A guide to the analysis of KASP genotyping data using cluster plots
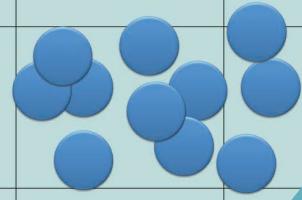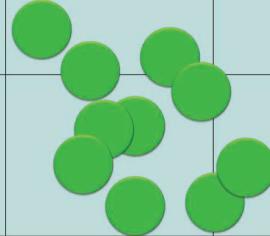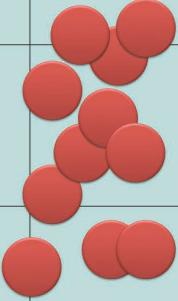
Contents of this guide

1. Introduction

KASP genotyping assays are based on competitive allele-specific PCR and enable bi-allelic scoring of single nucleotide polymorphisms (SNPs) and insertions and deletions (Indels) at specific loci. The SNP-specific KASP Assay mix and the universal KASP Master mix are added to DNA samples, a thermal cycling reaction is then performed, followed by an end-point fluorescent read. Bi-allelic discrimination is achieved through the competitive binding of two allele-specific forward primers, each with a unique tail sequence that corresponds with two universal FRET (fluorescence resonant energy transfer) cassettes; one labelled with FAM™ dye and the other with HEX™ dye.

Once the KASP reactions are complete, and the resulting fluorescence has been measured, the raw data must be interpreted to enable genotypes to be assigned to the DNA samples. This is typically achieved by viewing the fluorescence values for each sample on a Cartesian plot. These plots are often referred to as cluster plots.

The purpose of this document is to provide an explanation of how to analyse KASP genotyping data using a cluster plot, and to provide troubleshooting guidance for common issues that can arise.

For any queries about this guide please contact:
All locations except USA: email tech.support@lgcgroup.com or call +44 (0)1992 476 486
USA only: email us-support@lgcgroup.com or call +1 978 338 5317

3

2. How to visualise genotyping results using a cluster plot

The fluorescent signal from each individual DNA sample is represented as an independent data point on a cluster plot. One axis is used to plot the FAM fluorescence value (typically the X axis) and the second axis is used to plot the HEX fluorescence value (typically the Y axis) for each sample.

A sample that is homozygous for the allele reported by FAM will only generate FAM fluorescence during the KASP reaction. This data point will be plotted close to the X axis, representing high FAM signal and no HEX signal (blue data points in Figure 1).

A sample that is homozygous for the allele reported by HEX will only generate HEX fluorescence during the KASP reaction. This data point will be plotted close to the Y axis, representing high HEX signal and no FAM signal (red data points in Figure 1).

A sample that is heterozygous will contain both the allele reported by FAM and the allele reported by HEX. This sample will generate half as much FAM fluorescence and half as much HEX fluorescence as the samples that are homozygous for these alleles. This data point will be plotted in the centre of the plot, representing half FAM signal and half HEX signal (green data points in Figure 1).

To ensure the reliability of the results, a KASP reaction without any template DNA must be included as a negative control. This is typically referred to as a no template control or NTC. The NTC will not generate any fluorescence and the data point will therefore be plotted at the origin (black data points in Figure 1).

