Aspects of amplicon sequencing

Koen De Gelas[1], Jeroen Van Houdt[2],
Erik verheyen[1]

[1]Royal Belgian Institute for Natural Sciences

[2] Genomics Core KU Leuven, UZ Gasthuisberg

Before you start...

- What is the main problem / objective?
- Is NGS the optimal solution? Cost, time-efficiency, output
- Experimental design? # samples, # loci, coverage or read depth
- Platform to use? Read length, capacity, error rate
- Handling large datasets, bio-informatics

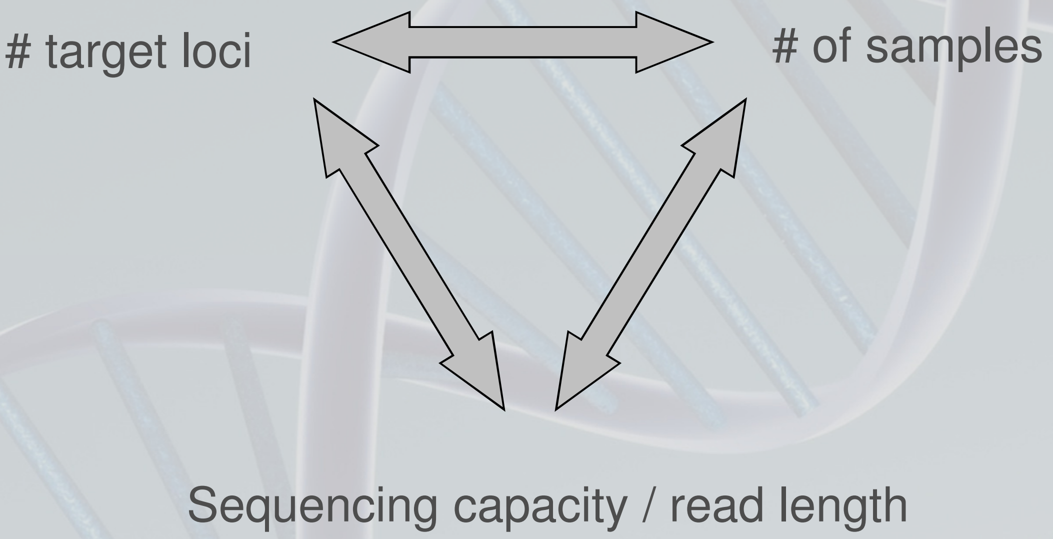
- No “standard” approach, rapid technological evolution, talk to your sequencing core facility

Comparison sequencing methods

- Sanger sequencing
  - Limited output
  - High per base cost
  - Rel. low error rate
  - Versatile (1 amplicon - 1 sample)
  - Fast

- NGS
  - Massive output
  - Low per base cost
  - Higher error rates
  - Inefficient for low number of amplicon – sample combinations
  - Slow

Designing NGS amplicon seq experiment



# target loci

# of samples

Sequencing capacity / read length

- Prep. steps, price, time of run, error rates

NGS instruments

- Originally designed for whole genome sequencing
- Very large seq capacity

NGS instruments

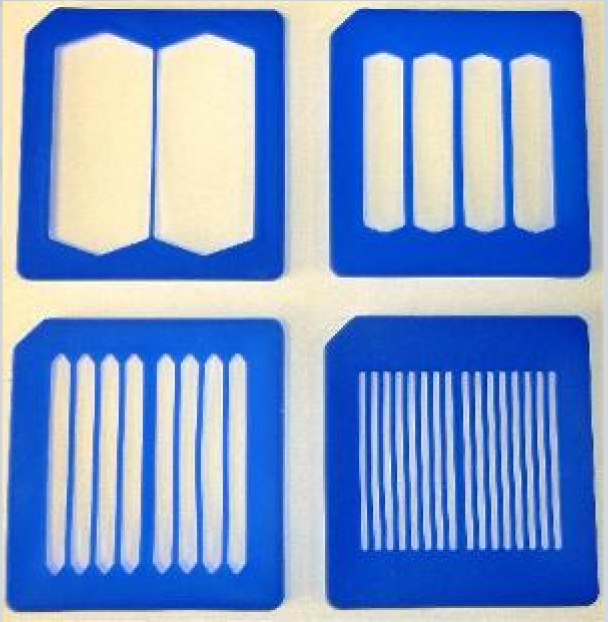
Instrument	Run time[a]	Millions of reads/ run	Bases / read[b]	Yield MB/run
3730xl (capillary)	2 hrs.	9.6 x 10^{-5}	650	0.06
PacBio RS	≤ 2 hrs.	0.03	> 3,000	100-150
454 GS Jr. Titanium	10 hrs.	0.1	400	50
Ion Torrent – ‘314’ chip	4 hrs.	0.1	400	40
454 FLX Titanium	10 hrs.	1	400	400
454 FLX+	20 hrs.	1	650	650
Ion Torrent – ‘316’ chip [e]	4 hrs.	1.6	400	400
Illumina MiSeq – version 1	26 hrs.	4	150+150	1,200
Ion Torrent – ‘318’ chip	7 hrs.	4	400	1,500
Ion Torrent – Proton I	≤ 4 hrs.	70	≤ 200	10,000
MiSeq – v. 2	39 hrs.	15	250+250	7,500
Illumina GAIIx	14 days	300	150+150	96,000
Ion Torrent – Proton III [e]	[> 4 hrs.]	[500]	[≤ 200]	[100,000]
Illumina HiSeq 2500 – rapid [e]	40 hrs.	≤ 600	150+150	≤ 180,000
SOLiD – 5500xl [e]	8 days	> 1,410[d]	75+35	155,100
Illumina HiSeq 2000	11.5 days	≤3000	100+100	≤600,000

http://www.molecularecologist.com/next-gen-table-2a-2013/

Multiplexing: sequencing multiple samples in parallel

- Physical separation of different samples
  - Space: different lanes for different samples
  - Time: different runs for different

- Waste of time and consumables

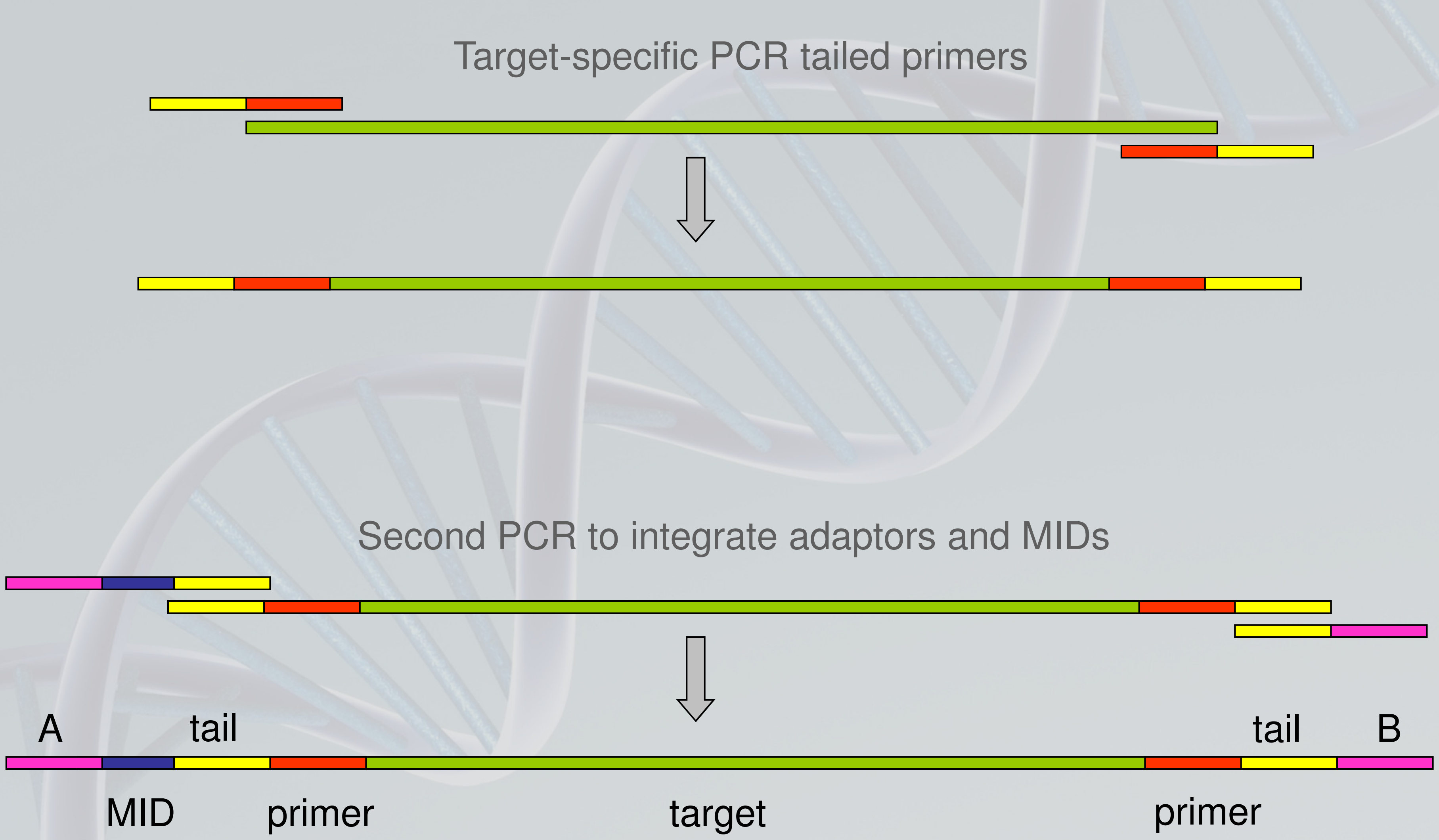


Roche 454 gaskets

Multiplexing: sequencing multiple samples in parallel

- Indexing / Barcoding / MID-tagging
- Use of Molecular Identifier tags (MIDS)
  - Short (3-16 nucleotide) fragments with known sequence
  - Different tag sequence for each sample
  - Inclusion in (amplicon) fragments by ligation/pcr

