

## **An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer**

David L. Rimm<sup>1</sup>, Samuel C.Y. Leung<sup>2</sup>, Lisa M. McShane<sup>3</sup> Yalai Bai<sup>1</sup>, Anita L. Bane<sup>4</sup>, John M.S. Bartlett<sup>5,16</sup>, Jane Bayani<sup>5</sup>, Martin C. Chang<sup>6</sup>, Michelle Dean<sup>7</sup>, Carsten Denkert<sup>8</sup>, Emeka K. Enwere<sup>7</sup>, Chad Galderisi<sup>9</sup>, Abhi Gholap<sup>10</sup>, Judith C. Hugh<sup>11</sup>, Anagha Jadhav<sup>10</sup>, Elizabeth N. Kornaga<sup>7</sup>, Arvydas Laurinavicius<sup>12</sup>, Richard Levenson<sup>13</sup>, Joema Lima<sup>5</sup>, Keith Miller<sup>14</sup>, Liron Pantanowitz<sup>15</sup>, Tammy Piper<sup>16</sup>, Jason Ruan<sup>13</sup>, Malini Srinivasan<sup>15</sup>, Shakeel Virk<sup>17</sup>, Ying Wu<sup>4</sup>, Hua Yang<sup>11</sup>, Daniel F. Hayes<sup>18</sup>, Torsten O. Nielsen<sup>2</sup> and Mitch Dowsett<sup>19</sup>.

**Institutional affiliations:** <sup>1</sup>Department of Pathology, Yale University School of Medicine, New Haven, Connecticut, United States; <sup>2</sup>Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada; <sup>3</sup>Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland, United States; <sup>4</sup>Department of Pathology and Molecular Medicine, Juravinski Hospital and Cancer Centre, McMaster University, Hamilton, Ontario, Canada; <sup>5</sup>Transformative Pathology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada; <sup>6</sup>Sinai Health System and University of Toronto, Toronto, Ontario, Canada; <sup>7</sup>Translational Laboratories, Tom Baker Cancer Centre, Alberta Health Services, Calgary, Alberta, Canada; <sup>8</sup>Institut für Pathologie and German Cancer Consortium (DKTK), Charité Campus Mitte, Berlin, Germany; <sup>9</sup>MolecularMD, Portland, Oregon, United States; <sup>10</sup>Optra Technologies, NeoPro SEZ, BlueRidge, Hinjewadi, India; <sup>11</sup>Department of Laboratory Medicine and Pathology, University of

Alberta, Edmonton, Alberta, Canada; <sup>12</sup>National Center of Pathology, Vilnius University Hospital Santara Clinics and Vilnius University, Vilnius, Lithuania; <sup>13</sup>Department of Medical Pathology and Laboratory Medicine, University of California Davis Medical Center, Sacramento, California, United States; <sup>14</sup>Cancer Diagnostic Quality Assurance Services CIC, Poundbury Cancer Institute, Poundbury, Dorset, United Kingdom; <sup>15</sup>Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States; <sup>16</sup>Biomarkers & Companion Diagnostics Group, Edinburgh Cancer Research Centre, Edinburgh, United Kingdom; <sup>17</sup>Department of Pathology and Molecular Medicine, Queen's University, Kingston, Ontario, Canada; <sup>18</sup>Breast Oncology Program, Department of Internal Medicine, University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan, United States and <sup>19</sup>Institute of Cancer Research, London, United Kingdom.

**Running Title:** Automated Ki67 scoring in breast cancer

**Corresponding Author:** David L. Rimm, M.D. / Ph.D.

Professor of Pathology

Director, Yale Pathology Tissue Services

Department of Pathology, BML 116

Yale University School of Medicine

310 Cedar Street P.O. Box 208023

New Haven, CT 06520-8023

Phone: 203-737-4204

FAX: 203-737-5089

Email: [david.rimm@yale.edu](mailto:david.rimm@yale.edu)

This study was previously presented at the 39th Annual San Antonio Breast Cancer Symposium Dec. 6-10, 2016 (abstr P1-03-01)

## Abstract:

The nuclear proliferation biomarker Ki67 has potential prognostic, predictive, and monitoring roles in breast cancer. Unacceptable between-laboratory variability has limited its clinical value. The International Ki67 in Breast Cancer Working Group investigated whether Ki67 immunohistochemistry can be analytically validated and standardized across laboratories using automated machine-based scoring. Sets of pre-stained core-cut biopsy sections of 30 breast tumors were circulated to 14 laboratories for scanning and automated assessment of the average and maximum percentage of tumor cells positive for Ki67. Seven unique scanners and 10 software platforms were involved in this study. Pre-specified analyses included evaluation of reproducibility between all laboratories (primary) as well as among those using scanners from a single vendor (secondary). The primary reproducibility metric was intraclass correlation coefficient between laboratories, with success considered to be **intraclass correlation coefficient** > 0.80. **Intraclass correlation coefficient** for automated average scores across 16 operators was 0.83 (95% **credible interval**: 0.73–0.91) and **intraclass correlation coefficient** for maximum scores across 10 operators was 0.63 (95% **credible interval**: 0.44–0.80). For the laboratories using scanners from a single vendor (8 score sets), **intraclass correlation coefficient** for average automated scores was 0.89 (95% **credible interval**: 0.81-0.96), which was similar to the **intraclass correlation coefficient** of 0.87 (95% **credible interval**: 0.81-0.93) achieved using these same slides in a prior visual-reading reproducibility study<sup>1</sup>. Automated machine assessment of average Ki67 has the potential to achieve between-laboratory reproducibility similar to that for a

rigorously-standardized pathologist-based visual assessment of Ki67. The observed **intraclass correlation coefficient** was worse for maximum compared to average scoring methods, suggesting that maximum score methods may be suboptimal for consistent measurement of proliferation. Automated average scoring methods show promise for assessment of Ki67 scoring, but requires further standardization and subsequent clinical validation.

## Introduction:

The Ki67 immunohistochemistry assay is widely performed to assess cellular proliferation in breast cancer<sup>2</sup>, yet its assessment has never been standardized. This has limited its value for both clinical trial and routine diagnostic usage. The International Ki67 in Breast Cancer Working Group was convened in 2010 to address this problem<sup>3</sup>. This group designed and executed several studies to first assess the problem and then to develop methods of standardization, beginning with visual assessment.

The **International Ki67 in Breast Cancer Working Group** (Supplemental Table 1) has previously demonstrated that, in the absence of standardized scoring, concordance for visual reading of Ki67 in previously stained sections was satisfactory within, but not between, different laboratories<sup>4</sup>. When training was instituted to standardize scoring, inter-laboratory reproducibility improved substantially<sup>5,6</sup>. However, these prior studies were performed using slides containing previously cut and stained tissue microarray sections, and no effort to associate the data with clinical outcomes was undertaken. Therefore, evidence remains insufficient to support Ki67 use in routine clinical care. Now the **International Ki67 in Breast Cancer Working Group** is conducting additional studies to determine whether similar reproducibility can be achieved using core cut and whole section biopsy specimens representative of materials used in clinical practice.

During the last decade, technological advances have permitted development of software for automated assessment of **immunohistochemical** expression. Although

digital image analysis algorithms might be expected to be superior to visual analysis, computational approaches have not been proven superior in recognizing cancer and consistently selecting the correct objects to score. Recent progress in generation of image capture platforms and software packages has raised the possibility that machine-based approaches might rival pathologist-based visual assessments for scoring Ki67.

To investigate this possibility, we undertook a study to assess reproducibility of multiple existing technologies for automated machine-measurement of Ki67 expression using slides from core-cut biopsies previously analyzed in the **International Ki67 in Breast Cancer Working Group** phase 3 study that evaluated reproducibility of visual Ki67 assessment. In that study, a standardized visual approach to scoring Ki67 met its pre-specified criterion for success. We now report results comparing 14 laboratories from 6 countries, using 7 different scanners and 10 software packages, in the absence of any prescribed scanner or software harmonization.