All samples that have the same genotype will have generated similar levels of fluorescence and will therefore all appear close together on the plot. Based on the relative position of these clusters, it is possible to determine the genotype of all the data points. It is important that a sufficient number of individual samples are included in a KASP reaction plate to ensure that there are enough data points on the Cartesian plot to allow cluster analysis.
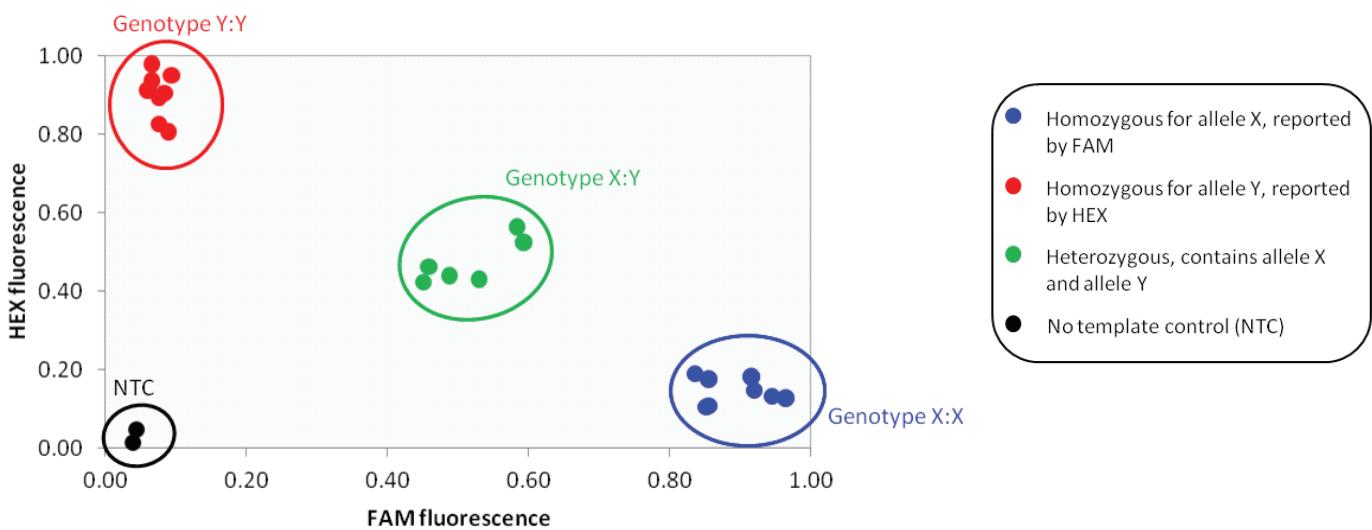

Figure 1. A typical genotyping cluster plot. Each data point represents the fluorescence signal of an individual DNA sample. Samples of the same genotype will have generated similar levels of fluorescence and will therefore cluster together on the plot.

For any queries about this guide please contact:
All locations except USA: email tech.support@lgcgroup.com or call +44 (0)1992 476 486
USA only: email us-support@lgcgroup.com or call +1 978 338 5317

4

3. Common issues with interpretation of genotyping cluster plots

3.1 Too few individual data points

LGC recommend that a minimum of 22 DNA samples plus 2 NTCs are run per assay per plate to enable cluster analysis to be performed effectively. Running fewer samples than the recommended 22 (plus 2 NTCs) can make it difficult to score genotypes reliably, and may cause misinterpretation of genotyping results.

Figure 2 provides an example of how data can be misinterpreted when too few samples are run; the limited number of data points on Figure 2A affects the scaling of the Y axis, and hence suggests that one of the data points is heterozygous (green triangle). When sufficient data points are included (Figure 2B), the presence of true heterozygote data points (green circles) make it clear that this potential heterozygous sample is actually homozygous for allele X (blue triangle).
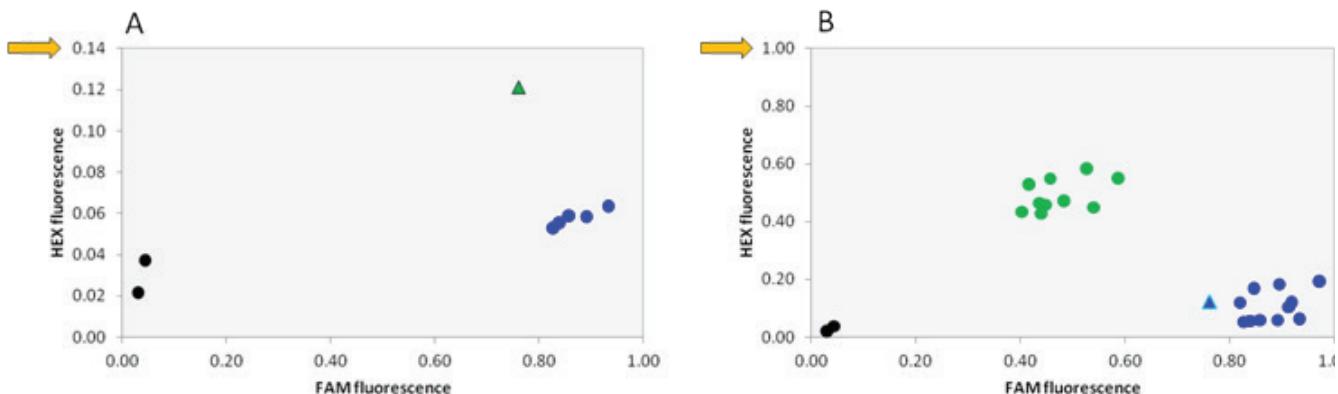

