ROSLIN

What can we learn from Whole-Genome-Sequencing?

Uncovering transmission patterns using sequence data and phylodynamics

Dr Samantha Lycett
4th September 2015
Bovine Tuberculosis Workshop 2015, Glasgow

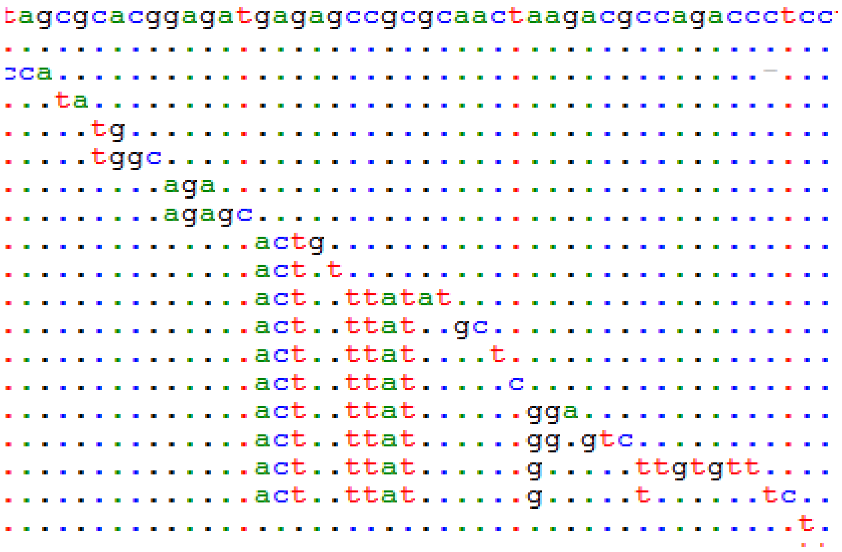
THE UNIVERSITY *of* EDINBURGH

BBSRC
20 Years of Pioneering
Great British Bioscience

Whole Genome Sequencing

- What phenotype/properties does the pathogen have ?
  - Host range and transmissibility ? Drug resistance ?

- Where did it come from ?
  - Which host species
  - Which locations ?
  - Who infected whom ?
  - Co-infection and mixing ?



```
tagcgcacggagatgagagccgcgcaactaagacgccagaccctcc
..............................................
cca...........................................
...ta.........................................
.....tg.......................................
.....tggc.....................................
.........aga..................................
.........agagc................................
..............actg............................
..............act.t...........................
..............act..ttatat.....................
..............act..ttat..gc...................
..............act..ttat....t..................
..............act..ttat.....c.................
..............act..ttat......gga..............
..............act..ttat......gg.gtc...........
..............act..ttat......g....ttgtgtt.....
..............act..ttat......g....t.......tc..
............................................t.
                                            ..
```

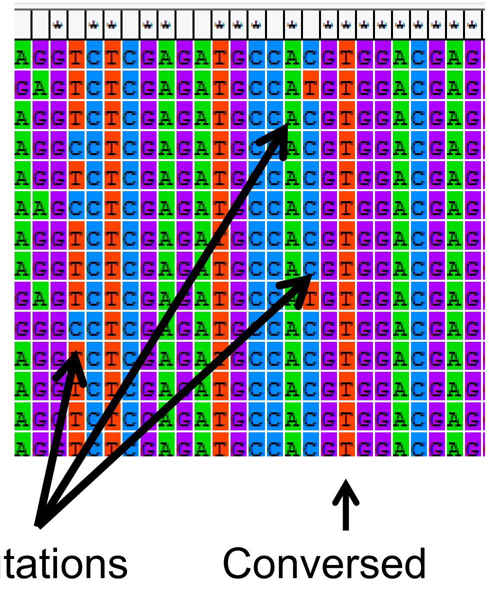
Fine scale properties of genome
Every SNP can be useful

Pathogen Sequence Data

- Pathogen sequence data provides richer information than strain type

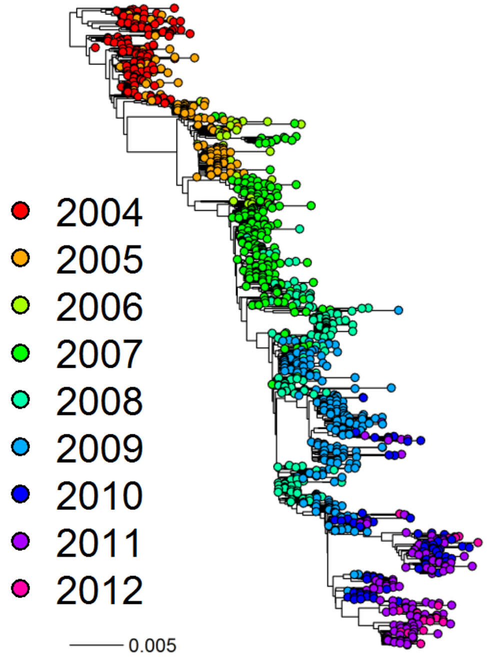
- Sequences accumulate mutations over time – classic picture (influenza)

Sequences, one per row

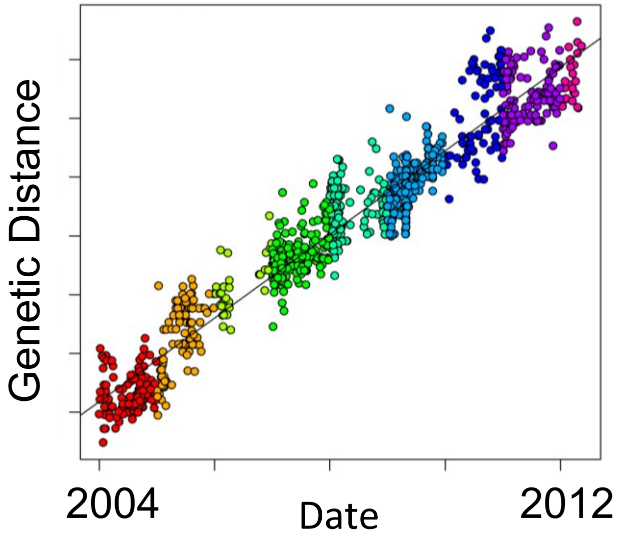


```
  *  ** **  *** * *********
AGGTCTCGAGATGCCACGTGGACGAG
GAGTCTCGAGATGCCATGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
AGGCCTCGAGATGCCACGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
AAGCCTCGAGATGCCACGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
GAGTCTCGAGATGCCATGTGGACGAG
GGGCCTCGAGATGCCACGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
AGGTCTCGAGATGCCACGTGGACGAG
```

Tree of Human Influenza



Genetic Distance from Root



ROSLIN

How much sequence variation ?

THE UNIVERSITY of EDINBURGH

BBSRC
20 Years of Pioneering
Great British Bioscience

Evolutionary Rates

	RNA Viruses	DNA Viruses	Bacteria
Replication & Evolution	Fast and error prone	Slower, more conserved	Slow
Genome size	8-14kb	20-200kb	4Mb
Mutations per year	10-100	1-20	0-1 ?