## **Materials and Methods:**

### Tissue:

This study was approved by the British Columbia Cancer Agency Clinical Research Ethics Board (protocol H10-03420). All samples used were donated by patients who signed a generic consent. All core-cut biopsy material used was excess to diagnostic requirements and ethically available for quality control studies.

The core-cut biopsy slides, all used in phase 3 of the International Ki67 in Breast Cancer Working Group initiatives<sup>1</sup> were from 30 cases of ER+ breast cancer (Supplemental Table 2).

Fourteen volunteer laboratories (3 of whom participated in the phase 3 visual scoring study) representing 6 countries, completed this automated image analysis study. Two laboratories contributed two sets of analysis results each, and these were treated as though they were independent laboratories for purposes of the analysis. Preparation of the Ki67, H&E and myoepithelial marker (p63) slides were as described<sup>1</sup>. Briefly, 5 adjacent sections from each of the 30 core-cut biopsy source blocks were centrally cut and stained for H&E (1 section), p63 (1 section) and Ki67 (3 sections), resulting in 3 groups of 30 Ki67 slides from 30 cases. One group of slides was damaged in the previous study, leaving only 2 sets available for this study. Participating laboratories were divided into 2 groups (7 laboratories in each group) and members within the same group were given the same set of glass slides to analyze (Supplemental Figure 1). Each laboratory had 2 weeks to scan the slides and then send them to the next laboratory on the list. Two sets of slides were circulated to expedite study completion, with the assumption that serial sections would be essentially identical with respect to Ki67 expression.

#### Slide Scanning and Analysis:

Image analysis systems selected by site principal investigators for use in this study covered a wide range (Table 1). The most common scanning platform was the Aperio which was used by 7 of the 14 laboratories. Once scanned images were generated, each site implemented its own choice of software packages for analysis,

with 10 different ones chosen by the 14 laboratories. While some groups used the same software packages, even then they were not used identically. Five laboratories scanned the slides at 40x while 9 scanned at 20x. Most systems required some human intervention in the image analysis process. Specifically, 9 systems required a human operator for an initial training step, 11 required human visual selection of “region of interest” and 2 had the user specify the number of “fields of view” for analysis. Five systems did not analyze by “field of view” methods, counting and averaging across the entire slide, and hence they were not able to generate maximum scores. One system analyzed images based solely on pixel colors, while all other systems included some notion of shape/size object selection. One system (laboratory D) did not use a slide scanner, but used a live microscope camera directly connected to the image analysis software. Two systems are based on open-source software.

#### Scoring instructions:

Participating laboratories were instructed to score the 30 core-cut biopsy slides using the image analysis system of their choice following their own standard operating procedure. No further instructions were given, no standardization slides were sent out and participants were unaware of others’ scores (including previous visual scores). All participating laboratories were given online access to the H&E and myoepithelial marker (p63) images for the 30 study cases. A Microsoft Excel spreadsheet was sent out to each laboratory for entering 1) the number of fields of view analyzed, 2) maximum score of the fields analyzed, 3) average scores across all the fields analyzed, 4) timing data, and 5) any comments they may have on the study slides. All laboratories provided details of their image analysis system by answering a set of questions. Two

laboratories (Lab H and Lab L) submitted scores using two image analysis approaches. Table 1 shows the details of the image analysis systems used in this study.

### *Ki67 score calculation*

The various image analysis systems used in this study have their own definition of Ki67 score. Most defined Ki67 score as the percentage of invasive tumor cells positively stained in the examined field(s). However, one measured simply the percentage of pixels with a certain color. Five (out of 14) image analysis systems did not analyze the scanned image by field of view; instead, the entire region of interest was analyzed and a single Ki67 score was reported.

### Statistical Design and Analysis:

#### *Intraclass correlation coefficient as the reproducibility metric*

Intraclass correlation coefficient estimates (ranging from 0 to 1, with 1 representing perfect reproducibility) were computed by variance component analysis previously described<sup>6</sup> (see statistics supplement). Analyses partitioned total variability in log-transformed Ki67 scores into variance contributions from scoring laboratory, patient tumor (biological variation – each core-cut biopsy block represents a unique patient), section (slide) of the core-cut biopsy block, and remaining variability absorbed in residual error. Same-section (laboratories scoring same set of slides) and different-section intraclass correlation coefficients (laboratories scoring different sections of same block) were computed, representing proportion of the total variation (biological + technical) attributable to biological variability between patients at the tumor section level and patient biopsy level, respectively.

Variance component and **intraclass correlation coefficient** estimates with 95% credible intervals were obtained using packages lme4 and MCMCglmm in R version 3.2.1<sup>7</sup>. Data were visualized using heat maps, boxplots and spaghetti plots.

#### *Pre-specified criteria for success*

Primary criteria for success used in the phase 3 visual scoring study<sup>1</sup> were also used here: achieving an **intraclass correlation coefficient** significantly greater than 0.80 for both same-section and different-section **intraclass correlation coefficient**. Significance was interpreted as the 95% credible interval for **intraclass correlation coefficient** lying completely above 0.80. (See statistics supplement for power analysis.)

#### *Handling of revised data from one laboratory*

After initial data analysis, data from one laboratory (Lab B) appeared markedly different than data from all other laboratories. Study leadership requested Lab B to quality review its data, without revealing to Lab B how its data differed from other laboratories' data. Lab B identified problems in its process and was permitted to submit revised data, acknowledging both in the study report (see statistics supplement). Summary statistics reported here are based on the revised data unless otherwise specified. However, for figures/plots showing individual data points, both the initial and revised results from Lab B are shown.

#### **Results:**

##### *Interlaboratory reproducibility of Ki67 according to score type.*

Participating laboratories were divided into Groups 1 and 2, with seven different laboratories in each group (Lab H and L submitted scores using two image analysis approaches and they were analyzed as though they were independent laboratories resulting in four sets of scores from these two laboratories, combined). A pre-stained set of 30 specimens, covering a representative range of Ki67 levels for ER+ breast cancer, was sent to an initial laboratory, scanned, and then sent to the next laboratory within each group. Figure 1 displays the side-by-side boxplots of Ki67 scores across laboratories, by group. Summary statistics for the Ki67 scores across the 14 laboratories are given in supplemental tables 3 and 4.

Variance components analysis produced estimates of the biological, laboratory, section, and residual variances for the average and maximum scoring methods (supplemental tables 5a-b). Estimates for different-section **intraclass correlation coefficient**, obtained without standardization across laboratories and using originally submitted data, were 0.83 (95% **credible interval**: 0.73–0.91) for automated average scores across 16 operators and 0.63 (95% **credible interval**: 0.44–0.80) for maximum scores across 10 operators (supplemental table 6). However, original data submissions were discovered to include outlier results from one laboratory that failed to follow its internal standard operating procedures. After quality review and correction of that laboratory's aberrant data, revised **intraclass correlation coefficient** estimates were 0.86 (95% **credible interval**: 0.79–0.93) for average scores and 0.76 (95% **credible interval**: 0.64–0.88) for maximum scores. The corresponding same-section **intraclass correlation coefficient** estimates for the average and maximum scores were 0.89 (95% **credible interval**: 0.83–0.95) and 0.77 (95% **credible interval**: 0.64–0.88) respectively.

This observation indicates excellent reproducibility for average score between automated image analysis systems scoring the same physical glass slides. Although the revised different-section **intraclass correlation coefficients** did not meet the pre-specified success criterion (lower bound of 95% credible interval did not exceed 0.80), the one for average score using corrected data from Lab B came very close.

When the secondary analysis was performed restricting to only the subgroup of laboratories using the Aperio platform (8 score sets), different-section **intraclass correlation coefficient** for automated average scores was 0.89 (95% **credible interval**: 0.81-0.96). (Only two laboratories in this subgroup reported maximum Ki67 scores, so variance components analysis was not conducted for that method.) A modest numerical increase compared to the analogous **intraclass correlation coefficient** for the full group of laboratories was observed. Although perhaps not a statistically significant increase, this result provides motivation to investigate whether standardization of automated scoring could further improve reproducibility.