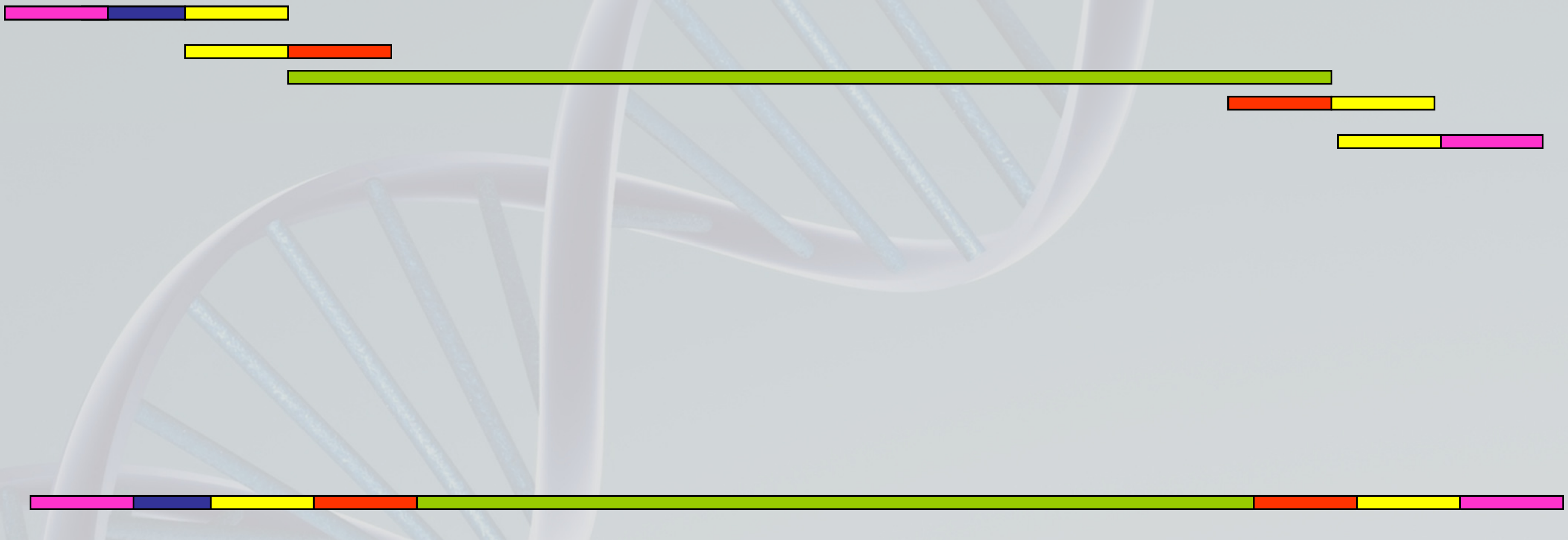
Integration of MIDs by pcr



Target-specific PCR tailed primers

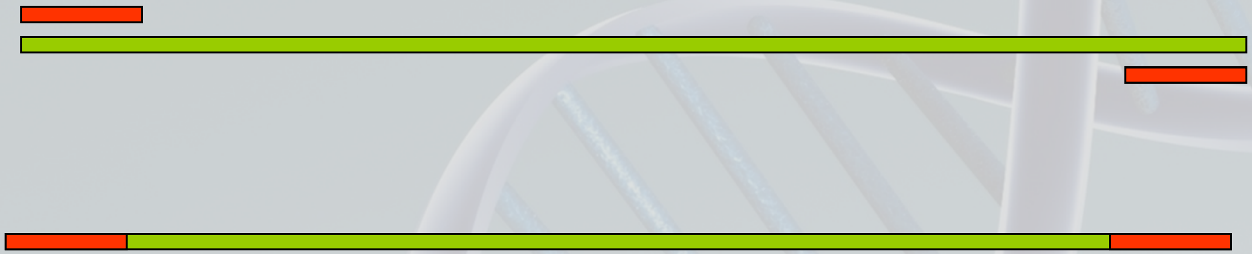
Second PCR to integrate adaptors and MIDs

4-primer pcr

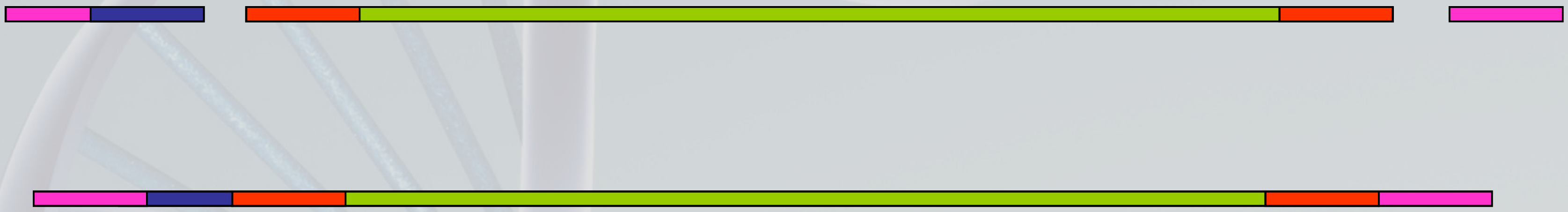


Integration of MIDs by ligation

Target-specific PCR



Integration of adaptors and MIDs by ligation



PCR amplification of amplicons

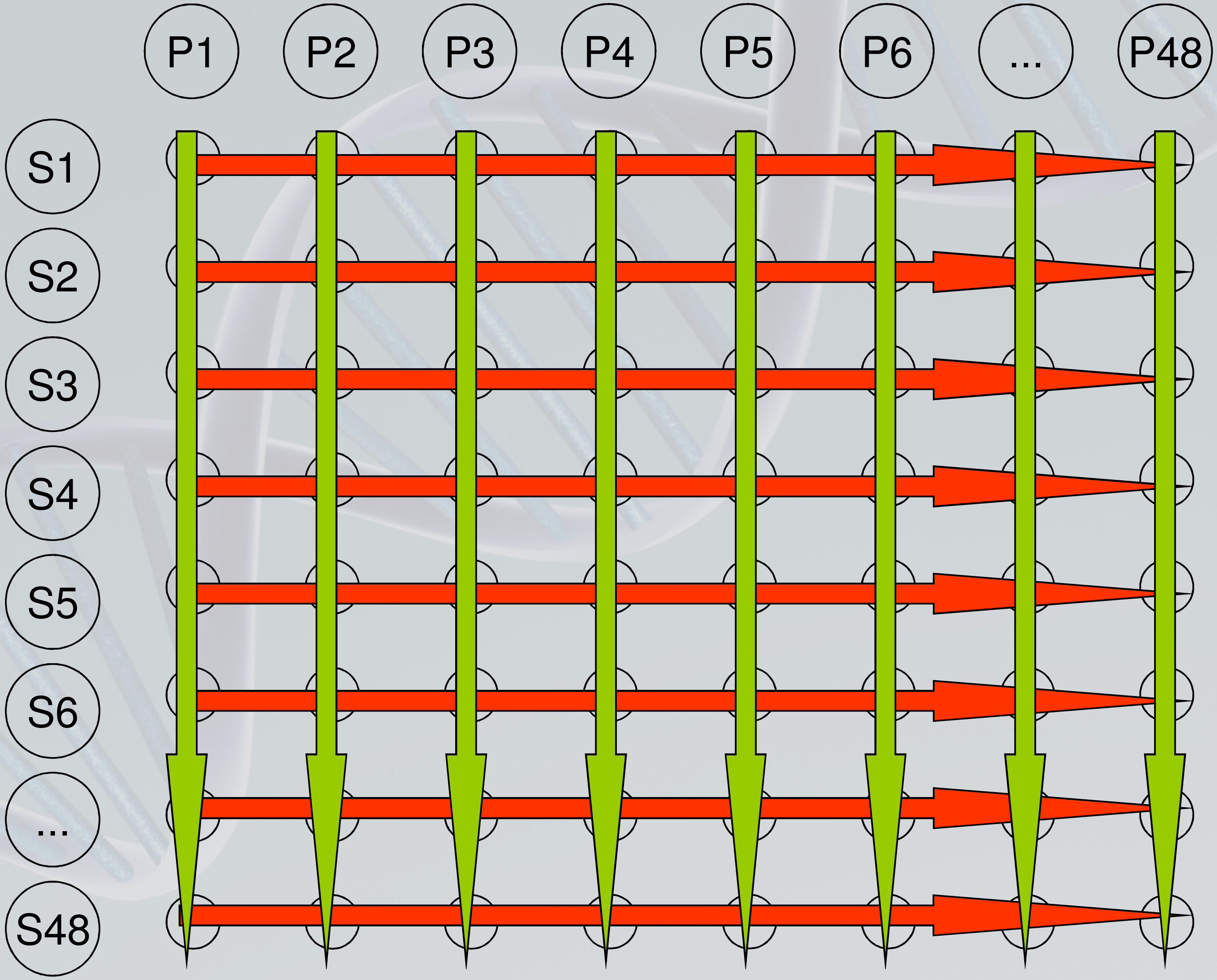
PCR: Fluidigm Access Array

- 48 samples x 48 primers = 2304 reactions in one pcr
- Medium - high throughput of samples

