INFERRING DIRECTION OF HIV TRANSMISSION USING NEXT GENERATION SEQUENCING

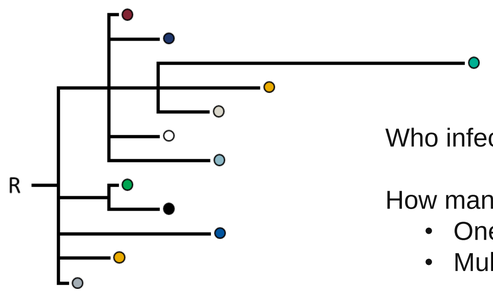
Bibby, DF. El Bouzidi, K. Mulka, L. Roy, S. Breuer, J. Mbisa JL. on behalf of COMPARE-HIV Study

Public Health England

Protecting and improving the nation's health

INTRODUCTION

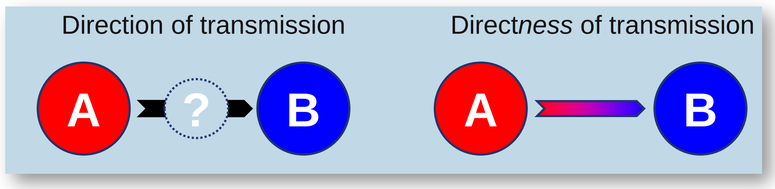
- Phylogenetic analysis of large sequence datasets has proven extremely valuable
- Yet within clusters of related sequences, information on transmission remains limited:



Who infected whom?

How many transmitters?
- One-to-many
- Multiple serial infections

- Romero-Severson *et al.*[1] have shown how the relationship between pairs of linked samples can be inferred from phyletic analysis of large numbers of clone sequences



CAN THE GENOME-LEVEL DATA FROM NEXT GENERATION SEQUENCING (NGS) PROVIDE SIMILAR INSIGHTS?

METHODS

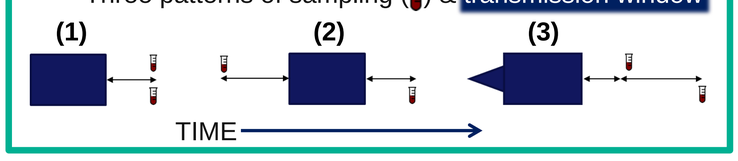
LINKED PAIRS COHORT

- Samples from 10x known, self-declared transmission pairs
  - Epidemiological, clinical and virological evidence
  - Diverse transmission routes and subtypes:

Transmission	
MSM	x3
HET	x5
MtCT	x2

Subtype	
B	x5
C	x2
G, 01, 02	x1 each

- Three patterns of sampling (🩸) & **transmission window**



WHOLE GENOME SEQUENCING

- RNA extracted from stored plasma samples
- SureSelect HIV-specific hybridization probes
- 2x MiSeq sequencing runs
  - Sample pairs split between runs
  - V3 chemistry (251nt paired-end reads)
  - x12 multiplexing

HAPLOTYPE ANALYSIS

1. Derive 'joint genome'
   - Combine FASTQs from paired samples
   - Trim, map to multi-reference file (BWA MEM, LANL)
   - Obtain consensus using QuasiBAM[2]
2. Map individual FASTQs to joint genome
   - Parse SAM files to obtain haplotypes
3. Bin haplotypes by tile (**Fig. 1**)
   - Minimum 2 occurrences per haplotype

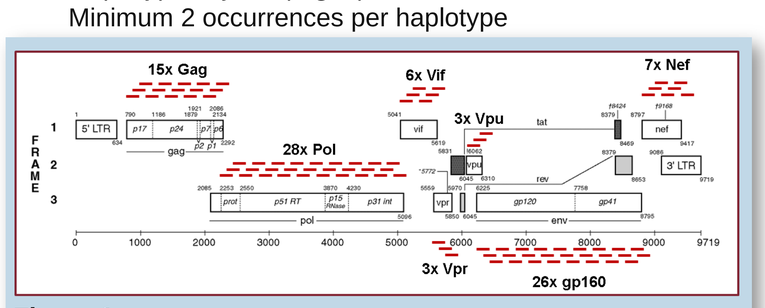


Figure 1. 88x 204nt tiles overlapping by 102nt, from HXB2 genome. HMMER used to derive corresponding joint genome co-ordinates