Variance component analyses show that, regardless of scoring method, biological variation among different patients was the largest component of the total variation, indicating that the Ki67 score is reflecting inherent properties of the tumor and that the variation in scores introduced by different laboratories' scanning and scoring is not obscuring biological signal (Figure 2, supplemental tables 5a-b).

#### *Comparisons of absolute Ki67 scores between laboratories.*

The variation in scores across laboratories is shown in Figure 3, in spaghetti plot format. Each line represents scores from one laboratory for each of the 30 core biopsy

cases. The between-laboratory reproducibility at the lower end of the range of Ki67 values appears to be particularly good for the average/global method using the automated approach but this good performance did not extend to the lower values of Ki67 using the automated maximum/hot-spot method.

#### *Agreement of categorical Ki67 scores.*

In routine clinical laboratory settings, some pathologists may provide categorical Ki67 scores rather than exact staining percentages. To reflect this, an analysis was performed on a categorical level (instead of continuous 0-100% scale), considering categories <10%, 10-20% and >20% (commonly interpreted as low, intermediate and high Ki67 indices). Concordance of these categorical scores across laboratories and cases can be appreciated in a heat map format with the columns (laboratories) sorted (within each group) by the median scores across cases, and the rows (cases) sorted by the median scores across laboratories (Figure 4). Each box (representing one laboratory's score for one case) is color-coded according to the three categories. Among the 30 breast cancer cases, 11 showed complete agreement across laboratories for categorized average scores, and 12 showed complete agreement using categorized maximum scores (Figures 3 and 4). This display also illustrates that laboratories measuring higher or lower than others did so fairly consistently, presumably influenced by thresholds set by the software each laboratory used.

#### *Comparison with standardized visual scoring.*

Standardized visual scores obtained previously on these same slides<sup>1</sup> and the non-standardized automated scores obtained in the current study show a high degree of

similarity across the spectrum of cases, although better for average score than for maximum score methodology (Figure 5). Although this study was not statistically designed to compare standardized visual to non-standardized automated scoring, observed score ranges and reproducibility appear similar: **intraclass correlation coefficient** for average standardized visual = 0.87 (95% credible interval: 0.81-0.93) compared to **intraclass correlation coefficient** for average non-standardized automated (using Lab B's revised results) = 0.86 (95% **credible interval**: 0.79-0.93).

## **Discussion:**

The analysis of Ki67, while often perceived as valuable, has not been widely adopted for directing routine breast cancer management, mostly due to lack of standardization across laboratories<sup>8</sup>. While other **International Ki67 in Breast Cancer Working Group** studies have focused on standardization of visual scoring, here we tested the hypothesis that simple adoption of an automated method could achieve standardization. The data support the hypothesis in that reproducibility across independent laboratories from around the world was observed to be much higher with non-standardized digital imaging analysis compared to what was seen in the first similar effort involving pathologist visual scoring without standardization<sup>5</sup>. Given this higher “starting point” for reproducibility, we are optimistic that the addition of standardization to the automated process may lead to a highly uniform and reproducible scoring method suitable for eventually achieving clinical validation and ultimately broad clinical adoption, but this remains to be assessed.

Arguably, establishing comparability of machine scores to human reads is an important step toward incorporating Ki67 results into routine clinical care. However, since we felt that the machine-based scoring should first be shown to have good reproducibility prior to comparison to visual scoring, this study was primarily designed to assess reproducibility among (unstandardized) automated methods. Although this study was not designed for a statistically well-powered comparison of the two approaches, we did conduct an exploratory comparison of reproducibility of automated scoring versus human visual scoring of these slides. In this regard, the difference between the non-standardized automated average score **intraclass correlation coefficient** (using Lab B's revised results) of 0.86 (95% **credible interval**: 0.79-0.93) and the standardized visual **intraclass correlation coefficient** of 0.87 (95% **credible interval**: 0.81-0.93) appears minimal. We propose that this observation suggests that a statistically powered formal comparison of the ability of Ki67 to predict clinical outcome when scored according to the various approaches should proceed only following standardization of the automated systems. Standardization should include an assessment of whether the apparent superior performance of the automated average/global method of scoring at lower levels of Ki67 can be confirmed.

Another key observation was provided, perhaps inadvertently, by Lab B. Upon assessment of its initially submitted data, it was clear that Lab B was an outlier compared to all other laboratories (Figure 1). In consultation with the director of Lab B, it became evident that deviations from the laboratory's standard operating procedures had occurred. Submission of revised scores was permitted with agreement to report both sets of results. Although Lab B's second data set remained somewhat high

compared to data from other laboratories, differences were less dramatic. This illustrates the importance of both careful human oversight of machine data and also standardization across laboratory sites, whether for pathologist reads or for machine calibration.

Limitations to this work, besides inadequate power to compare automated to visual scoring, are important to appreciate. There was heterogeneity in the scanners and software used by the laboratories, but insufficient numbers using each platform for formal comparison. All sections were cut and pre-stained in a single laboratory using a uniform method, but these factors would contribute additional variability in Ki67 determinations across clinical laboratories in practice. Further, some laboratories batched scanning before analysis while others scanned and analyzed individual cases in succession. These aspects could affect results, but the impact was not quantifiable to this level of detail.

The relative prognostic performance of hot spot (i.e., determining score based on most mitotically active area of tumor) versus average scoring of Ki67 expression has been a longstanding and still unresolved issue<sup>9,10</sup>. We used image analysis-determined maximum scores to attempt to reflect the human concept of hot spot, but these assessments relied on selection of a FOV without standardized hot spot sampling criteria. FOV size may be another important aspect of standardization as suggested by reports that larger FOV sizes for hot spot determination are associated with decreased Ki67 scores<sup>11</sup>. Further studies are needed to define optimal criteria for hot spot analysis to improve reproducibility of both visual and machine measurement. Future studies of Ki67 that include clinical outcome data are also needed to determine which of average,

hot spot or other score quantification can deliver Ki67 values most predictive of outcome when analytically standardized.

**Acknowledgments:**

This work was supported by a generous grant from the Breast Cancer Research Foundation (D.F.H.). Additional funding for the UK laboratories was received from Breakthrough Breast Cancer and the National Institute for Health Research Biomedical Research Centre at the Royal Marsden Hospital. Funding for the Ontario Institute for Cancer Research is provided by the Government of Ontario. Judith Hugh is the Lilian McCullough Chair in Breast Cancer Surgery Research and the CBCF Prairies/NWT Chapter. We are grateful to the Breast International Group and North American Breast Cancer Group (BIG-NABCG) collaboration, including the leadership of Nancy Davidson, Thomas Buchholz, Martine Piccart, and Larry Norton.

**Disclosure/Conflict of Interest:**

David Rimm works or has worked as a consultant to Astra Zeneca, Agendia, Agilent, Biocept, BMS, Cell Signaling Technology, Cepheid, Merck, OptraScan, Perkin Elmer, and Ultivue; has equity in PixelGear; and received research funding from Astra Zeneca, Cepheid, Navigate/Novartis, Gilead Sciences, Ultivue, and Perkin Elmer .

John Bartlett received honorarium from Oncology Education. John Bartlett have a consulting or advisory role with Insight Genetics, BioNTech, Due North and Biotheranostics.

Carsten Denkert received honoraria from Novartis, Pfizer, Amgen, MSD, Roche, Celgene and Teva. Carsten Denkert has been cofounder and shareholder of Sividon Diagnostics, Cologne. Carsten Denkert has a patent or intellectual property interest for VmScope Digital Pathology Software.

Chad Galderisi is the executive vice president, chief medical officer and laboratory director of Molecular MD.

Abhi Gholap is the chief executive office of Optra Technologies.

Anagha Jadhav is the director of digital pathology of Optra Technologies.

Richard Levenson is the co-founder of MUSE Microscopy Inc.

Keith Miller has a consulting or advisory role with Visiopharm.

Liron Pantanowitz has a consulting or advisory role with Hamamatsu, Leica, Ibex and Cambridge Healthtech Institute.

Mitch Dowsett has received lecture fees from Myriad.

Supplementary information is available at Modern Pathology's website.