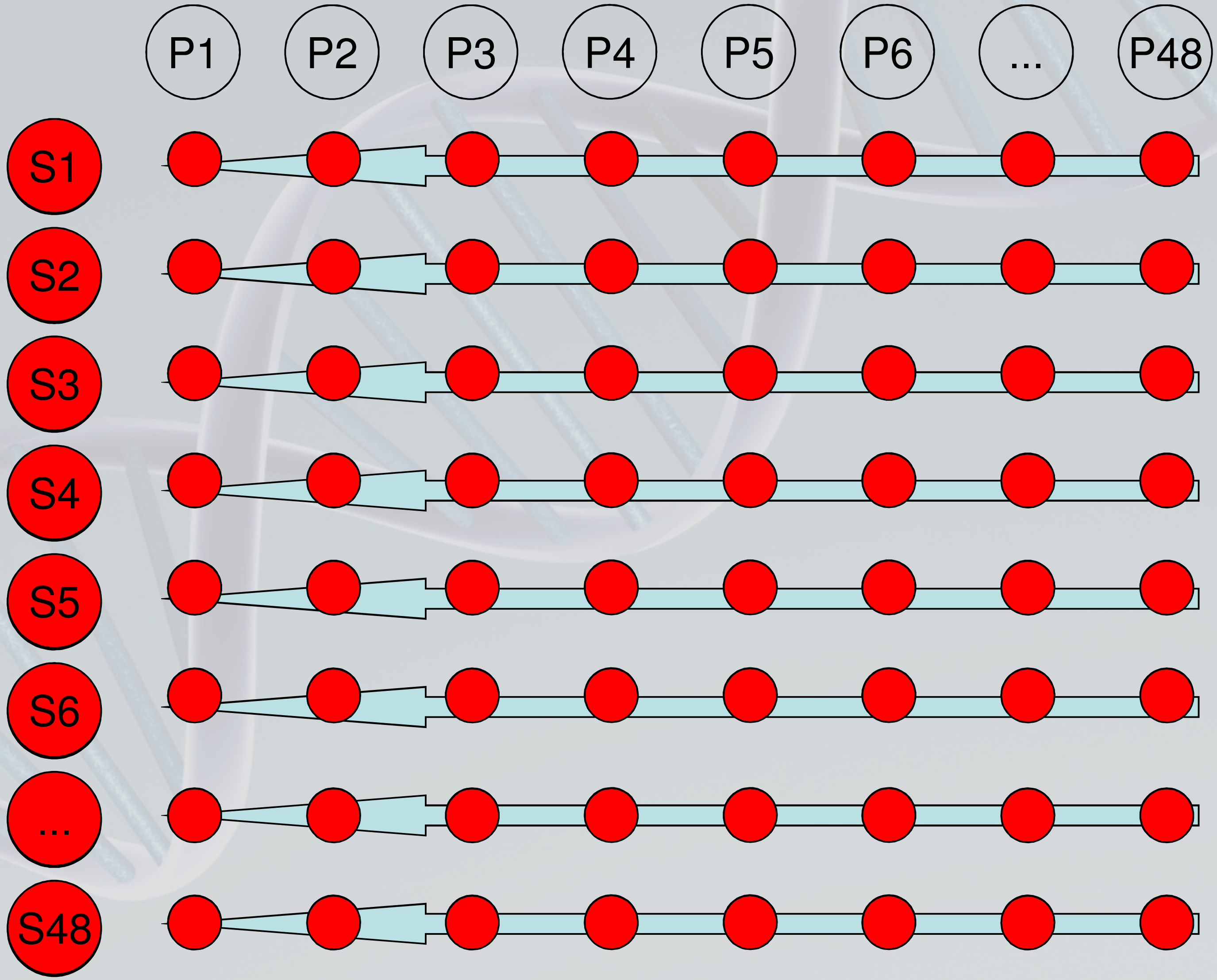
Primers

Samples

Barcode-primer

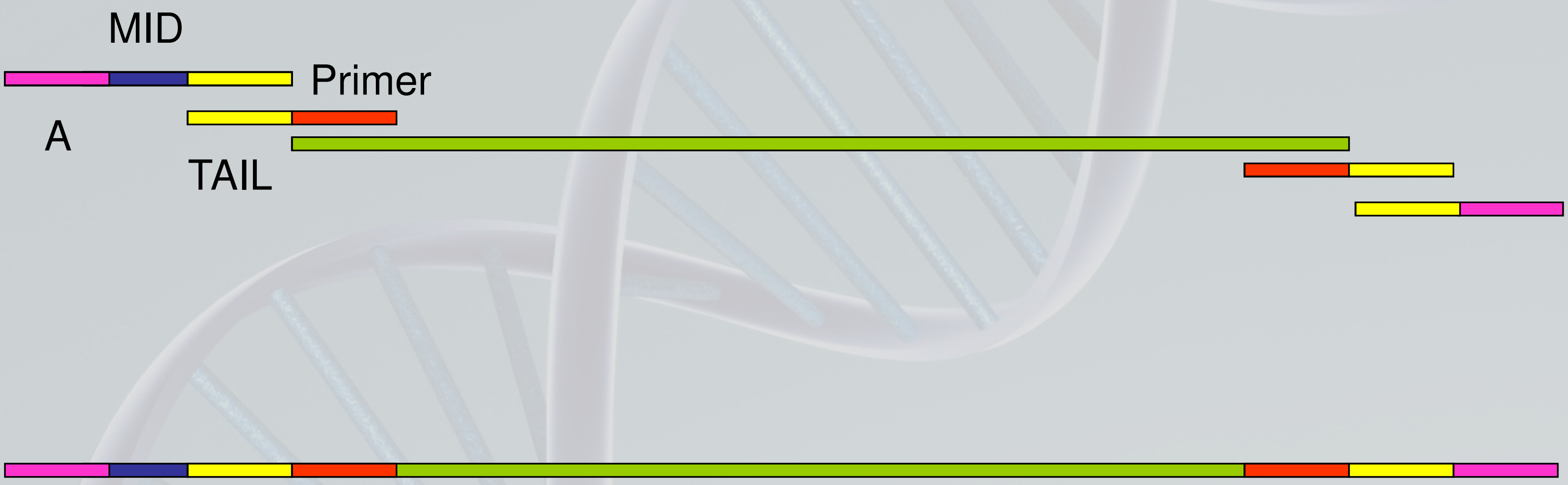


Primers



PCR thermocycling

4-primer pcr



Fluidigm Access Array

- Low quantities of DNA required (50 ng template / 48 pcr reactions)
- Efficient use of reagents
- Time-efficient, limited pipeting
- Avoid multiplexing problems (i.e. primer-primer interactions)
- Equal yield/reaction (within 2-fold)

- Restrictive primer conditions (Ta = 60°C), one PCR protocol

Number of different tags vs tag length ?

- Number of tags = 4^x (x length of tag)

- 1 nt = 4 tags (A,C,T,G)
- 2 nt = 16 tags (AA, AC, AT, AG, CC, CA,...)
- 3 nt = 64 tags
- 4 nt = 256 tags
- 5 nt = 1024 tags
- ...
- 10 nt = 1 048 576 tags

Not all sequence tags are created equal

- No tags with long homopolymer stretches
- No tags that are self-complementary (hairpin)
- Balanced GC content ($40\% < GC < 60\%$)
- Tags should be robust against insertions – deletions – pcr/sequencing errors

Faircloth & Glenn (2012). Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels.

Platform error rates

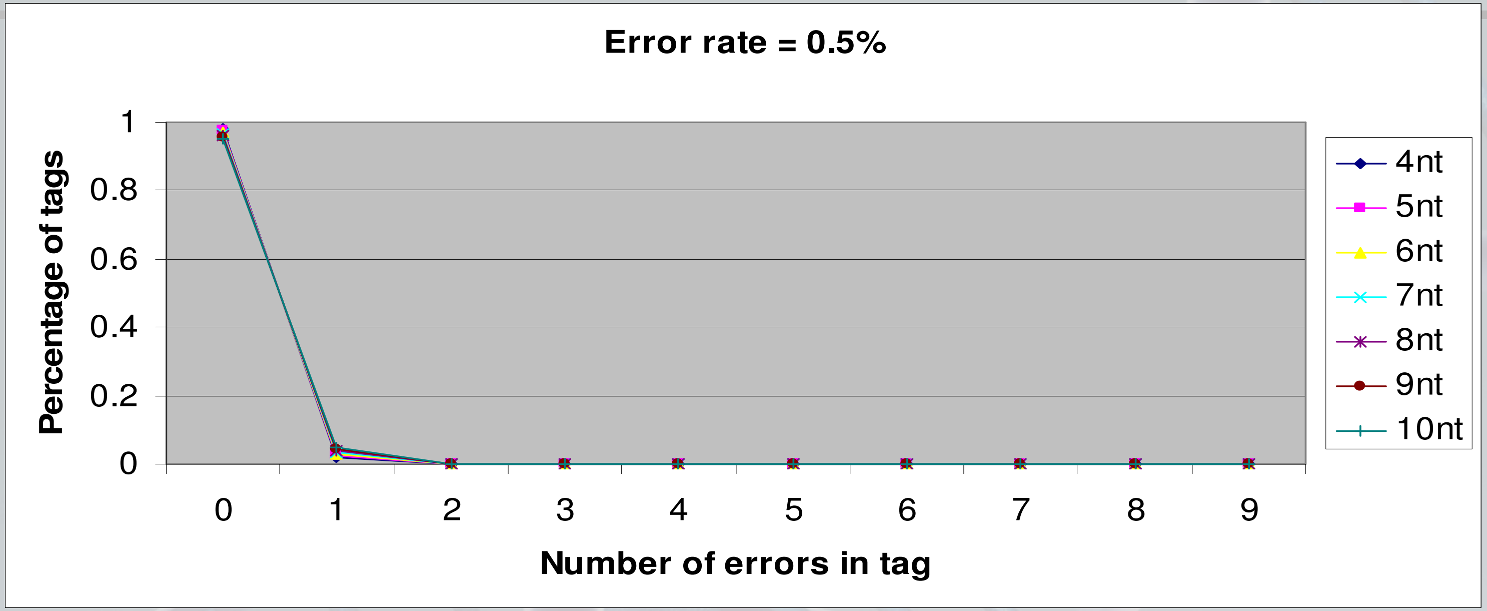
Instrument	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
3730xl (capillary)	Substitution	0.1-1	0.1-1
454 All models	Indel	1	1
Illumina All Models	Substitution	~0.1	~0.1
Ion Torrent – all chips	Indel	~1	~1
SOLiD – 5500xl	A-T bias	~5	≤0.1
Oxford Nanopore	Deletions	≥4*	4*
PacBio RS	CG deletions	~13	≤1