4. Select the 10 most populous tiles
   - Incorporate close references for rooting tree
   - Align haplotypes (Clustalw2)
   - Generate tree (FastTree)
   - Collapse nodes with <0.7 SH branch support
   - Examine tree topology for suggested transmission pattern
5. Collate patterns to determine direction and/or directness

RESULTS

NGS AND HAPLOTYPE BINNING

Per sample:
- ~400,000 reads (range 225,000 – 685,000)
- Median >100 haplotypes per tile (**Fig. 2**)

Abundances evenly distributed across genome

PHYLOGENETIC ANALYSIS (Fig. 3)

Diverse topologies obtained, suggesting

DIRECTION (a) & (d)
and / or
DIRECTNESS (b) & (d)

A dual monophyletic pattern (c) was predominant in one sample pair, precluding inference of either directionality or directness

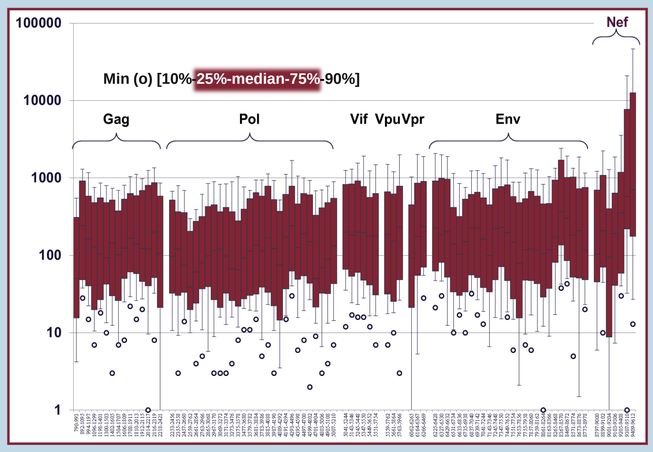


Figure 2. Haplotype counts by tile across 20 FASTQ sets (2x10 linked pairs)

Pair	A → B		A ↔ B	B → A		A ? B	Transmission & sampling pattern	Expected result
1			2		8		MSM 1	
2			3	6		1	HET 2	
3	10						HET 1	
4	1	6	1	1		1	HET 3	?
5				5	5		HET 3	?
6	3		5		2		MSM 2	
7	3		6	1			MSM 2	
8	4					6	HET 3	?
9			10				MtCT 1	
10	6	3	1				MtCT 1	

Table 1. Summary of tree topologies from 10 most populous tiles for the ten linked pairs. Colours describe the relationship with known / unknown transmission histories: Consistent / Inconsistent / Suggestive