## References:

1. Leung SCY, Nielsen TO, Zabaglo L, et al. Analytical validation of a standardized scoring protocol for Ki67: Phase 3 of an international multicenter collaboration. *NPJ Breast Cancer*. 2016;2:16014.
2. Yerushalmi R, Woods R, Ravdin PM, et al. Ki67 in breast cancer: Prognostic and predictive potential. *Lancet Oncol*. 2010;11:174-183.
3. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: Recommendations from the international Ki67 in breast cancer working group. *J Natl Cancer Inst*. 2011;103:1656-1664.
4. Nielsen T, Polley M, Leung S, et al. An international Ki67 reproducibility study. *Cancer Res*. 2012 (SABCS abstr S4-6);72(24 Suppl).
5. Polley MY, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. *J Natl Cancer Inst*. 2013;105:1897-1906.
6. Polley MY, Leung SC, Gao D, et al. An international study to increase concordance in Ki67 scoring. *Mod Pathol*. 2015;28:778-786.
7. R Core Team. *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. 2017.

8. Harris LN, Ismaila N, McShane LM, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American society of clinical oncology clinical practice guideline. *J Clin Oncol*. 2016;34:1134-1150.
9. Jang MH, Kim HJ, Chung YR, et al. A comparison of ki-67 counting methods in luminal breast cancer: The average method vs. the hot spot method. *PLoS One*. 2017;12:e0172031.
10. Brown JR, DiGiovanna MP, Killelea B, et al. Quantitative assessment ki-67 score for prediction of response to neoadjuvant chemotherapy in breast cancer. *Lab Invest*. 2014;94:98-106.
11. Christgen M, von Ahsen S, Christgen H, et al. The region-of-interest size impacts on Ki67 quantification by computer-assisted image analysis in breast cancer. *Hum Pathol*. 2015;46:1341-1349.
12. Schuffler PJ, Fuchs TJ, Ong CS, et al. TMARKER: A free software toolkit for histopathological cell counting and staining estimation. *J Pathol Inform*. 2013;4(Suppl):S2-3539.109804. Print 2013.
13. Klauschen F, Wienert S, Schmitt WD, et al. Standardized Ki67 diagnostics using automated scoring--clinical validation in the GeparTrio breast cancer study. *Clin Cancer Res*. 2015;21:3651-3657.
14. Wienert S, Heim D, Kotani M, et al. CognitionMaster: An object-based image analysis framework. *Diagn Pathol*. 2013;8:34-1596-8-34.

15. Wienert S, Heim D, Saeger K, et al. Detection and segmentation of cell nuclei in virtual microscopy images: A minimum-model approach. *Sci Rep.* 2012;2:503.

## Figure Legends:

**Figure 1.** Ki67 scores (a: average; b: maximum) of all 14 laboratories (by group): light grey for Group 1 and black for Group 2. Laboratories are ordered (within each group) by the median scores. The bottom/top of the box in each box plot represent the first (Q1)/third (Q3) quartiles, the bold line inside the box represents the median and the two bars outside the box represent the lowest/highest datum still within  $1.5 \times$  the inter-quartile range (Q3-Q1). Outliers are represented with empty circles. Lab B, H and L contributed two sets of scores each and they are indicated by “.1” and “.2”. Boxplot of B.1 is striped to indicate that this data was subsequently corrected by Lab B. The revised data from Lab B is B.2.

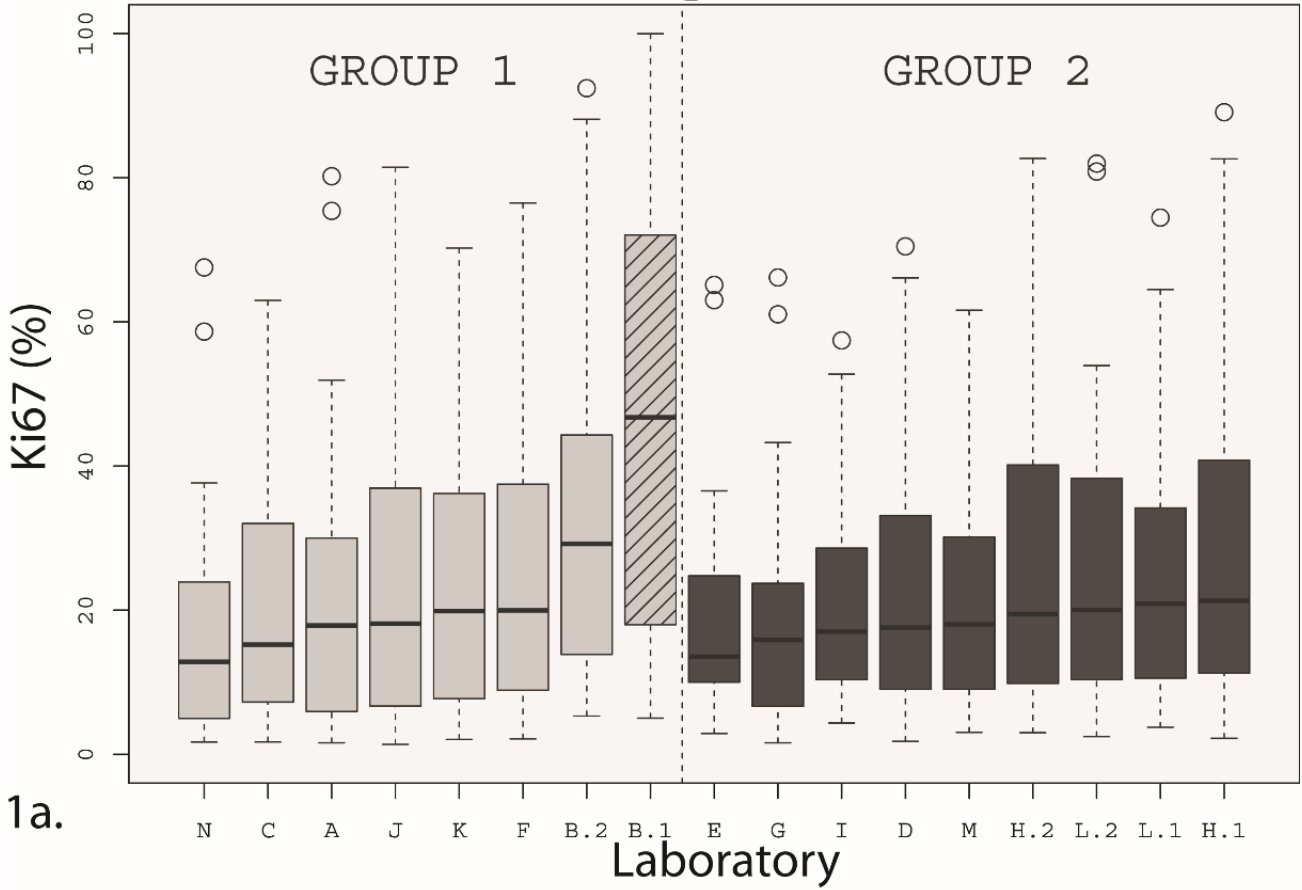
**Figures 2.** Variance component analysis. Variation due to different components are presented in a bar plot to show the relative magnitude differences between them. Numeric values of the variance component estimates and the corresponding confidence intervals are shown in supplemental table 5a.

**Figure 3.** Variability in Ki67 scores (a and c correspond to Group 1; b and d correspond to Group 2). Each line represents Ki67 scores from one laboratory, across the 30 cases (labelled by study code below the x-axis, where they have been ordered by mean score across laboratories). Shaded region indicates Ki67 scores between 10-20%. Original data from Lab B (B.1) are represented by a dotted line to indicate that these data were subsequently revised by Lab B.