Figure 2. Plot A shows a cluster plot where only six DNA samples have been genotyped. There is separation between the blue circles and the green triangle, and the results suggest that the green triangle is a heterozygote. Plot B shows the same data, but including the results for an additional sixteen DNA samples. The data point represented with a green triangle in A is identified with a blue triangle in B. By including the required number of DNA samples, the genotyping clusters are sufficiently distinct and the triangular data point now clearly belongs to the homozygous cluster (blue). The orange arrows (on A and B) indicate the effect that the number of plotted data points has on the automatic scaling of the Y axis.

3.2 Scaling of X and Y axes is not comparable

Of the two reporting fluorophores used in KASP genotyping chemistry, it is known that FAM will typically produce a greater signal than HEX. This difference in relative fluorescence does not negatively affect the genotyping data when viewed as a cluster plot as the X and Y axes of the plot will be scaled independently to adjust for the different maximum signal produced by each of the fluorophores.

The software supplied with a number of widely available qPCR instruments can sometimes automatically scale the X and Y axes so that they are not comparable. This can affect the view of the genotyping clusters and may cause misinterpretation of genotyping results. To prevent this issue, manually inspect the minimum and maximum values for the X and Y axes. If they differ considerably, rescale the axes so that both have the same minimum and maximum value.

Figure 3 provides an example of how data can be misinterpreted when the X and Y axes are not scaled comparably; on Figure 3A the two clusters present (red and blue) can be interpreted as the two homozygote genotypes, yet once the X axis is scaled correctly to match the Y axis (Figure 3B) it is clear that the blue homozygous cluster (genotype X:X) actually represents heterozygote samples (green, genotype X:Y).
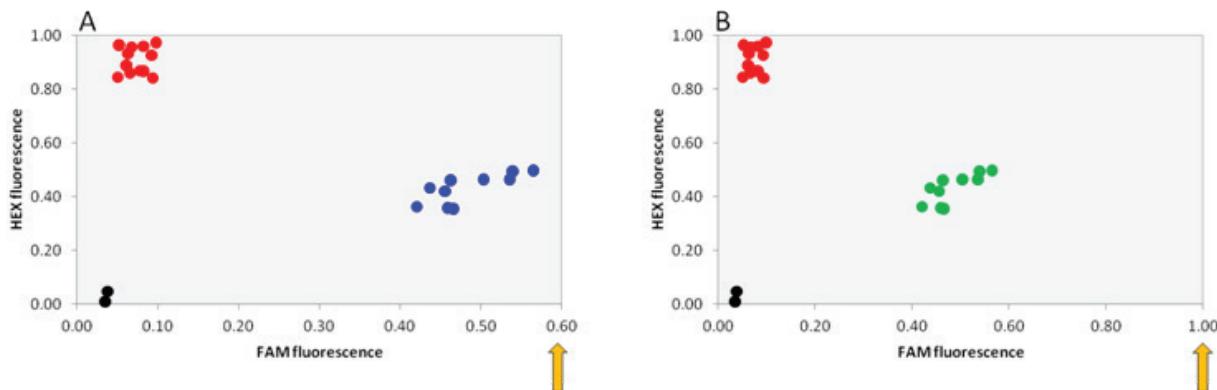

Figure 3. Plot A shows a cluster plot with two clear genotyping clusters; these have been assigned as homozygous for allele X (blue data points) and homozygous for allele Y (red data points). In plot A, the X and Y axes are not scaled comparably; X has a maximum of 0.60 whilst Y has a maximum of 1.00. Plot B shows the same data as plot A, but with the X axis scaled correctly to match the Y axis. It is then clear that the blue cluster actually represents samples with a heterozygous genotype (green data points).

