Performance Characteristics of the BluePrint® Breast Cancer Diagnostic Test☆

Lorenza Mittempergher [a,1], Leonie JMJ Delahaye [a,1], Anke T Witteveen [a], Mireille HJ Snel [a], Sammy Mee [b], Bob Y Chan [b], Christa Dreezen [a], Naomi Besseling [a], Ernest JT Luiten [c], Annuska M Glas [a] *

[a] Research and Development, Agendia N.V., Science Park 406, 1098 XH Amsterdam, The Netherlands
[b] Product Support, Agendia Inc., 22 Morgan, Irvine, CA 92780, USA
[c] Department of Surgery, Amphia Hospital, Molengracht 21, 4818 CK, Breda, The Netherlands

ARTICLE INFO

ABSTRACT

The analytical performance of a multi-gene diagnostic signature depends on many parameters, including precision, repeatability, reproducibility and intra-tumor heterogeneity. Here we study the analytical performance of the BluePrint 80-gene breast cancer molecular subtyping test through determination of these performance characteristics. BluePrint measures the expression of 80 genes that assess functional pathways which determine the intrinsic breast cancer molecular subtypes (i.e. Luminal-type, HER2-type, Basal-type). Knowing a tumor's dominant functional pathway can help allocate effective treatment to appropriate patients.

Here we show that BluePrint is a highly precise and highly reproducible test with correlations above 98% based on the generated index and subtype concordance above 99%. Therefore, BluePrint can be used as a robust and reliable tool to identify breast cancer molecular subtypes.

Introduction

Breast cancer (BC) is a heterogenous disease in nearly all aspects of disease progression from genetic risk and environmental accelerators, to growth and metastatic potential, and finally to treatment response. Multiple methods have been utilized to categorize the heterogeneity into distinct subgroups, be it with immunohistochemistry (IHC) for hormone receptor (HR) protein status or more recently with RNA-based assays for molecular subtypes [1–3]. Estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are well established as the gold standards in IHC testing and used to classify patients into what are now generally accepted clinical subtypes. The majority of BC patients are classified as HR-positive/HER2-negative, steadily comprising about 70% of all BC cases [2–4]. The less abundant subgroups include patients with HR-negative/HER2-negative (i.e. triple negative) and HER2-positive clinical subtypes. These clinical subtypes have closely correlated molecular subtype equivalents where HR-positive patients are mostly Luminal-type, triple-negative are Basal-type, and HER2-positive are HER2-type. While there is still not a gold standard for molecular subtyping, these three subgroups are the most universally recognized across different subtyping methods.

The BluePrint 80-gene breast cancer molecular subtyping test is unique in that it was developed using the IHC-based clinical subtype as a guide [2]. Conversely to other assays for breast cancer molecular subtyping [1,3], BluePrint was developed to bridge clinical pathology and research molecular subtyping, resulting in a molecular diagnostic assay with predictive value. The predictive status of IHC biomarkers were leveraged by using HR and HER2 status as way to group patients with the same expected treatment response before choosing the best set of genes for each subtype signature. For example, only tumors that were HR-positive by both IHC and single-gene RNA gene expression were used for choosing Luminal-type signature genes. The BluePrint 80-gene assay is composed of three signatures, each measuring the similarity of a tumor to a Luminal-type (58 genes), Basal-type (28 genes), and HER2-type (4 genes) representative profile. For each tumor, the similarity to all three representative profiles is calculated and the subtype with the most positive magnitude is determined to be the test result. Several prospective studies have shown that tumors of different BluePrint subtype exhibit differences in long term survival and

response to neoadjuvant therapy; Luminal-type tumors have more favorable distant metastasis (DM) free survival but less pathological complete response to neoadjuvant therapy, whereas Basal-type and HER-type tumors have less favorable DM free survival but are more sensitive to chemotherapy [2,5–7]. By expanding from a single gene (i.e. ER-IHC) to a multigene signature, the RNA-based BluePrint assay can more robustly classify tumors, and with better response stratification than IHC-based clinical subtyping [2,5,8–10].

While the clinical validity and utility of BluePrint have been discussed previously, here we describe the technical performance characteristic of this multi-gene diagnostic signature proving its analytical validity. Reliable analytical results are important in diagnostics as the test results may be used in making informed decisions on treatment options, hence the quality and safety need to meet very high standards. The precision, repeatability, and reproducibility were evaluated using Agendia control samples. Intra-tumor heterogeneity was evaluated using serial sections of tumor samples and it shows that BluePrint consistently gives high quality and accurate results.

Methods

BluePrint Test

BluePrint is a microarray-based test that measures the expression of RNA extracted from formalin-fixed paraffin embedded (FFPE) breast tumor tissue. The test uses a custom-designed array chip manufactured under good manufacturing practices (GMP) by Agilent Technologies and the Agilent oligonucleotide microarray platform, which assesses the mRNA expression of the replicates of the 80 genes included in the BluePrint profile. The diagnostic microarray features eight subarrays per glass slide which are each individually hybridized. Each subarray

includes 465 normalization genes and over 500 probes for hybridization and printing quality control. All probes are printed in replicates on the array with a maximum of 15 replicates per probe per gene.

Briefly, total RNA is isolated from FFPE tissue with the RNeasy FFPE kit (Qiagen) in accordance with the manufacturer's instructions. Total RNA is DNase treated and amplified using a TransPLEX C-WTA whole-transcriptome amplification kit (Rubicon Genomics, Ann Arbor, MI). Amplified cDNA is labeled using the Genomic DNA Enzymatic Labeling Kit (Agilent Technologies, Santa Clara, CA) and hybridized onto Agendia's diagnostic arrays (custom-designed, Agilent Technologies), both according to the manufacturer's instructions. The BluePrint indices are calculated by taking the expression of the 80 BluePrint genes and comparing them to the three different subtype profiles (Luminal-type, HER2-type, and Basal-type) [2]. For each sample, three indices are generated and the subtype with the highest index of the three is the categorical subtype reported for the tumor.

All steps of the laboratory process and bioinformatics have quality control measurements and all steps and instruments are evaluated according to quality system regulation (QSR) design control defined by US Food and Drug Administration (FDA, https://www.fda.gov/medical-devices/) (21 CFR Part 820), the International Organization for Standardization (ISO, https://www.iso.org/) (13,485:2016 certification) and the National Committee for Clinical Laboratory Standards (NCCLS) guidelines (https://clsi.org/standards/).