Classical Swine Fever
Bovine Viral Diarrhoea
Foot-and-Mouth

Segmented ssRNA

Avian influenza
Schmallenberg

African Swine Fever

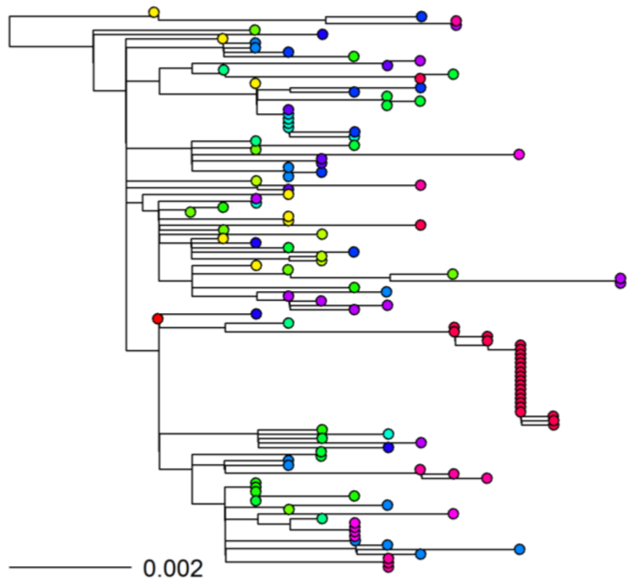
Segmented dsRNA

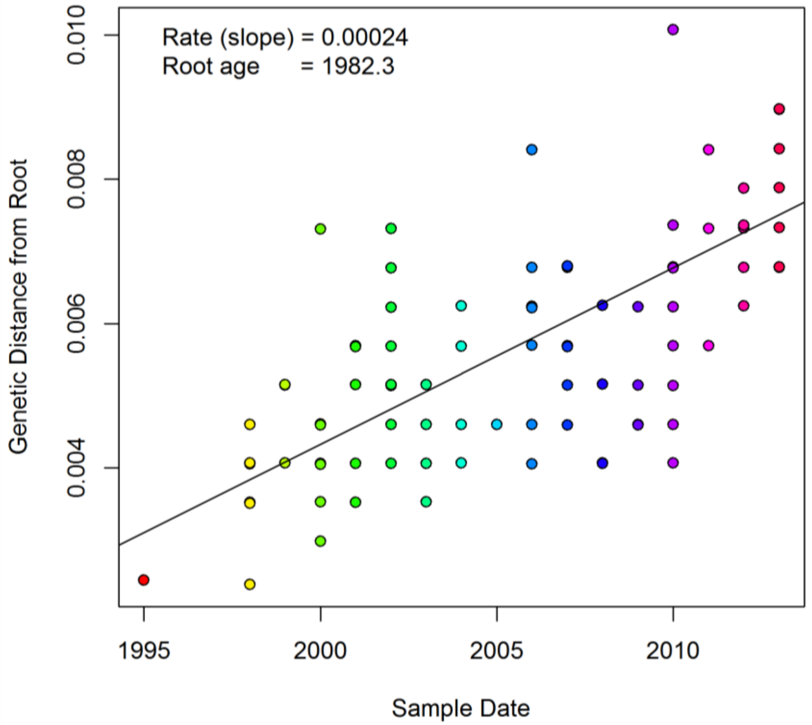
Blue Tongue
African Horse Sickness

Bovine Tb

Mbovis example – one clade one region

- True rate estimates ~ 0.1 SNP per lineage per year
- Example:
  - 126 samples, date range = 18 years
  - Number of variable sites = 309 (concat. SNPs)





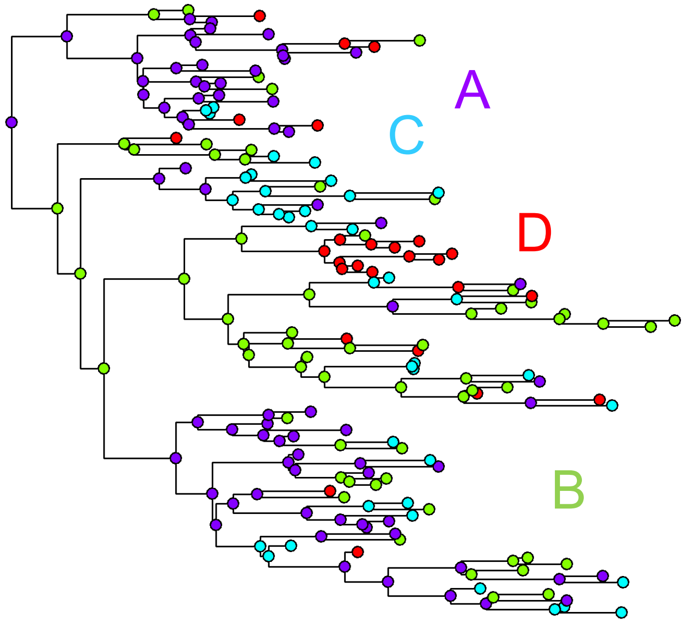
Inferring Transmission Patterns

THE UNIVERSITY *of* EDINBURGH

BBSRC 20 Years of Pioneering Great British Bioscience

Transmission Patterns and Discrete Trait Models

Tree with Location Traits



Transition Rate Matrix (M)

	A	B	C	D
A	-	B->A	C->A	D->A
B	A->B	-	C->B	D->B
C	A->C	B->C	-	D->C
D	A->D	B->D	C->D	-

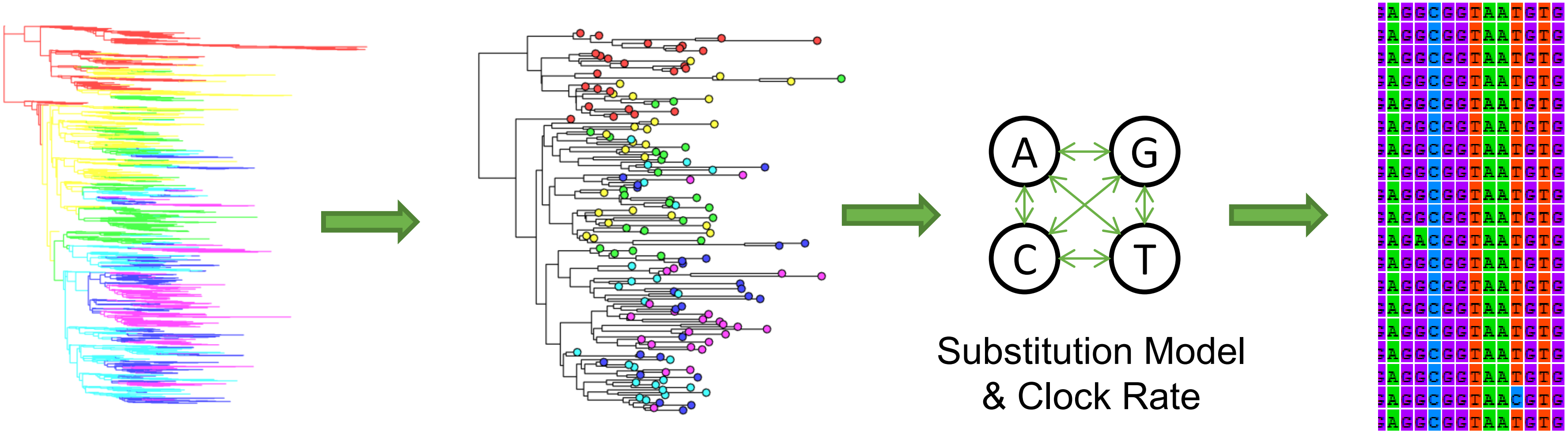
Probability of Ancestral state (x'), given branch length t and child state x:

$$p(x'|t) \sim e^{Mt}x$$

- Add locations to phylogenetic tree
- Estimate transition rates between locations along branches
- Transmission pattern represented by rate matrix

