# Detection of genetic manipulation in thoroughbred racehorses – a new frontier in doping control

Jillian Maniego, James Scarth, and Edward Ryder
Sport and Specialised Analytical Services, LGC, Newmarket Road, Fordham, Cambridgeshire CB7 5WW, UK
Edward.Ryder@LGCgroup.com

## Introduction

Throughout the history of horse racing, doping techniques used to suppress or enhance performance have expanded to match the technology available. Examples range from simple chemicals such as caffeine, to misuse of steroids and peptides used in human and veterinary medicine.

The next frontier in doping, both in the equine and human sports areas, is predicted to be genetic manipulation (Wilkin *et al*, 2017), either by germline editing or gene therapy. Detection will not only ensure fairness in racing but act as a deterrent against initialising such activities, contributing to the health status and continued welfare of the animals.

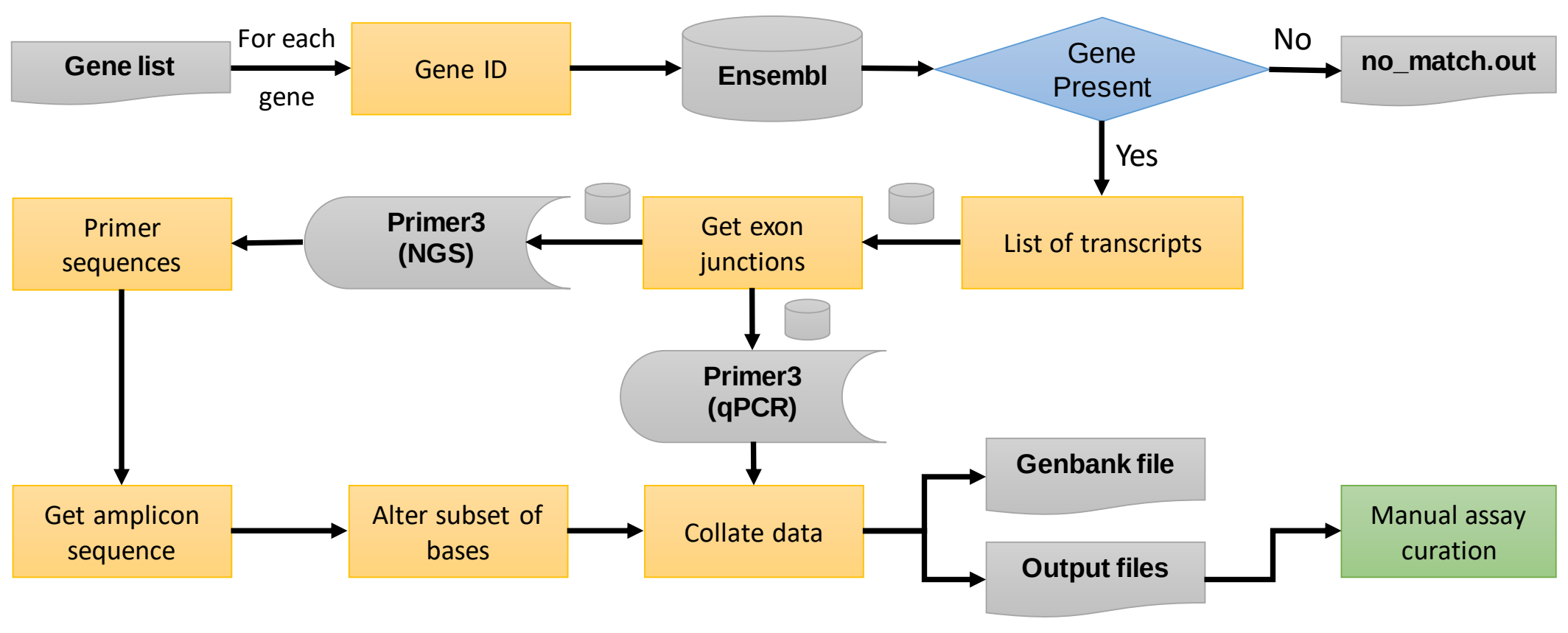
We highlight here our research, funded by the British Horseracing Authority (BHA), to screen for the presence of exogenous transgenes by use of automated assay design, PCR and Illumina parallel sequencing. This system has the potential benefits of high sensitivity and scalability to a large number of samples and targets.