For this study, only data and not samples were collected. All data and analyses used for this study comply with the current ethical laws of the Netherlands. All patient sample data were anonymized in accordance with national ethical guidelines ('Code for Proper Secondary Use of Human Tissues' Dutch Federation of Medical Scientific Societies), study samples had Institutional Review Board approvals.
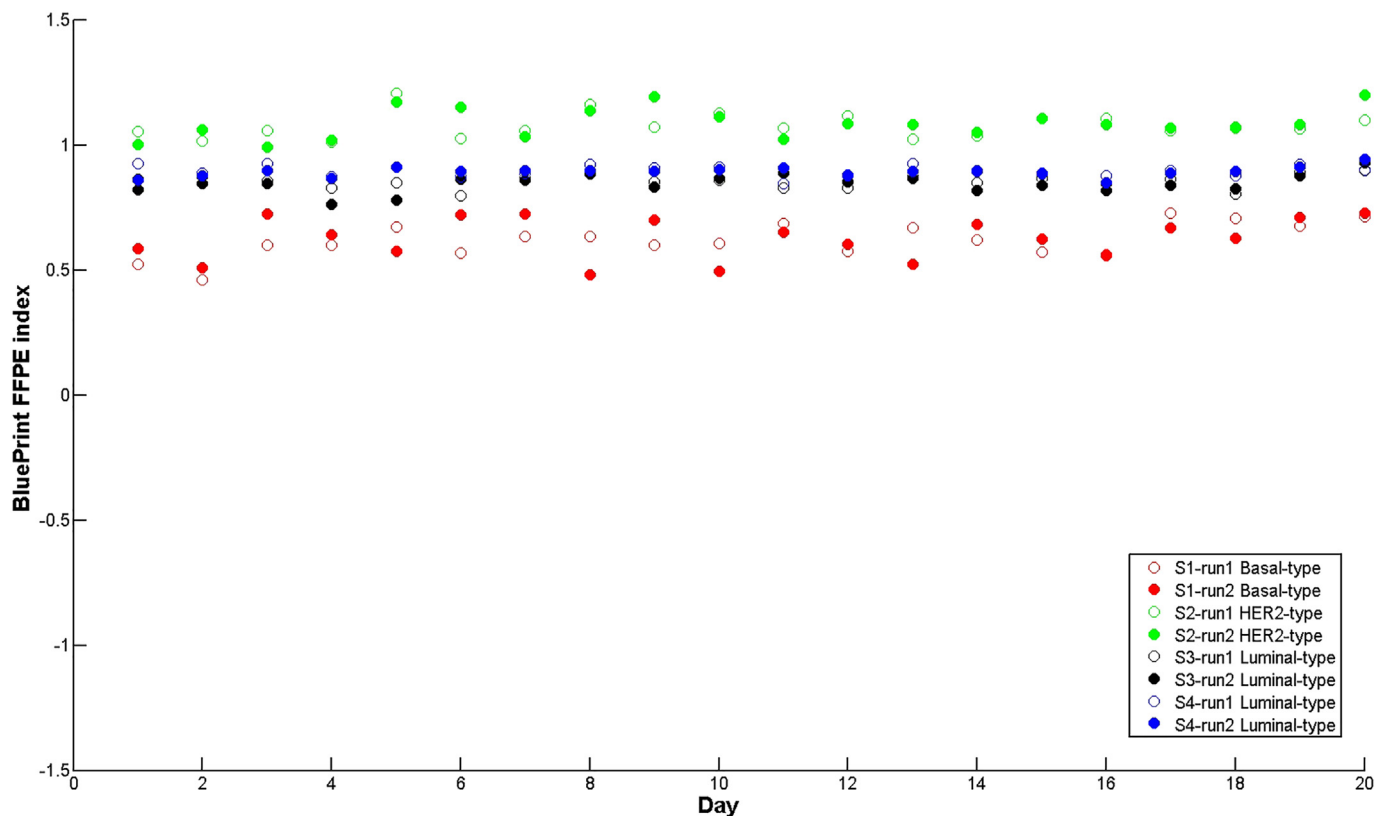

Figure 1. Repeatability and precision of BluePrint FFPE indices. Chart showing BluePrint indices (y-axis) of four clinical samples (S1, red, BluePrint Basal-type; S2 green, BluePrint HER2-type; S3 black, BluePrint Luminal-type; S4, blue, BluePrint Luminal-type) in duplicate (run1, run2) over a 20-day period (x-axis). Each dot represents a single breast cancer sample for which total RNA underwent BluePrint microarray laboratory processing and analysis. FFPE = formalin-fixed paraffin embedded. Categorical concordance was 100%.
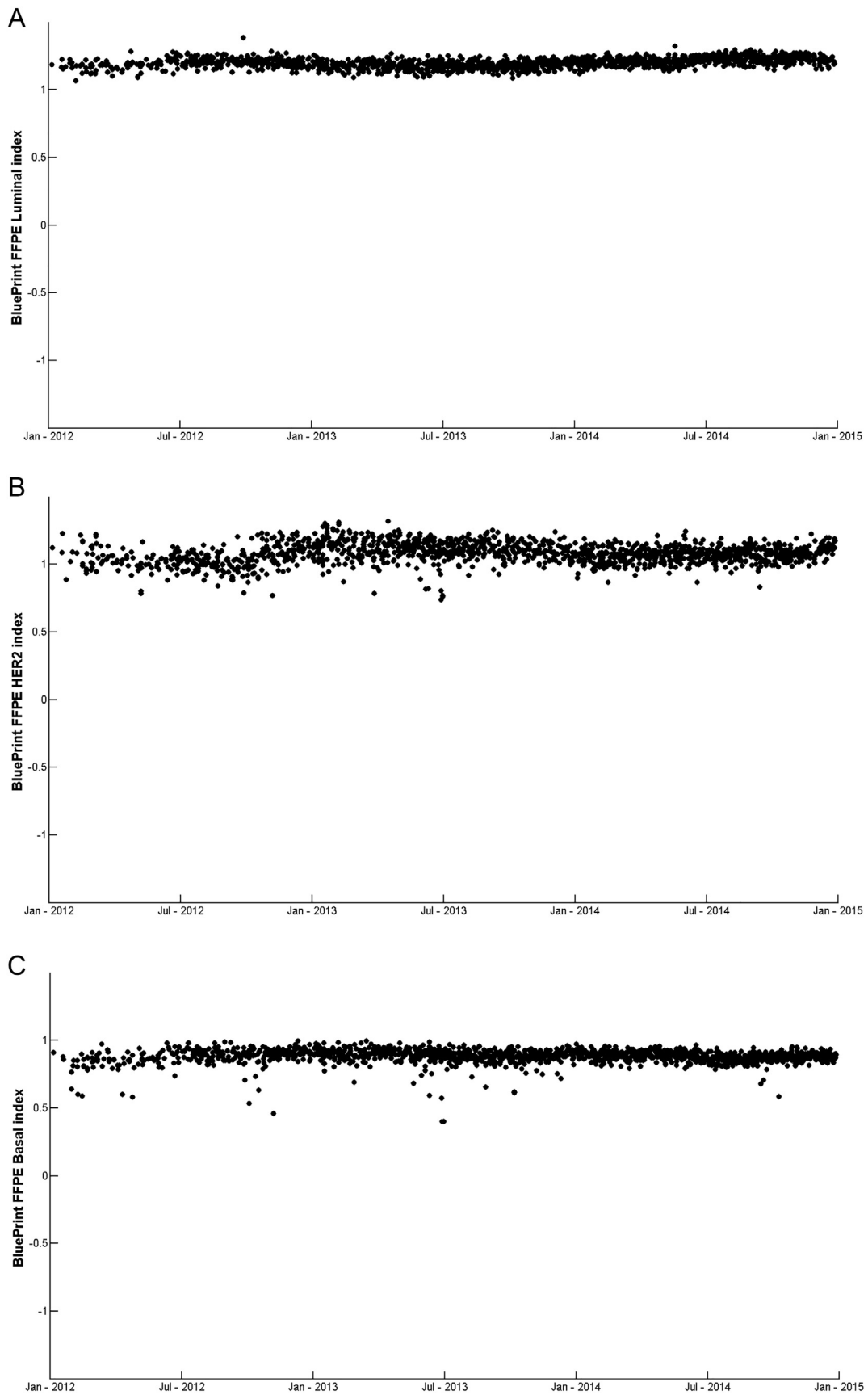
Figure 2. Over-time reproducibility of BluePrint FFPE indices. Chart showing over-time index measurements from January 2012 through January 2015 of three clinical control samples covering the three subtypes. **A**. C1, measuring Luminal-type index n = 1639, **B**. C2, measuring HER2-type index n = 1534, **C**. C3, measuring Basal-type index n = 1594. Concordance for all samples overtime was 100%. Each dot represents a single breast cancer sample for which total RNA underwent BluePrint microarray laboratory processing and analysis.