Simulate Sequences

- Examine phylodynamic situations by simulation
  - output true transmission tree and phylogenetic tree
- Simulate sequences down tree
  - Use different mutation rates and lengths, equivalent to:
  - 0.025 – 2 substitutions per genome per year
  - samples spanning 15 years
- Total number of SNPs in data: 10 - 750



Full Tree
(2370 tips)

Rescaled & subsampled tree
(25 per deme = 150 tips)

Substitution Model
& Clock Rate

150 Sequences

Simulate Trees

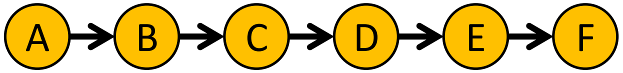
- DiscreteSpatialPhyloSimulator (DSPS) to simulate infection over structured population
- Individual based model, individuals are farms
- 6 regions (random mixing within demes)
  - 500 farms per deme
  - Each farm is SIR; beta = 0.1, gamma = 0.05
  - Infection between demes = 0.1
  - Demes connect in LINE or STAR network

LINE

From

To

	A	B	C	D	E	F
A	-	r2	r2	r2	r2	r2
B	r1	-	r2	r2	r2	r2
C	r2	r1	-	r2	r2	r2
D	r2	r2	r1	-	r2	r2
E	r2	r2	r2	r1	-	r2
F	r2	r2	r2	r2	r1	-

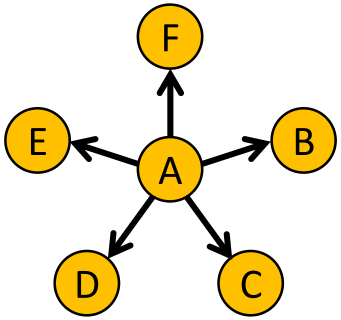


STAR

From

To

	A	B	C	D	E	F
A	-	r2	r2	r2	r2	r2
B	r1	-	r2	r2	r2	r2
C	r1	r2	-	r2	r2	r2
D	r1	r2	r2	-	r2	r2
E	r1	r2	r2	r2	-	r2
F	r1	r2	r2	r2	r2	-

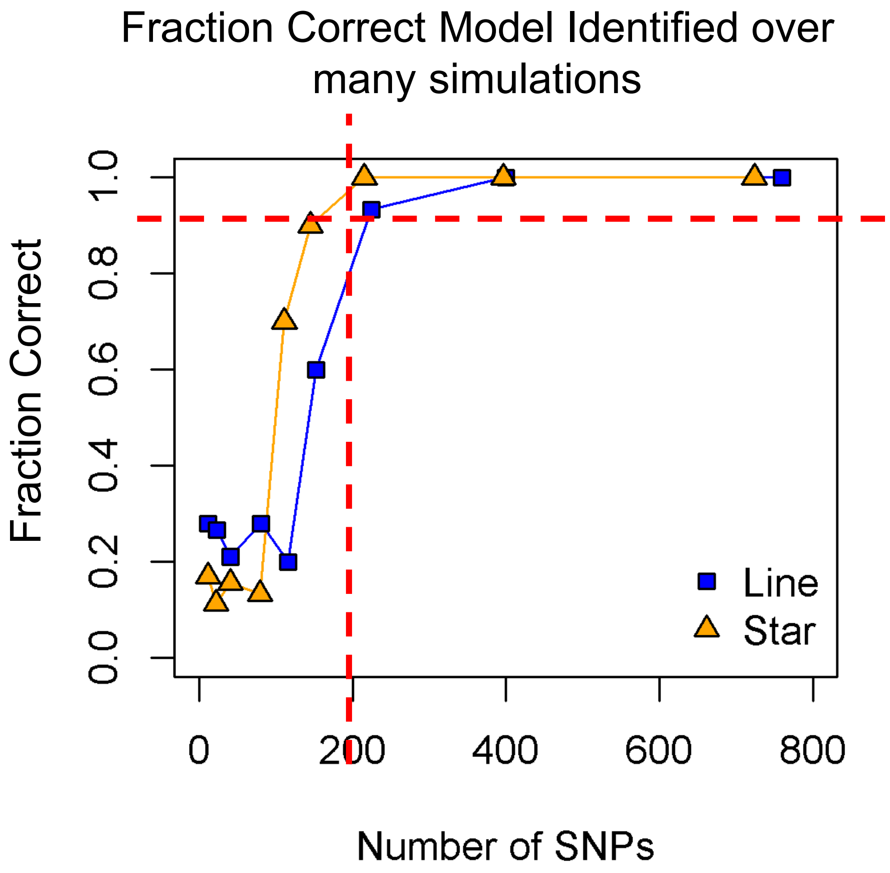


Code available – currently tidying & validating…

https://github.com/hxnx-sam/DiscreteSpatialPhyloSimulator

Finding the correct pattern

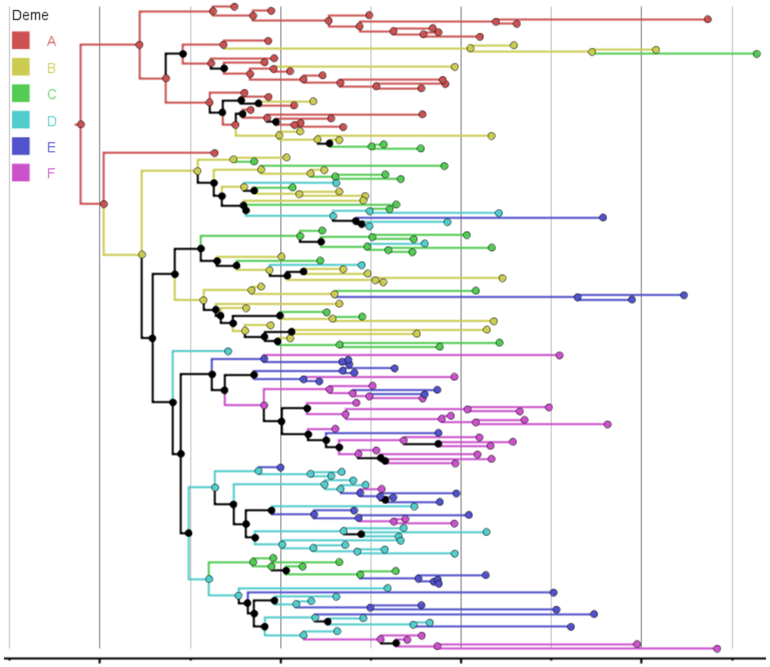
- Reconstruct trees from simulated sequences using Neighbour Joining
- Calculate likelihood of line, reverse line, star, reverse star models upon reconstructed trees
- Fraction of simulations where correct transmission pattern found is a function of number of mutations