## Gene choice and assay design

Genes with a potential role in performance or injury recovery were identified by literature search or analysis of RNAseq data (Parks *et al*, 2001).

Due to packaging size limits of AAV and plasmid vectors commonly used in gene therapy experiments, transgenes are usually constructed with the introns (non-coding regions) of genes removed. Discriminatory assays designed across the exon boundaries should, therefore, not detect the endogenous gene. To this end, we created a Perl script (Fig. 1) to interrogate the Ensembl database, retrieve cDNA sequences and design primers across exon-exon boundaries.

Figure 1. Automated assay design. The Perl script makes use of Primer 3 and is species agnostic.



## Altered-sequence reference standards

To validate tests for routine use, reference standards are used as positive controls. The use of such controls, however, has the potential to produce unintended amplification in the test samples.

Figure 2. Example reference sequence design. Altering a subset of bases conserves the CG contents and should reduce biasing of amplification with the transgene sequence.

| | |
|---|---|
| TGCAGGGCCAGGCCCTGTTGGCCAAC | **Endogenous exon sequence** |
| AGCTGGGCCTGGCCCAGAAGGCCAAC | **Internal reference sequence** |

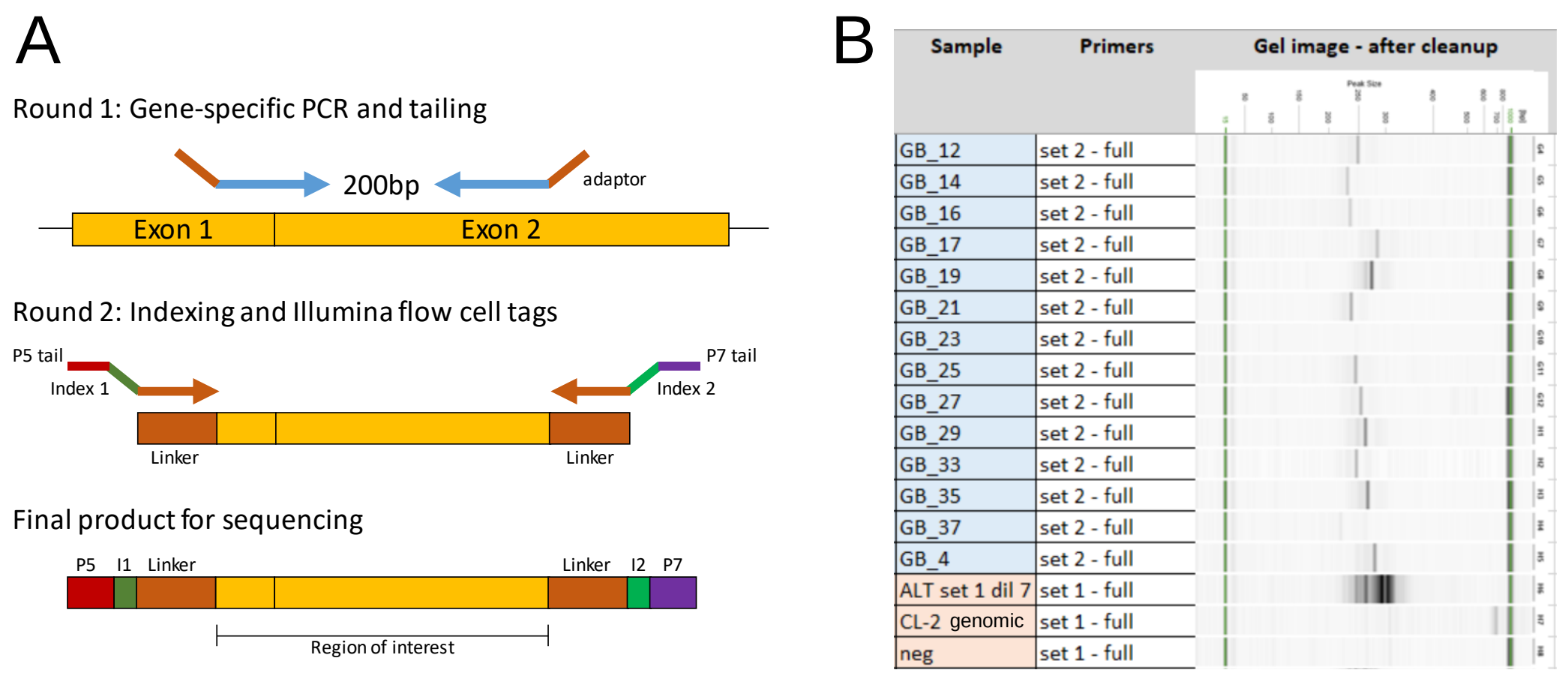
A major advantage of parallel sequencing is the ability to sequence individual strands of a PCR product. This allows us to design standards which differ by only a few bases of DNA and can be discriminated from transgene sequences in the analysis (Fig 2.). The Perl script designs these sequences automatically, and is easily configurable if requirements change. Sequences can be ordered as artificial gene constructs in plasmids or gene blocks.