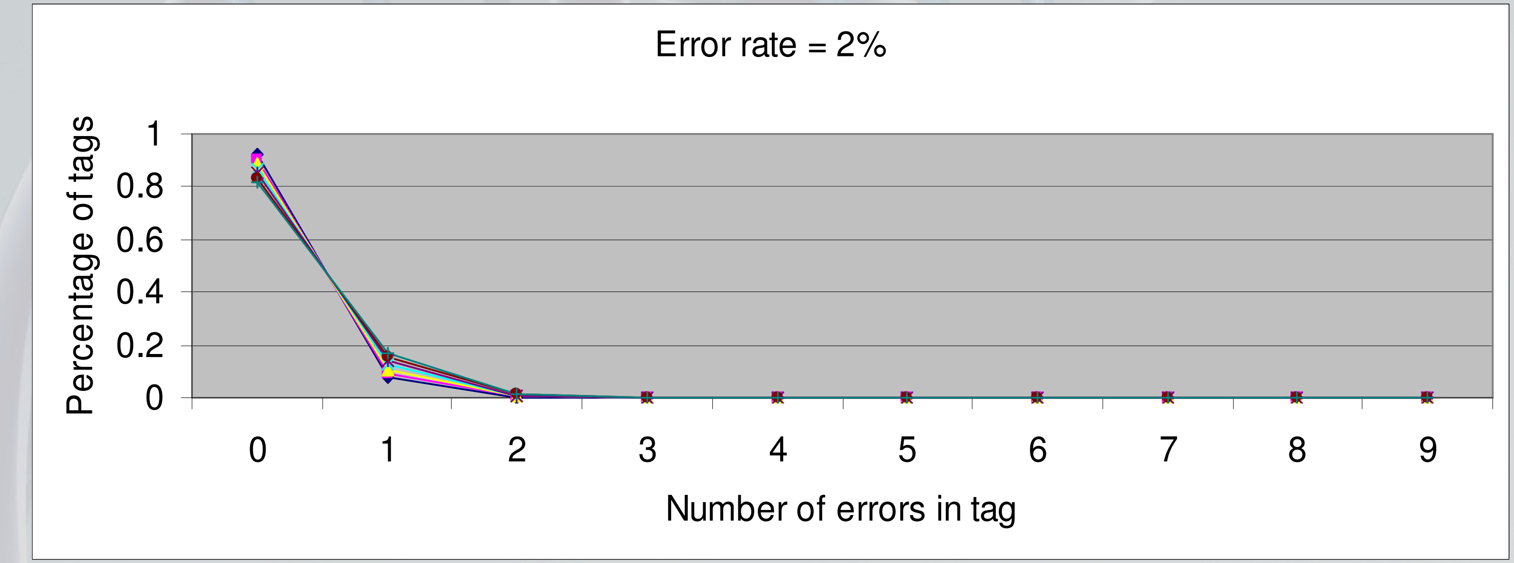
http://www.molecularecologist.com/next-gen-table-3c-2013/

Platform error rates

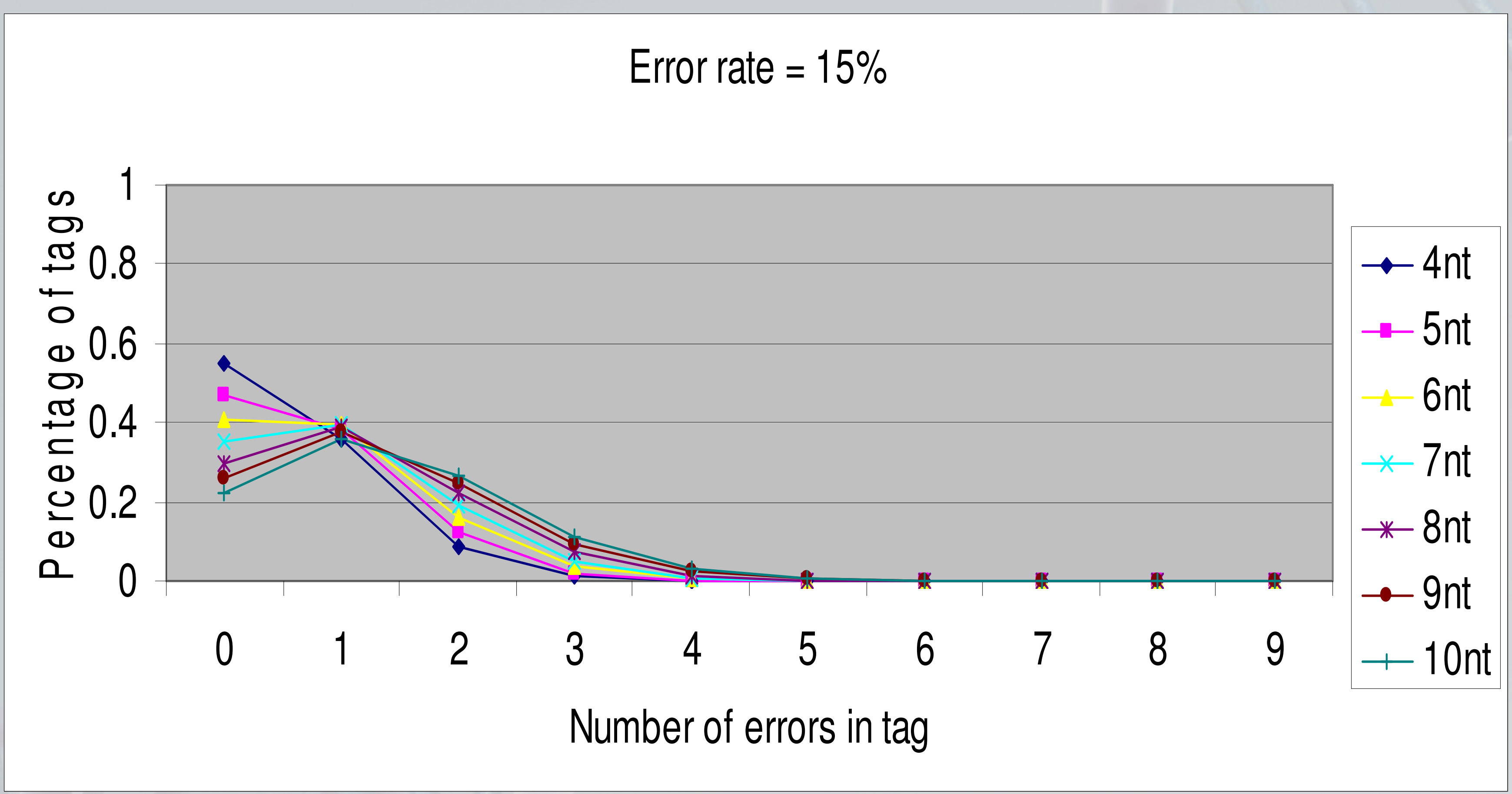
- Read depth or (high) coverage to counter error rates
- Multiplexing: sorting or demultiplexing samples before counting coverage

Influence of error rate on tags

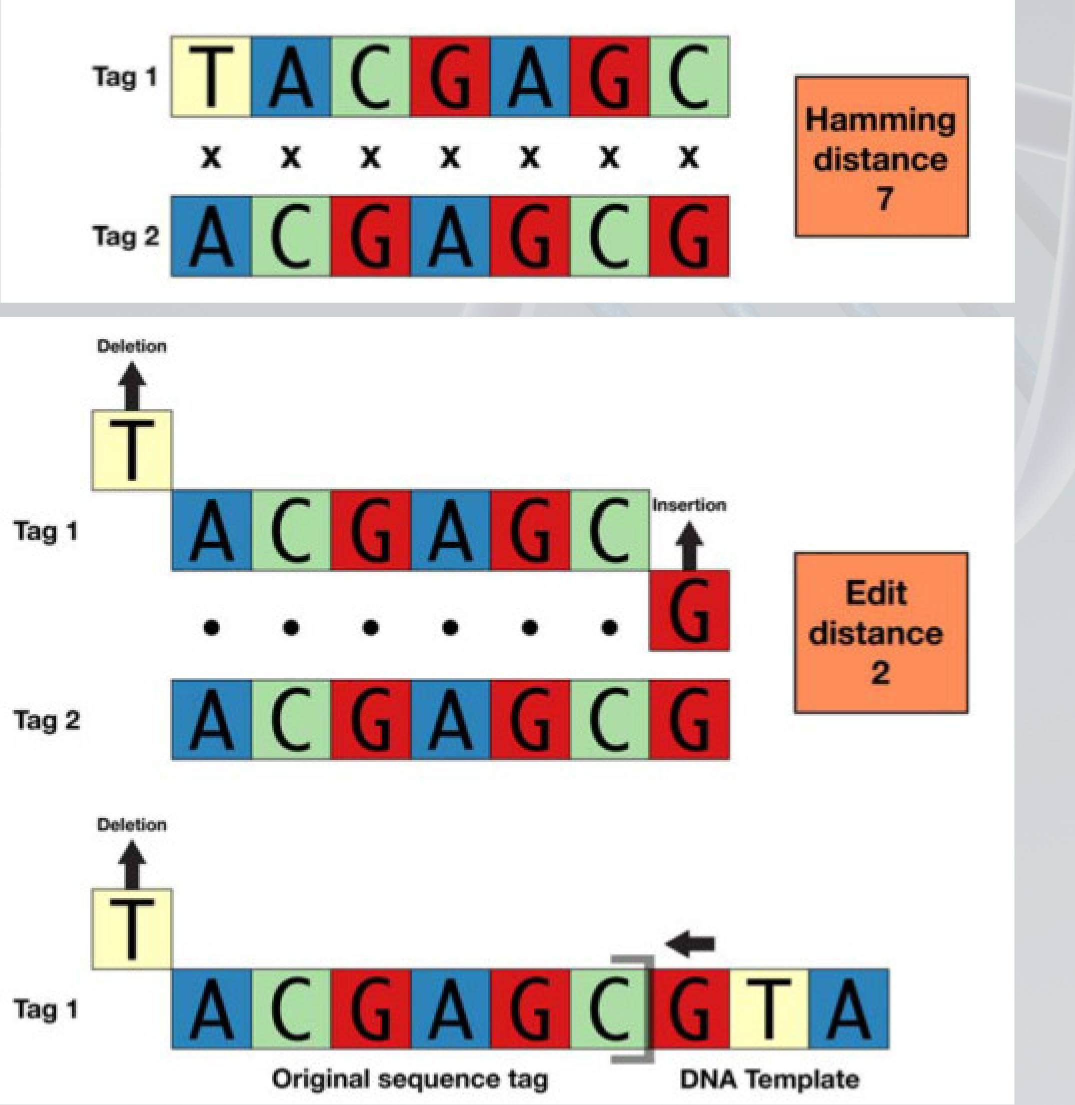




Influence of error rate on tags



Risk of mixing samples



Faircloth & Glenn (2012)

Tag sets with edit distance

TAG length	EDIT distance						
	3	4	5	6	7	8	9
4	7						
5	25	7					
6	61	15	5				
7	211	41	11	4			
8	531	103	24	8	3		
9	1936	301	62	18	6	3	
10	7198	971	164	40	14	5	3

Faircloth & Glenn (2012)

Normalisation of pcr products

- Avoid excessive presence of reads of one or a few reactions with high yield
- Balance the number of reads over sample x locus combinations

