2nd generation sequencing – a tool for exploring arbovirus distribution

Steve Kemp & Phil Toye, ILRI, Nairobi

AVID - Arbovirus Incident & Diversity

icipe - International Centre for Insect Physiology and Ecology

KEMRI – Kenya Medical Research Institute

KARI - Kenya Agricultural Research Institute

KWS - Kenya Wild life Service

MoH – Ministry of Health, DVBD, Kenya

MPH – Ministry of Public Health, Kenya

DVS – Department of Veterinary Services, Kenya

ILRI – International Livestock Research Institute

AVID - Arbovirus Incident & Diversity

ILRI's component – Sequencing and data management

THE UNIVERSITY
of LIVERPOOL

The Wellcome Trust Funded
Host-Pathogen Project

Questions

- "Where" is the virus (between "outbreaks") ?
  - Environment
  - Vectors
  - Reservoirs
- What is the diversity of ?
  - Virus
  - Vector
  - Reservoir
- And how do these interact ?
- Distribution of other pathogens ?
- Novel pathogens and variants ?

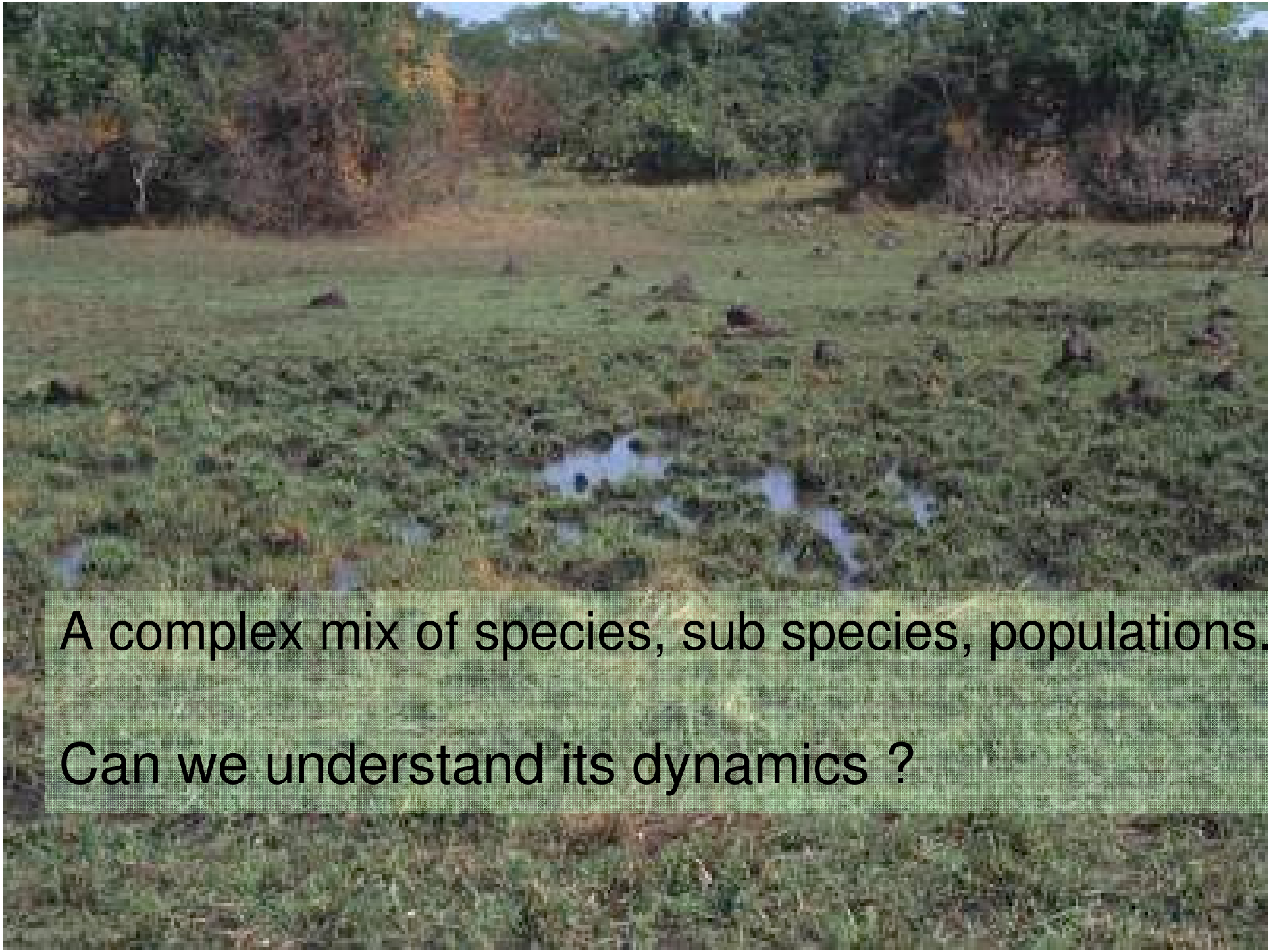
For example:
Does a particular virus variant occur in a particular vector variant associated with a particular mammalian variant ?

