SOLiD-SAGE analysis of the Deer Transcriptome

agresearch Farming, Food and Health. First™

JOHNE'S DISEASE RESEARCH CONSORTIUM

Rudiger Brauning (speaker), Paul Maclean, Nauman Maqbool, Jo Stanton, Anar Khan and Colin Mackintosh.

Transcriptomics

- Basically “what is expressed”
- Need to work out what you want from it:
  - Differential expression of known genes?
  - Find new/novel transcripts?
  - Accuracy?

- Three options are considered here

Summary of SOLiD-SAGE

Advantages:

- Gives 1 tag (25 bp fragment) per transcript (in theory).
  - No transcript length bias
  - Less sequence real estate being taken up by longer sequences
  - These facilitate statistical evaluation greatly
- Large dynamic range (in theory)

Disadvantages:

- Need a good reference sequence dataset
- Short tags reduce unique mappings
- Can’t detect new transcripts/genes/MiRNAs

Summary of RNA-seq

Advantages:

- Full length transcripts can be assembled from reads
- Longer reads make mapping relatively straightforward
- Lots of tools and support available
- Is the favorite protocol of the research community
- Sequencing of MicroRNAs

Disadvantages:

- Many reads representing one transcript
  - Length bias: Longer transcripts get more reads than shorter transcripts
  - Less sequencing real—estate to detect genes expressed at a low level

Summary of Microarray

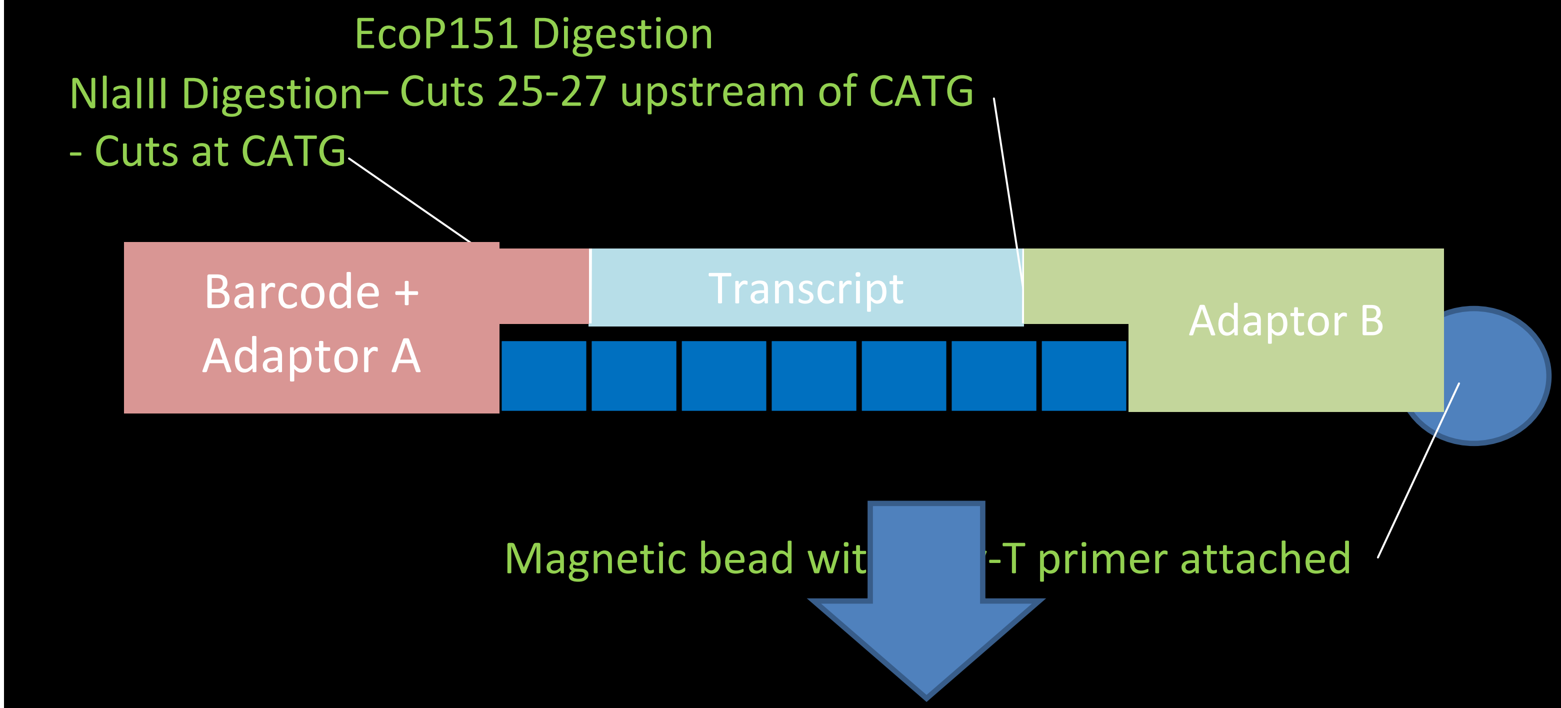
Advantages:

- Robust, well-established methods for analysis
- Ease of analysis
- Cheap: many replicates easily affordable

Disadvantages:

- Limited to what is on the array
  - Limited number of species
  - Can’t detect new transcripts/genes/MiRNAs
- Low dynamic range

SOLiD-SAGE: How it works



EcoP151 Digestion

NlaIII Digestion– Cuts 25-27 upstream of CATG

- Cuts at CATG

Barcode + Adaptor A

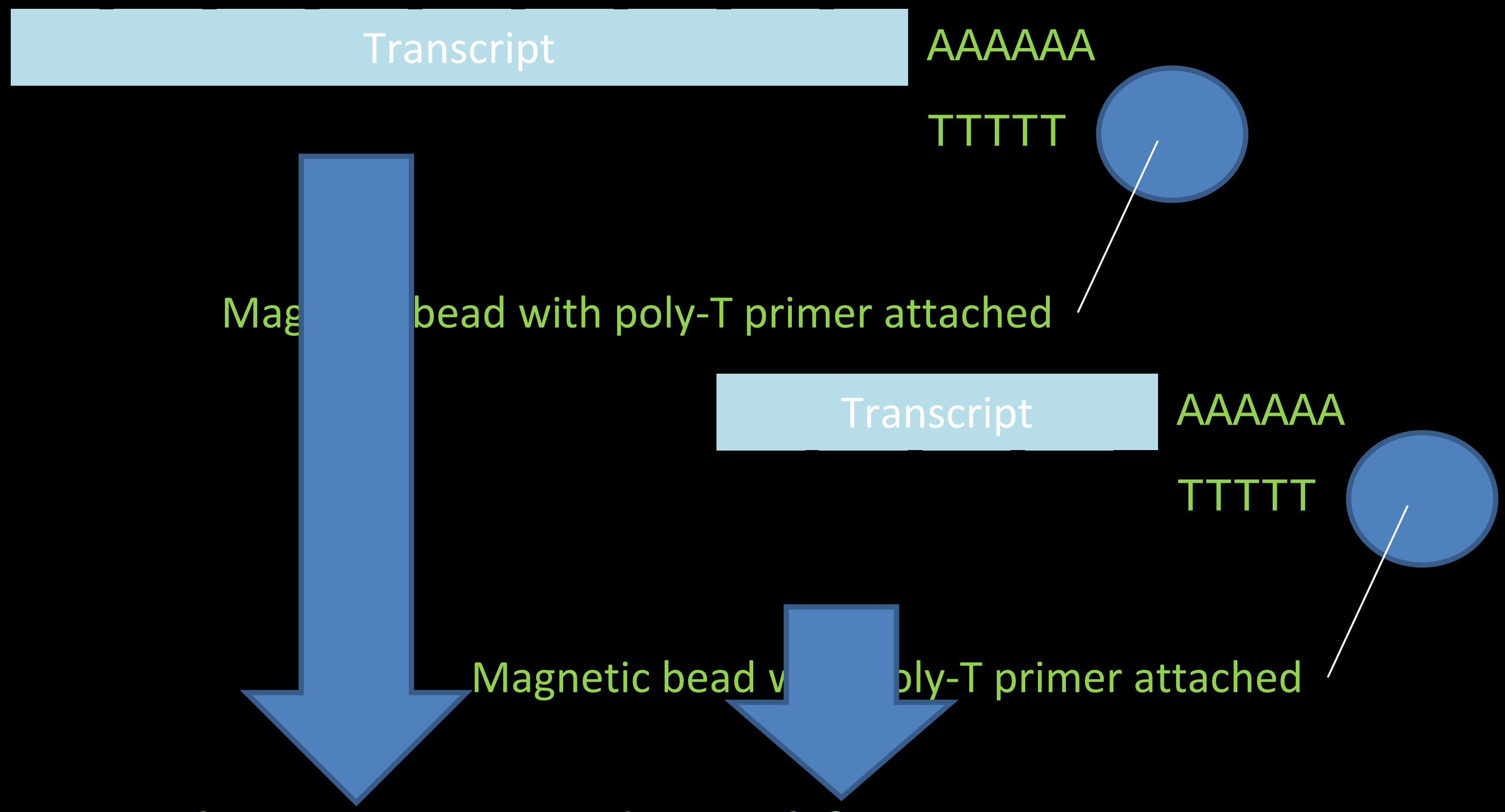
Transcript

Adaptor B