- Agilent Bioanalyser
- Picogreen measurement
- Normalisation plates (Sequalprep, ABI)

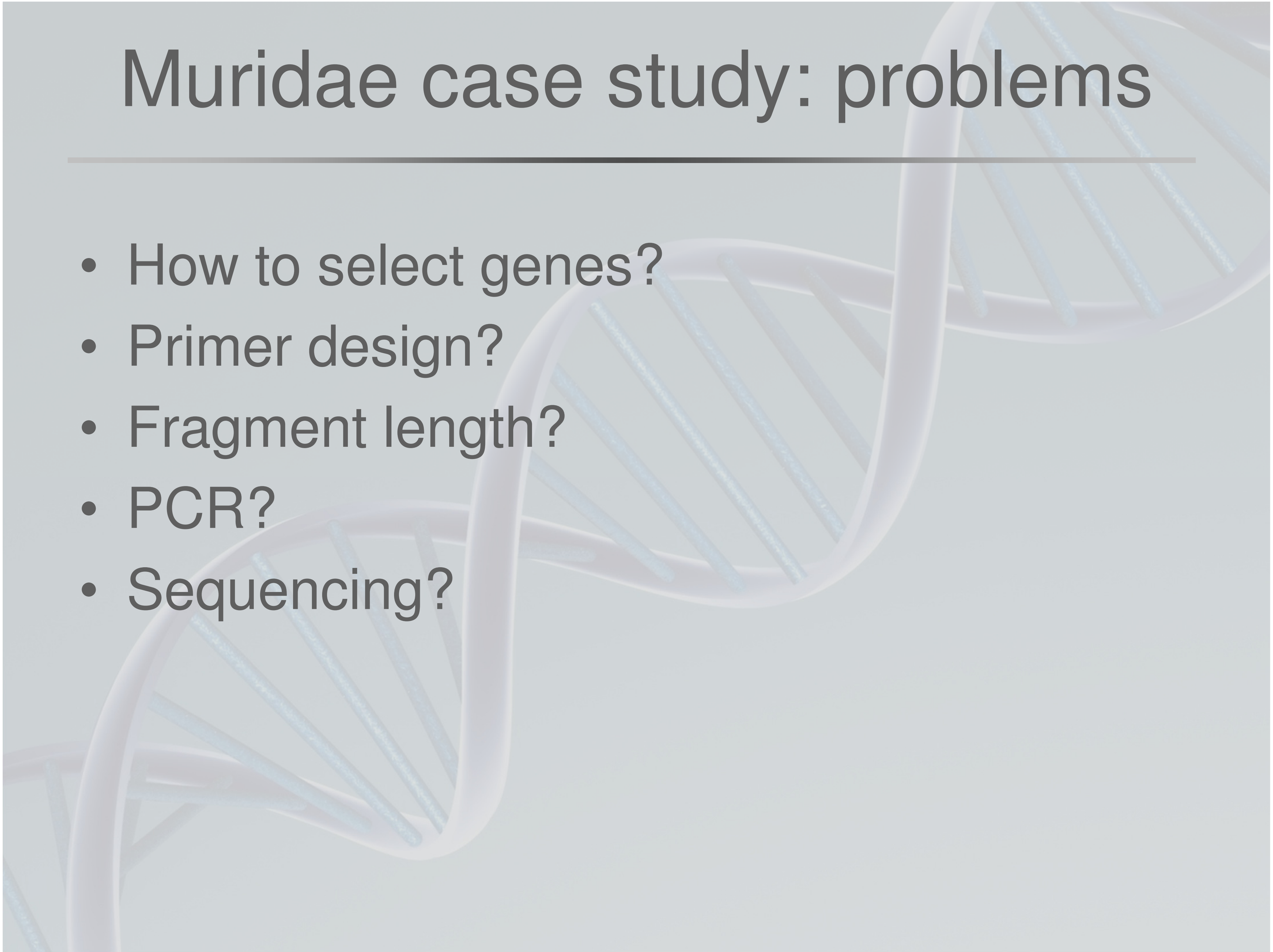
Muridae case study

- Objective:

  Construct a genetic phylogeny of the Muridae using 48 loci , randomly distributed over the genome

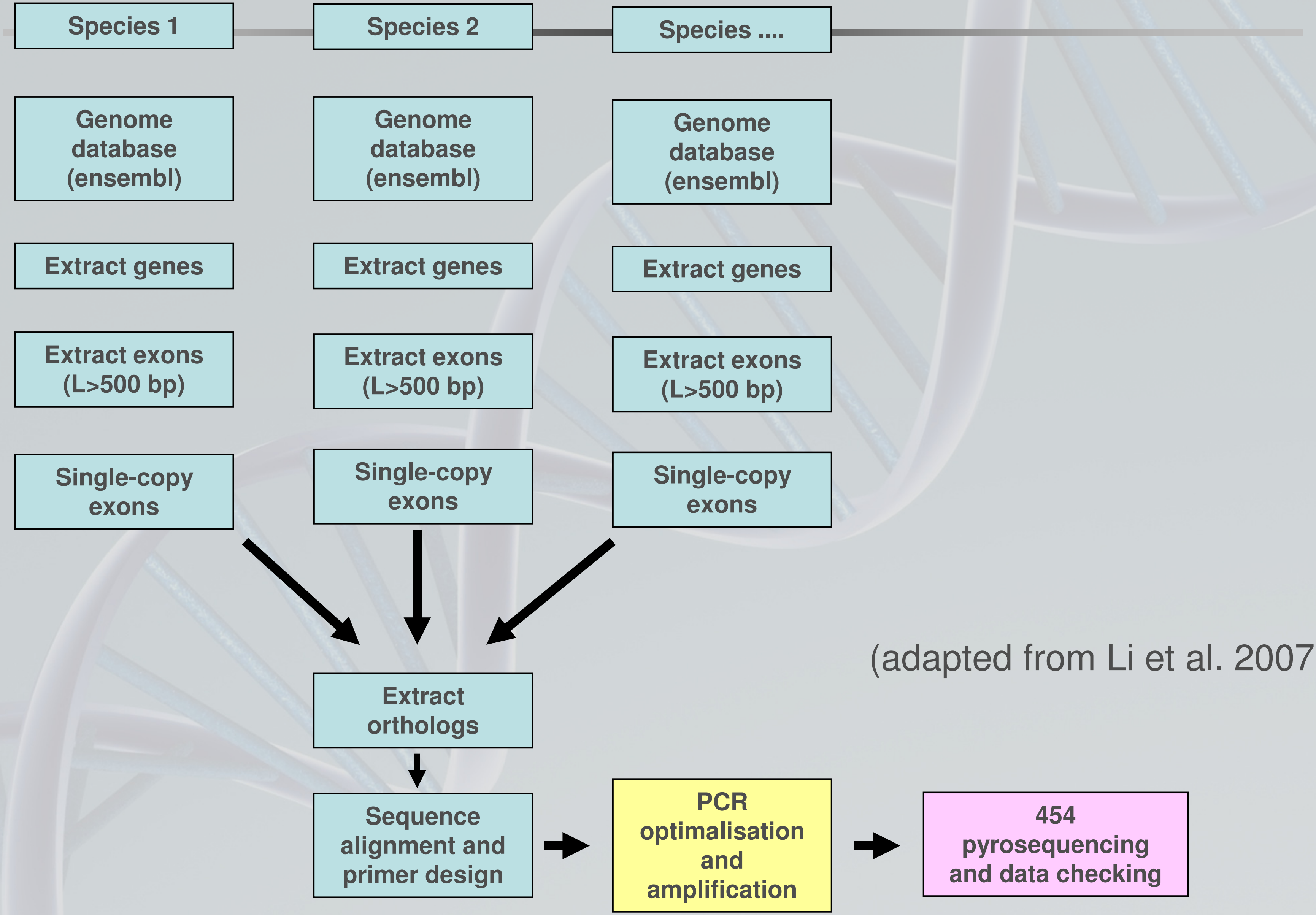
  Test case for further use on museum samples

Muridae case study: problems



- How to select genes?
- Primer design?
- Fragment length?
- PCR?
- Sequencing?

Workflow: visual overview



Species 1

Genome database (ensembl)

Extract genes

Extract exons (L>500 bp)

Single-copy exons

Species 2

Genome database (ensembl)

Extract genes

Extract exons (L>500 bp)

Single-copy exons

Species ....

Genome database (ensembl)

Extract genes

Extract exons (L>500 bp)

Single-copy exons

Extract orthologs

Sequence alignment and primer design

PCR optimalisation and amplification

454 pyrosequencing and data checking

(adapted from Li et al. 2007)

Bio-informatics

- Selection of single copy orthologs
  - Mouse vs rat
  - Mouse vs rat vs Guinea pig (more conserved)

Species	*Mus musculus*	*Rattus norvegicus*	*Cavia porcellus*
# genes	34 221	28 111	23 824
# exons	375 777	240 318	196 401
# exons >500 nt	45 463	18 151	7 141
# single copy exons	7576	5029	2567
# orthologs mouse vs. rat	1371		
# orthologs mouse vs. rat vs. Guinea pig	160		