Viral Geneflow

Questions

- "Where" is the virus (between "outbreaks") ?
  - Environment
  - Vectors
  - Reservoirs
- What is the diversity of ?
  - Virus
  - Vector
  - Reservoir
- And how do these interact ?
- Distribution of other pathogens ?
- Novel pathogens and variants ?

For example:
Does a particular virus variant occur in a particular vector variant associated with a particular mammalian variant ?



A complex mix of species, sub species, populations.

Can we understand its dynamics ?

Another complex mix of species, sub species & populations.

Can we understand its dynamics ?

Can we understand the dynamics of complex mixes of species, sub species & populations?

- (RT)-PCR Alone cannot do this.
  - It is a yes/no test (or a collection of yes/no tests).
  - It will not tell you about unknown sequences or sequence variants.
  - It *will* tell you about known variants - but each requires its own PCR
  - It is very slow, laborious and time consuming when scaled-up.

2nd Generation DNA sequencers

They are just DNA sequencers.

But their massively increased throughput means they provide a completely new way to do biology

They are tools for:

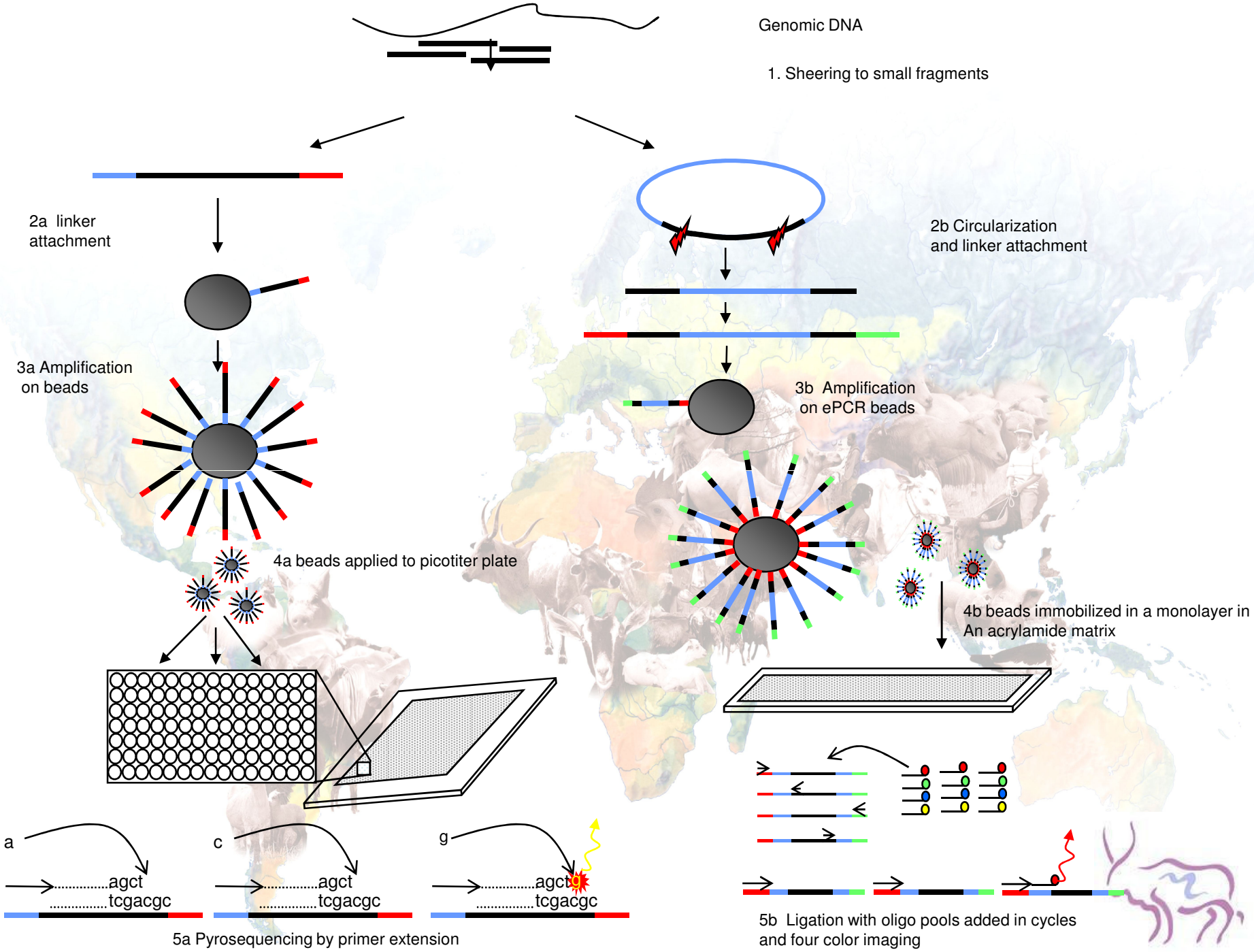
Probing complex biological mixes

Uncovering diversity at species, sub-species, individual level

2 types

- 454 GSFLX
  - 500 Mbases in 7 hours
  - £6,000 per run
  - 500bp reads

- ABI Solid
  - 30Gbases in 5 days (->90Gb)
  - £10,000 per run
  - 35 bp reads


Genomic DNA
1. Sheering to small fragments
2a linker attachment
2b Circularization and linker attachment
3a Amplification on beads
3b Amplification on ePCR beads
4a beads applied to picotiter plate
4b beads immobilized in a monolayer in An acrylamide matrix
a
c
g
agct
tcgacgc
5a Pyrosequencing by primer extension
5b Ligation with oligo pools added in cycles and four color imaging

2 broad approaches

1. Sequence everything in your sample