The scaling of the axes can be a particular issue if all of the data points are the same genotype e.g. monomorphic for the FAM allele. In this scenario, only FAM signal will have been generated from each sample and hence the X (FAM) axis may be scaled to have a much higher maximum than the Y (HEX). This can result in monomorphic results appearing as very scattered across the plot (Figure 4A, overleaf) – rescaling the axes can improve data visualisation and hence the ability to confidently assign a monomorphic genotype (Figure 4B, overleaf).
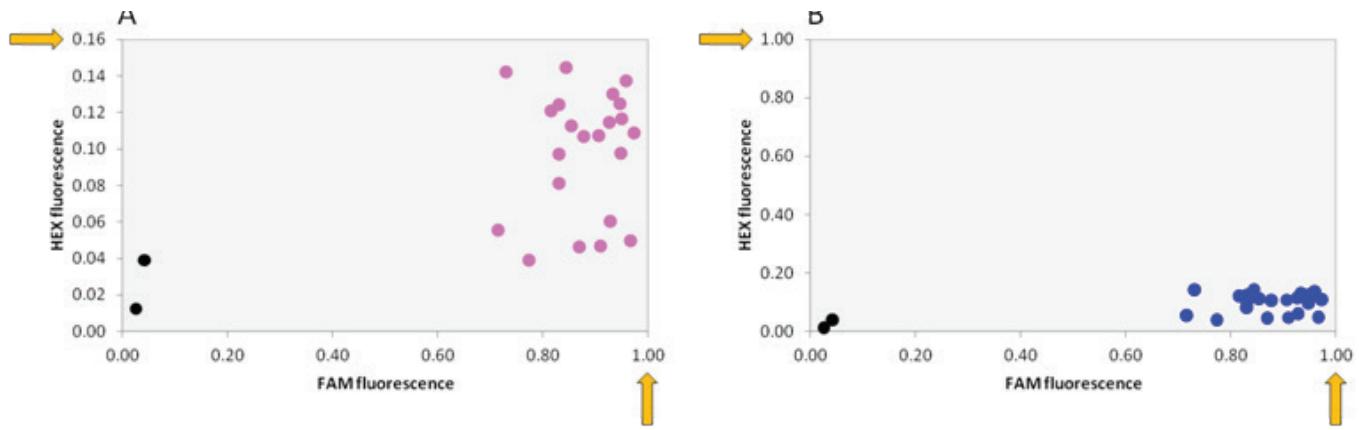
Figure 4. Plot A shows a cluster plot where the data points do not appear to be forming into clusters (pink, no genotype assigned). The orange arrows on plot A indicate the maximum fluorescence values for the X and Y axes; in this case the X axis has a maximum of 1.00 whilst the Y axis has much smaller maximum of 0.16. Plot B shows the same data as plot A, but with the Y axis re-scaled to have the same maximum fluorescence value as the X axis (1.00, as indicated by the orange arrow). By re-scaling the X axis to match the Y axis, the genotyping data points are positioned correctly and it is then clear that all of the data points belong to one cluster and represent samples with a homozygous genotype (blue cluster).

3.3 Reliance on the software's in-built autocalling option

The software packages supplied with a number of widely available qPCR instruments typically have the ability to automatically assign genotypes to data points based on their position on the cluster plot. Please be aware that the algorithms used by these packages for data analysis will differ in robustness, and no software should be considered as 100% reliable. Reliance on the autocalled genotypes with no manual review may result in misinterpretation of genotyping results.

