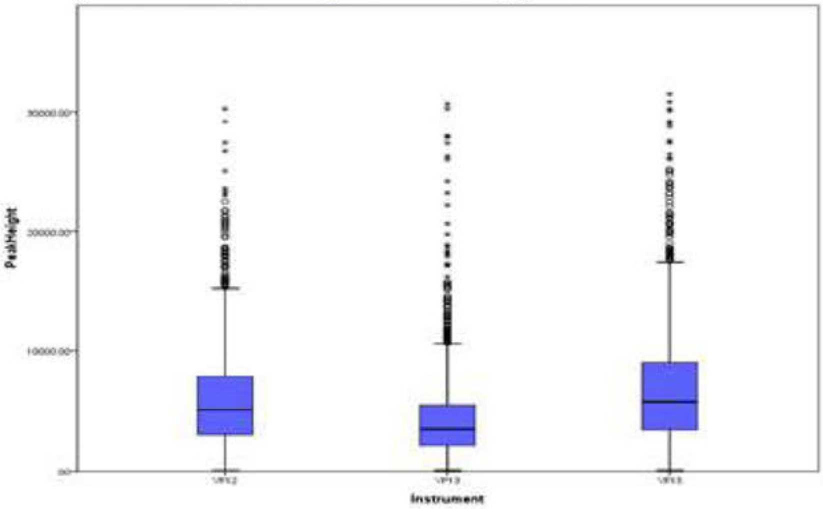
## Cathie Allen

| | |
|---|---|
| **From:** | Ballantyne, Kaye [redacted] |
| **Sent:** | Monday, 24 September 2018 12:48 PM |
| **To:** | Justin Howes |
| **Cc:** | Paula Brisotto; Sharon Johnstone |
| **Subject:** | RE: Validation question [SEC=UNCLASSIFIED] |

**UNCLASSIFIED**

Hi Justin,

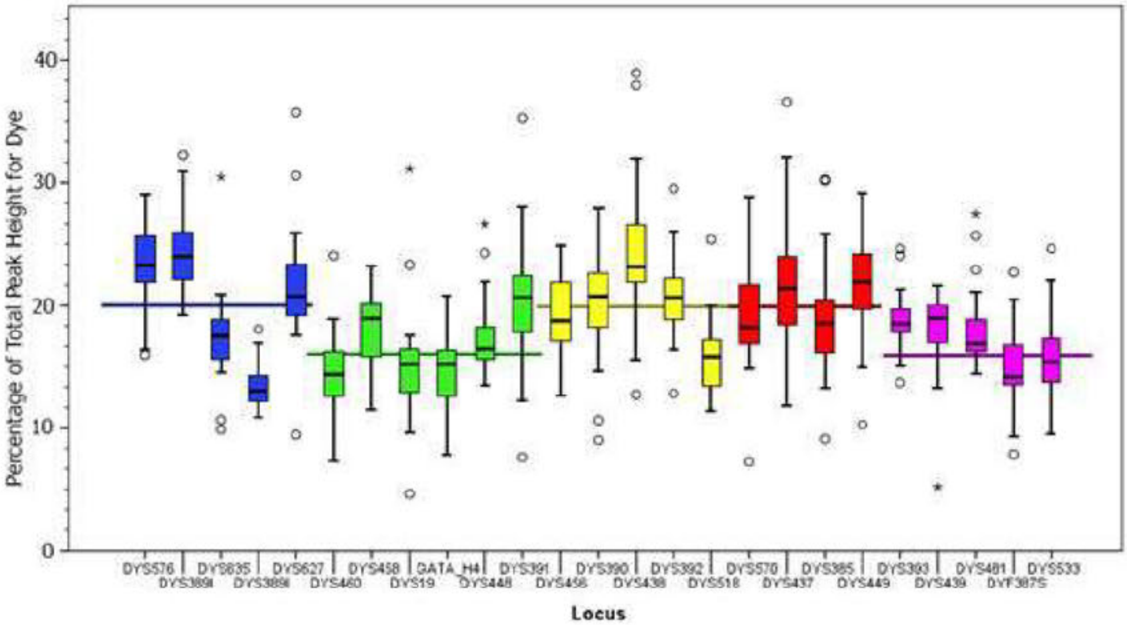
Nice to hear from you. Those are some tricky questions, and ones we have been debating recently here. We don't generally do any formal statistical testing for verifications, partly because we don't have a statistician in the DNA team (I'm not part of that group, but a separate research group), and partly because for most verification questions statistical measures of similarity/difference wouldn't be particularly valuable. For example, it's self-evident that different amounts of DNA will result in different numbers of alleles, or that different body fluids will have different quantities of DNA. A p-value won't provide any useful information for these. In some cases it can be useful to do some comparative statistics, such as when comparing the same samples run with two different methods or looking at the variation between different instruments running the same methods. However, there is a question about what a 'significant' result means operationally – if one thermal cycler gives significantly lower peak heights than another, but the end profiles are the same – does it matter? We recently did some work comparing three 3500 instruments, and found that stats testing actually complicated the picture. We ran the same plates across all three instruments, and looked at the variation in peak heights between instruments. Statistical testing indicated highly significant differences ($p<0.000001$), while simple graphing (below) indicated that although one instrument was slightly less sensitive than the others, there was overlap between them, and that it probably wasn't as big a deal as the p value indicated. At the end of the process, the number of alleles recovered was the same, so it was determined that the peak height variation probably didn't matter, given STRMix can account for it.