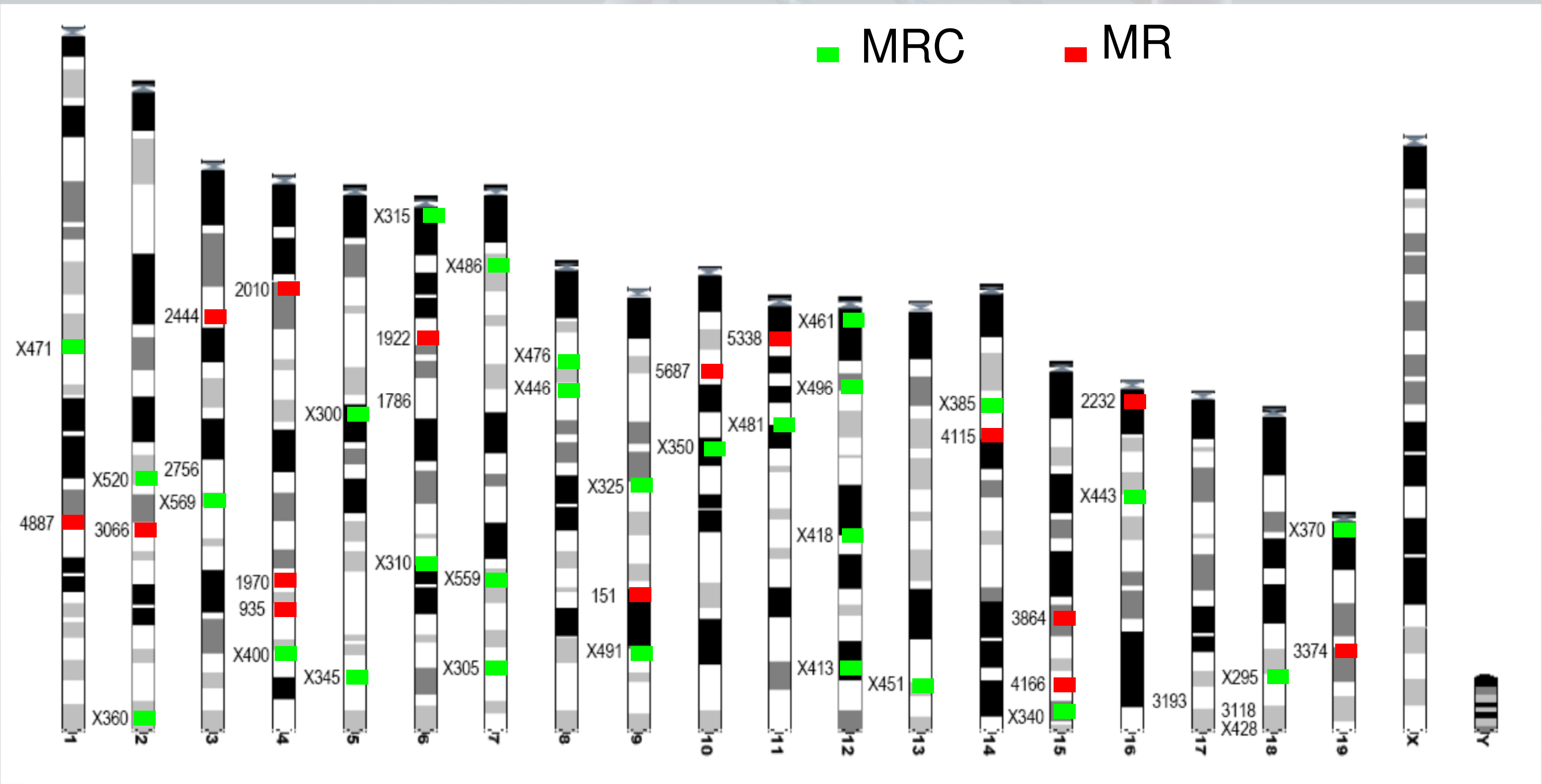
Primer design

- Primer3
- Melting temperature 60°C
- Fragment length 180-250 nt

- 19 primers Mouse Rat (MR)
- 29 primers Mouse Rat Guinea Pig (MRC)

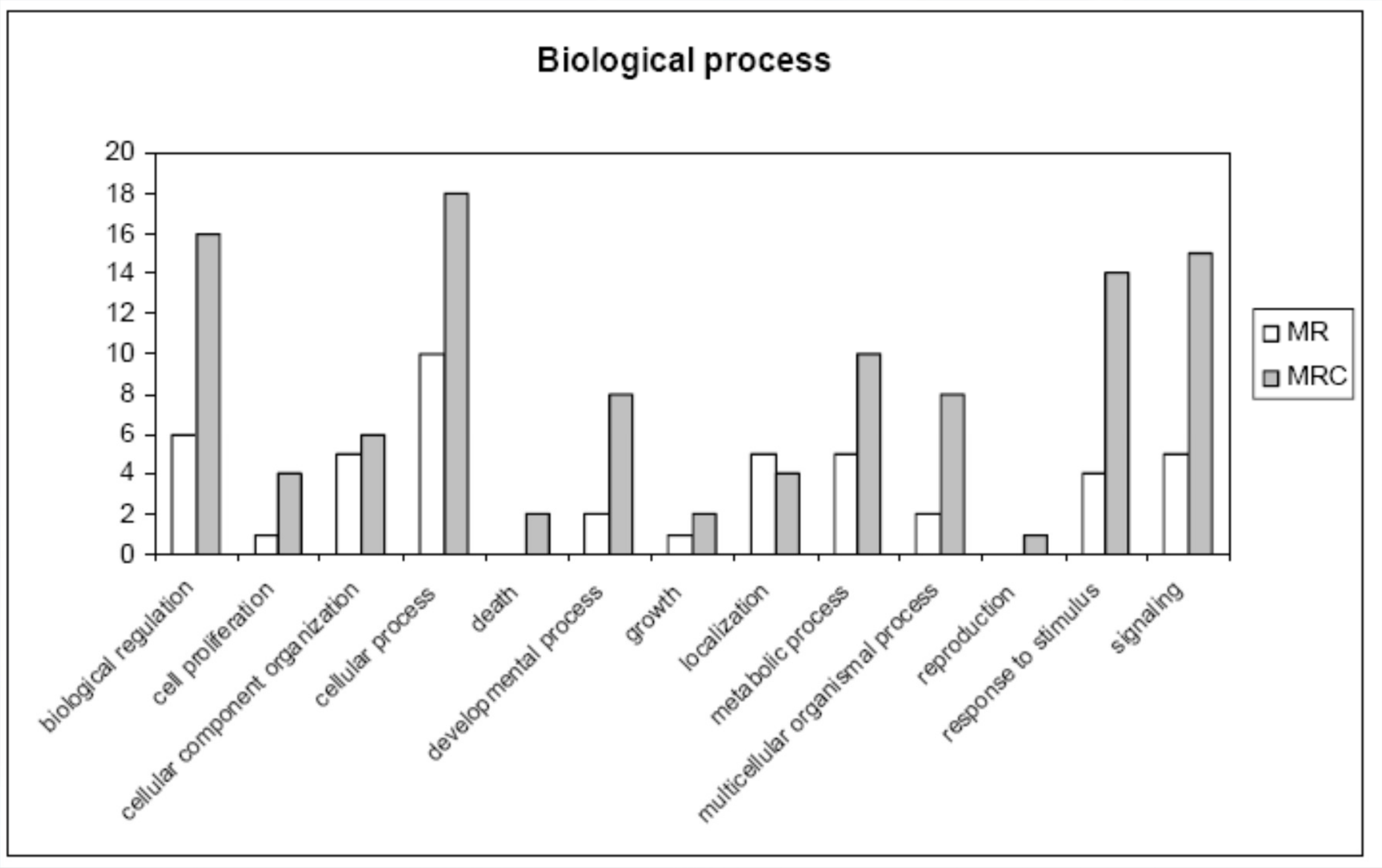
Distribution of selected exons

- Mouse reference genome



Function of selected exons

- Gene Ontology - database



Selection of samples

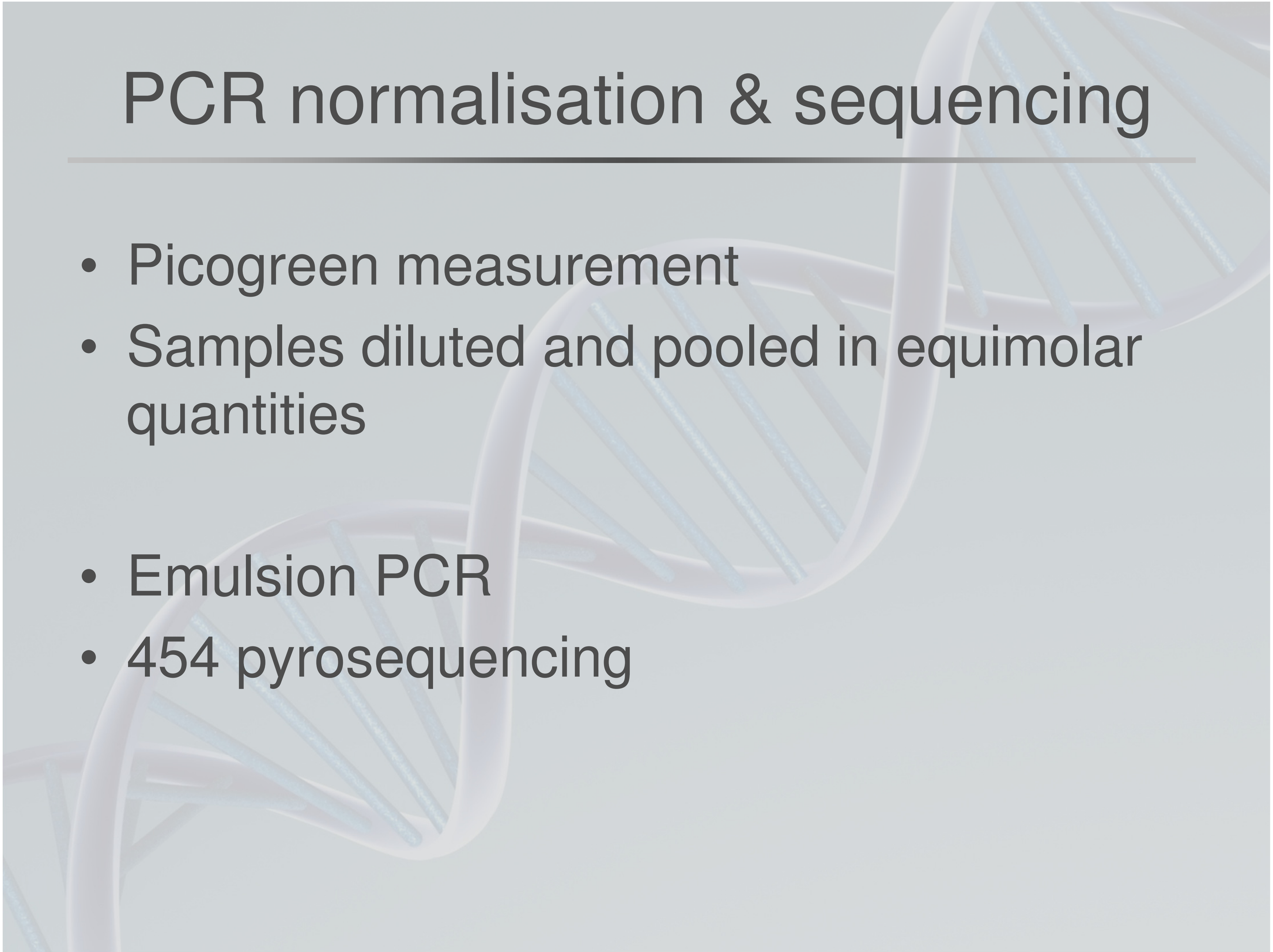
- 37 frozen Muridae samples (5-30 years old)
  - 23 Murinae (Mouse – Rat)
  - 14 Broad selection within Muridae