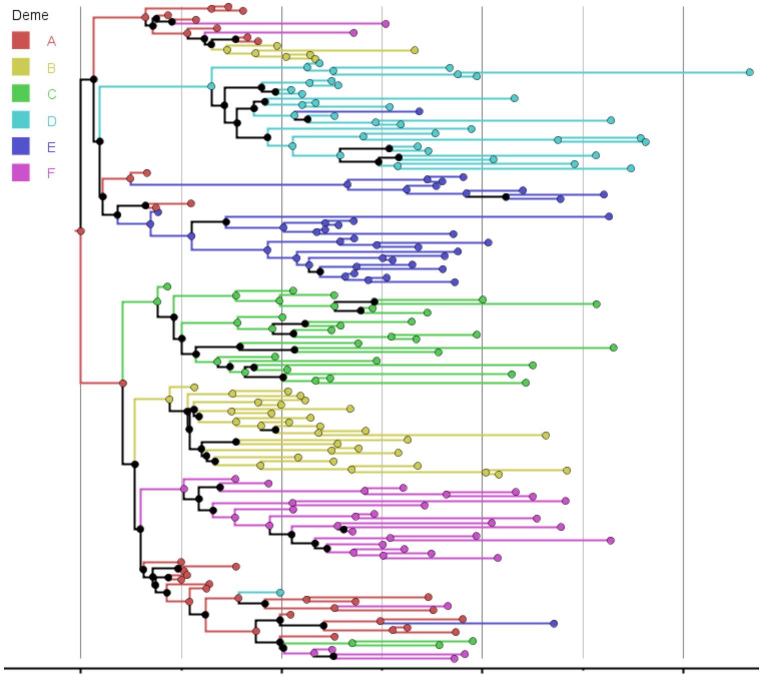
200 SNPs => 90% correct
(~0.5 mutations per year)

Phylodynamics using BEAST

- Infer trees and transition rate matrix with BEAST
- Use Line and Star scenarios with differing sequence lengths and mutation rates (slow & short, moderate, fast & long)

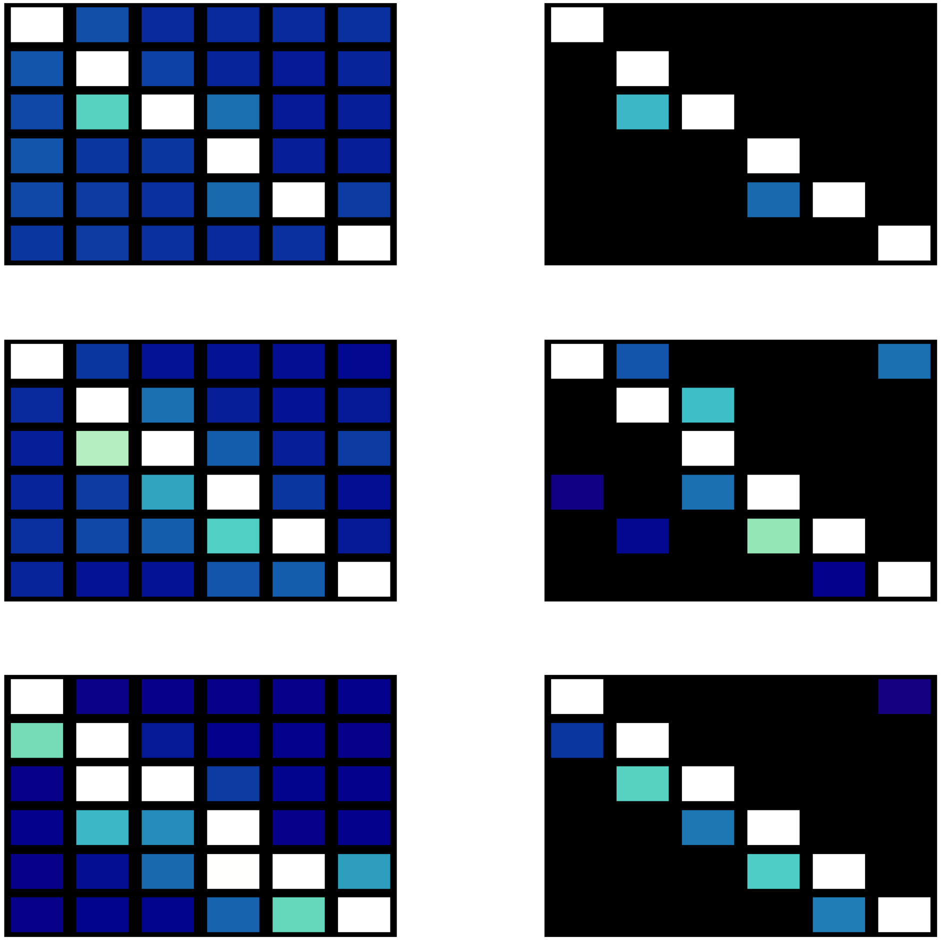


LINE



STAR

Detecting Line Population Structure



Full Rate Matrix

Significant Rates

(i) Short and slow
22 SNPs

LINE
Population
Structure

(ii) Moderate
80 SNPs

BVD single region
(3 years)

TB WGS
(10-20 years)

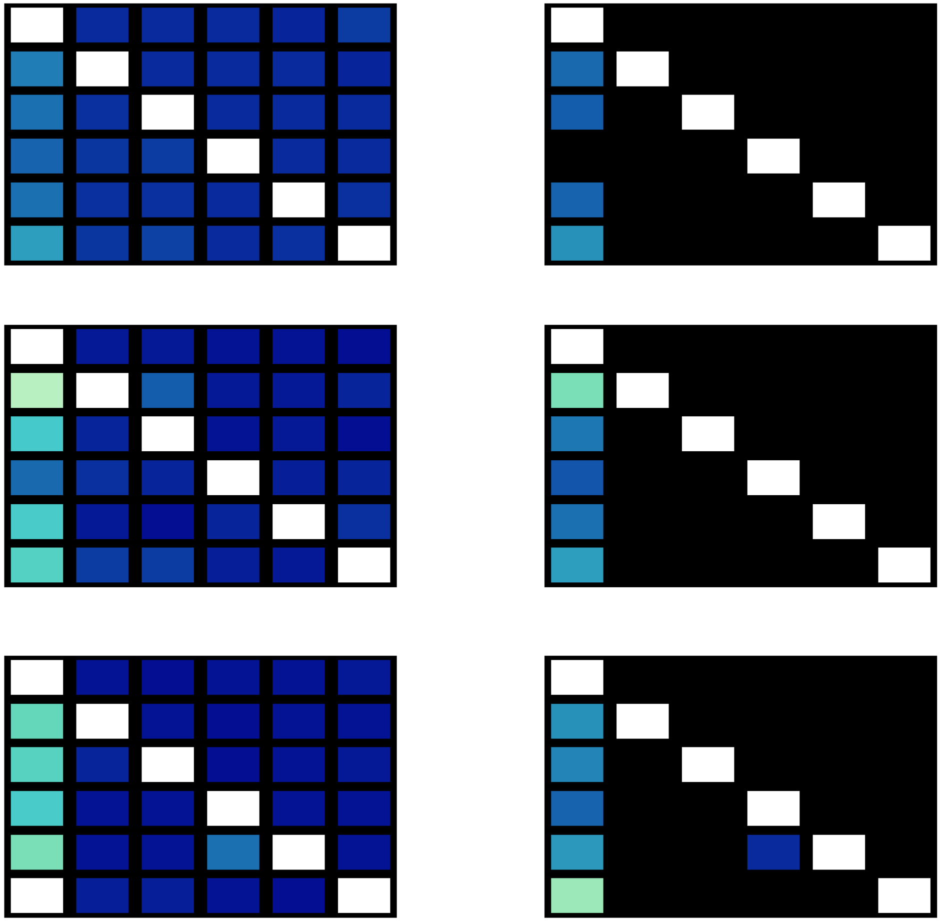
(iii) Long and fast
400 SNPs

Flu segment
(3 years)