For most applications, I'm a big fan of using boxplots to look at the variation present in a system, and use a rule of thumb of 'if the boxes overlap, the difference will generally be functionally non-significant (i.e. $p>0.05$)'. To me, they display what's happening much better than a bar chart or scatter graph, so I tend to use them for all my validation/verification work. Generally, I won't then follow it up with a statistical test, particularly if a difference between the groups won't have an operational meaning, or significantly impact on the end result – the LR in STRMix. If a formal test is needed, I would normally use an ANOVA or Kruskal-Wallis test to compare independent groups of samples, or a paired sample t-test to compare repeated samples (i.e. same sample run through two different methods). If the validation report makes a strong claim (i.e. decrease in allele number showed a linear trend, or there were significant differences in peak heights), then I normally expect it to be followed up with a statistical test – so in practice, people just temper their words ("there appeared to be a linear trend, or differences were seen"), rather than bothering with formal testing.

In terms of sample numbers, it's fairly difficult to say how many is enough. A minimum of three replicates (biological, not technical) is always required to enable variation to be calculated, but it will depend on what is being measured, and how much variation is expected. For example, if you were looking at amplification success in different sample types, three blood samples would be all that is required to show that you get a full profile – that isn't going to vary. Getting an accurate indication of the variation of peak height ratios in low quantity samples might take hundreds, as there is a lot of variation expected. If I have no idea what to expect, I use a sequential testing program – start with 3 samples for each variable/level (i.e. three blood, saliva, semen, trace etc), then graph the variation/scrutinise the results, and decide if I need to do more testing or not. If I expect a lot of variation then I will perform more testing up-front – so for mixtures using 5-10 replicates might be appropriate, particularly for the difficult 3-4 person mixes. No validation can ever cover every situation, or show every possible result – my aim is generally to do enough to allow the expected variation to be modelled for each relevant factor, so that our methods/thresholds/guidelines encompass *most* situations. There are particular statistical methods of estimating how many samples need to be used (called power tests), but to get an accurate indication you need to know what the expected variation is already, and what power you want (the probability that you reject Ho when it is false, normally set at 80%, but can go higher). However, in practice they aren't particularly useful, largely because you need to know a lot about the variation already – and if you did, you wouldn't need to do the test!

An example of where I used a lot of samples, and did both graphing and formal stats testing was in the validation of Yfiler Plus, looking at inter-locus peak height balance. Different loci have different amplification efficiencies, and it's not a simple linear situation like with PP21/Globalfiler. Knowing the balance expected in a single source sample, and the types of imbalance seen mixtures, would be important for interpretation downstream. So I used ~250 population samples (which were being run anyway) to get the average inter-locus balance, and graphed it:



Then, to show that the peak heights really were different between loci, and that caseworkers should not expect a linear relationship, I did a simple chi-square test for average peak height within each dye layer – a p value of less than 0.05 indicated that there were significant differences between loci. It was significant for the blue loci, and not for any other layer. Doing a more sophisticated clustering analysis indicated that users should be aware that peak heights for particular loci would be lower (with $p<0.05$, for example [redacted], and so when interpreting mixtures they should take that into account. Statistical testing was important here, because I needed to know that the differences weren't just random variation, but due to an actual trend that would carry through to casework.