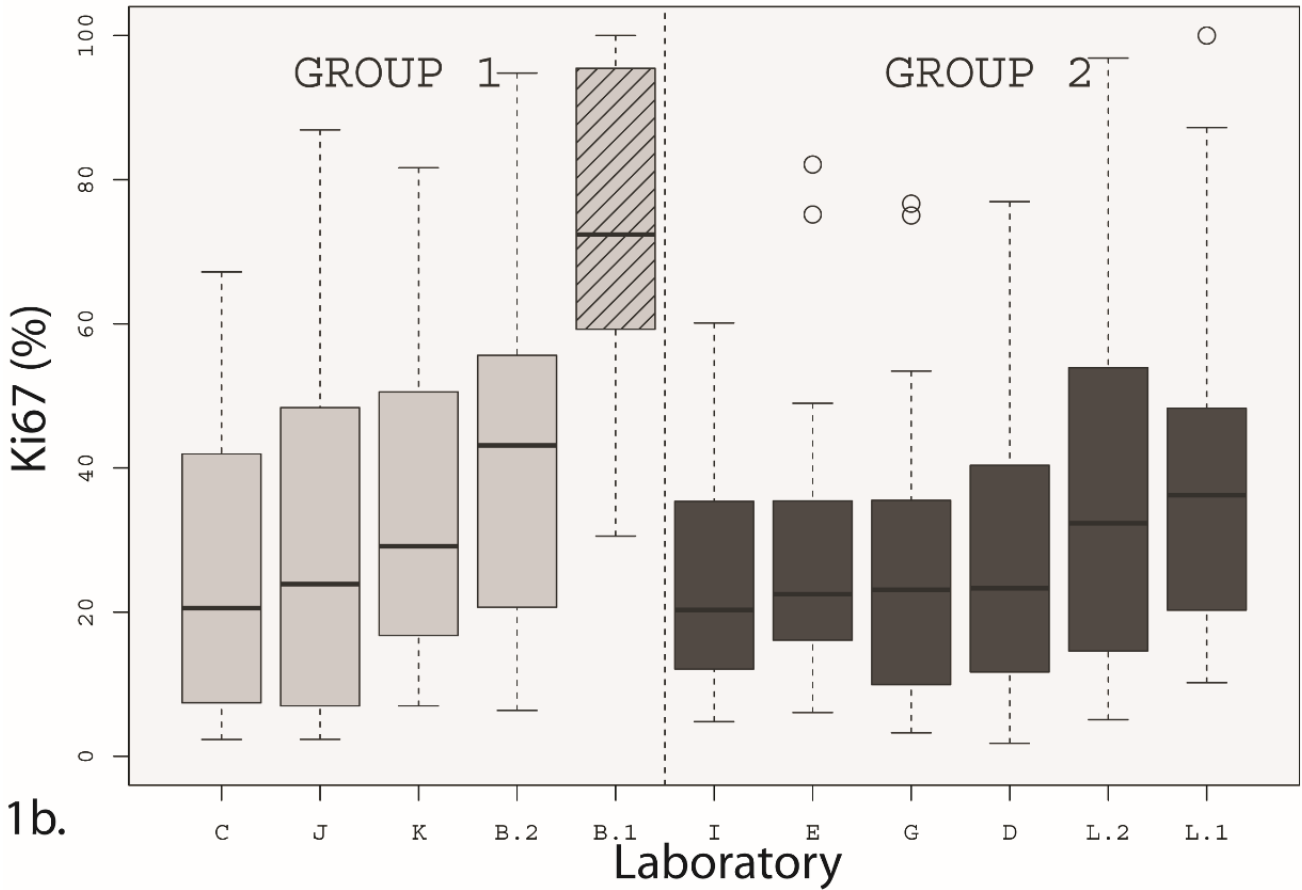
**Figure 4.** Heat map of Ki67 scores (a: average; b: maximum). Rows represent cases and columns represent laboratories. Green color indicates that the score is <10%, yellow 10-20% and red >20%. Grey indicates missing data. Cases are ordered by the median scores (across laboratories). Laboratories are ordered (within each group) by the median scores (across cases). B.1 is striped to indicate that these data were subsequently revised by Lab B. The revised data from Lab B are labeled B.2. The concordance between categorized (<10%, 10-20% and >20%) scores was assessed by Kappa statistics: 0.67 for average score and 0.46 for maximum score (using revised data by Lab B). The corresponding Kappa statistics for categorized visual scores was 0.60 for unweighted global score and 0.54 for hot-spot score<sup>1</sup>.

**Figure 5.** Comparison between standardized visual scoring and non-standardized automated scoring (a and c correspond to Group 1; b and d correspond to Group 2; a and b correspond to average/global score; c and d correspond to maximum/hot-spot score). Highlighted area indicates 10-20%. Original data from Lab B (B.1) have been removed from this analysis.