Detecting Star Population Structure

Full Rate Matrix

Significant Rates



(i) Short and slow
22 SNPs

(ii) Moderate
80 SNPs

(iii) Long and fast
400 SNPs

STAR
Population
Structure

BVD single region
(3 years)

TB WGS
(10-20 years)

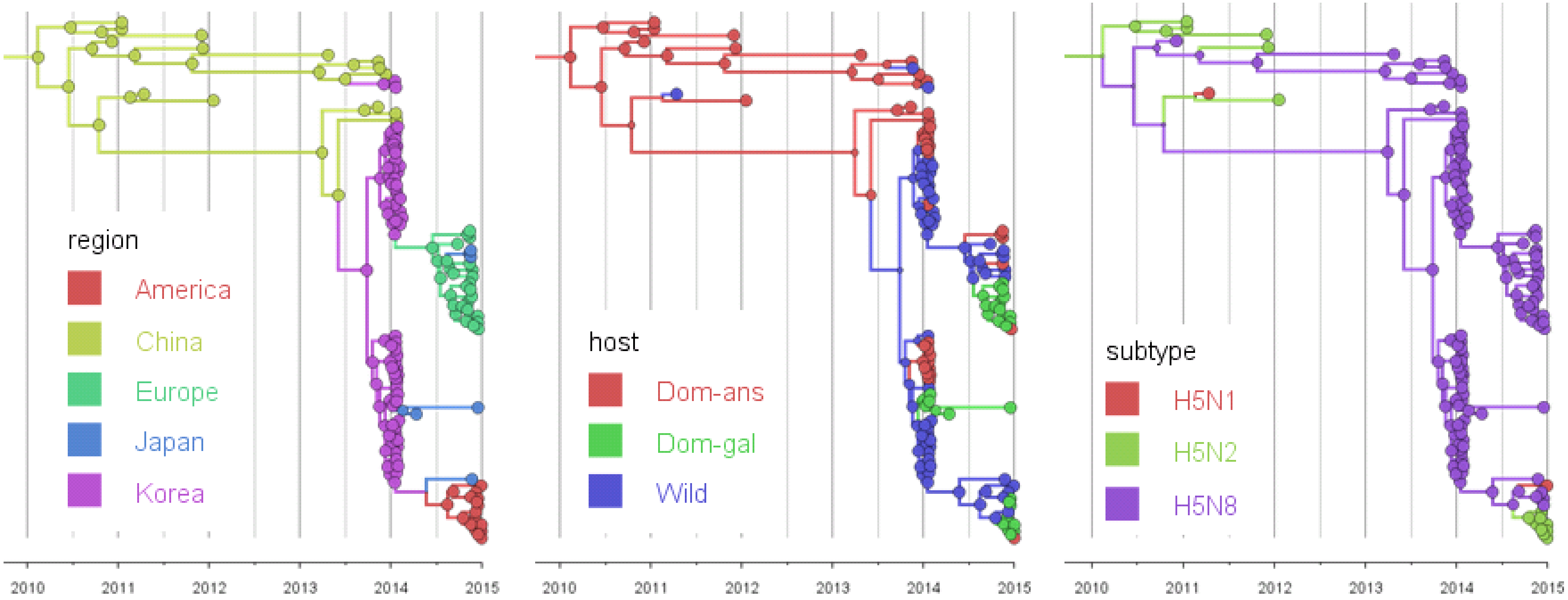
Flu segment
(3 years)

Example

With multiple traits !

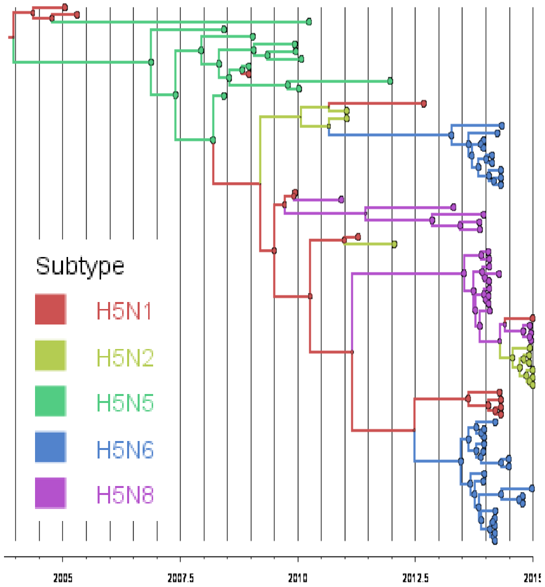
Phylodynamics with multiple traits

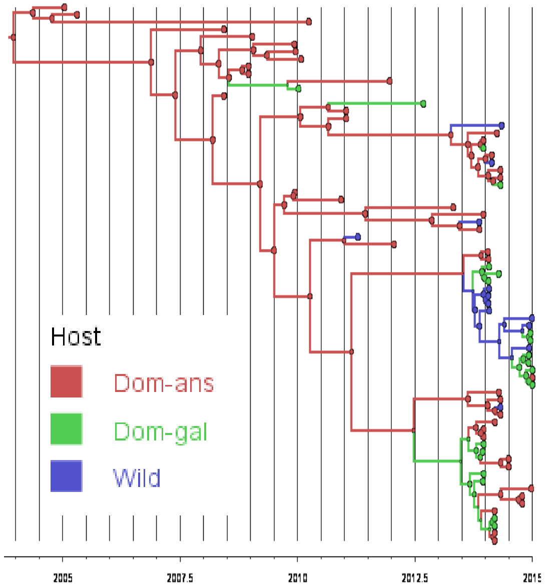
- Recent HPAI Avian influenza in UK and North America
- Where did it come from ?
- Generate time resolved trees from HA sequences
- Include region, host and subtype as discrete traits

