UNIVERSITY OF LIVERPOOL

Professor Sarah Coupland MBBS, PhD, FRCPath

George Holt Chair of Pathology
North West Cancer Research Centre Director-UoL
Department of Molecular and Clinical Cancer Medicine
Institute of Translational Medicine
William Henry Duncan Building,
6 West Derby Street
Liverpool

July 10th 2018

Professor Göran Sandberg
Chairman,
Board for Investigation of Scientific Misconduct
Lund University
PO Box 118, 221 00 Lund
Tel: +46 46 2228146

Re: Suspected scientific misconduct in the evaluation of retina of mice and pig experiments.

Dear Prof. Sandberg

Prof. Steffen Heegaard and myself (Prof. Sarah Coupland) have spent a large amount of time and effort over the last 8 months evaluating the case of OM versus LT, with respect to suspected scientific misconduct. We enlisted the help of a post-doctoral scientist (OBS), who has considerable expertise in image analysis, and who has been allocated time from usual duties to analyse the scanned slides.

Essentially, an automated programme as well as manual analysis was used to evaluate the digitalised slides of the following research projects:

- **Mouse Run Alz Project (151014) _ RBPMS_dapi (Early)**
- **Mouse Run Alz Project (151014) _ RBPMS_dapi (Late)**
- **Mouse Run Alz Project (151014) _ neuN_dapi (Early)**
- **Pig Archive (Experiment 1 and Experiment 2)**

OBS compared the scores with those of OM and LT, highlighted any differences in the scores, and also undertook statistical analyses. This is provided in detail in the document labelled "Final report – June", which we spoke through on the last telephone conference (June 14th, 2018). All of the raw data and analyses are to be found in Excel sheets, labelled appropriately. **These documents are most important as they highlight the detail of the assessment undertaken.**

You asked specific questions in your last E-mail following the above-mentioned TC, which we will go through in turn:

Is the difference between the OM and the LT assessments outside normal assessment variability?

Partially. For the mouse project studies, OBS identified that there is a similar trend in the scores obtained in all quantifications (OBS, OM, LT, and an additional automated control method (Auto) performed for Mouse Run Alz Project RBPMS (151014)_dapi_ (Early). Some assessments of LT are outside normal assessment variability, which could partially influence the results of Mouse projects (Early) and Pig archive (Gal-inhibitor vs. Kontroll), but not in the Mouse Run Alz Project RBPMS (Late) project.

How does LT's assessment compare to your estimation based on the microscopic images?

From our experience, some of the assessments scored by LT would not be possible to be obtained in specific images and are inconsistent with the quantifications of OM, OBS and Auto. For example, regarding the assessments of Mouse Run Alz Project_(151014)_RBPMS_dapi_(Early) project, for the image file 151014_12_RBPMSms2000_ CALBms2000_DAPIms40_mid1.jpeg), LT scored 18 counts, compared with Auto (30 counts), OBS (28 counts) and OM (27 counts), and for image 151014_1_RBPMSms2000_CALBms2000_DAPIms40_mid2.jpeg, LT scored 19 counts, compared with Auto (7 counts), OBS (3 counts) and OM (8 counts). Other examples are highlighted in a linear graph comparing all quantifications, and in a separated spreadsheet of the corresponding project (File: Mouse Project (151014) RBPMS (Early) + Stats).

Regarding the quantification of Mouse Run Alz Project (151014)_NeuN_dapi (Early) project, some of LT's assessments are inconsistent with those of OBS and OM. For example, in the quantification of image 151014_7b_NeuN_ab_dapi_20x_mid, LT scored 27 counts, compared with OBS (44 counts) and OM (43 counts). Importantly, other inconsistencies were observed in the quantification of other mice (151014.7, 151014.11, 151014.12, 151014.15, 151014.16, 151014.19, 151014.20, 151014.23, 151014.24, 151014.27 and 151014.28). A specific spreadsheet of the corresponding project (File: Mouse Run Alz Project (151014)_NeuN_dapi (Early) + Stats) shows some examples.

Regarding the quantification of Pig archive, periodic changes in the thickness of ONL were documented by LT throughout Pigs/biological replicates when assessing Gal-3 inhibitor vs. Kontroll. These differences were not observed by OBS, where the thickness assessments showed a reasonable level of similarity among different files, Pigs, and groups. It is important to mention that OBS aimed to quantify the entire ONL of each image file to achieve the highest level of accuracy. So, each figure was quantified in sextuplicate, and each pig is represented by four quantified figures (where file names end in _1, _10, _20 and _25).

For the Galectin-3 vs. PBS experiments, no relevant differences were found between OBS and LT image assessments.

Does LT's assessment of images lead to the result of the study becoming more interesting, i.e. having achieved statistical significance?

Probably yes. In the Mouse Run Alz Project (151014) RBPMS_dapi (Early), translated images were grouped to generate averages of each mice per group (Run since - 2mo and Sedentary). Detected inconsistencies could affect the average of the corresponding mouse score, and consequently, the average of the experimental group. Indeed, among all quantifications, there is a slight decrease in the averages of positive RBPMS cells for the sedentary group compared with the run since 2mo group. However, these differences are not significant for OBS, OM, and Auto quantifications; however, are significantly different in LT's quantification. Most inconsistencies are related with the mice 151014_11, 151014_12 and 151014_16, contributing to the significant decrease observed in the Sedentary group of LT's quantifications. If specific inconsistent mouse averages among LT's assessments are changed by those obtained by OM, OBS, and Auto, this statistical difference disappears.