LGC strongly recommend that genotyping data is always visually inspected and that results are sense checked to ensure that they meet expectations. Ensuring that the X and Y axes on the cluster plot are scaled comparably can aid data interpretation. The inclusion of positive controls (samples of known genotype) on each reaction plate can help to provide confidence in results that are autocalled by the software, and will ensure that there are always three clusters present on the cluster plot. Figures 5A and 5B demonstrate the effect of axes scaling, and Figure 5C shows the benefits of including positive controls on each reaction plate.
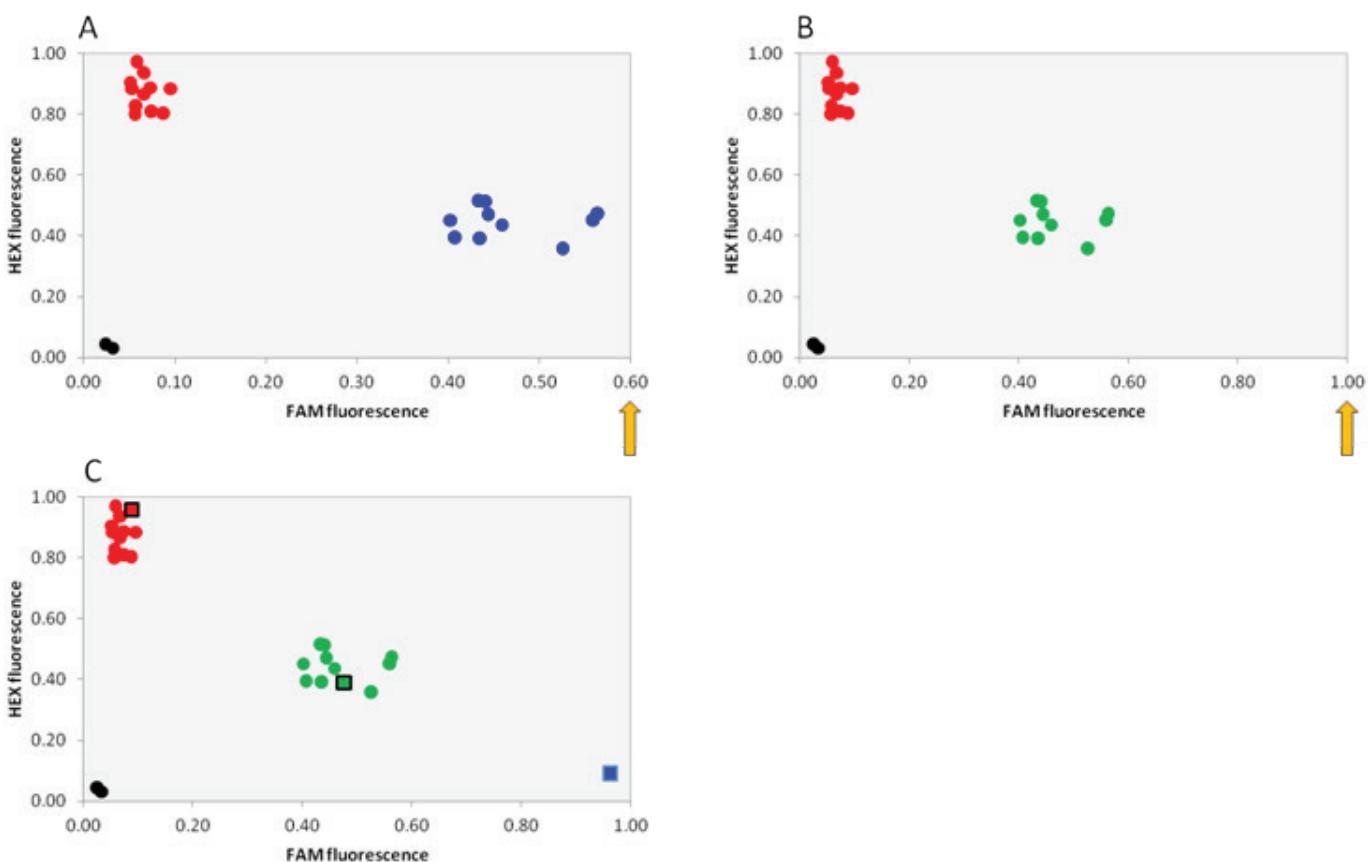

Figure 5. Plot A shows a cluster plot with two clear genotyping clusters that have been autocalled using qPCR instrument software. The orange arrow on plot A indicates the maximum fluorescence value for the X axis, in this case 0.60. Plot B shows the same data as plot A, but with the X axis re-scaled to have the same maximum fluorescence value as the X axis (1.00, as indicated by the orange arrow). Rescaling the axes ensures that the genotyping clusters are positioned correctly – in this case, the blue cluster from plot A is now in the centre of the plot and has been manually assigned as a heterozygote cluster (green data points). Plot C shows the same data as plot A and plot B, but with three additional data points, indicated by square markers. These represent a positive control sample for each genotype. Inclusion of these positive controls ensures that the three possible genotypes are represented and can hence improve the accuracy of autocalling features that are in-built in qPCR instrument software.