- 10 Museum preserved specimens (15-100 years old)
- One blanco sample

Fluidigm 4-primer PCR

- Fluidigm system
- 48 samples x 48 markers = 2304 combinations
- 4-primer pcr, MID tags
- Simplex reactions: microdroplets
- Pooled per sample

PCR normalisation & sequencing



- Picogreen measurement
- Samples diluted and pooled in equimolar quantities

- Emulsion PCR
- 454 pyrosequencing

454 run: expected coverage

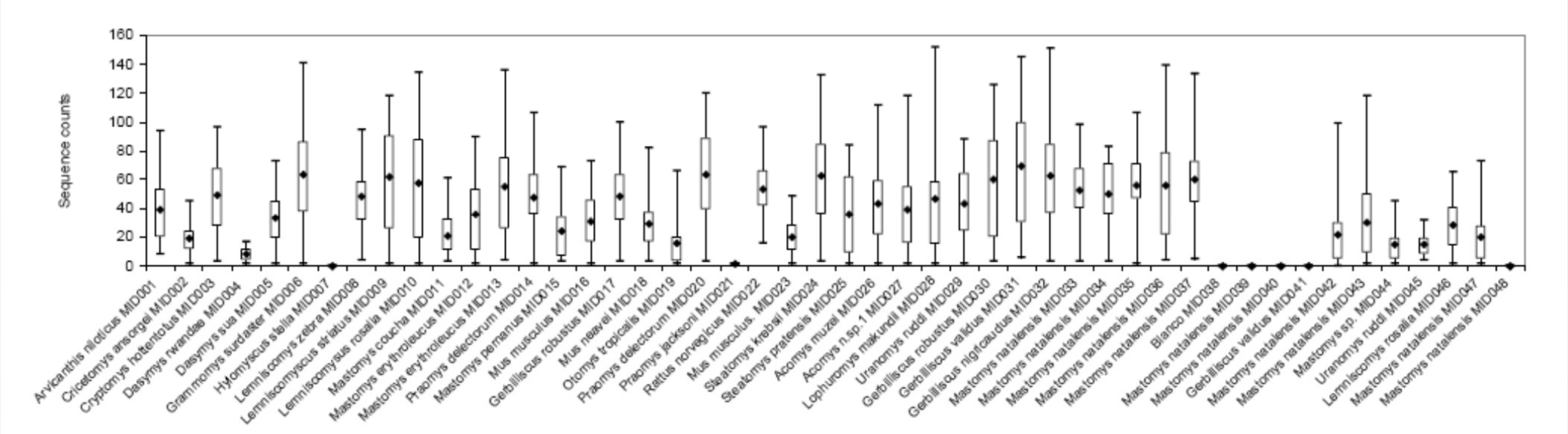
Size	Expected reads (x 1000)	Expected different fragments (48x48)	Expected coverage / fragment
Full plate	900 - 1300	2304	390 - 564
1/2	450 - 650	2304	195 - 282
1/4	160 – 250	2304	69 – 108
1/8	80 – 120	2304	35 – 52
1/16	25 – 40	2304	11 - 17

Sequencing results

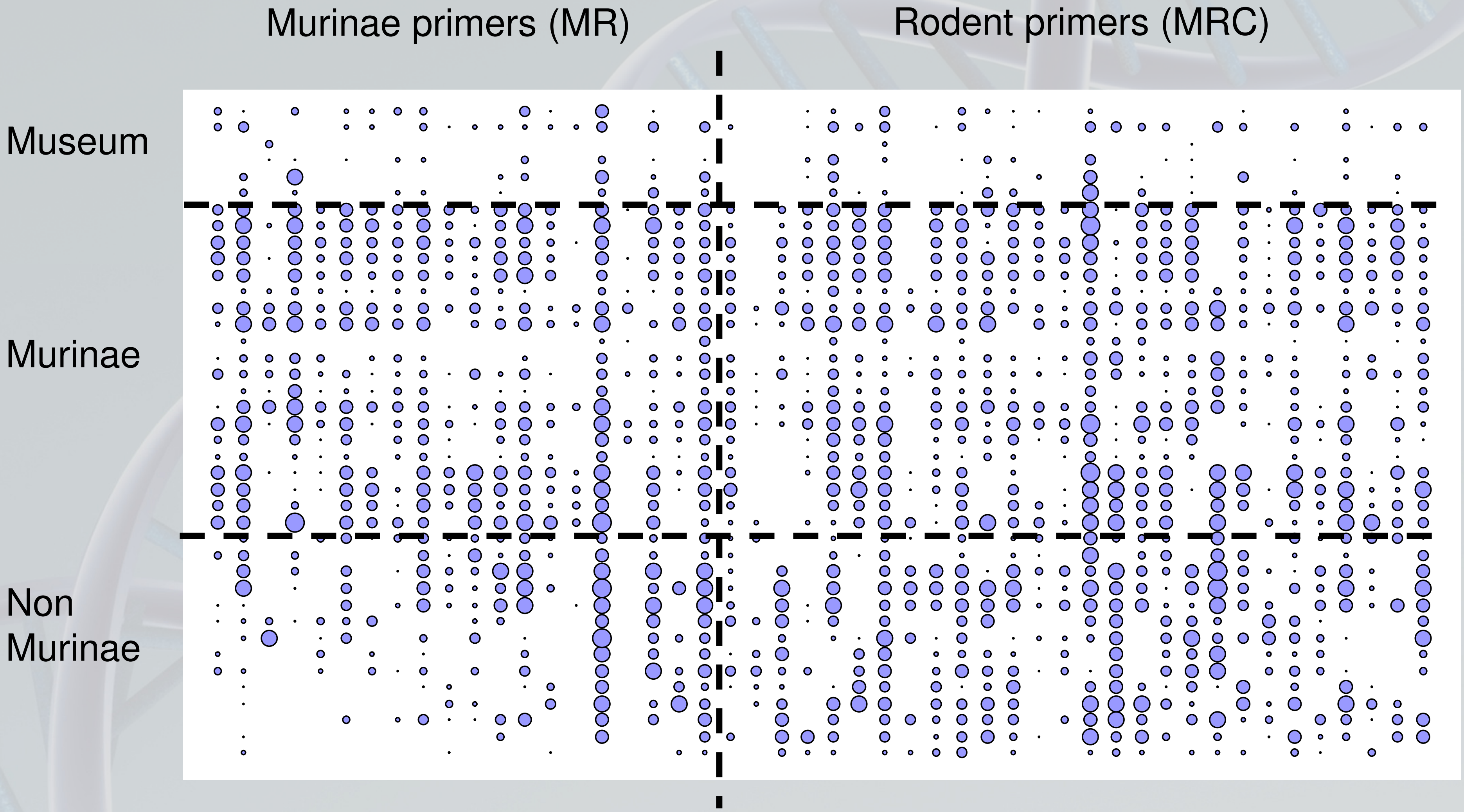
- 96 339 reads
- 62 500 retained after sorting by sample and locus
- 1399 positive sample-locus combinations
- 905 negative sample-locus combinations

Coverage / sequencing depth

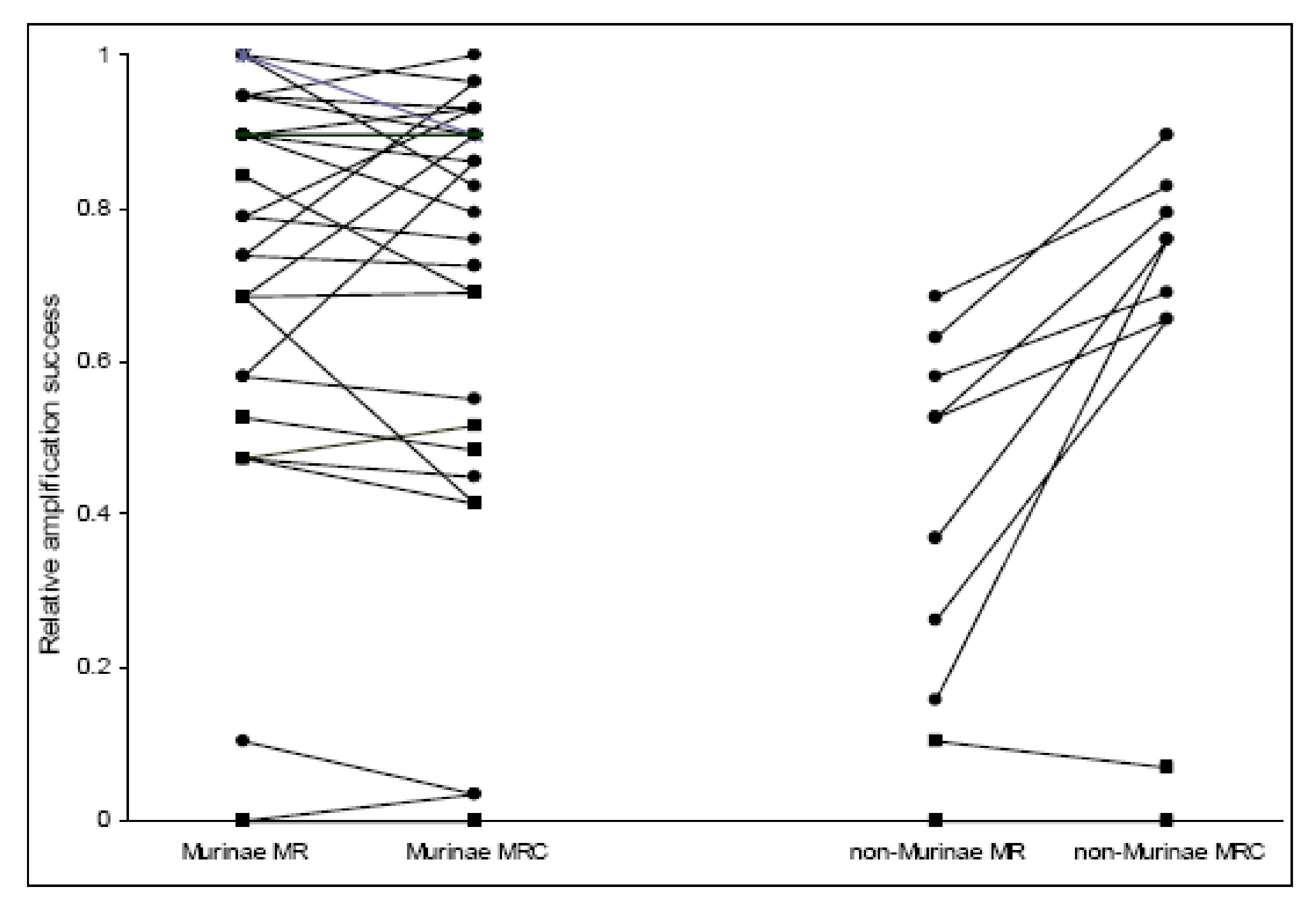
- Mean coverage/seq depth 42x
- Max read count = 147 (within 4x of mean)
- Museum samples lower number of reads