## Multiplex PCR and Illumina sequencing

### Tailed PCR and indexing

Our method uses a 2-step PCR, similar to metagenomics studies (Pichler *et al*, 2018). An initial, transgene-specific PCR is followed by a second round which adds flow cell adaptors and Nextera indexes for subsequent pooling of up to 384 reactions (Fig. 3A). Products are analysed on an Illumina MiniSeq sequencer using the Mid Output kit and a read length of 2x150bp.

Figure 3. A) PCR tailing and index strategy. B) Amplification of ~100 copies of gene block in a pooled primer set.


A
Round 1: Gene-specific PCR and tailing
200bp
adaptor
Exon 1
Exon 2
Round 2: Indexing and Illumina flow cell tags
P5 tail
Index 1
P7 tail
Index 2
Linker
Linker
Final product for sequencing
P5
I1
Linker
Linker
I2
P7
Region of interest
B
Gel image - after cleanup
Peak Size

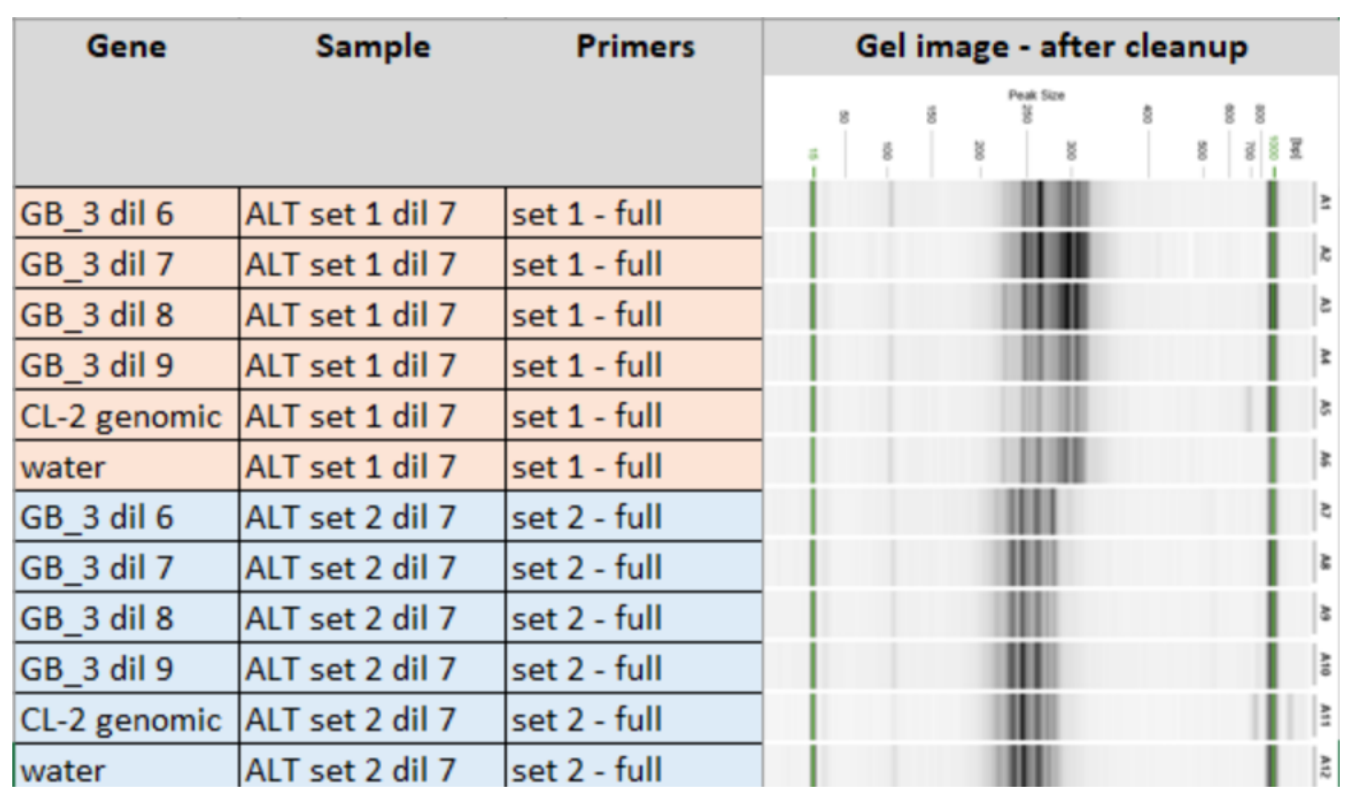
| Sample | Primers |
|---|---|
| GB_12 | set 2 - full |
| GB_14 | set 2 - full |
| GB_16 | set 2 - full |
| GB_17 | set 2 - full |
| GB_19 | set 2 - full |
| GB_21 | set 2 - full |
| GB_23 | set 2 - full |
| GB_25 | set 2 - full |
| GB_27 | set 2 - full |
| GB_29 | set 2 - full |
| GB_33 | set 2 - full |
| GB_35 | set 2 - full |
| GB_37 | set 2 - full |
| GB_4 | set 2 - full |
| ALT set 1 dil 7 | set 1 - full |
| CL-2 genomic | set 1 - full |
| neg | set 1 - full |

### Primer pooling and multiplex PCR

Fifteen genes with two assays per gene (selected to cover the maximum number of transcripts on the Equcab 3.0 genome) were chosen as a pilot,. After determining the specificity of the assays and a lack of amplification in genomic controls, four experiments were initially performed with ~100 copies of target per reaction:

- Ratios of transgene vs altered sequences.
- Amplification of a single gene block within a pooled primer set (Fig. 3B).
- Multiplex PCR of all reference gene targets
- Spiking of a single transgene block at different copy numbers within a multiplex PCR of altered gene block sequences (Fig. 4).

Figure 4. Multiplex PCR of gene blocks and primer pools, spiked with different quantities of transgene sequence GB_3 ranging from 500 to 4 copies in 1:5 dilutions. Gene identifiers have been redacted to preserve the integrity of any future diagnostic tests