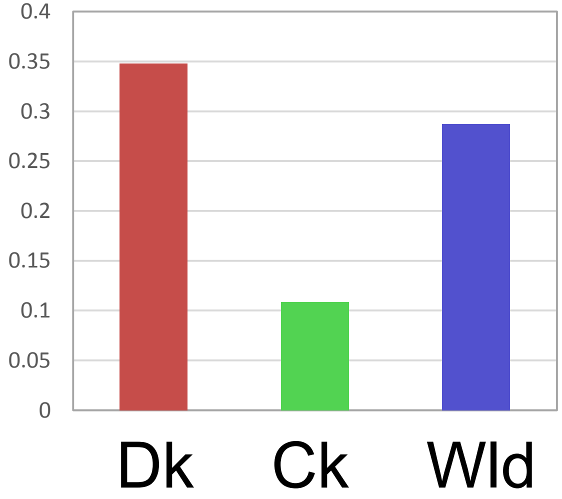


Correlation of traits

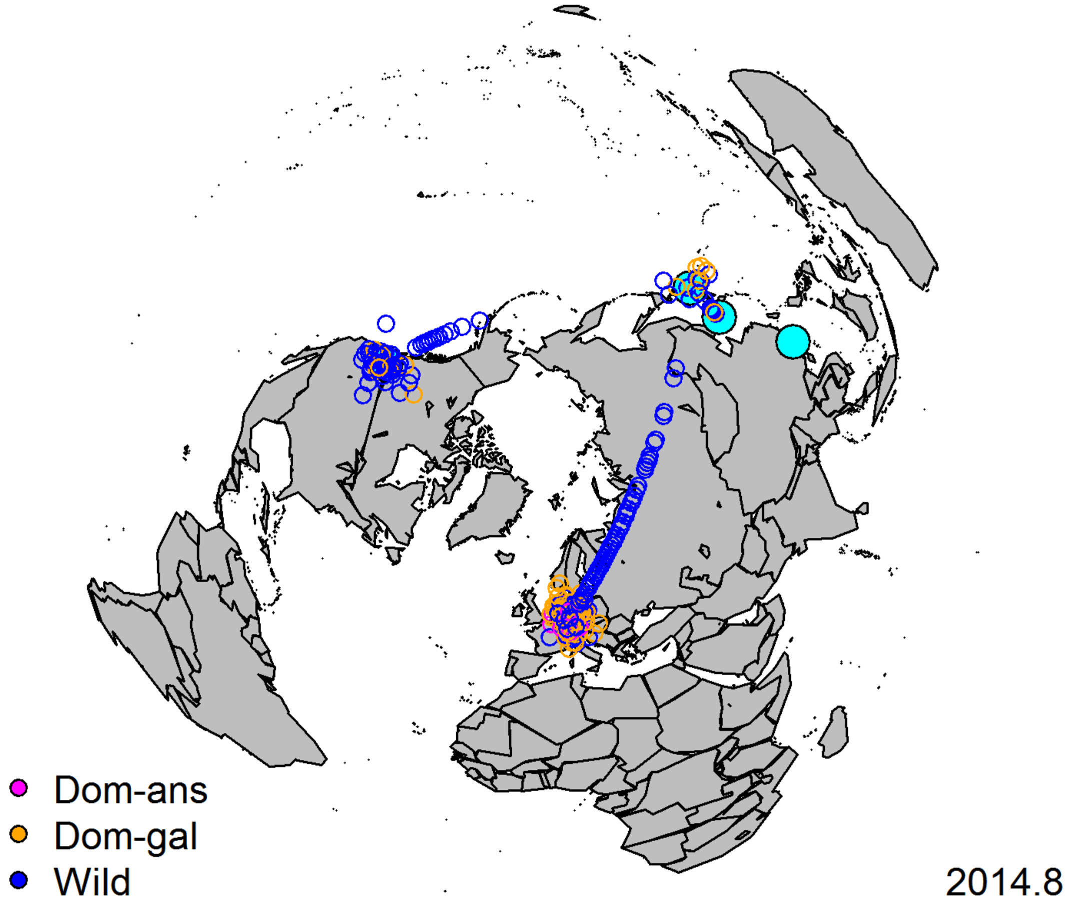
- Map host and subtype on same set of trees
- Count subtype changes on duck, chicken or wild birds only branches
- Find more reassortment in ducks and wild birds (anseriformes)







Phylodynamics with host and spatial information
Shows dispersion by one host species



(one image of movie)

ROSLIN

Distinguishing transmission patterns

THE UNIVERSITY *of* EDINBURGH

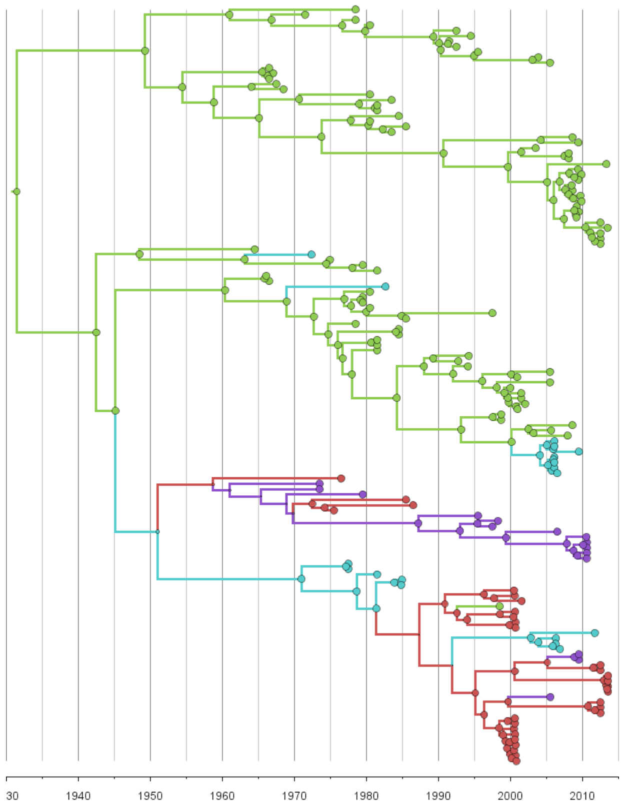
BBSRC
20 Years of Pioneering
Great British Bioscience

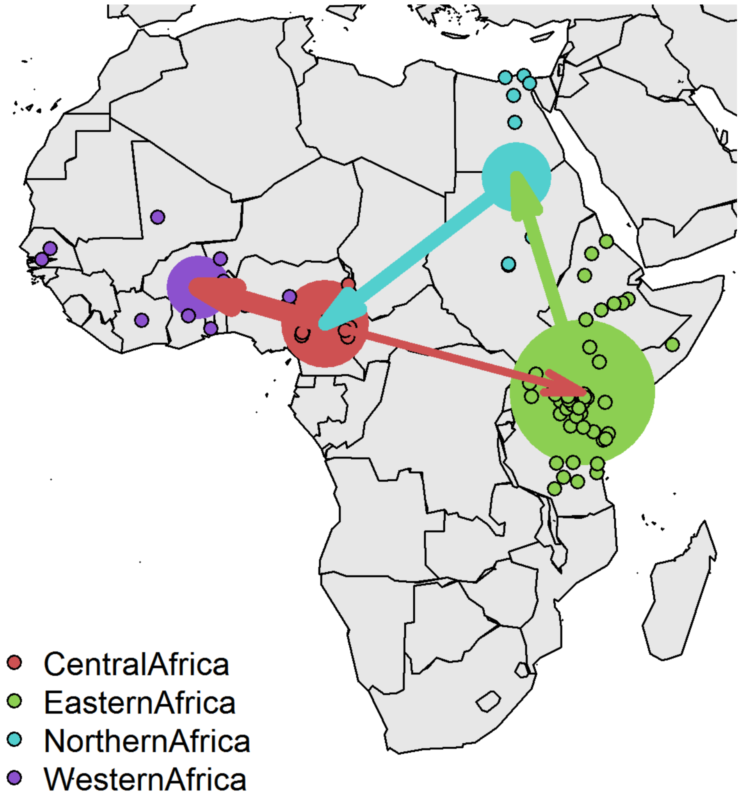
Detecting transmission patterns

- Are the transmission patterns due to known animal movements ?
- Or is there something else ?
- Method 1:
  - Infer rate matrix from discrete locations and calculate significant links between places using BSSVS
- Method 2:
  - use Latitude & Longitude and infer routes taken
- (Both) Compare to known movements (manually)