1. Reduce complexity by enrichment or amplification. Then sequence.

It all comes down to

1. how much complexity you can handle informatically and financially.
2. How much 'depth' you want.

2nd Generation DNA sequencers

Applications:

- Metagenomics
- Genome diversity
- Genome function
- Digital expression analysis
- Pathogen discovery
- Epidemiology
- etc etc etc

Actually, this list is meaningless; applications run into each other.

Strategy

Changing as technology improves !

- Sample potential vectors
- Sample people
- Sample wildlife

From 3 eco-climatic zones in kenya


Eldoret
Kenya
Kisumu
Nakuru
Nairobi
Moshi
Mombasa
Garissa (outbreak 2006-7)
Baringo (outbreak 2006-7)
Naivasha (reported endemicity)

Strategy

- Each *sample* collected with a full meta data description (location, date/time, eco-geo-socio descriptors)
- Simultaneously amplify and tag signature sequences from multiple points on multiple possible genomes – virus, insect, mammal, others.
- Sequence these *amplicons* simultaneously from 1,000s of samples.

Strategy

- Analyse sequence – look for distribution and co-occurrence.
- (Refine primers for a ‘simple’ (RT)-PCR approach.)

Strategy

- This is already becoming obsolete. As capacity increases can we consider total sequence combined with sample fractionation? Eg low pass whole genome *sequencing* of all RNA viruses

- We can fully sequence approx.10 viruses on 1 machine run. [1st candidate - RVF vaccines?]

Outputs

- Information on diversity, dynamics, gene-flow, interaction to inform modeling and epidemiology

[Also ??

- A different kind of diagnostics ?????
  - Very high throughput
  - Not yes/no, but providing a signature for each virus
  - Can examine multiple pathogens simultaneously
  - Very cheap per sample, very expensive per run
  - NOT ‘pen-side]

- NOT confounded by vaccination

When you have a hammer, every problem looks like a nail.

Thank you

International Livestock Research Institute