Magnetic bead wit -T primer attached

Emulsion PCR and Sequencing by ligation

mRNA-SEQ: How it works



Transcript

AAAAAA

TTTTT

Mag bead with poly-T primer attached

Transcript

AAAAAA

TTTTT

Magnetic bead v oly-T primer attached

Library prep and amplification

SOLiD Reads

Two files:

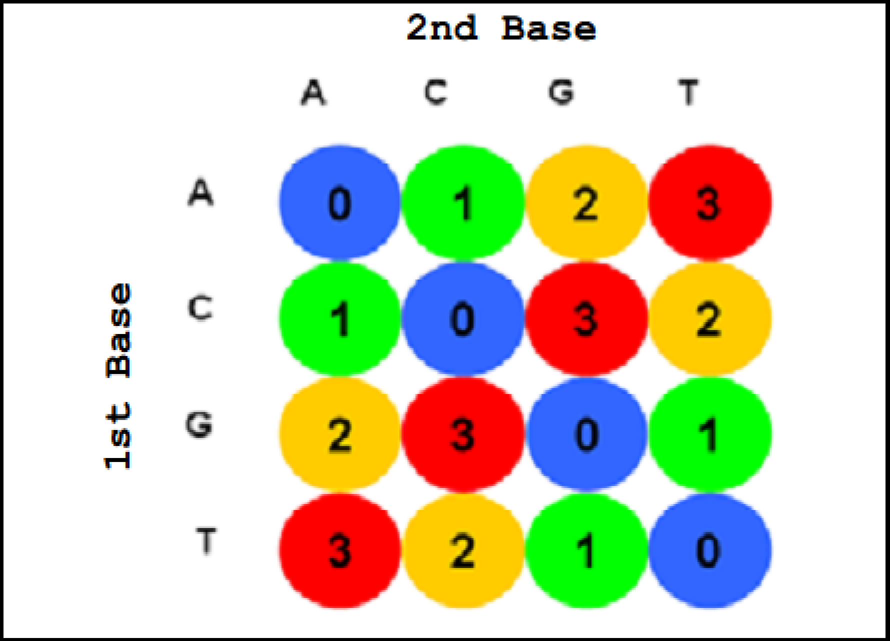
- Csfasta
  - Format:

    ```
    >559_31_72_F3
    T2303..22231..2332113.322
    ```

  - Di-base encoding – see next slide
- Qual
  - Format:

    ```
    >559_31_72_F3
    26 18 18 5 -1 -1 8 22 9 18 21 -1 -1 5 11 16 19 2 11 17 -1 19 13 3
    ```

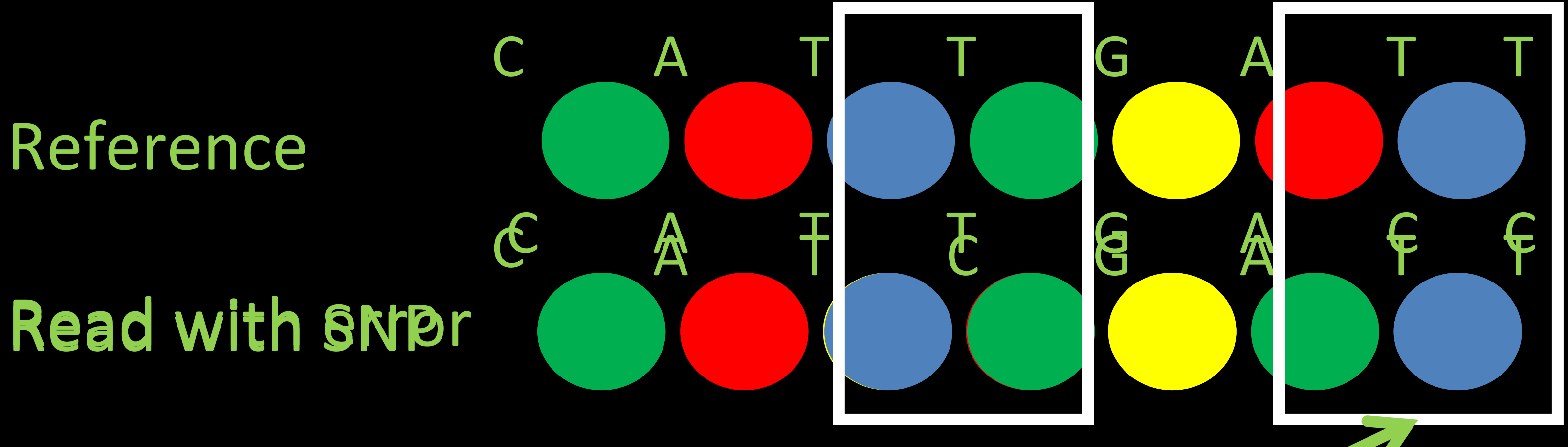
SOLiD di-base encoding



		2nd Base			
		A	C	G	T
1st Base	A	0	1	2	3
	C	1	0	3	2
	G	2	3	0	1
	T	3	2	1	0

```
T 2 3 1 3 0 0 1 0 1 0 0 0 3 3 3 1
T  C G T A A  A A C C A A A A T A T
```