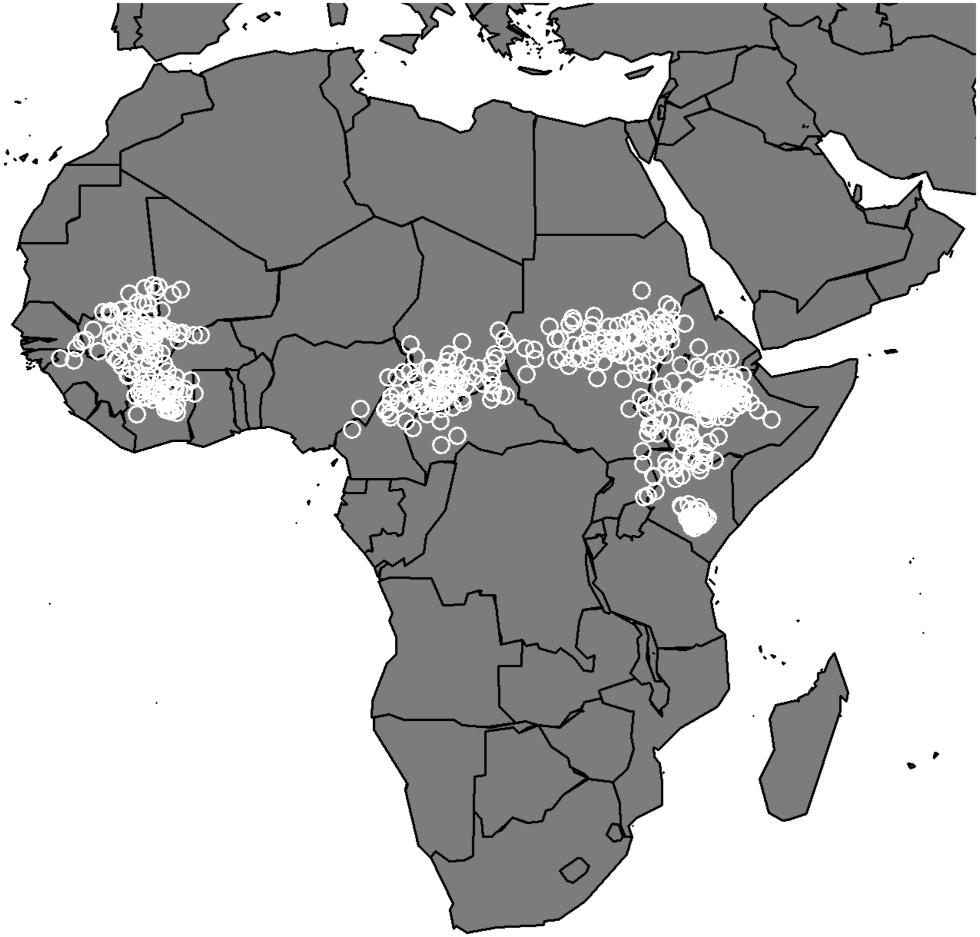
FMD – Serotype A in Africa

- Sequences ~600 bases long of VP1
- 444 SNPs for 142 sequences in time scale 1964 – 2013 (49 years)
- Using regional groupings => 4 discrete states





Dispersion in space and time
Using Latitude & Longitude as continuous trait (lower clade)



1993 A

(one image of movie)

Method 3

- Problems with Methods 1 & 2
  - Too many rates not enough different transmission events ?
  - Distances too far / diffusion not working ?
  - Why is it those rates or diffusion c/e anyway ?
- Use a Generalised Linear Model to parameterise the rates:

Transition Rate Matrix

For i=1 to n predictors

Predictor Rate Matrix

$$\log \Lambda = \sum_{i} \partial_i \beta_i \log R_i$$

(for each MCMC step, propose βi's and δi's)

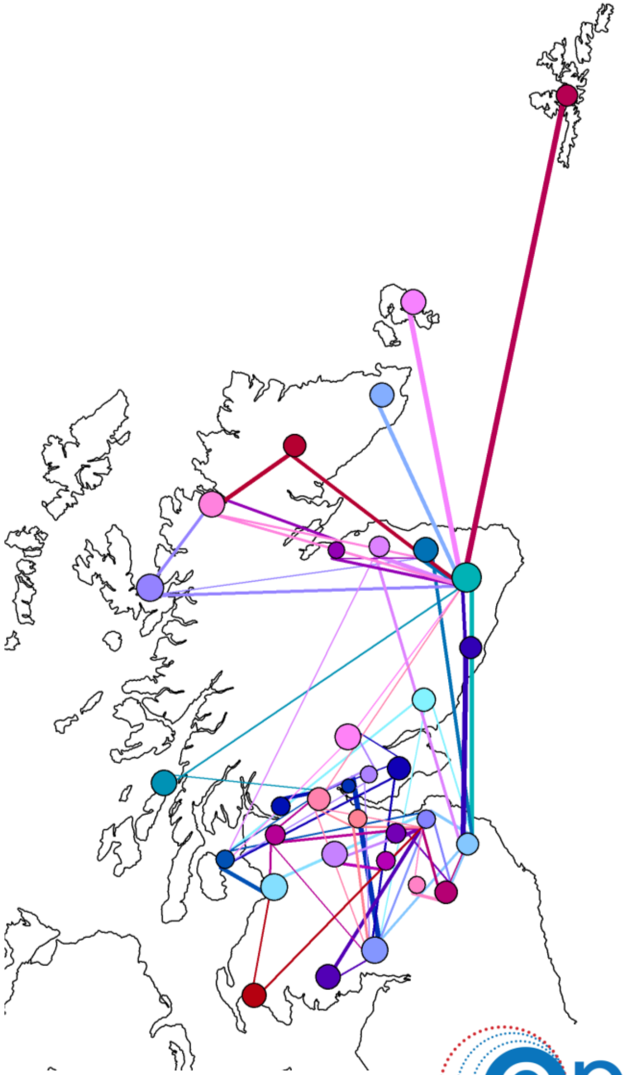
Indicator variable, delta dirac (0 or 1)