The same issue seems to occur in the Mouse Run Alz Project (151014) Neu_N_dapi (Early) analysis. Inconsistencies arise among LT's assessments compared with those obtained by OM and OBS, and those seem to be key in decreasing the average of the following mice (151014.7, 151014.11, 151014.12, 151014.15, 151014.16, 151014.19, 151014.20, 151014.23, 151014.24, 151014.27 and 151014.28), contributing to a significant difference among the run since 2mo and sedentary groups. This is not seen in OM and OBS quantifications. Again, by changing specific inconsistent mouse averages amongst LT's assessments by the average of scores obtained by OM and OBS, led to LT's statistical difference disappearing, suggesting that these changes are essential to affect the final analysis to obtain higher differences.

Regarding the quantification of Mouse Run Alz Project (151014) _ RBPMS_dapi (Late) project, there are some inconsistencies in LT's assessments compared to those obtained by OBS and OM. However, the inconsistencies are <u>insufficient</u> to result in statistical differences amongst the analysed groups, i.e.,

there are no significant differences between run since 2mo and sedentary groups following all quantifications of this specific project.

Regarding the quantification of Pig archive project, the Pig retina were quantified by OBS and LT individually and pooled together for each corresponding condition (Gal-inhibitor vs. Kontroll and Galectin-3 vs. PBS). Periodic changes in the thickness of the ONL occur in LT's quantification in a linear graph showing LT's and OBS' scores throughout the average of each quantified Pig. The changes resulted in the increase of the thickness of the treated group (Gal-Inhibitor) compared with the corresponding 'Kontroll' among LT's assessments. This result does not occur in OBS' quantification where the ONL thickness is quite similar amongst the Gal-3 inhibitor and Kontroll images. The difference between OBS and LT's values may be explained by the number of thickness values assessed by OBS and LT, as OBS quantified 6 measurements for the ONL in each specimen. From the values provided to us, it would appear that LT quantified a small number of measurements, potentially leading to inaccuracy.

By pooling values together from all biological replicates into one scatter plot, the thickness profile does not change between the Gal-inhibitor and Kontroll groups in OBS quantification, but significantly increased at least in the Gal-3 inhibitor 100uM group in LT's measurements.

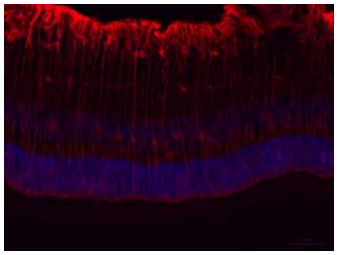
In the combined analysis of all biological replicates of Galactin-3 vs. PBS control approach, there is a significant decrease on the thickness following Galectin-3 treatment compared with PBS in both LT and OBS quantifications where both observed statistical differences (for LT and *** for OBS). It is important to note that it is not clear which statistical test was used by LT. The differences in LT's quantification do not appear to be critical to alter the final results of this particular experiment but may influence the power of statistical significance observed (perhaps because some measurements were missing in LT's control and treatment groups, which were Identified as 'Infection').*

Have the missing observations in the pig project resulted in a (better) statistical significance and thus a more interesting result?

This is relative. For the Galectin-3 vs. PBS experiments, no relevant differences are seen between LT and OBS assessments. However, LT did not quantify some images that were made available for OBS to quantify – these were labelled "Infection" cases. Therefore, if the infection' cases were truly 'infected' and therefore appropriately removed from the evaluation, then the better statistical significance is justified and can be thereby explained (three statistical stars for LT and one star for OBS); however, this needs to be part of the described results for any final published data. However, if the labelling of the slides with 'infection' is just as a control and does not justify the exclusion of these scores, then this would not be justified and would lead to an improved statistically relevant result.

Could you please outline the Imaging and Statistical methods used?

OBS quantified and analysed the project's images in a masked manner using the Nis Elements Advanced Research Image Analysis Solution (Nikon). Fluorescence background was subtracted using the provided unstained controls as reference. The statistical tests used by OBS was a t-test, two-tailed and unpaired test, with confidence intervals of 95%, where $p > 0.05$ is considered not significant (n.s). Scores and corresponding averages are in the excel spreadsheets. GraphPad-5 Prism was used for statistical analysis of these projects. For the Pig Archive, metrics calibration was performed using the image below.



In summary, in the above-mentioned mice experiments, there are noticeable differences between some of LT's scores when compared to those of the other observers. This is particularly the case in the so-called "early" experiments (RBPMS and neuN). It is difficult to understand how some of these values were achieved by LT.

With respect to the pig experiments, as mentioned above it is noticeable that there are variations in the scores between OBS and LT in Experiment 1 (i.e. Galectin inhibitor vs control). This is likely to be due to the number of measurements assessed by OBS compared to LT: OBS assessed 6 measurements in the ONL of all pig specimens, whilst it would appear that LT examined fewer measurements, leading to inaccuracy.

Finally, OBS also undertook statistical analyses of the values obtained by LT in Experiment 2. Whilst OBS also observed a statistical significance in the Galectin-3 treatment group, OBS' value was not as strong in statistical power compared to that observed by LT. This difference may be due to the smaller number of cases quantified by LT, as it would appear that she excluded the 'infection' cases.

We hope our review of these experiments and our conclusions will enable you, the research Board and the scientists themselves to move forward on this issue.

Kind regards,

Professor Sarah Coupland
University of Liverpool, UK

Professor Steffen Heegaard
University of Copenhagen
Denmark