Analytical Performance Analysis

The BluePrint numerical index was used to evaluate the following performance characteristics: precision and repeatability, reproducibility (over-time, between different laboratories, between different RNA isolations of the same tumor tissue block), equivalence to the previous versions of the BluePrint test following technical adjustments over-time and equivalence to the initially developed BluePrint version using RNA isolated from fresh frozen (FF) tissue [2].

Precision and repeatability were assessed using a precision evaluation (PE) experiment using four FFPE clinical samples (two Luminal-type, one HER2-type and one Basal-type) that were repeatedly measured over a 20-day period. A duplicate run was performed for each sample each day to determine repeatability of the assay. This experiment includes all variables that occur in the diagnostic setting, such as different operators, equipment and reagent batches which are all part of the precision calculation.

The relative standard deviation is the standard deviation measured as a percentile of the total BluePrint range. The results are reported as relative stability (repeatability) and relative precision and are calculated by 100 minus the relative standard deviation.

Over-time reproducibility was tested by evaluating FFPE diagnostic control samples over 3 years to assess nearly all potential sources of variation in over 4700 measurements. In total three control samples (C1, C2 and C3) were used that cover the three BluePrint subtypes. (Luminal-type, HER2-type and Basal-type) Reproducibility is reported as percent relative range and is calculated by 100 minus the relative standard deviation.

Inter-laboratory reproducibility (i.e. between different laboratories) was evaluated by assessing the concordance of the BluePrint subtype between two laboratories situated in different locations (location 1, location 2). In total 97 FFPE samples were used for this assessment.

Reproducibility between different RNA isolations was assessed using two samplings derived from the same FFPE tumor tissue block. For 46 tumor samples, 4 FFPE sections of 5 μm each were used. A second sampling (4 sections of 5 μm) from the same tumor tissue was performed approximately 1 year later. Since the samples were stored at room temperature between samplings one should note that not only tumor heterogeneity but also tissue conservation might influence results. However, this effect is likely to be very small and therefore negligible.

Comparative analyses were performed to validate technical improvements (i.e. adjustments over-time) implemented in the BluePrint test compared to previously-released BluePrint test versions used as a gold standard. For any technical improvement substantial equivalence had to be shown versus this gold standard prior to implementation. Equivalence was determined by comparing BluePrint subtype outcome and indices between different versions. Acceptable limits were defined a priori, based on microarray data previously generated at Agendia between 2012 and 2016.

Agreement (equivalence) analysis between BluePrint results for FF and FFPE tissue was performed on the categorical outcome level in terms of concordance on a set of 413 clinical samples for which matched FFPE and FF microarray data were available at the time of the study. Out of the 413
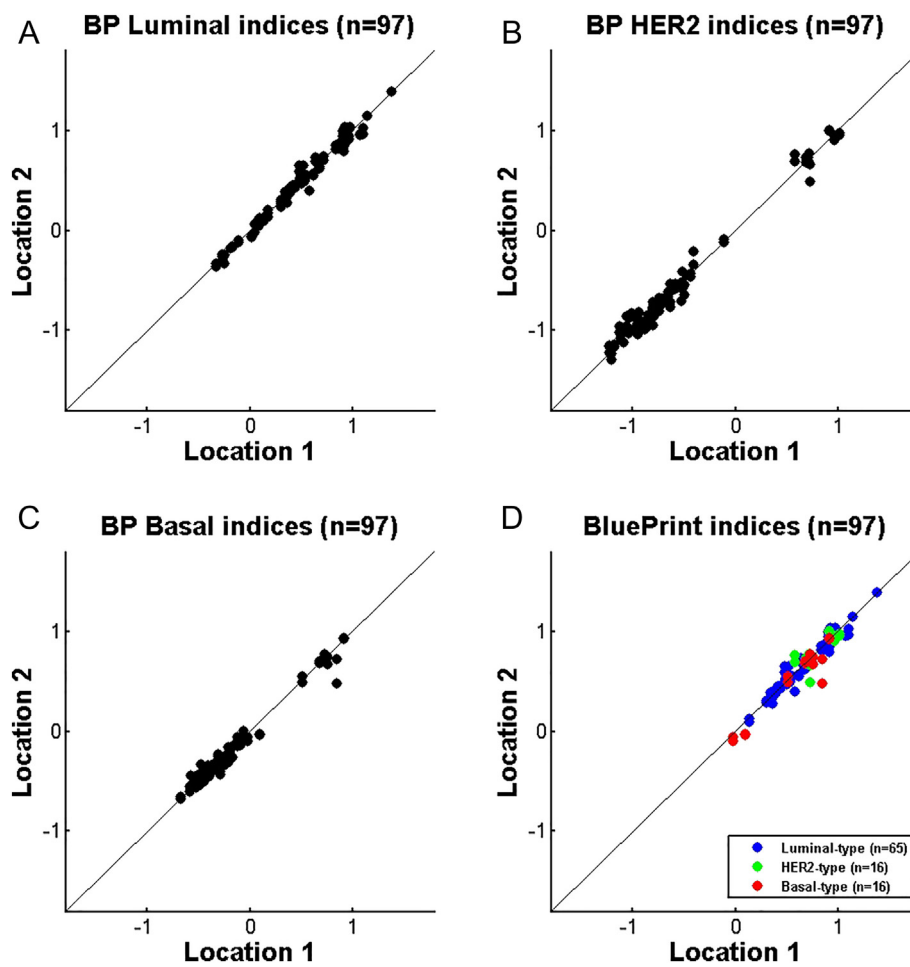
Figure 3. Inter-laboratory reproducibility of BluePrint FFPE – comparison of BluePrint indices between two Agendia sites Amsterdam (location 1) and Irvine (location 2) (N = 97). **A.** BluePrint Luminal-type indices, **B.** BluePrint HER2-type indices, **C.** BluePrint Basal-type indices, **D.** BluePrint indices comparison for the final assigned subtype (Luminal-, HER2- and or Basal-type). BluePrint subtype concordance was 100%. Each dot represents a single breast cancer sample for which total RNA underwent BluePrint microarray laboratory processing and analysis.

samples, 345 belong to the microarRAy-prognoSTics-in-breast-cancER (RASTER) study, for which clinicopathological and 5-year outcome data are available and published previously [11–13]. All patients were aged 18–61 years and had a histologically confirmed invasive adenocarcinoma of the breast (cT1–3N0M0). Survival analyses were performed to compare clinical performance obtained using results from FF and FFPE matched tissues. Kaplan–Meier analyses were used to compare the survival distributions of BluePrint subtypes for FF and FFPE for distant recurrence-free interval (DRFI). Analyses and visualization of data were performed using the MATLAB (The MathWorks) software version R2012a. Scatterplots and Bland–Altman analysis [14] were used to examine the existence of any constant bias in the difference of measurements between paired samples. Equivalence of BluePrint indices was determined by the Pearson correlation for assessment of the degree of linear correlation. Clinicopathological data and clinical outcome data were analyzed using the statistical package SPSS 22.0 for Windows (SPSS Inc., Chicago, US).

Results

Analytical Performance

Samples classified by BluePrint are reported as one of three subtypes: Luminal-type, Basal-type or HER2- type. The classification is based on the index that is calculated for each of the three subtypes, the subtype belonging to the highest index is the reported result. The numeric ranges or indices of the subtypes are used to evaluate the analytical performance.