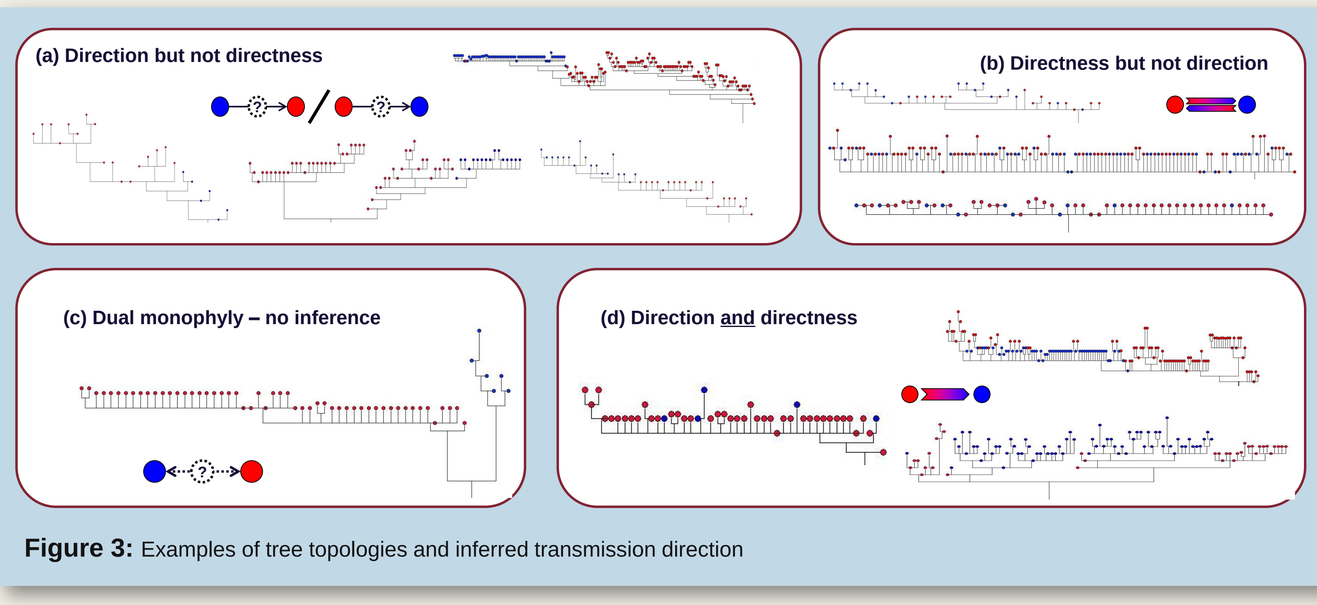


Figure 3: Examples of tree topologies and inferred transmission direction

TREE-DERIVED TRANSMISSION INFERENCES (Table 1)

- Seven of ten linked pairs have known or strongly-suspected transmission histories:
  - ✓ Pairs 1 & 10 have tree patterns consistent with **directness** as well as **direction**
  - ✓ Pairs 3, 6 & 7 are consistent with the **direction** of their associated tree patterns
  - = Pair 9 is *equivocal* : MtCT, all 10 trees dually polyphyletic
  - × Pair. 2 strongly contradicts the predicted direction

    5 years between samples, or possible sample switch?
- Three lack an established transmission history:
  - ✓ Pairs 4 & 5 have patterns strongly suggesting **directness** as well as **direction**
  - ✓ Pair 8 shows a pattern suggesting **direction** but not directness

    Both samples taken >8 years after the most recent possible transmission date

DISCUSSION

Despite often lengthy intervals between samples and/or transmission windows, directionality of transmission was supported by generalised phyletic inference, with directness suggested in several cases.

Inference of direction and directness limited by several factors
- Sequencing depth of least frequent / diverse sample
- Timings of sampling and transmission
- Diversity within tiles – 204nt is relatively short
- FastTree does not deal well with gaps in haplotypes

Lack of an evolutionary model may compromise accuracy of transmission inferences, with dually polyphyletic trees being otherwise uninterpretable

CONCLUSIONS

- **Whole genome sequencing by NGS is a powerful tool to obtain quasispecies information about an HIV sample**
- **Comparison of haplotypes between paired samples allows inference of transmission direction and directness**
- **Current sample set presents challenges to interpretation, owing to temporal separation of sampling times and transmission windows**
- **Next steps to enhance signal**
  1. **Expand tile range to scan entire genome (almost complete)**
  2. **Develop metrics to select best tiles, i.e. most informative**
  3. **Incorporate coalescence-based evolutionary model to improve tree definition**
  4. **Test 'related unlinked' controls from clustered samples**

ACKNOWLEDGEMENTS

We are grateful to the British HIV Association and Public Health England for funding

We would also like to thank Brighton and Sussex University Hospitals NHS Trust for sponsorship of the COMPARE-HIV Study (Comparison of Molecular & Phylogenetic Approaches to Reconstruct an Epidemic of HIV).

Sequencing was performed at the Antiviral Unit, Public Health England and the UCL Pathogen Genomics Unit.

REFERENCES

1. Romero-Severson E *et al.*, (2016) PNAS 113(10): 2690-5.
2. Penedos A *et al.*, (2015) PLoS One 10(11): e0143081.

British HIV Association BHIVA

Brighton and Sussex University Hospitals NHS Trust

PGU Pathogen Genomics Unit