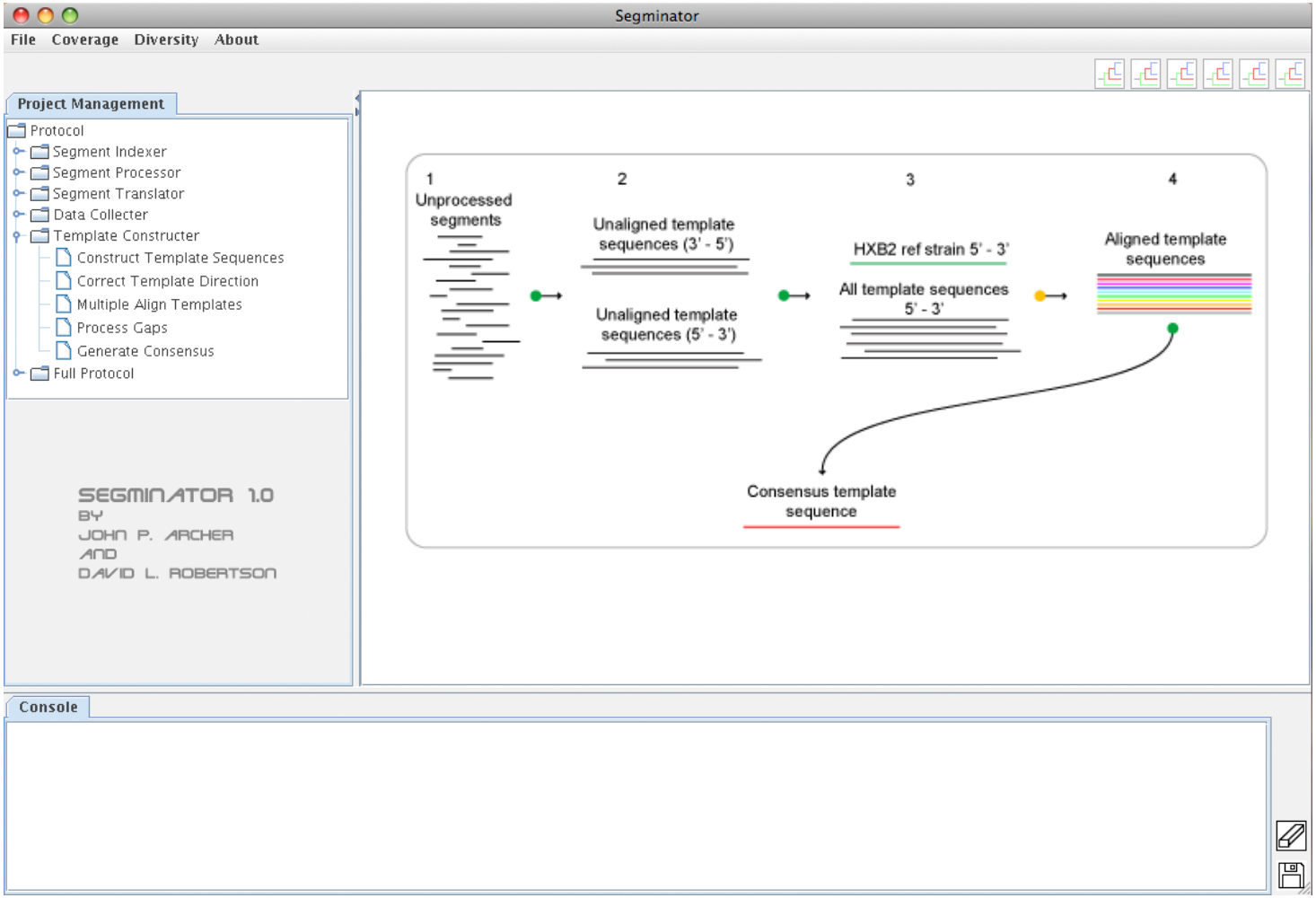
The additional positively charged sites 7, 8, 22 and 30 do not appear have a major influence on increasing the number of CXCR4-viruses detected.

With the additional sites the potential error rate rises to 32.