Precision and Repeatability Evaluation

Precision and repeatability were assessed using four clinical samples (one Basal-type (S1), one HER2-type (S2) and two Luminal-types (S3 and S4)) that were repeatedly measured over a 20-day period (Figure 1). This was performed according to the FDA recommended guideline issued by the National Committee for Clinical laboratory standards (NCCLS) [15,16]. In this experiment all variables were included that occur in the clinical setting, such as different operators, equipment and reagent lots. Predefined acceptance criteria for maximum allowed variation were established and documented in a validation plan. BluePrint showed very stable results for all subtype indices, with a median standard deviation for precision of 0.044, that results in a relative standard deviation of 1.4% and a precision of 98.6%.

A duplicate run was performed for each sample each day to determine repeatability of the assay, expressed as the standard deviation between the duplicate runs per day over all 20 days, which was equal to 0.032. This corresponds to a relative standard deviation for repeatability of 1.0% and a stability of 99.0%.

Reproducibility Over Time

To get real life data on reproducibility that is not restricted to a limited time frame, the BluePrint indices for different clinical control samples were measured over 3 years.

These control samples are continuously monitored as technical and experimental controls in the clinical diagnostic setting to ensure quality and

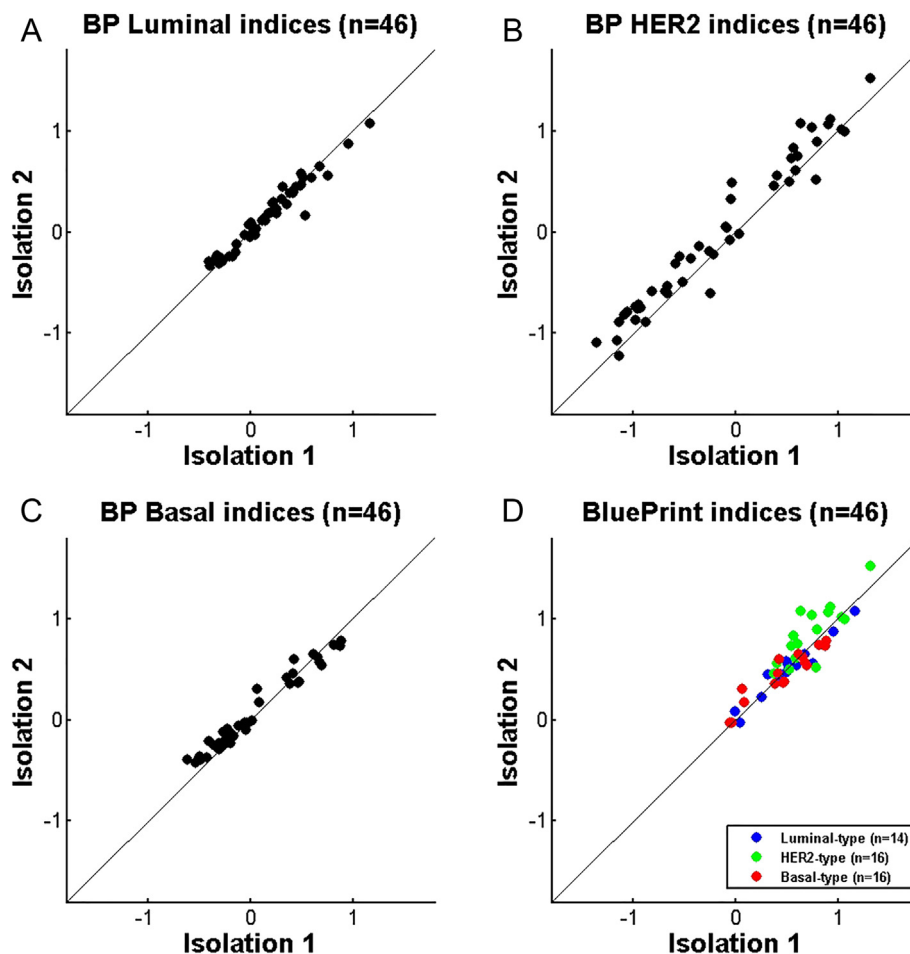**Figure 4.** Reproducibility of BluePrint between two independent isolations, comparison of indices between isolation 1 and isolation 2 using different sections of the tumor (N = 46). **A.** BluePrint Luminal-type indices, **B.** BluePrint HER2-type indices, **C.** BluePrint Basal-type indices, **D.** BluePrint indices comparison for the final assigned subtype (Luminal-, HER2- and or Basal-type). BluePrint subtype concordance was 100%.

safety. These samples are processed within each batch of samples. Each control sample consists of a pool of clinically representative breast cancer tumors [13,17]. Figure 2, A–C show the BluePrint index assessment for three clinical control samples C1 (Luminal-type, n = 1639), C2 (HER2-type, n = 1534) and C3 (Basal-type, n = 1594). BluePrint indices showed very stable results over-time for all three control samples with reproducibility of 98.9% (standard deviation of 0.036), 97.6% (standard deviation of 0.076) and 98.3% (standard deviation of 0.053) for C1, C2 and C3 respectively, which resulted in a median reproducibility of 98.3% (median standard deviation of 0.054). In all cases the categorical result was 100% accurate.

Interlaboratory Reproducibility

BluePrint FFPE testing is carried out in two centralized laboratories, one in Amsterdam (The Netherlands, Location 1) and one in Irvine (CA, USA, Location 2). Both laboratories are Clinical Laboratory Improvement Amendments (CLIA) certified. Inter-laboratory reproducibility was tested using 97 samples that were processed in both locations starting from isolated RNA. Figure 3, A–C show the comparison scatterplots of BluePrint Luminal-, HER2- and Basal-type indices respectively generated at the two sites for all 97 samples, irrespective of their subtype (for each subtype an index is generated). Figure 3D shows the BluePrint index comparison for the final subtype. The categorical concordance of BluePrint subtype between Location 1 and 2 was 100%. The bias was estimated using the BluePrint indices and a Bland Altman analysis was performed by calculating the mean difference and limits of agreement. The mean difference was 0.3% of the reported index range and the limits of agreement were within the predefined acceptance criteria.

Reproducibility of BluePrint from Independent Isolations

It is known that heterogeneity within the same primary tumor can occur [18–20]. Gene expression analysis performed on different parts of the same tumor can therefore vary. To evaluate this effect on BluePrint, we compared BluePrint results obtained from two independent isolations of the same tumor tissue block using 46 samples. The concordance in subtypes was 100%.; the BluePrint indices were similar between two isolations (Luminal-type Pearson's r = 0.975, HER2-type Pearson's r = 0.976, Basal-type Pearson's r = 0.982) (Figure 4).

Updates Over Time of the BluePrint Test: Comparative Analyses

Over the years FDA/QSR compliant updates and technical improvements of the BluePrint test have been implemented. Such adjustments include additional qualifications (i.e. certifications) to perform the Blue-Print in different laboratories, software updates, reagent and equipment changes (array design, scanner updates and other tissue types (FF and FFPE)). All these adjustments were assessed by comparing the BluePrint results to results obtained using previous BluePrint versions including the initial developed version described by Krijgsman and colleagues [2].