For any queries about this guide please contact:
All locations except USA: email tech.support@lgcgroup.com or call +44 (0)1992 476 486
USA only: email us-support@lgcgroup.com or call +1 978 338 5317

8

3.4 Combining data from multiple reaction plates on the same plot

LGC do not recommend combining data from multiple reaction plates. This is because the fluorescence generated, and the resultant scaling of the cluster plots, will differ from plate to plate. Viewing combined data on one plot is likely to make the genotyping clusters less defined, and will affect interpretation of the results.

Figure 6 shows data for the same KASP assay run on two separate reaction plates; viewing each set of data independently (Figures 6A and 6B) gives three tight genotyping clusters. When the data are plotted together on the same plot, however, the clusters become difficult to interpret (Figure 6C).
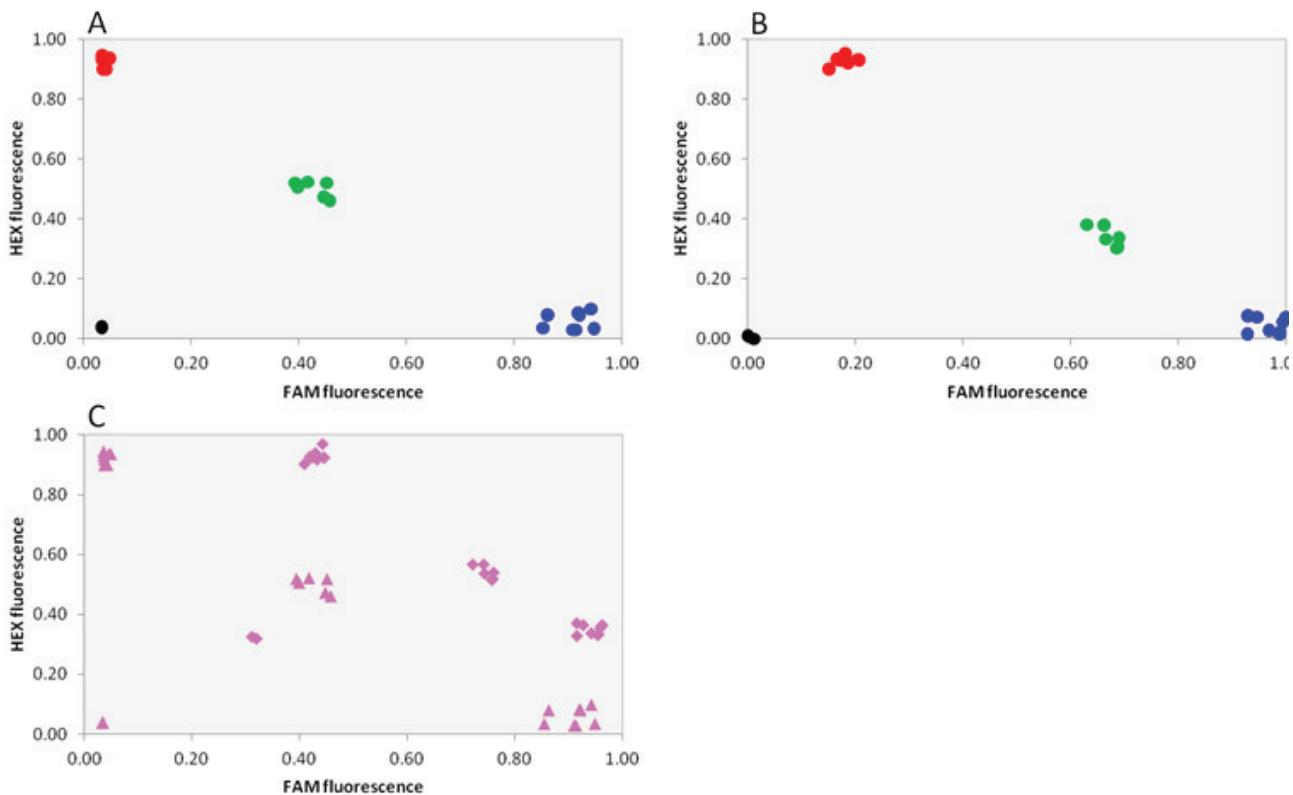

Figure 6. Plot A shows the data for an individual KASP assay, generated on reaction plate A. Plot B shows the data for the same KASP assay, generated on reaction plate B. In each case, the cluster plot axes have been scaled appropriately for the signal values generated on each reaction plate and, on both plots, clear clusters are visible and can be assigned genotyping calls. Plot C illustrates the effect of plotting data from two different reaction plates on the same cluster plot; the axes have been scaled based on all the fluorescent signal values generated across the two reaction plates. Variations in the fluorescence generated between plates have resulted in multiple clusters being present on the plot. Consequently, it is no longer possible to confidently assign genotyping calls.