SOLiD di-base encoding Part 2

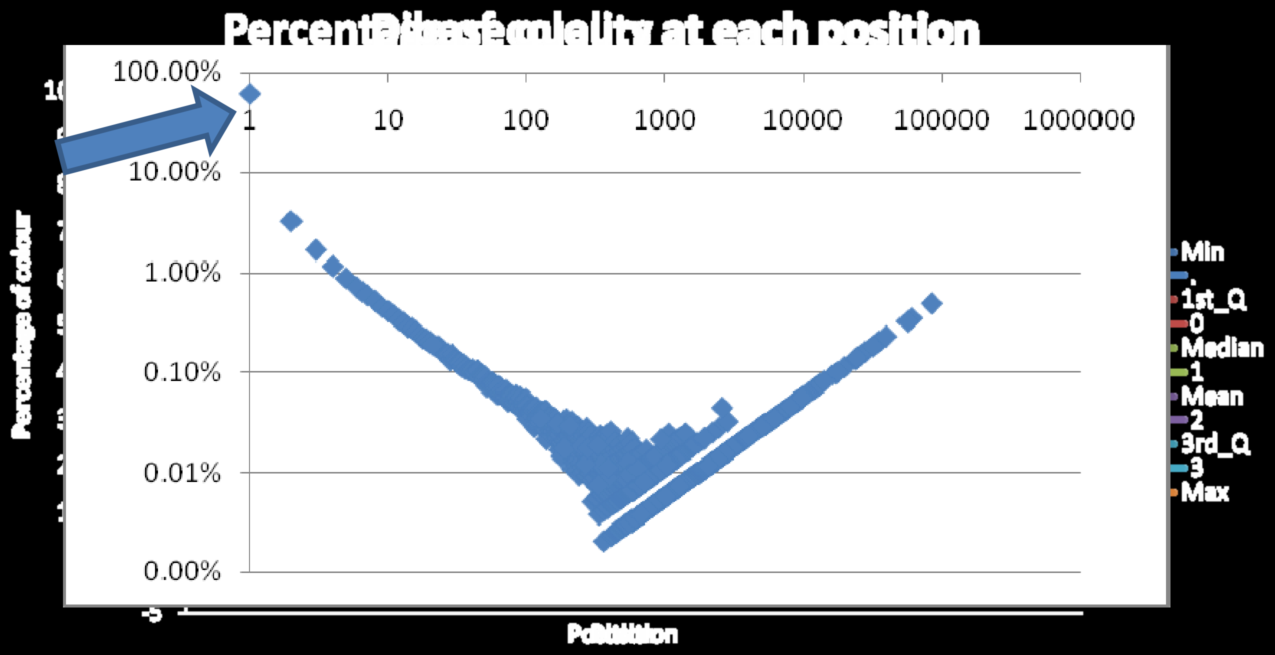


Keep alignments in colourspace!

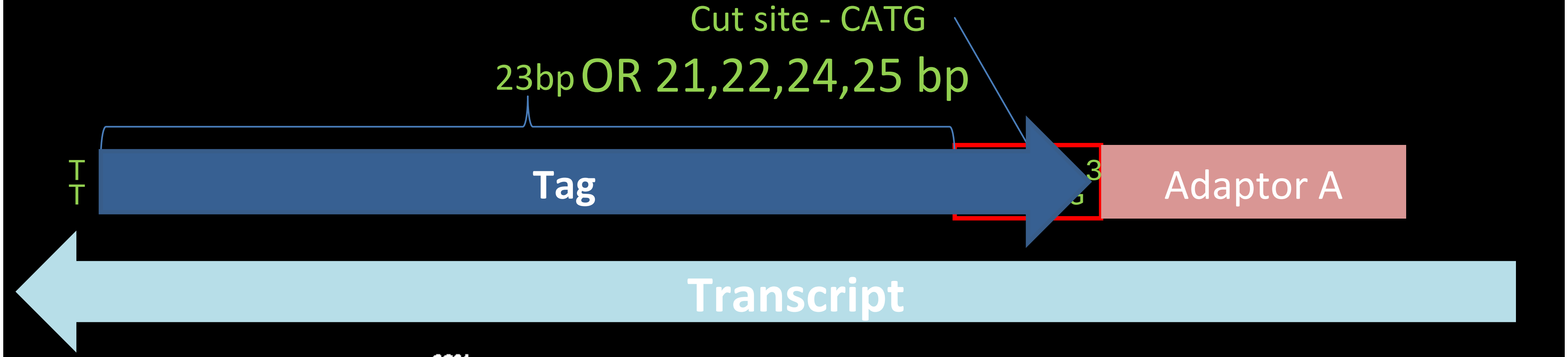
This makes it difficult to check alignments