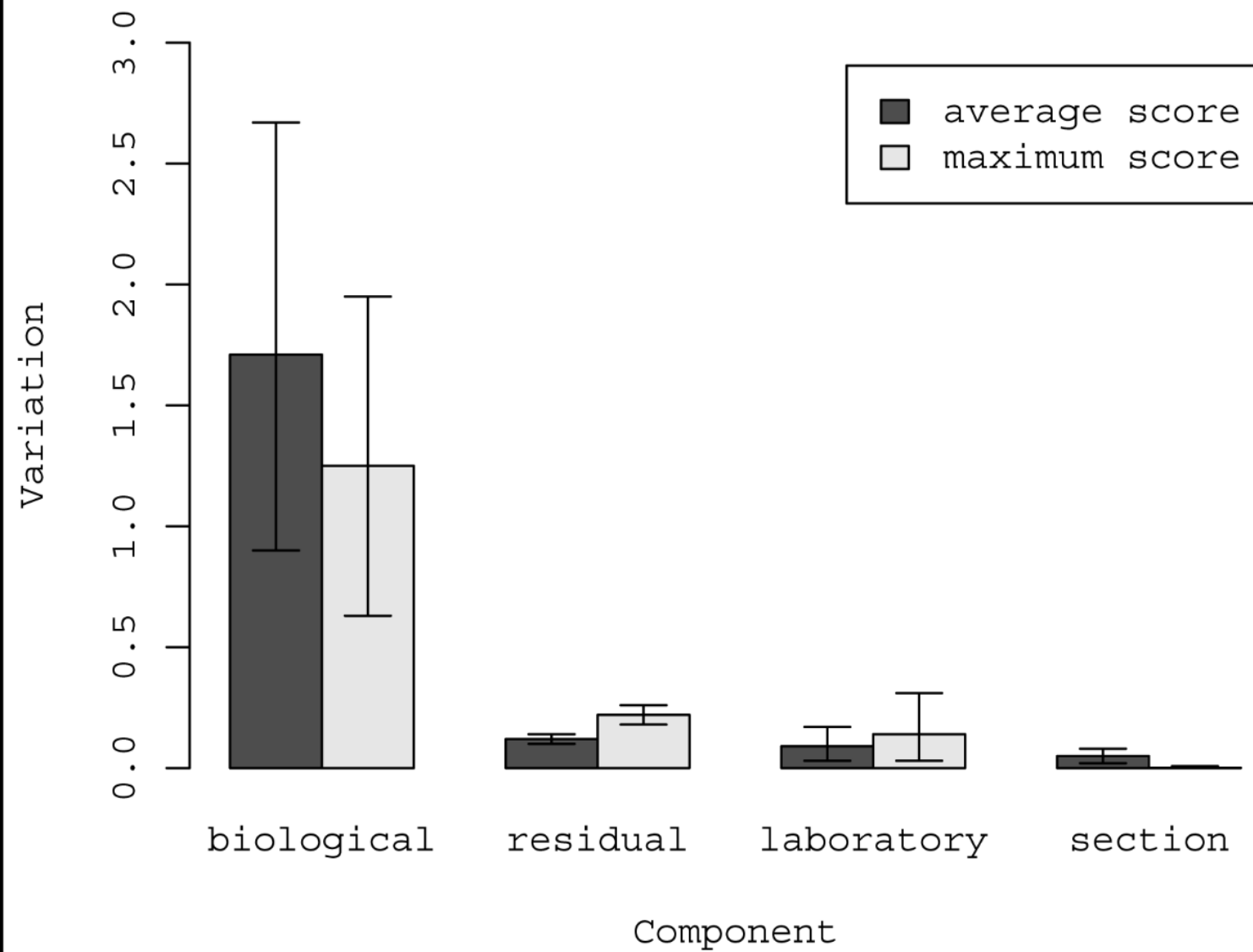
# Average score

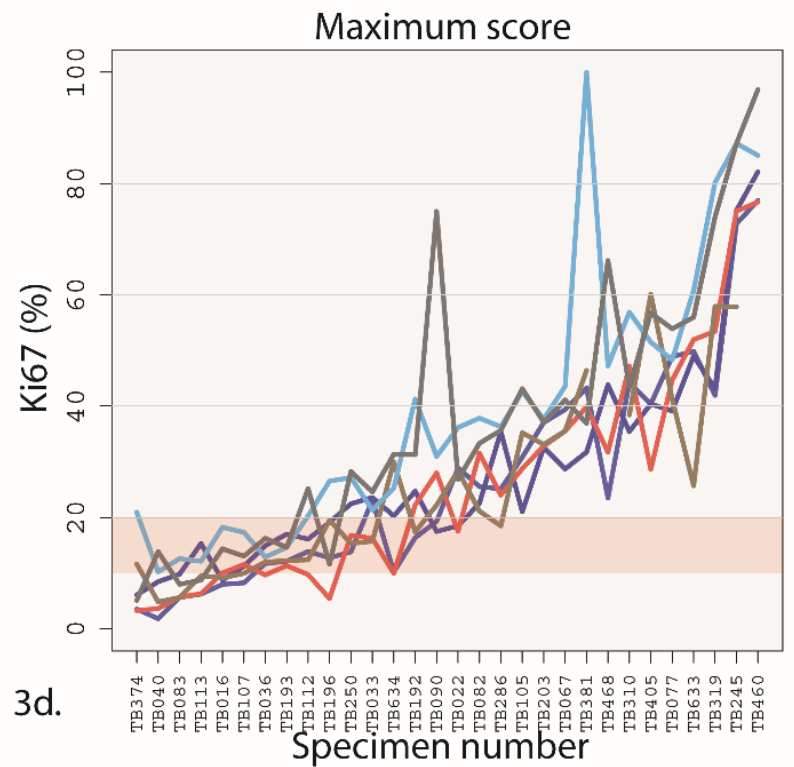
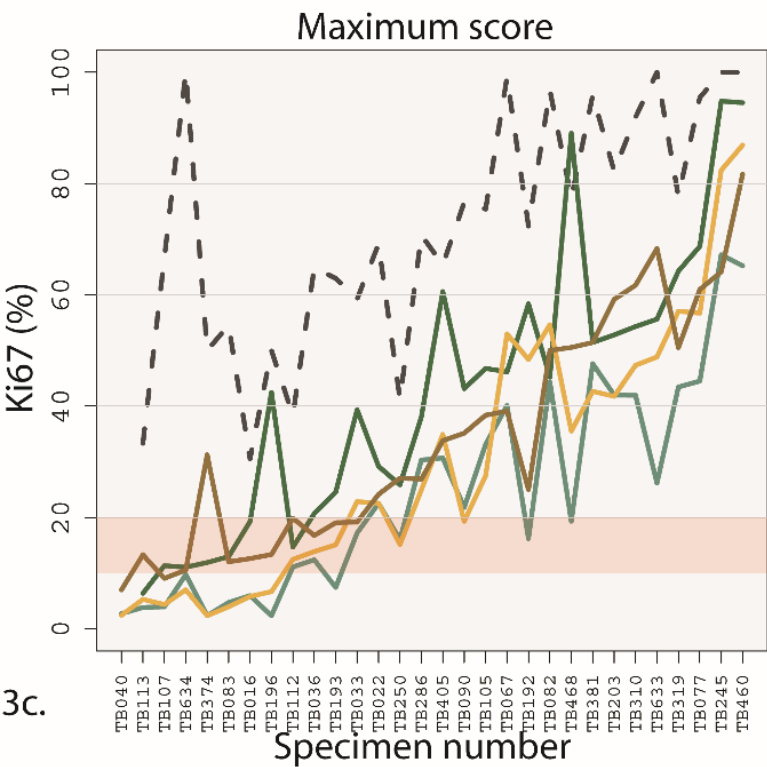
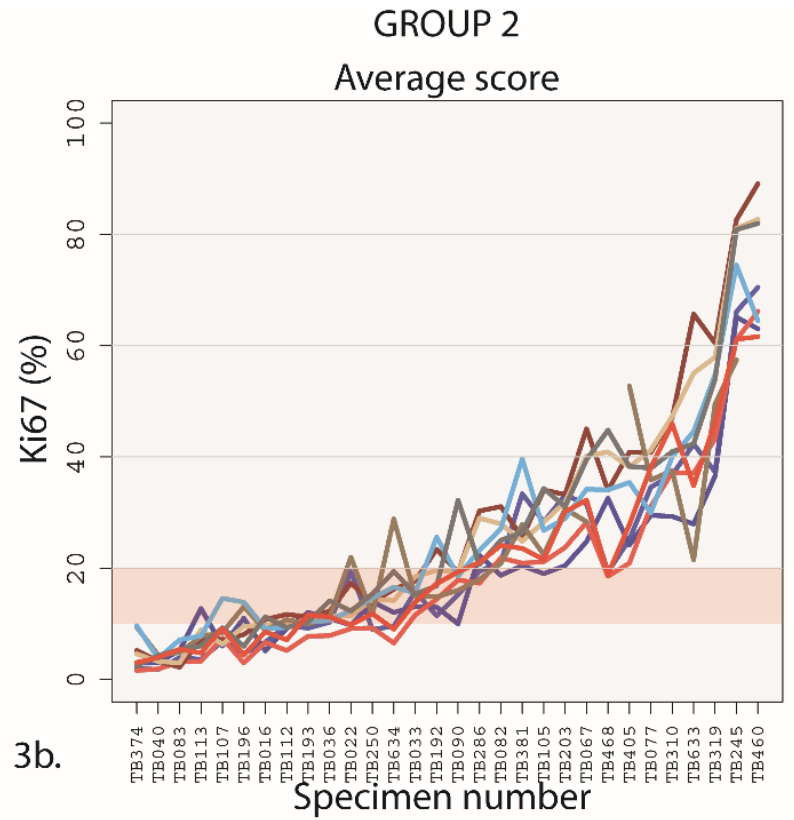
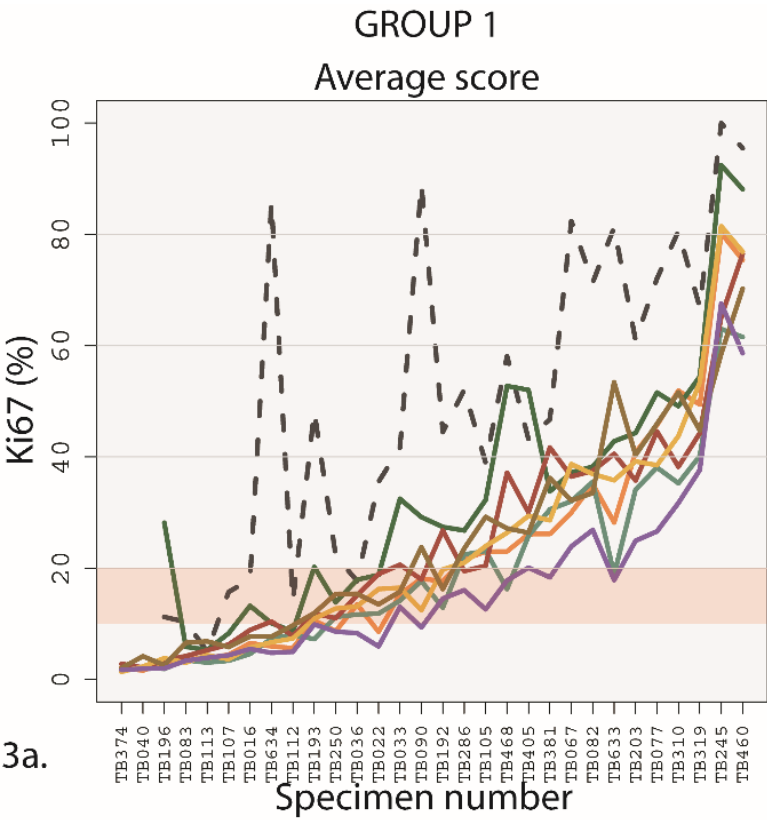