3.5 Combining data from multiple KASP assays on one plot

LGC do not recommend combining data from different KASP assays on the same cluster plot. Each individual KASP assay will amplify template DNA at a slightly different rate, and will subsequently generate differing levels of fluorescent signal. Plotting data from more than one assay on the same plot does not provide any meaningful comparison, and will result in the data being very difficult to interpret
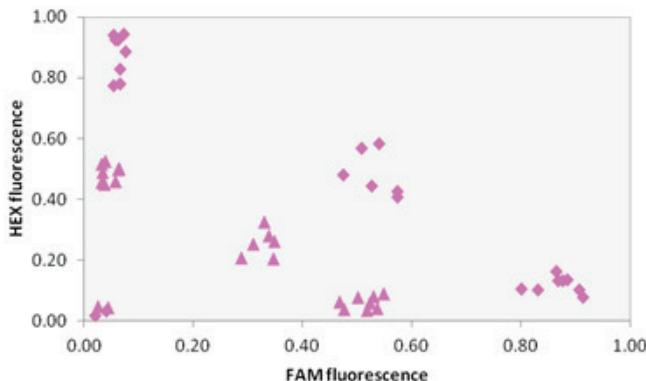

Figure 7. This cluster plot contains data points from two different KASP assays. One assay has generated greater fluorescent signal values than the other assay and, as a result, there are multiple clusters present on the plot. Consequently, data interpretation and assigning of genotyping calls is much more difficult.

Figure 8 (overleaf) illustrates the effect of viewing data from multiple assays on the same cluster plot; both data quality and ease of interpretation are drastically affected. In Figure 8A, the pink smear of data points represents data from four different KASP assays and the ability to confidently assign genotyping calls is reduced. Figures 8B – 8E identify the data points for each of the individual assays and, by viewing this data individually, the three genotyping clusters are much more apparent. All four assays have generated good genotyping data, with three clear clusters, and if the data points for each assay were viewed on an individual plot then it would be straightforward to confidently assign genotyping calls.

For any queries about this guide please contact:
All locations except USA: email tech.support@lgcgroup.com or call +44 (0)1992 476 486
USA only: email us-support@lgcgroup.com or call +1 978 338 5317