Sample/locus amplification success



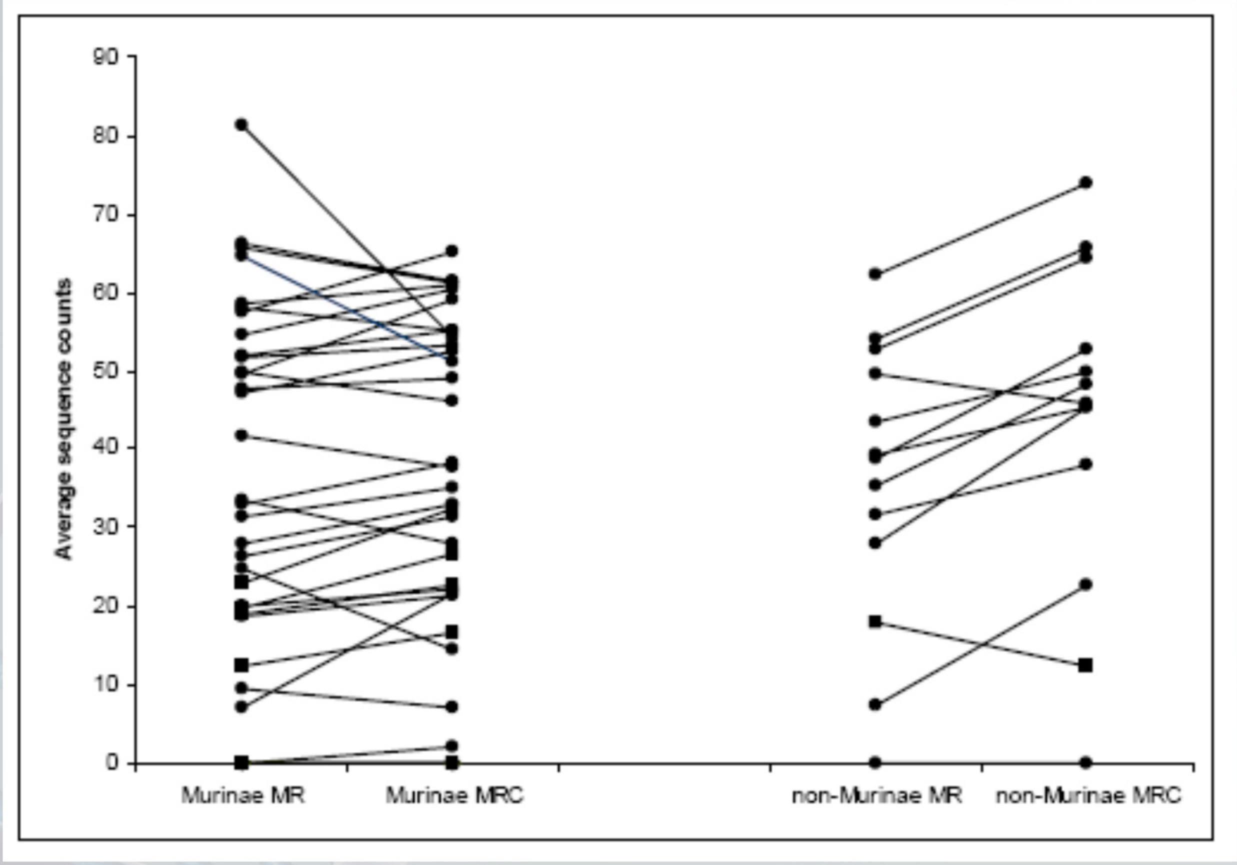
Relative amplification success



Sign test: p = 0.360

Sign test: p = 0.001

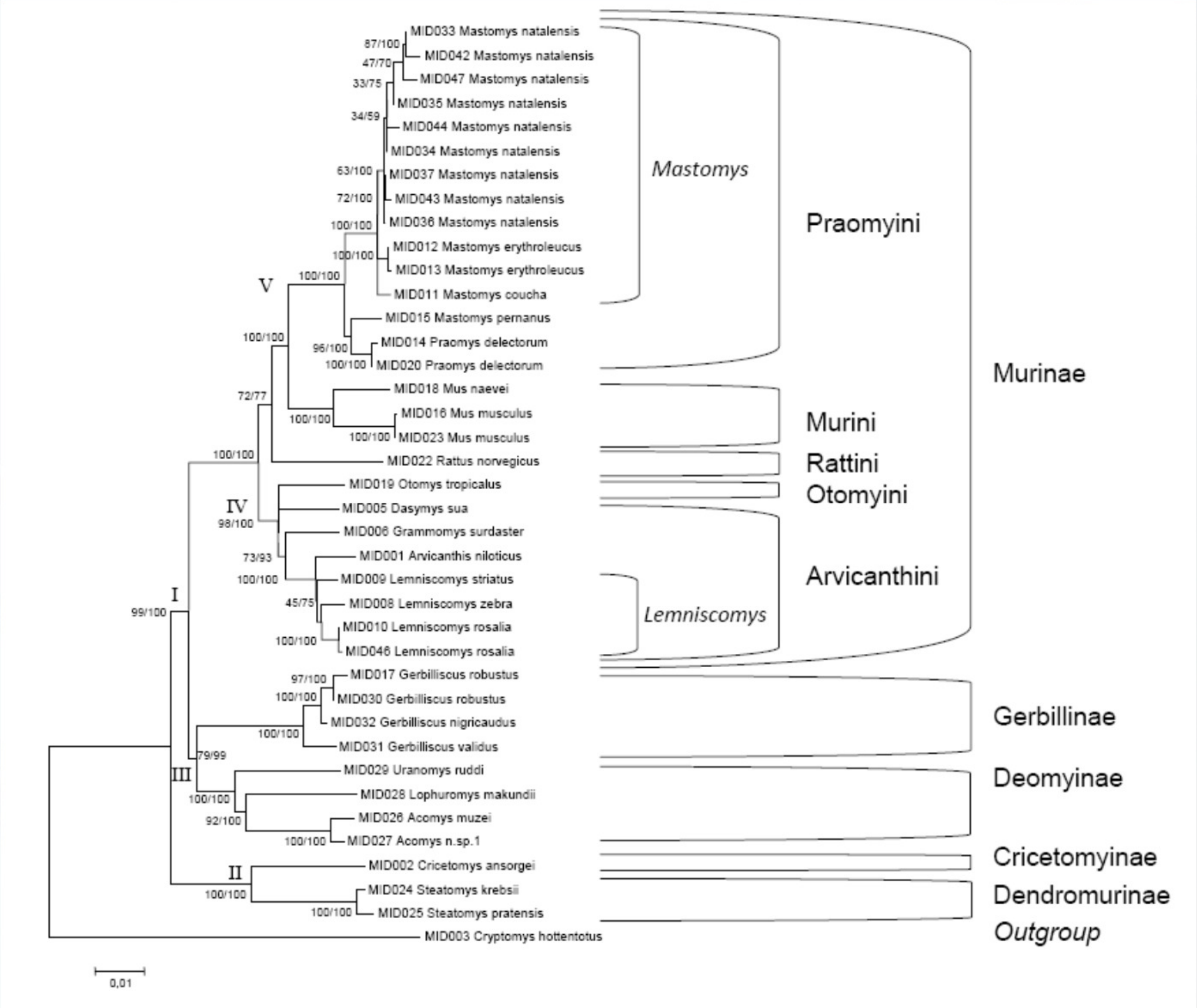
Average read counts



Sign test: p = 0.09

Sign test: p = 0.037

Muridae phylogeny


MID033 Mastomys natalensis
87/100
MID042 Mastomys natalensis
47/70
MID047 Mastomys natalensis
33/75
MID035 Mastomys natalensis
34/59
MID044 Mastomys natalensis
MID034 Mastomys natalensis
63/100
MID037 Mastomys natalensis
72/100
MID043 Mastomys natalensis
100/100
MID036 Mastomys natalensis
MID012 Mastomys erythroleucus
100/100
MID013 Mastomys erythroleucus
V
100/100
MID011 Mastomys coucha
MID015 Mastomys pernanus
100/100
96/100
MID014 Praomys delectorum
100/100
MID020 Praomys delectorum
MID018 Mus naevei
72/77
MID016 Mus musculus
100/100
100/100
MID023 Mus musculus
100/100
MID022 Rattus norvegicus
MID019 Otomys tropicalus
IV
MID005 Dasymys sua
98/100
MID006 Grammomys surdaster
73/93
MID001 Arvicanthis niloticus
100/100
MID009 Lemniscomys striatus
I
45/75
MID008 Lemniscomys zebra
99/100
MID010 Lemniscomys rosalia
100/100
MID046 Lemniscomys rosalia
97/100
MID017 Gerbilliscus robustus
100/100
MID030 Gerbilliscus robustus
MID032 Gerbilliscus nigricaudus
100/100
MID031 Gerbilliscus validus
79/99
III
MID029 Uranomys ruddi
100/100
MID028 Lophuromys makundii
92/100
MID026 Acomys muzei
100/100
MID027 Acomys n.sp.1
II
MID002 Cricetomys ansorgei
100/100
MID024 Steatomys krebsii
100/100
MID025 Steatomys pratensis
MID003 Cryptomys hottentotus
0,01
Mastomys
Praomyini
Murinae
Murini
Rattini
Otomyini
Arvicanthini
Lemniscomys
Gerbillinae
Deomyinae
Cricetomyinae
Dendromurinae
Outgroup