# Maximum score



## Variance component analysis



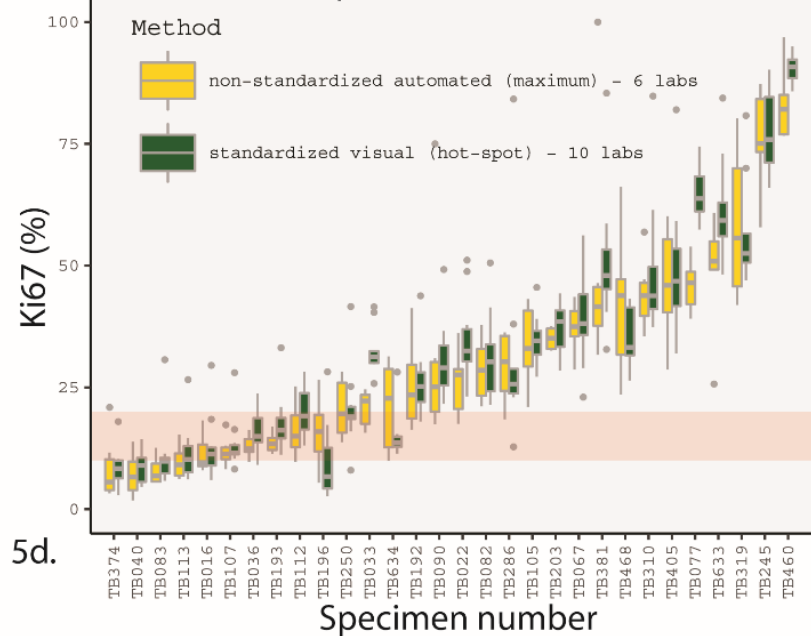
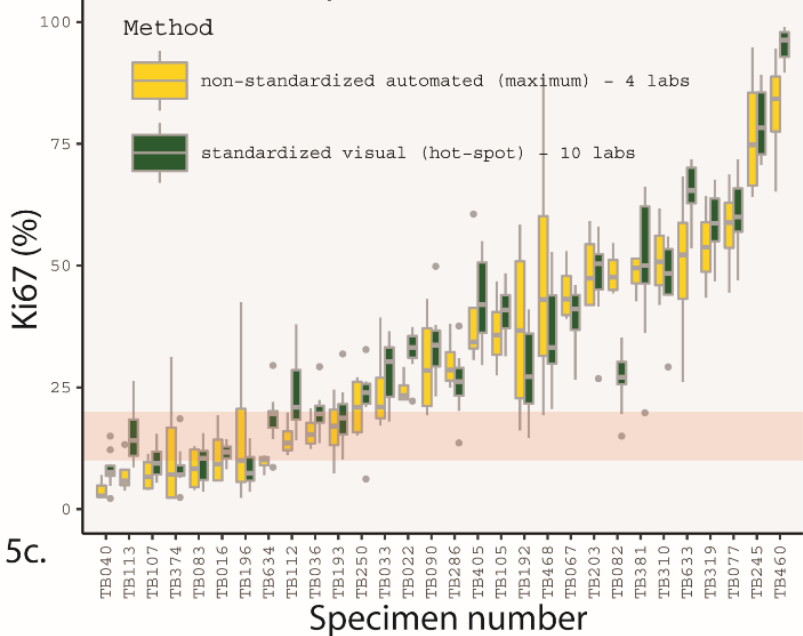
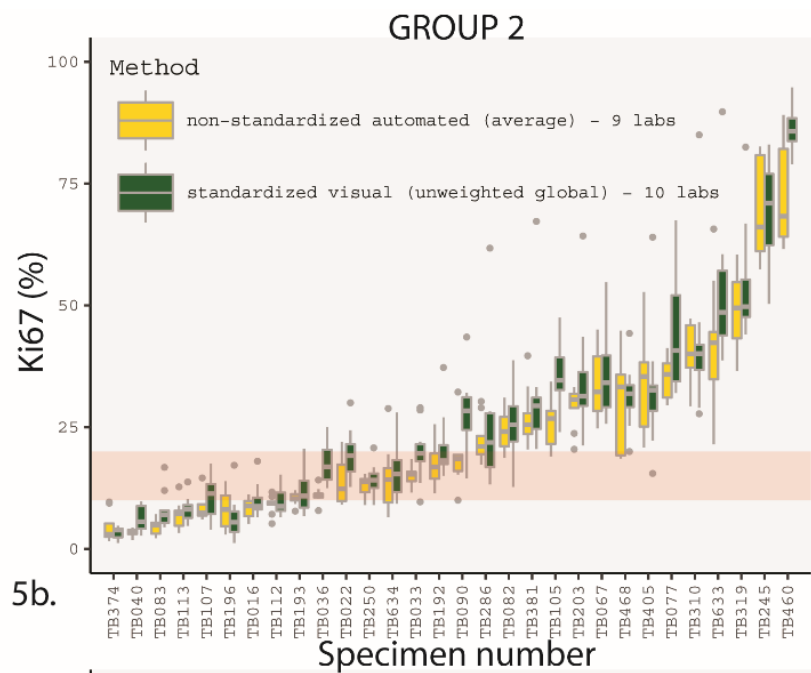
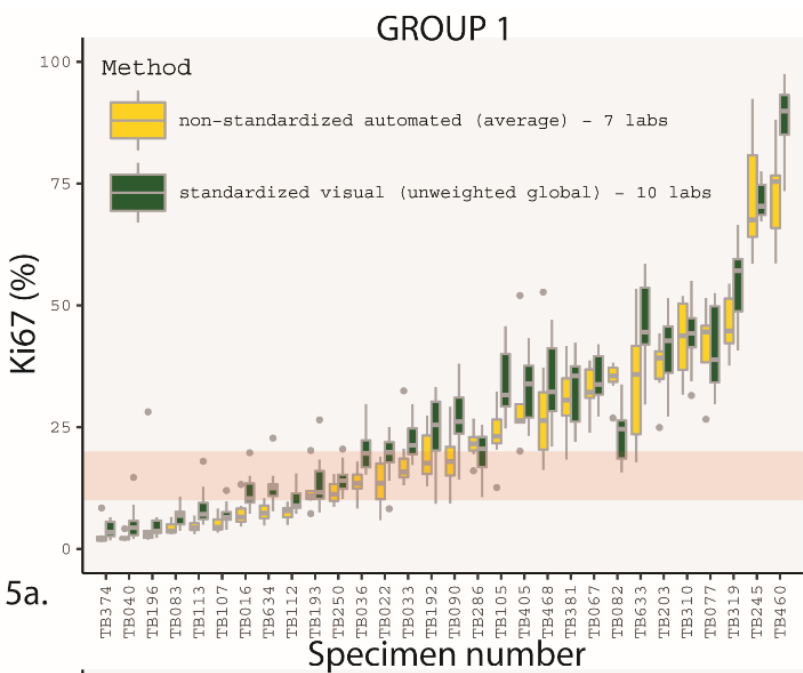


4a.