Gel image - after cleanup
Peak Size

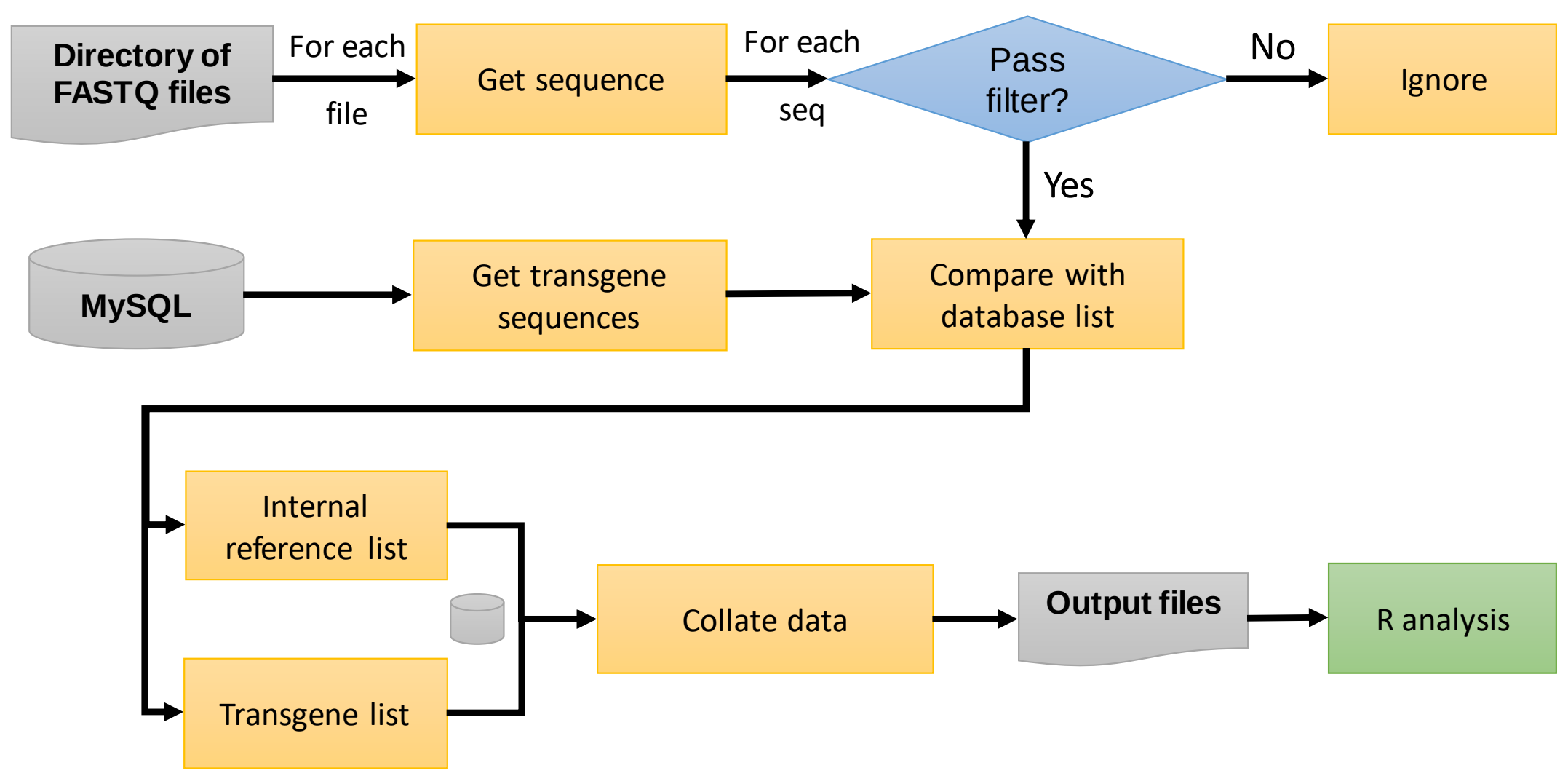
| Gene | Sample | Primers |
|---|---|---|
| GB_3 dil 6 | ALT set 1 dil 7 | set 1 - full |
| GB_3 dil 7 | ALT set 1 dil 7 | set 1 - full |
| GB_3 dil 8 | ALT set 1 dil 7 | set 1 - full |
| GB_3 dil 9 | ALT set 1 dil 7 | set 1 - full |
| CL-2 genomic | ALT set 1 dil 7 | set 1 - full |
| water | ALT set 1 dil 7 | set 1 - full |
| GB_3 dil 6 | ALT set 2 dil 7 | set 2 - full |
| GB_3 dil 7 | ALT set 2 dil 7 | set 2 - full |
| GB_3 dil 8 | ALT set 2 dil 7 | set 2 - full |
| GB_3 dil 9 | ALT set 2 dil 7 | set 2 - full |
| CL-2 genomic | ALT set 2 dil 7 | set 2 - full |
| water | ALT set 2 dil 7 | set 2 - full |

### Sequence analysis

To facilitate the analysis of the FASTQ files produced in sequencing, a MariaDB database and Perl script were produced. The workflow is outlined in Fig. 5. Each sequence is quality filtered and compared to a database of known motifs. Any matches are assigned to bins and then visualised in R.