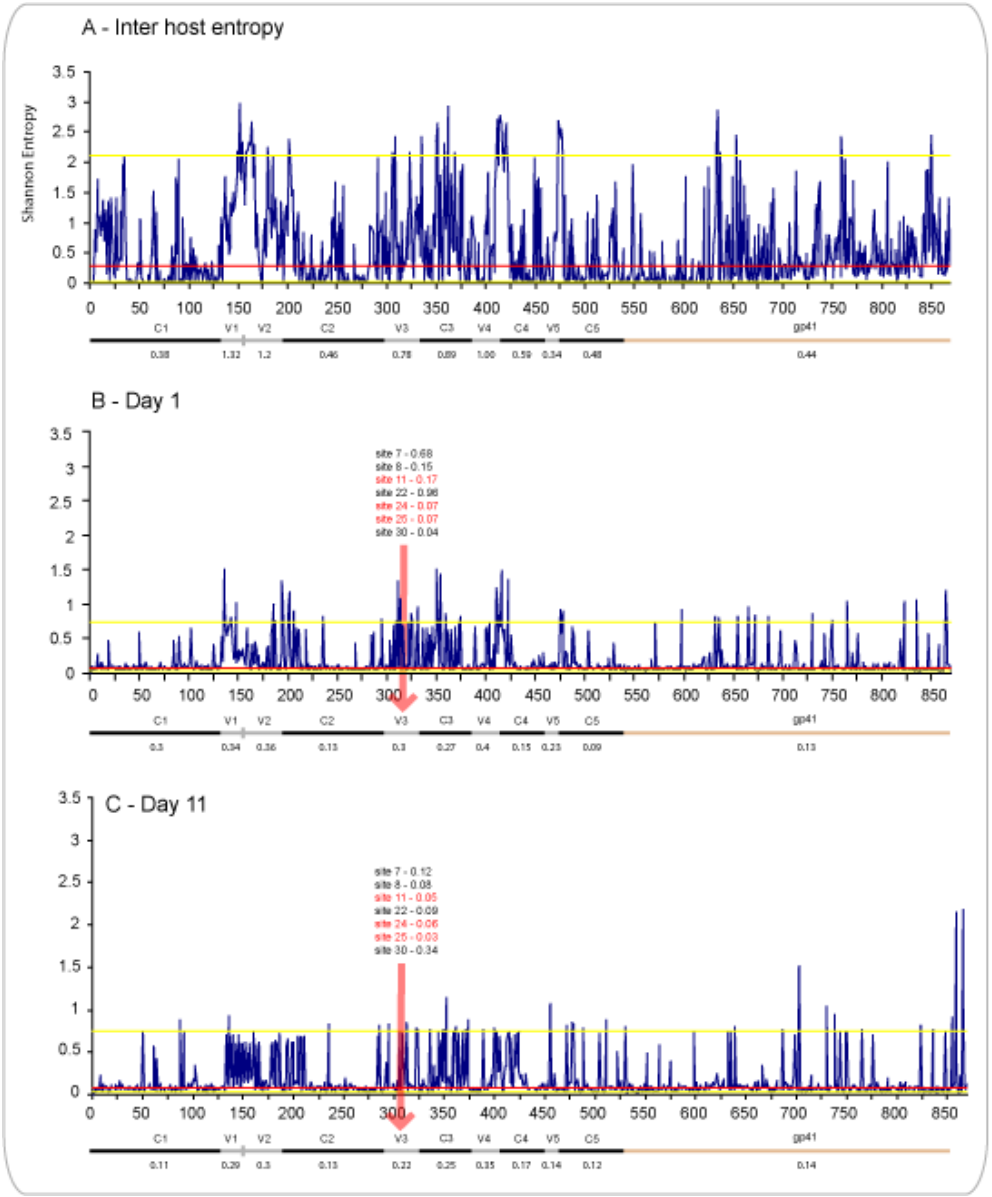
Results


0.05
day 1, R5
day 1, X4
day 11, R5
day 11, X4
day 1, R5 clone
day 1, X4 clone
day 11, R5 clone
day 11, X4 clone

The software: Segminator



Results


A - Inter host entropy
Shannon Entropy
B - Day 1
site 7 - 0.68
site 8 - 0.15
site 11 - 0.17
site 22 - 0.98
site 24 - 0.07
site 25 - 0.07
site 30 - 0.04
C - Day 11
site 7 - 0.12
site 8 - 0.08
site 11 - 0.05
site 22 - 0.09
site 24 - 0.06
site 25 - 0.03
site 30 - 0.34
C1 V1 V2 C2 V3 C3 V4 C4 V5 C5 gp41

Results

	Day 1		Day 11	
Insertion/Gap Size	Insertions	Deletions	Insertions	Deletions
1	152652	84731	383587	230072
2	3847	3009	6060	3897
3	243	36	11694	9089
4	50	1	180	12
5	25	0	65	1
6	17	0	55	0
7	11	0	10	0
8	14	0	9	0
9	17	0	13	0
10	5	0	6	0
11	2	0	2	0
12	4	0	2	0
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
Total	156887	87777	401683	243071
% of Data	0.8549	0.4796	1.2107	0.7423
% forming complete codons	0.0054	0.0006	0.0995	0.0763
% other	0.8495	0.4790	1.1112	0.6659

Large numbers of single base pair insertions and deletions are present in the pyrosequencing data.

The insertions are removed during alignment to the reference sequence.

The codons with one or two base pair deletions are removed at the translation step.

Real insertions and deletions will also be removed during this process thus a small proportion (0.00597 and 0.1758%) of the data is potentially lost.