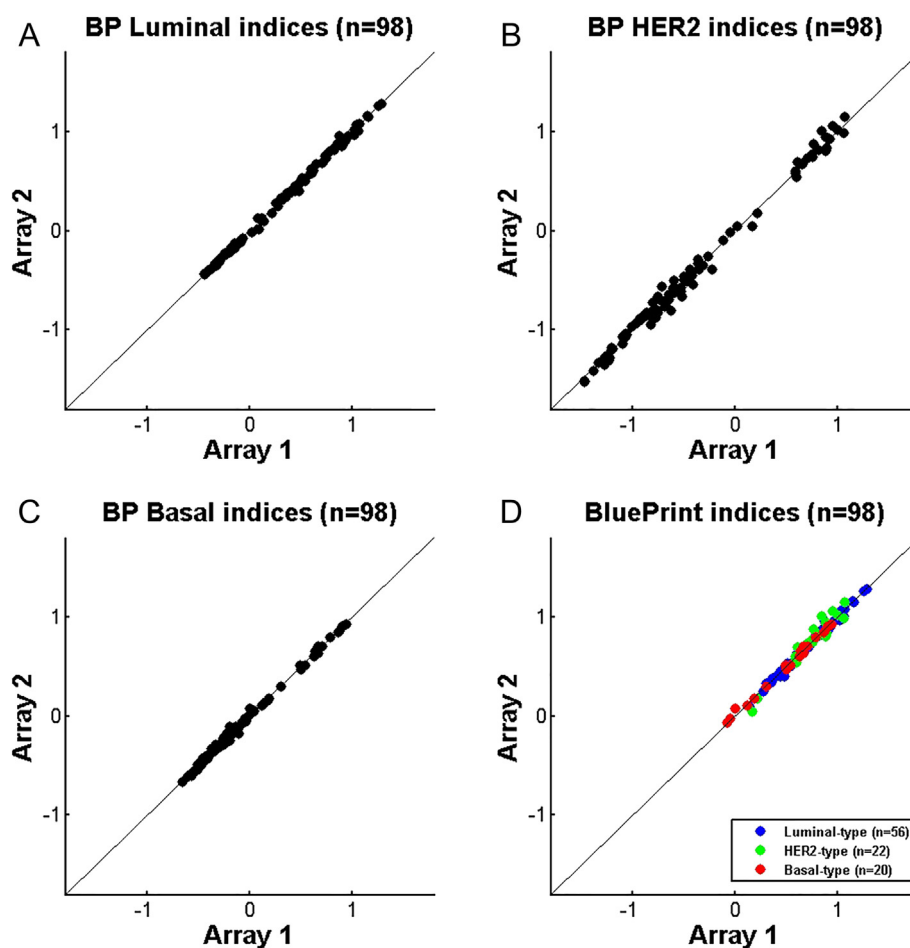The equivalence was measured as BluePrint subtype concordance of the two versions.

Figure 5. Comparison study: agreement of BluePrint FFPE between the gold standard array (array 1) and the new array (array 2) (N = 98). **A.** BluePrint Luminal-type indices, **B.** BluePrint HER2-type indices, **C.** BluePrint Basal-type indices, **D.** BluePrint indices comparison for the final assigned subtype (Luminal-, HER2- and or Basal-type). BluePrint subtype concordance was 99%. Each dot represents a single breast cancer sample for which total RNA underwent BluePrint microarray laboratory processing and analysis.

Comparison Between Different Array Types

Figure 5 shows the comparison of the BluePrint indices obtained using the original array type, similar to the customized mini-array described by Glas and colleagues [21] (called Array 1) with those obtained using a customized whole-genome microarray as previously described [13] (called Array 2). In total 98 samples were compared and indices for all samples and all subtypes were similar (Luminal-type Pearson's r = 0.999, HER2-type Pearson's r = 0.996, Basal-type Pearson's r = 0.998). Even though the correlations are nearly perfect, the categorical concordance is 99%. One patient sample was classified as BluePrint Luminal-type with array 1 and as BluePrint HER2-type with array 2. For this sample, the Basal and the Luminal BluePrint indices were close to each other with a difference smaller than the technical variance of the BluePrint FFPE test (average control standard deviation 0.052). The Bland–Altman showed no bias towards one of the array platforms with a mean difference of 0.04% of the reported dynamic range.

Comparison Between Different Microarray Scanner Systems

Figure 6 shows the BluePrint index comparison between two microarray scanner versions, where Scanner 1 (Agilent G2565BA microarray scanner system) represents the gold standard and Scanner 2 (Agilent G2565CA microarray scanner system) the upgraded scanner. A total of 80 samples were processed and hybridized to the microarray. The hybridized microarray was scanned multiple times on the two scanners (first Scanner 1, then Scanner 2 and lastly Scanner 1 again). This was to determine possible signal loss due to scanning. As the first and the last scan were similar, the signal loss due to scanning was determined as none (data not shown). The scatterplots for BluePrint indices generated using Scanner 1 and Scanner 2 showed nearly perfect correlations (Luminal-type Pearson's r = 0.999, HER2-type Pearson's r = 0.999, Basal-type Pearson's r = 0.999). The concordance by subtype was 100%.

Comparison Between FFPE and FF Tissue Types

Currently, the BluePrint microarray test is performed using RNA isolated from FFPE tissue. Initially BluePrint was developed using RNA isolated from fresh frozen (FF) tissue [2] and this version was used in the MINDACT (ClinicalTrials.gov identifier NCT00433589) and in the I-SPY2 prospective clinical trials (ClinicalTrials.gov Identifier: NCT01042379), as recently published [22–24]. Therefore, we assessed here the equivalence of the BluePrint FFPE and FF results on a large series of 413 matched FFPE and FF samples available at Agendia. Overall concordance of categorical BluePrint Luminal-type, HER-type and Basal-type classification was 97.1% (Supplementary Table 1A). Such a value of concordance indicates that BluePrint results obtained from FFPE and FF tissues are substantially equivalent. There were twelve samples that switched BluePrint subtype result when comparing the FFPE and their matched FF samples. The majority of the discordant samples (7 of 12) had the subtype result indices close to each other with a difference between the indices of the discordant subtypes ranging from a maximum of 4.3% to a minimum of 0.7% of the BluePrint index dynamic range.

In order to assess how the FF and the FFPE BluePrint test compare with respect to their clinical outcome, we performed a survival analysis on a
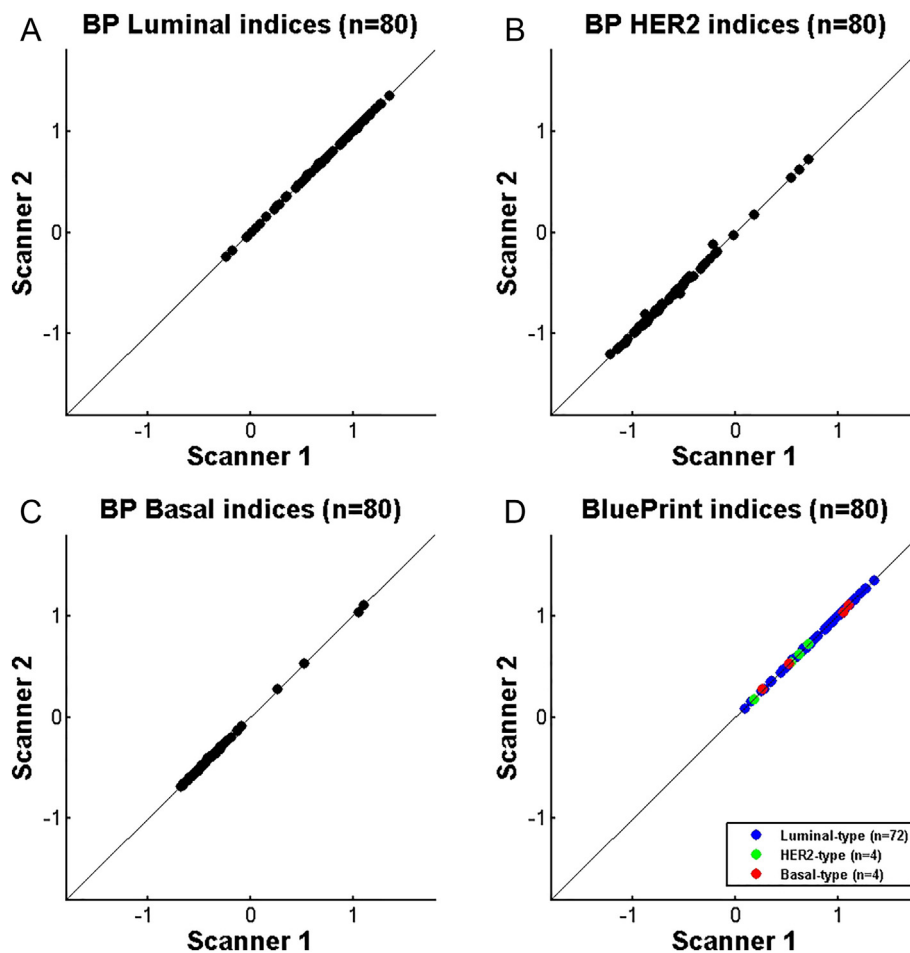