Fuzzy matching is employed to account for sequencing / PCR errors, or external attempts to alter the gene sequence.

Figure 5. Sequence analysis workflow

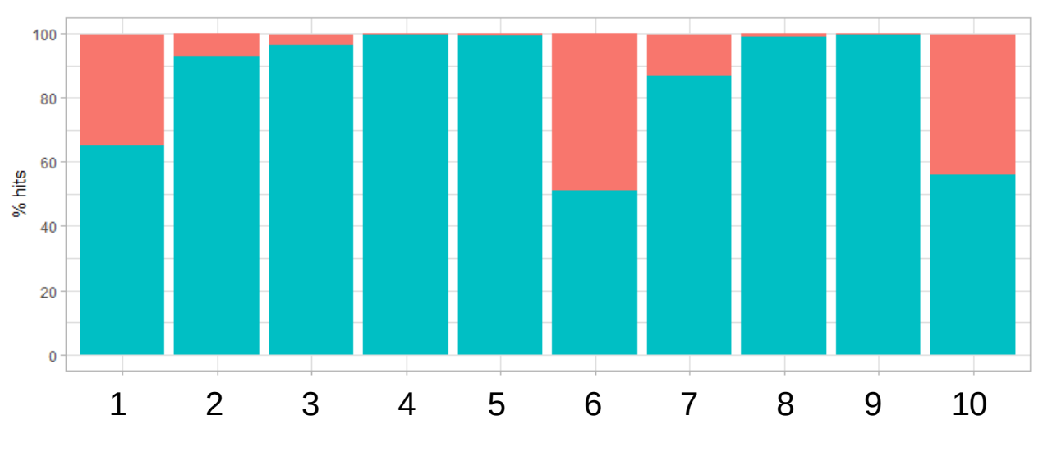


## Results

### Target ratios

Transgene and altered gene blocks were pooled at different ratios. Results for one gene are shown in Fig. 6. Even at very low starting copy numbers, both targets were amplified in the expected manner.

Figure 6. PCR reactions spiked with varying starting copy numbers of target



| Lane | Altered c.n | Transgene c.n |
|---|---|---|
| 1 | 1416 | 1542 |
| 2 | 1416 | 309 |
| 3 | 1416 | 62 |
| 4 | 1416 | 12 |
| 5 | 1416 | 2 |
| 6 | 283 | 309 |
| 7 | 283 | 62 |
| 8 | 283 | 12 |
| 9 | 283 | 2 |
| 10 | 11 | 12 |