Data QC

- Important for all sequencing platforms and methods
- Quality Colour Uniqueness



Tag Extraction



Cut site - CATG

23bp OR 21,22,24,25 bp

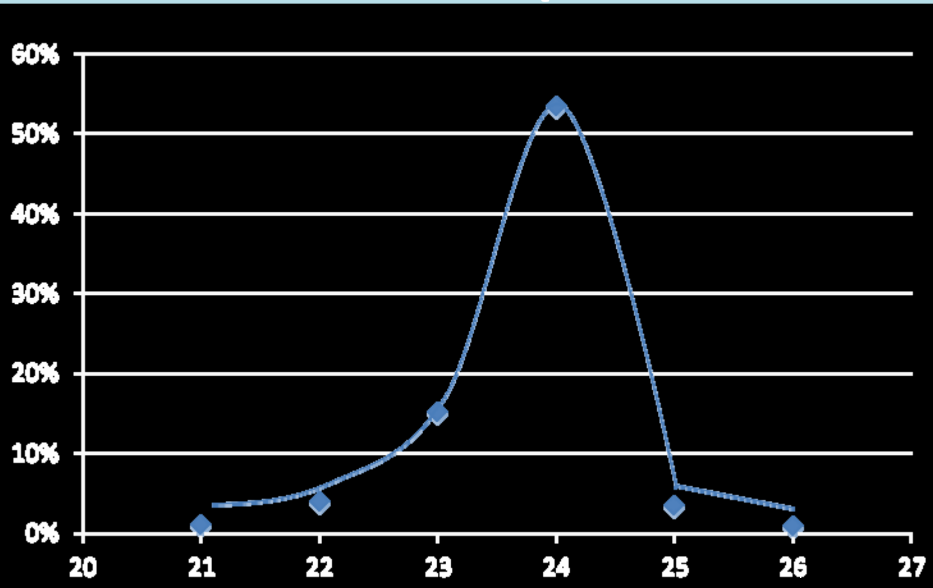
T
T

Tag

3
G

Adaptor A

Transcript



Reference databases

- AgResearch Deer and Elk contigs
- ENSEMBL Cow transcripts
- Cow genome version 4.2
- Deer genome version 1