Coefficient of predictor matrix

Now estimate the δ and β instead of each rate matrix element

Detecting Transmission Patterns

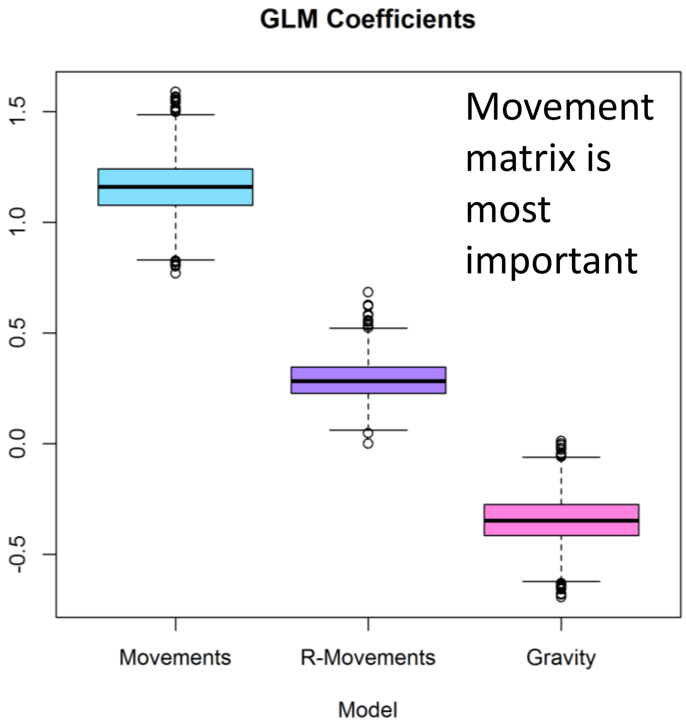
- Simulate infection over population using "DiscreteSpatialPhyloSimulator"
- Simulate individual farms within 33 counties in Scotland
- Probability of infection between counties proportional to averaged movements of Cattle Tracing System
- Generate who-infected-who, but subsample to 10 sequences per county (massively undersampled!)
- Simulate sequences as before



Detecting Transmission Patterns

- Infer tree with Discrete traits model in BEAST
  - Model is 33 x 33 matrix (1056 rates)
  - Far too many individual rates !
- Use Generalised Linear Model
  - Predictor 1: Movement matrix
  - Predictor 2: Reversed movement
  - Predictor 3: Gravity Model

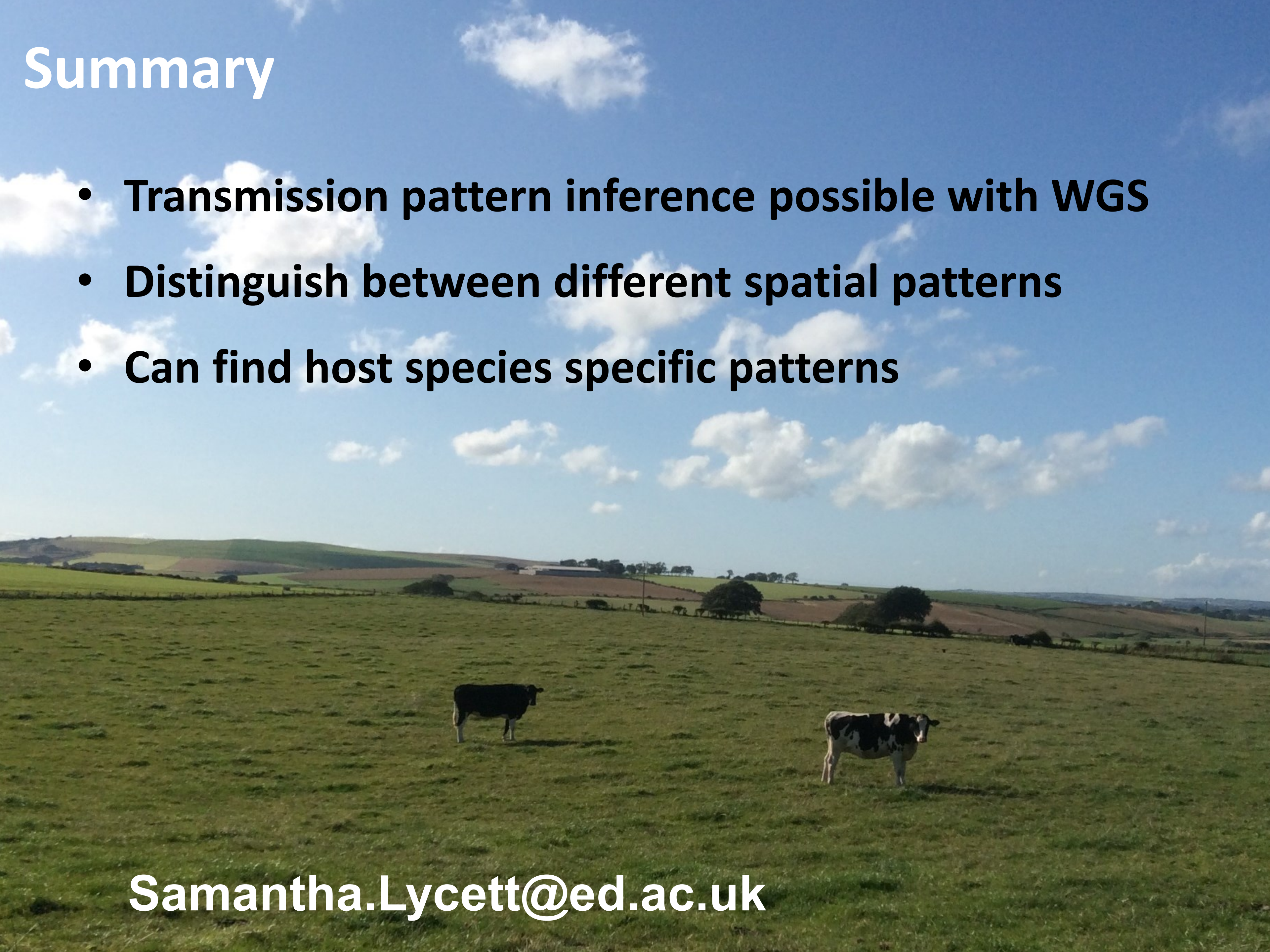
$$\frac{\text{Source size x Dest. size}}{\text{distance}^2}$$



Movement matrix is most important

- Can distinguish between possible transmission patterns in principle

Summary



- Transmission pattern inference possible with WGS
- Distinguish between different spatial patterns
- Can find host species specific patterns

Samantha.Lycett@ed.ac.uk

Acknowledgements

ROSLIN

SRUC

BioSS

The James Hutton Institute

Moredun

- Rowland Kao
- Jessica Enright
- Ruth Zadoks
- George Russell
- George Gunn
- Mark Bronsvoort
- Andrew Rambaut
- Mark Woolhouse
- Andrew J. Leigh Brown
- Lu Lu
- Matthew Hall
- Emma Hodcroft

University of Glasgow

UNIVERSITY OF STIRLING

ROSLIN

epic
Centre of Expertise on Animal Disease Outbreaks

Supported by wellcome trust

BBSRC
20 Years of Pioneering Great British Bioscience

Thank you !

Samantha.Lycett@ed.ac.uk