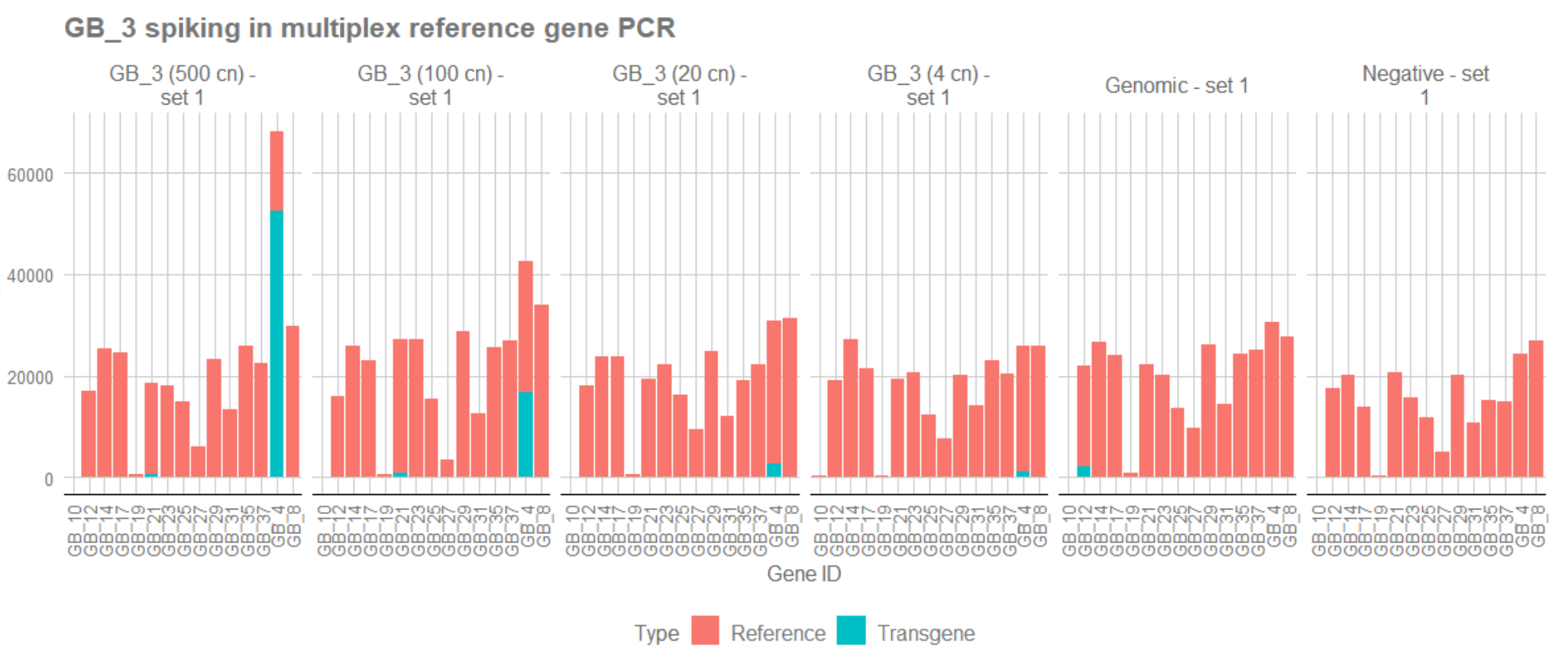
### Multiplex PCR and spiked transgenes

Amplification of a single target from pooled primer sets were clearly detected in the majority of reactions. Fainter bands produced correspondingly lower sequence counts on the MiniSeq.

Multiplexed PCR reactions of ~100 copies of pooled, altered sequence blocks were spiked with the transgene sequence of one gene (Fig. 7).

We were able to detect the presence of the spiked sequence down to below 10 starting copies. A few unexpected transgene amplifications were also noted in some reactions (sequences used to optimise the initial PCRs), which highlights the need for strict protocols when using such a sensitive technology.

Figure 7. Spiking of transgene sequence GB_3 in a multiplexed PCR of pooled primers and altered reference gene blocks. GB_4 is the altered form of GB_3 in this example



## Conclusions

The pilot study presented here shows the potential of using Illumina-based amplicon sequencing for rapid and scalable detection of gene doping in horses. We are currently re-optimising the amplicons that failed to amplify well in multiplex, expanding the gene set to include common vector features, and exploring extraction methods from spiked matrices.

Recent work on human gene doping detection using next-generation sequencing has also recently been described (de Boer *et al*, 2019), highlighting the utility of this method.

## References

de Boer, E *et al* A next-generation sequencing method for gene doping detection that distinguishes low levels of plasmid DNA against a background of genomic DNA, Gene Ther. (2019)
Park, K *et al*. Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq, BMC Genomics. 13 (2012) 473.
Pichler, M *et al*. A 16S rRNA gene sequencing and analysis protocol for the Illumina MiniSeq platform, Microbiologyopen. 7 (2018) e00611.
Wilkin, T et al. Equine performance genes and the future of doping in horseracing, Drug Test. Anal. 9 (2017) 1456–1471.