Reduced Reference

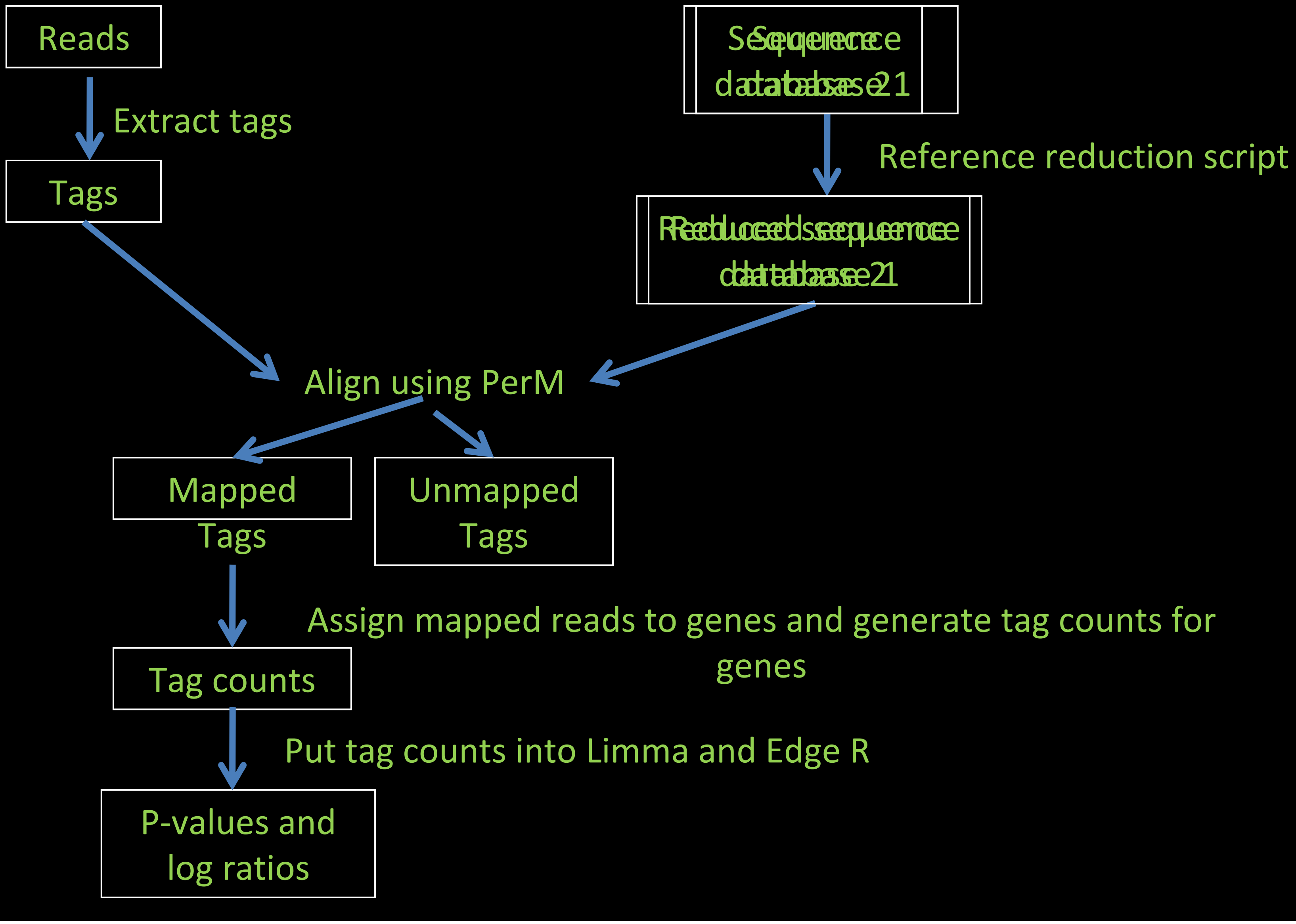
- What?
  - >Example_sequence

    CTGTGTGCAGCAGCTCAAGGACTTTGATGGCAAGCAGGCAAGGCCAAGTTT
    GAACGTTTTGAACCTCTGCATGAAAGAGATCTTGGATAAGAAGGTGGAGA
    AGGTGTCTAGTCCTTTATTTCTTTCCTTCTAAGGCAGGATACATGGGCCTC
    TTCCACGTTTGCATGGGGTTTGAATGTTGTGTGTTGTGTGTTTTGTTTTTT

    >Example_sequence_64
    CATGAAAGAGATCTTGGATAAGAAGG

    >Example_sequence_166
    CATGGGGTTTGAATGTTGTGTGTTGTG
    G
- Why?
  - only sequences that, in theory, result from the DGE-tag protocol using the DpnII restriction enzyme.

Mapping strategy



Reads

Extract tags

Tags

Sequence database 1

Sequence database 2

Reference reduction script

Reduced sequence database 1

Reduced sequence database 2

Align using PerM

Mapped Tags

Unmapped Tags

Assign mapped reads to genes and generate tag counts for genes

Tag counts

Put tag counts into Limma and Edge R

P-values and log ratios

Tag Mapping

PerM: Efficient mapping of short reads accomplished with periodic full sensitive spaced seeds

- Fast
- Easy to use
- Designed for SOLiD data
- Better results than default SOLiD software

Tag Aggregation

- In theory, the SOLiD-SAGE protocol gives 1 tag (25 bp fragment) per transcript
- This doesn’t quite always happen.
  - SNPs may prevent enzymes from cutting
  - Different alleles from deer (2N)
- Solution: Aggregate all tags that map to the reduced reference of a gene.

Mapping Results

- ~60% mapping, mostly unique
  - Reasons:
    - Reference sequences not perfect
    - Most reads were unique (searching of unmapped reads returned nothing)
- Significant differential gene expression of expected genes and gene pathways
- Insights into Johne's disease in deer
- To follow up with Illumina sequencing
  - Transcriptome assembly
  - Comparison of differential gene expression

Summary

- Different sequencing protocols require different analysis approaches
- There is much to consider in any NGS experiment
- All sequencing protocols have advantages and disadvantages

Acknowledgements:

JOHNE'S DISEASE
RESEARCH CONSORTIUM

PerM

The authors of PerM (Yangho Chen, Tade Souaiaia and Ting Chen)

Thanks!