Figure 6. Comparison study: agreement of BluePrint FFPE between gold standard scanner (Scanner 1) and the upgraded scanner (Scanner 2) (N = 80). **A.** BluePrint Luminal-type indices, **B.** BluePrint HER2-type indices, **C.** BluePrint Basal-type indices, **D.** BluePrint indices comparison for the final assigned subtype (Luminal-, HER2- and or Basal-type). BluePrint subtype concordance was 100%. Each dot represents a single breast cancer sample for which total RNA underwent BluePrint microarray laboratory processing and analysis and was scanned on both scanners.
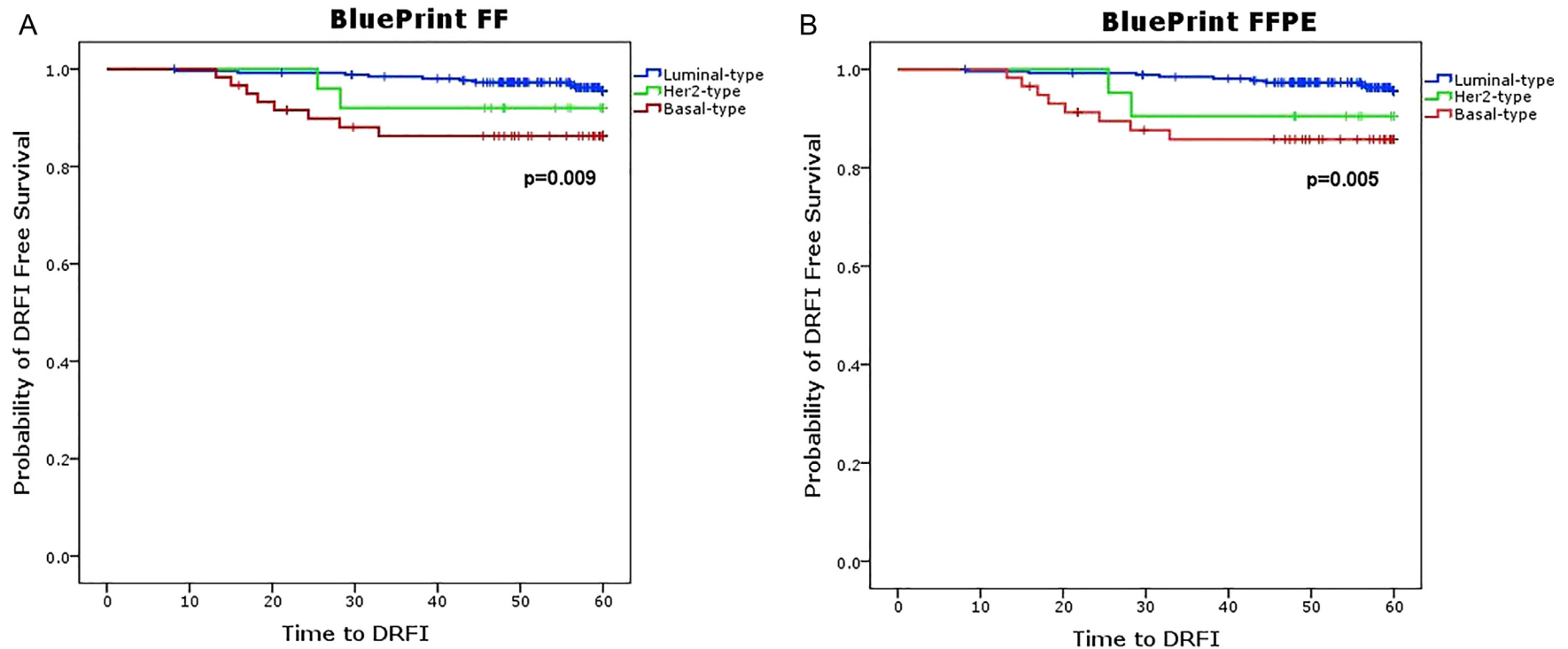
Figure 7. Kaplan–Meier curve of 345 microarRAy-prognoSTics-in-breast-cancER (RASTER) study patients for 5-year distant recurrence-free interval (DRFI) assessed using BluePrint FF (**A**) and FFPE (**B**) tissues. Kaplan–Meier curves plotted for DRFI show comparable clinical performance of BluePrint Luminal-type (blue lines), HER2-type (green lines) and Basal-type (red lines) patients in matched FF and FFPE in a series of 345 early-stage breast cancer patients. P-values of significance are assessed using log-rank test and showed on the graphs. FF = Fresh Frozen, FFPE = formalin-fixed paraffin embedded.

subset of 345 (of the 413 samples used for the technical equivalence assessment) for which 5-year DRFI outcome data were available. Figure 7 shows that the Kaplan–Meier curves are similar between the FF and the FFPE test results and that BluePrint Luminal-, HER2- and Basal-type groups significantly differ in 5-year DRFI for both FF and FFPE (FF log-rank P = .009 versus FFPE log-rank P = .005). Moreover, 5-year survival percentages of distant recurrence-free interval for the three BluePrint subtype patient groups were comparable between FF and FFPE (Supplementary Table 1B).

Discussion

Molecular classification of breast tumors by conventional IHC relies only on the presence of the protein receptor and it shows limitations in predicting whether a tumor is truly positive for functional ER, PR, or HER2 protein. Instead, multigene expression based-tests, such as BluePrint, enable the assessment of the activation status of the pathways where ER, PR and HER2 play a fundamental role. As previously reported [2,5,6,8,25], BluePrint showed to be an excellent tool to identify breast cancer functional molecular subtypes and to guide treatment selection for patients, confirming the clinical utility and validity of this test. Here we focused on the analytical performance of the BluePrint test and proved its analytical validity. In medical diagnostic research, extensive clinical validation studies are often performed to prove the clinical validity and utility of a diagnostic test, while relatively little attention is given to the many factors that could contribute to its analytical validity. Any source of variation could affect the outcome of the diagnostic test, with direct consequence for the patient treatment recommendations. Here we investigate several factors that could influence the BluePrint test result and we highlight the robustness and reproducibility of this test. We show that repeatability and precision of the BluePrint test on the same RNA sample are very high, with concordance values above 98%. Similarly, when we assess reproducibility over a long time period (from 2012 to 2015) BluePrint indices appear very stable with an average standard deviation of 0.054. The strengths of this result are both the large sample size (over 4700 measurements) and that in this long-term validation, nearly all potential sources of variation are included by default (operator, reagent lot, day, scanner used etc.). Concordance of BluePrint results under different conditions (such as different array type or scanner system) showed to be nearly perfect as well as when we compared BluePrint results from independent isolations of the same tissue sample (100%). This is in line with what was previously reported by Mittempergher and colleagues when comparing BluePrint results from

different RNA isolations and different platforms (microarray and next generation sequencing) [26]. Intra-tumor heterogeneity is a known contributor to the variation in the outcome of multigene diagnostic tests [27–29], nevertheless the results from our study could suggest that the BluePrint test captures the expression of genes that tend to be less susceptible to this factor. Moreover, it should be noted that this analysis is an approximation of tumor heterogeneity, as a true tumor heterogeneity study would analyze two different parts of the tumor. This experiment aims instead to assess the cellular heterogeneity of the same primary tumor, rather than a so-called spatial heterogeneity [30].