Conclusion

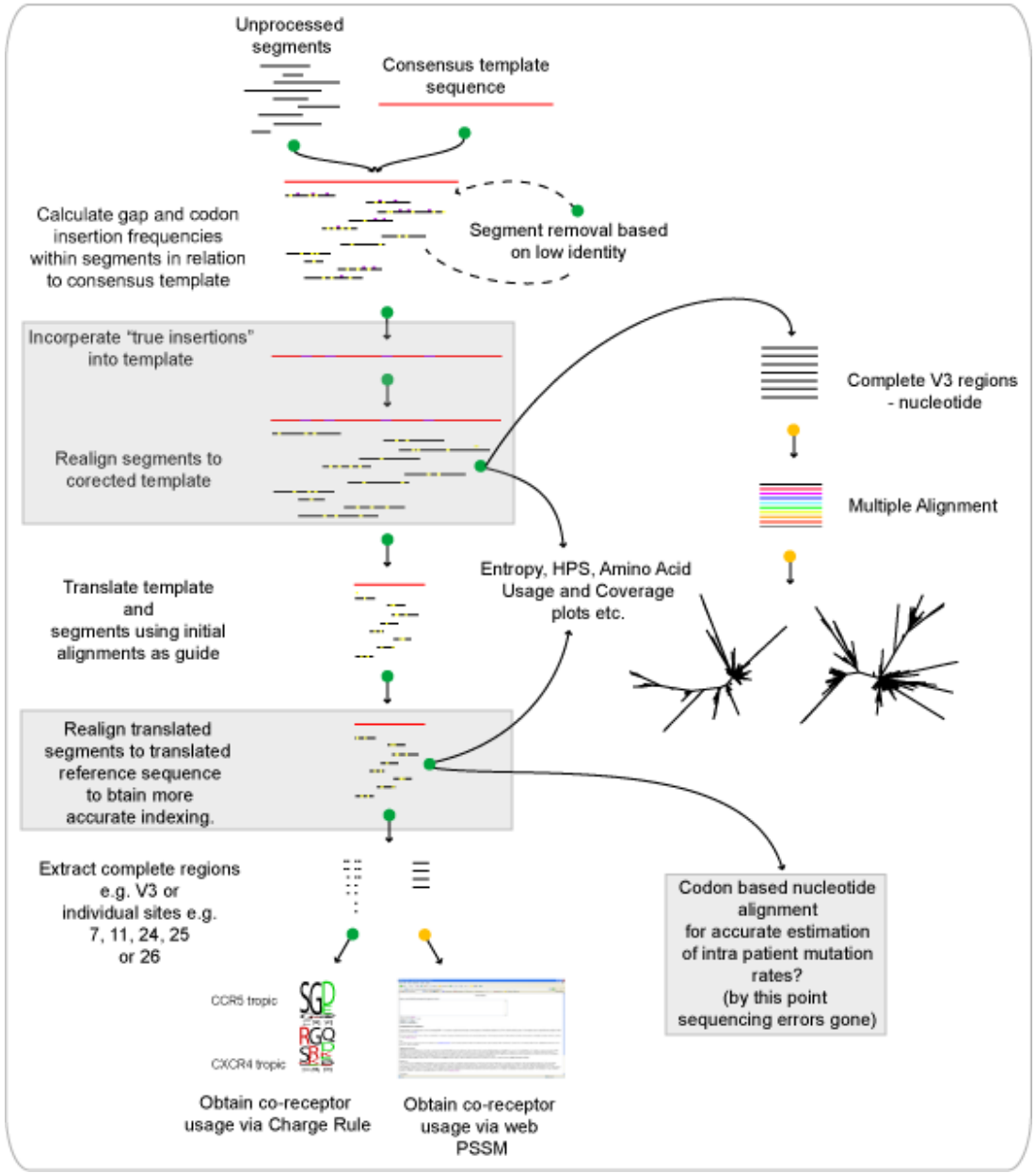
CXCR4-using strains were present within the day 1 dataset at low numbers

Pyrosequencing data can be used to detect these low frequency variants.

Analysis supports the previous finding (Westby *et al*., 2007; *Journal of Virology* **80**:4909-4920) that the day 11 CXCR4-using virus cluster emerged from pre-existing CXCR4-using strains present at day 1.

Segminator is a useful diagnostic software for genotypic testing as well as a general software for the analysis of pyrosequenced data.

Future Work


Unprocessed segments
Consensus template sequence
Calculate gap and codon insertion frequencies within segments in relation to consensus template
Segment removal based on low identity
Incorperate "true insertions" into template
Realign segments to corected template
Complete V3 regions - nucleotide
Multiple Alignment
Translate template and segments using initial alignments as guide
Entropy, HPS, Amino Acid Usage and Coverage plots etc.
Realign translated segments to translated reference sequence to btain more accurate indexing.
Extract complete regions e.g. V3 or individual sites e.g. 7, 11, 24, 25 or 26
Codon based nucleotide alignment for accurate estimation of intra patient mutation rates? (by this point sequencing errors gone)
CCR5 tropic
CXCR4 tropic
Obtain co-receptor usage via Charge Rule
Obtain co-receptor usage via web PSSM

Acknowledgements

David Robertson, the University of Manchester.

Marilyn Lewis, Pfizer Global R&D, Sandwich.

454 Life Sciences.

Investigators, study-site staff, the Pfizer Maraviroc development team and the patients who participated in the Maraviroc studies.

Alex Thielen for help with geno2pheno.

The BBSRC and Pfizer Global R&D for funding.