I do all my statistical testing and graphing in SPSS – I can share some setup files/scripts with you to do some testing if you have this program, but it does cost a lot for a non-university business licence (~$20,000). It is possible to do basic stats testing in Excel, but I'm afraid I've never done it, so can't give any guidance there! I'm afraid I haven't been a lot of help, despite writing a lot of words. I guess my summary on statistical testing would be: It depends. If you are making strong claims, or there is the possibility of a functional difference that will affect operational methods or outcomes, do a statistical test to confirm what your gut thinks. If a difference (or lack of difference) between groups will not cause any change in operation, and/or the difference is expected and explainable, then I wouldn't bother – graphing the data and making a statement like "data is comparable" is sufficient in my view. The summary on sample numbers is: Do as few as you can to model the variation you expect to see – you can always do more later if needed.

If it's easier, or you require any more information, I'm happy to have a chat over the phone – I'm easiest to reach on [redacted] I hope some of the information has helped, or at least furthered some of your discussions!

Kind regards,
Kaye

**Dr Kaye Ballantyne** | Senior Research & Development Officer
Office of the Chief Forensic Scientist | Forensic Services Centre | Victoria Police

Adjunct Associate Professor | School of Psychology and Public Health, La Trobe University
Adjunct Senior Researcher | College of Arts and Law, University of Tasmania

[redacted] | [redacted] | [redacted]
[redacted]

**From:** Justin Howes [redacted]
**Sent:** Thursday, 20 September 2018 16:05
**To:** Ballantyne, Kaye
**Cc:** Paula Brisotto; Sharon Johnstone
**Subject:** Validation question

Hi Kaye
I don't think we have met, but I am one of the two Team Leaders in QLD and at the last BSAG meeting, someone (I think Pam Scott) recommended I send you an email.

We have multiple discussions here every time we do a validation/verification on how to best statistically compare data in repeatability/reproducibility/sensitivity etc. experiments. I have consulted other labs and found some may not compare statistically and may just provide a qualitative statement eg. 'peak heights are comparable'. Others have indicated they just say something like the instrument is demonstrating it functions and is fit for purpose.

Essentially, the discussions are focussed on what statistical method should/could be used to compare two data sets and the number of samples to be statistically significant.

I was wondering if you would have time to provide some advice for us on how many samples to use in experiments, and how to compare if I was to provide an example?

I am about to take a week of leave, so if you have time, could you please reply all and we can forward an example?

Kind Regards
Justin

**Justin Howes**
Team Leader - Forensic Reporting and Intelligence Team

**Forensic DNA Analysis, Forensic & Scientific Services**
Health Support Queensland, Queensland Health

[redacted]

Integrity | Customers and patients first | Accountability | Respect | Engagement

*Queensland Health acknowledges the Traditional Owners of the land, and pays respect to Elders past, present and future.*

********************************************************************************

This email, including any attachments sent with it, is confidential and for the sole use of the intended recipient(s). This confidentiality is not waived or lost, if you receive it and you are not the intended recipient(s), or if it is transmitted/received in error.

Any unauthorised use, alteration, disclosure, distribution or review of this email is strictly prohibited. The information contained in this email, including any attachment sent with it, may be subject to a statutory duty of confidentiality if it relates to health service matters.

If you are not the intended recipient(s), or if you have received this email in error, you are asked to immediately notify the sender by telephone collect on Australia +61 1800 198 175 or by return email. You should also delete this email, and any copies, from your computer system network and destroy any hard copies produced.

If not an intended recipient of this email, you must not copy, distribute or take any action(s) that relies on it; any form of disclosure, modification, distribution and/or publication of this email is also prohibited.

Although Queensland Health takes all reasonable steps to ensure this email does not contain malicious software, Queensland Health does not accept responsibility for the consequences if any person's computer inadvertently suffers any disruption to services, loss of information, harm or is infected with a virus, other malicious computer programme or code that may occur as a consequence of receiving this email.

Unless stated otherwise, this email represents only the views of the sender and not the views of the Queensland Government.

**********************************************************************************

==============================

EMAIL DISCLAIMER
This email and any attachments are the property of Victoria Police and should not be disclosed. They may also be subject to copyright.
If you are not an intended recipient of this email please immediately contact us by replying to this email and then delete this email. You must not read, use,copy, retain, forward or disclose this email or any attachment.
We do not accept any liability arising from or in connection with unauthorised use or disclosure of the information contained in this email or any attachment.
We make reasonable efforts to protect against computer viruses but we do not accept liability for any liability, loss or damage caused by any computer virus contained in this email.