Another major contributor of assay variation is the type of tissue source of the RNA used for the diagnostic assessment, either fresh frozen (FF) or formalin-fixed and paraffin embedded (FFPE). Compared with RNA isolated from FFPE tissue, FF RNA is of higher quality and generally considered the most suitable for biomarker identification and gene expression profiling [17,31,32]. Nonetheless, FFPE tissue is the most widely available source of tissue material in the routine diagnostics testing and for which long-term clinical follow-up data are available. In recent years, technological advances made possible the analysis of such degraded RNAs and gene expression profiling of FFPE RNA showed to be largely comparable to the FF matched counterpart [32,33]. The initial BluePrint test was developed using FF tissue and later on the test applicability was extended to FFPE. Our results show that the BluePrint test generates equivalent results when using either FF or FFPE RNA with concordance above 97%. The ~3% discordance rate observed in this study is in line with what was previously observed by others when comparing microarray diagnostic results of matched FF and FFPE tissues [13,17]. This can be ascribed to pre-analytical differences that occur when different parts of the tumor tissues are taken (tumor heterogeneity) and undergo different preservation processing, such as for FF and FFPE samples. Samples that switch between the different tissue source (FF vs. FFPE) may affect the clinical performance of the test, however when we assess the clinical performance of the BluePrint test for FF and FFPE samples, we observed that survival curves of BluePrint Luminal-, HER2- and Basal-type groups based on FF results are equivalent to those based on FFPE results with both FF and FFPE having a significant difference in 5-year DRFI (FF vs. FFPE, log-rank p-value of 0.009 vs. 0.003) for the three subtypes. This result indicates that in nearly all of the cases molecular subtyping based on BluePrint is the same irrespective to the tissue source, and the same patient group will be identified as Luminal-type, HER2-type or Basal-type after diagnostic testing.

Classification by molecular subtype has been recommended as a guide for the selection of therapy for breast cancer patients. At present, the most widely

Table 1

Analytical performance of BluePrint – summary table; overview of all performance characteristics assessed for BluePrint

Performance characteristic	Definition	BP FFPE performance
Repeatability	Closeness of agreement between results of successive measurements of the same sample (same operators/batches) according to the Clinical Laboratory Standards document EP5-A2 [18]	Median stdev = 0.032 Relative precision: 99.0%
Precision	Closeness of agreement between results of successive measurements of the same sample (different operators/batches) according to the Clinical Laboratory Standards document EP5-A2 [18]	Median stdev = 0.044 Relative precision: 98.6%
Reproducibility over-time	Closeness of agreement between results of measurements of the same sample (control sample) carried out under changed conditions	Median stdev = 0.054 Reproducibility: 98.3% Site-to-site comparison: Concordance = 100% Between different isolations: Concordance = 100%
Reproducibility under different conditions (categorical classification based)	Closeness of agreement between categorical results of measurements of paired samples processed under different conditions	Array comparison Concordance = 99% Scanner comparison: Concordance = 100% FFPE vs. FF comparison Concordance = 97%

BP = BluePrint, FFPE = formalin-fixed paraffin embedded, FF = Fresh Frozen, stdev = standard deviation.

adopted methodology in the clinical setting for molecular subtyping is based on the assessment of ER, PR, HER2 and the Ki67 proliferation marker using conventional IHC and fluorescence in situ hybridization (FISH)/ Chromogenic in situ hybridization (CISH) or silver in situ hybridization (SISH) in case of the HER2 [34–38]. However, standardization of IHC-based clinical subtyping shows to be challenging because experimental procedures and interpretation of results vary between laboratories [39–41]. As recently highlighted by Varga and colleagues [40], for ER, PR and HER2, several studies have reported discrepancies of up to 20% and for Ki67 discrepancies are even more prominent. An international Ki67 reproducibility study showed that interlaboratory reproducibility for Ki67 was moderate with a so-called intraclass correlation coefficient (ICC) of 0.71 when considering local scoring methods and central staining (and even lower for local staining) [42]. A follow-up study of the same group showed that ICC could increase up to 0.92 when a standardized scoring methodology was used, pointing out that Ki67 values and cutoffs for clinical-decision making are not always transferable between different laboratories due to limited analytical validity [43]. Standardized and reproducible assays are mandatory in routine diagnostics and the BluePrint test, based on the results described here, fulfills these needs.

Taken together, we validate here the analytical performance of the BluePrint FFPE test by showing repeatability, precision and reproducibility over-time above 98% based on the generated index (Table 1). Moreover, what is more important are the categorical results and, with this respect, the current FFPE BluePrint test showed to be robust under different conditions, reaching subtype outcome concordances above 99%.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tranon.2020.100756.

Author Contributions

LM and LJMJD designed the research study, analyzed and interpreted the data, and wrote the manuscript; ATW analyzed and interpreted the data, MHJS and SM conducted experiments; BYC analyzed and interpreted the data, CD analyzed the data; NB and EJTL interpreted the data; AMG designed and supervised the research study, interpreted the data, and wrote the manuscript.

References

[1] X. Dai, et al., Breast cancer intrinsic subtype classification, clinical use and future trends, Am. J, Cancer Res, 2015.

[2] O. Krijgsman, et al., A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response, Breast Cancer Res Treat 133 (2012) 37–47.

[3] T. Sørlie, et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, Proc Natl Acad Sci U S A (2001) https://doi.org/10.1073/pnas.191367098.

[4] Cho, N. Molecular subtypes and imaging phenotypes of breast cancer. *Ultrasonography* (2016). doi:10.14366/usg.16030

[5] P. Whitworth et al., Chemosensitivity predicted by BluePrint 80-Gene functional subtype and MammaPrint in the prospective Neoadjuvant Breast Registry Symphony Trial (NBRST), Ann Surg Oncol 21 (2014) 3261–3267.

[6] P. Whitworth, et al., Chemosensitivity and endocrine sensitivity in clinical luminal breast cancer patients in the prospective Neoadjuvant Breast Registry Symphony Trial (NBRST) predicted by molecular subtyping, Ann Surg Oncol 24 (2017) 669–675.

[7] Wolf, D. M. *et al.* DNA repair deficiency biomarkers and the 70-gene ultra-high risk signature as predictors of veliparib/carboplatin response in the I-SPY 2 breast cancer trial. *npj Breast Cancer* 3, 1–8 (2017).

[8] P. Beitsch, et al., Pertuzumab/trastuzumab/CT versus trastuzumab/CT therapy for HER2+ breast cancer: results from the prospective Neoadjuvant Breast Registry Symphony Trial (NBRST), Ann Surg Oncol 24 (2017) 2539–2546.

[9] F. Cardoso, et al., 70-Gene signature as an aid to treatment decisions in early-stage breast cancer, N Engl J Med 375 (2016) 717–729.