Specimen number	GROUP 1									GROUP 2								
	N	C	A	J	K	F	B.2	B.1	E	G	I	D	M	H.2	L.2	L.1	H.1	
TB374	2	2	2	1	2	3	8	12	3	2	9	2	3	5	2	10	5	
TB040	2	2	2	2	4	2			3	2	4	2	4	3	4	4	3	
TB083	3	3	4	3	7	4	6	10	5	3	5	4	5	3	5	7	2	
TB196	2	2	3	4	3	3	28	11	11	3	13	5	4	10	6	14	8	
TB113	4	3	4	4	7	5	5	5	13	3	8	4	5	9	6	8	7	
TB107	4	3	4	4	6	6	8	16	6	7	8	7	9	6	9	15	7	
TB016	5	5	7	6	8	9	13	18	5	7	9	7	9	10	11	9	11	
TB112	5	8	6	7	10	8	9	14	9	5	11	10	7	10	9	9	12	
TB634	5	7	6	7	8	10	10	86	12	7	29	10	9	14	19	17	16	
TB193	10	7	11	11	12	12	20	48	12	8	10	9	12	11	10	10	11	
TB036	8	12	13	13	15	15	18	18	11	8	11	10	11	11	14	11	12	
TB022	6	12	9	16	13	19	19	36	10	9	22	19	10	11	12	12	17	
TB250	9	11	9	13	15	11	14	22	14	9	12	9	12	14	15	15	14	
TB033	13	14	15	16	16	21	32	42	13	12	15	16	14	18	15	16	18	
TB192	15	13	18	20	16	27	27	44	13	14	15	11	17	20	17	26	23	
TB090	9	18	18	12	24	18	29	89	10	18	16	15	19	19	32	19	19	
TB286	16	22	22	21	24	19	27	52	22	17	18	19	21	29	21	23	30	
TB105	13	23	23	24	29	20	32	39	19	21	22	28	22	28	34	27	34	
TB381	18	31	26	29	36	42	34	47	20	21	28	33	24	25	26	40	25	
TB082	27	36	35	37	34	37	38	71	19	22	21	21	24	28	25	27	31	
TB405	20	26	26	29	26	30	52	43	24	21	53	25	28	38	38	35	41	
TB468	18	16	23	26	27	37	53	58	33	19		19	19	41	45	34	34	
TB067	24	32	30	39	32	36	37	82	25	28	28	32	32	40	40	34	45	
TB203	25	34	41	39	40	36	44	61	20	24	31	33	30	32	31	29	33	
TB077	27	38	46	39	46	45	52	72	30	31	36	35	38	41	38	30	41	
TB633	18	19	28	36	53	41	43	81	28	37	22	42	35	55	42	45	66	
TB310	32	35	52	44	52	38	49	81	29	37	38	37	46	47	41	40	47	
TB319	38	40	49	53	45	44	54	87	37	43	50	37	46	58	54	55	60	
TB245	68	63	80	81	59	65	92	100	65	61	57	66	61	81	81	74	83	
TB460	59	62	75	77	70	76	88	95	63	66		70	62	83	82	64	89	

4b.

Specimen number	GROUP 1					GROUP 2					
	C	J	K	B.2	B.1	I	E	G	D	L.2	L.1
TB040	3	2	7			5	8	4	2	14	10
TB374	2	2	31	12	50	12	6	3	4	5	21
TB083	5	4	12	13	55	6	10	6	6	8	13
TB113	4	5	13	6	33	9	15	6	6	9	12
TB016	6	6	13	19	31	9	9	10	8	14	18
TB634	10	7	11	11	100	30	20	10	10	31	25
TB107	4	4	9	11	87	10	11	12	8	13	17
TB196	2	7	13	42	80	19	19	5	13	12	27
TB036	12	14	17	21	65	12	15	10	12	16	13
TB193	7	15	19	25	63	12	17	11	12	15	15
TB112	11	12	20	15	38	12	16	10	14	25	20
TB250	16	15	27	26	41	15	22	17	14	28	27
TB033	17	23	19	39	59	16	24	16	23	25	21
TB192	16	48	25	58	72	17	25	22	16	31	41
TB022	23	23	24	29	69	28	18	18	29	27	36
TB090	22	19	35	43	77	22	17	28	19	75	31
TB286	30	25	27	38	71	18	35	24	25	36	36
TB105	33	27	38	47	75	35	21	29	31	43	43
TB203	42	42	59	53	82	33	33	33	37	37	38
TB082	44	55	50	45	97	21	23	32	26	33	38
TB067	40	53	39	46	99	36	29	36	39	41	44
TB405	31	35	34	61	65	60	40	29	40	57	52
TB468	19	36	51	89	77		44	32	23	66	47
TB381	48	43	51	51	96	46	32	40	43	37	100
TB310	42	47	62	54	92	38	35	47	44	43	57
TB077	44	57	61	69	95	41	39	45	49	54	48
TB633	26	49	68	56	100	26	49	52	50	56	61
TB319	43	57	51	64	78	58	43	53	42	74	80
TB245	67	82	64	95	100	58	75	75	73	87	87
TB460	65	87	82	95	100	82	77	77	97	85	85



DEID (GROUP #)	SCANNER/ CAMERA (SCAN RESOLUTION)	ANALYSIS SYSTEM	TRAINING (TIME USED IN MINUTES)	SELECT ROI	SPECIFY NUMBER OF FOVS	SCORE TYPE	APPROX. SCAN TIME PER SLIDE (MINUTES)	APPROX. ANALYSIS TIME PER SLIDE (MINUTES)	OPEN- SOURCE SOFT- WARE	YES = PIXEL BASED ONLY NO = INVOLVES OBJECT SELECTION BASED ON SIZE/ SHAPE
A (1)	Aperio ScanScope XT (40x)	Aperio ePathology image analysis (Spectrum image nuclear algorithm analysis software)	yes (60)	yes	no	avg. only	10	3.5	no	no
B (1)	Ariol SL50 version 4.0.053 (20x)	Kisight nuclear staining analysis (Ariol)	no	yes	yes	both	20	10	no	no
C (1)	Ventana iScan Coreo (20x)	TMARKER (ETH Zurich NEXUS Personalized Health Technologies) <sup>1</sup>	yes (15)	yes	no	both	4	2.5	yes	no
D (2)	Jenoptik microscope camera ProgRes SpeedXT core 5 (20x)	Cognition Master Professional Suite Ki67 Quantifier <sup>2,3,4</sup>	no	yes	no	both	NA <sup>5</sup>	5	no	no
E (2)	OPTRASCAN (20x)	Optra Assays Ki67 image analysis	yes (30)	yes	no	both	3	5	no	no
F (1)	Aperio ScanScope XT (20x)	Definiens Tissue Studio 4.1	no	yes	no	avg. only	not reported	not reported	no	no
G (2)	Aperio ScanScope XT (20x)	Indica Labs HALO image analysis (CytoNuclear v1.4)	yes (90)	yes	no	both	1	5	no	no
H (2)	Aperio AT2 (20x)	CellVigene (version 1.034)	yes (15)	no	no	avg. only	3	1.5	no	yes
I (2)	Menarini D- sight Fluo (40x)	D-sight (version 6)	yes (not reported)	yes	no	both	5	1	no	no
J (1)	Ventana iScan Coreo (20x)	Virtuoso (Ventana Ki-67 algorithm software version 5.3)	no	yes	no	both	5	4	no	no
K (1)	Aperio ScanScope XT (40x)	Aperio ePathology image analysis (Brightfield ToolBox)	yes (30)	yes	no	both	5	8	no	no
L (2)	Hamamatsu NanoZoomer (40x)	Definiens Tissue Studio 4.1	yes (180)	no	optional	both	1	4.5	no	no
M (2)	Aperio CS2 (40x)	Aperio ePathology image analysis (FDA cleared package for Estrogen Receptor nuclear staining; version 9)	no	yes	no	avg. only	10	7	no	no
N (1)	Aperio CS (20x)	Indica Labs HALO image analysis (CytoNuclear v1.4)	yes (10)	no	no	avg. only	2	15	no	no

**Table 1.** Details of image analysis systems.

<sup>1</sup> Schüffler *et al.* J Pathol Inform 2013<sup>12</sup>

<sup>2</sup> Klauschen *et al.* Clin Cancer Res. 2015<sup>13</sup>

<sup>3</sup> Wienert *et al.* Diagn Pathol. 2013<sup>14</sup>

<sup>4</sup> Wienert *et al.* Sci Rep. 2012<sup>15</sup>

<sup>5</sup> for system D no scanning was performed, the image analysis was performed directly with a live camera image from the microscope

ROI = region of interest

FOV = field of view