10
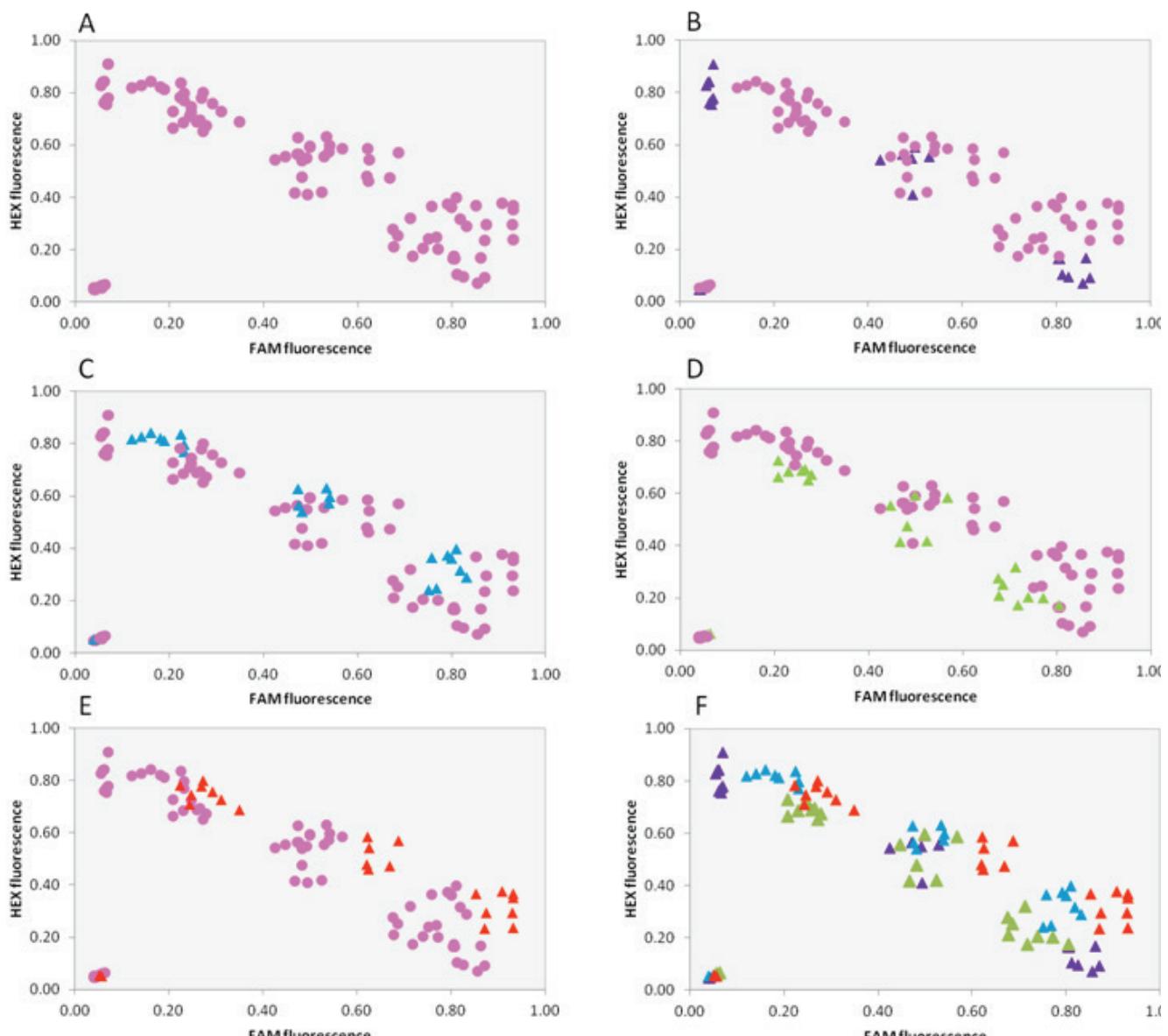
Figure 8. Plot A shows the data for multiple KASP assays on a single cluster plot. The data points form a smear across the plot and consequently it is not possible to confidently assign genotyping calls. Plots B, C, D and E identify the data for an individual assay using a unique colour for the assay-specific data points (purple, blue, green and red respectively). In each case, three distinct clusters are clearly visible and it would be possible to confidently assign genotyping calls if these data points were viewed on their own cluster plot. Plot F shows the same data as plot A, but with the data points for each KASP assay identified by their unique colour.

For any queries about this guide please contact:
All locations except USA: email tech.support@lgcgroup.com or call +44 (0)1992 476 486
USA only: email us-support@lgcgroup.com or call +1 978 338 5317

11

4. Summary

This document has outlined the key factors to consider when viewing KASP genotyping data on a cluster plot. When viewing and analysing your data, ensure that you follow these guidelines:

• Include a minimum of 22 data points (plus 2 NTCs) on each plot to facilitate cluster analysis.

• Check the scaling of the X and Y axes and, if required, rescale so that they are comparable. This will ensure that clusters are correctly interpreted.

• Where possible, include positive controls for each genotype as this will facilitate data analysis.

• Do not view results for more than one KASP assay on an individual cluster plot – always view results assay by assay.

• Always sense check genotypes that are automatically assigned by the instrument software as they may need to be manually adjusted.

For any queries about this guide please contact:
All locations except USA: email tech.support@lgcgroup.com or call +44 (0)1992 476 486
USA only: email us-support@lgcgroup.com or call +1 978 338 5317

www.lgcgroup.com/genomics • genomics@lgcgroup.com

Science for a safer world

**Brazil • Bulgaria • China • France • Germany • Hungary • India • Ireland • Italy • Netherlands
Nordic countries • Poland • Romania • Russia • South Africa • Spain • Turkey • United Kingdom • USA**