[10] Groenendijk, F. H. *et al.* Estrogen receptor variants in ER-positive basal-type breast cancers responding to therapy like ER-negative breast cancers. *npj Breast Cancer* 5, (2019).

[11] C.A. Drukker, et al., A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study, Int J Cancer 133 (2013) 929–936.

[12] J.M. Bueno-de-Mesquita, et al., Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER), Lancet Oncol 8 (2007) 1079–1087.

[13] I. Beumer, et al., Equivalence of MammaPrint array types in clinical trials and diagnostics, Breast Cancer Res Treat 156 (2016) 279–287.

[14] K. Roepe, A Bayesian approach to investigating age-at-death of subadults in a forensic context, Biochem Med 25 (2014) 141–151.

[15] CLSI. Evaluation of precision of quantitative measurement procedures ; approved guideline—third edition. *Wayne CLSI* CLSI document EP05-A3 (2014).

[16] D. Chesher, Evaluating assay precision, Clin Biochem Rev 29 (Suppl. 1) (2008) S23–S26.

[17] A. Sapino, et al., MammaPrint molecular diagnostics on formalin fixed paraffin-embedded tissue, J Mol Diagnostics 16 (2014) 190–197.

[18] L.G. Martelotto, C.K. Ng, S. Piscuoglio, B. Weigelt, J.S. Reis-Filho, Breast cancer intra-tumor heterogeneity, Breast Cancer Res 16 (2014) 210.

[19] A. Tabchy, C.X. Ma, R. Bose, M.J. Ellis, Incorporating genomics into breast cancer clinical trials and care, Clin Cancer Res 19 (2013) 6371–6379.

[20] S. Asioli, et al., Approaching heterogeneity of human epidermal growth factor receptor 2 in surgical specimens of gastric cancer, Hum Pathol 43 (2012) 2070–2079.

[21] A.M. Glas, et al., Converting a breast cancer microarray signature into a high-throughput diagnostic test, BMC Genomics 7 (2006) 278.

[22] S. Glück, F. De Snoo, J. Peeters, L. Stork-Sloots, G. Somlo, Molecular subtyping of early-stage breast cancer identifies a group of patients who do not benefit from neoadjuvant chemotherapy, Breast Cancer Res Treat (2013) https://doi.org/10.1007/s10549-013-2572-4.

[23] L.J. van 't Veer, D. Wolf, C. Yau, Z. Zhu, A. Glas, W. Audeh, L. Brown Swigart, S. Asare, G. Hirst, I. Investigators, A. Demichele, D. Yee, L. EssermanL.J. van 't Veer, D. Wolf, C. Yau, Z. Zhu, A. Glas, W. Audeh, L. Brown Swigart, S. Asare, G. Hirst, I. I, L. E. BluePrint basal subtype predicts neoadjuvant therapy response in ~ 400 HR + HER2− patients across 8 arms in the I-SPY 2 TRIALNo Title. in *EORTC – NCI – AACR Symposium- Abstract Book* (ed. Elsevier) e15 (European Journal of Cancer, 2018).

[24] G. Viale, et al., Immunohistochemical versus molecular (BluePrint and MammaPrint) subtyping of breast carcinoma. Outcome results from the EORTC 10041/BIG 3-04 MINDACT trial, Breast Cancer Res Treat (2018) https://doi.org/10.1007/s10549-017-4509-9.

[25] P. Beitsch, et al., Genomic impact of neoadjuvant therapy on breast cancer: incomplete response is associated with altered diagnostic gene signatures, Ann Surg Oncol (2016) https://doi.org/10.1245/s10434-016-5329-6.

[26] Mittempergher, L. *et al.* MammaPrint and BluePrint molecular diagnostics using targeted RNA next-generation sequencing technology. *J. Mol. Diagnostics* 1–16 (2019). doi:https://doi.org/10.1016/j.jmoldx.2019.04.007

[27] L.J. Delahaye, et al., Performance characteristics of the MammaPrint breast cancer diagnostic gene signature, Per Med 10 (2013) 801–811.

[28] C. Kim, S. Paik, Gene-expression-based prognostic assays for breast cancer, Nat Rev Clin Oncol (2010) https://doi.org/10.1038/nrclinonc.2010.61.

[29] K.M. Jochumsen, Q. Tan, B. Hølund, T.A. Kruse, O. Mogensen, Gene expression in epithelial ovarian cancer: a study of intratumor heterogeneity, Int J Gynecol Cancer (2007) https://doi.org/10.1111/j.1525-1438.2007.00908.x.

[30] Y. Yuan, Spatial heterogeneity in the tumor microenvironment, Cold Spring Harb Perspect Med (2016) https://doi.org/10.1101/cshperspect.a026583.

[31] F. Medeiros, C.T. Rigl, G.G. Anderson, S.H. Becker, K.C. Halling, Tissue handling for genome-wide expression analysis: a review of the issues, evidence, and opportunities, Archives of Pathology and Laboratory Medicine 131 (12) (2007) 1805–1816.

[32] L. Mittempergher, et al., Gene expression profiles from formalin fixed paraffin embedded breast cancer tissue are largely comparable to fresh frozen matched tissue, PLoS One 6 (2011).

[33] M. Iddawela, et al., Reliable gene expression profiling of formalin-fixed paraffin-embedded breast cancer tissue (FFPE) using cDNA-mediated annealing, extension, selection, and ligation whole-genome (DASL WG) assay, BMC Med Genomics (2016) https://doi.org/10.1186/s12920-016-0215-4.

[34] M. van de Vijver, Emerging technologies for HER2 testing, Oncology (2002) https://doi.org/10.1159/000066199.

[35] U. Krishnamurti, J.F. Silverman, HER2 in breast cancer: a review and update, Adv Anat Pathol (2014) https://doi.org/10.1097/PAP.0000000000000015.

[36] B.C. Calhoun, L.C. Collins, Predictive markers in breast cancer: an update on ER and HER2 testing and reporting, Semin Diagn Pathol (2015) https://doi.org/10.1053/j.semdp.2015.02.011.

[37] A. Prat, et al., Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal a breast cancer, J Clin Oncol (2013) https://doi.org/10.1200/JCO.2012.43.4134.

[38] P. Maisonneuve, et al., Proposed new clinicopathological surrogate definitions of luminal A and luminal B (HER2-negative) intrinsic breast cancer subtypes, Breast Cancer Res (2014) https://doi.org/10.1186/bcr3679.

[39] Jorns, J. M. Breast cancer biomarkers: challenges in routine estrogen receptor, progesterone receptor, and HER2/neu evaluation. *Arch. Pathol. Lab. Med.* arpa.2019–0205-RA (2019). doi:https://doi.org/10.5858/arpa.2019-0205-RA

[40] Z. Varga, et al., An international reproducibility study validating quantitative determination of ERBB2, ESR1, PGR, and MKI67 mRNA in breast cancer using MammaTyper®, Breast Cancer Res (2017) https://doi.org/10.1186/s13058-017-0848-z.

[41] Laenkholm, A. V. *et al.* An inter-observer Ki67 reproducibility study applying two different assessment methods: on behalf of the Danish Scientific Committee of Pathology, Danish breast cancer cooperative group (DBCG). *Acta Oncol. (Madr).* (2018). doi: https://doi.org/10.1080/0284186X.2017.1404127

[42] M.Y.C. Polley, et al., An international ki67 reproducibility study, J Natl Cancer Inst (2013) https://doi.org/10.1093/jnci/djt306.

[43] M.Y.C. Polley, et al., An international study to increase concordance in Ki67 scoring, Mod Pathol (2015) https://doi.org/10.1038/modpathol.2015.38.