Title

Tissue Sources for Accurate Measurement of Germline DNA Genotypes in Prostate Cancer Patients Treated With Radical Prostatectomy

Permalink

Authors

Emami, Nima C
Leong, Lancelote
Wan, Eunice
et al.

Publication Date

2016-11-01

Tissue Sources for Accurate Measurement of Germline DNA Genotypes in Prostate Cancer Patients Treated With Radical Prostatectomy

Authors and Affiliations

Nima C. Emami,[1] Lancelote Leong,[2] Eunice Wan,[3] Erin L. Van Blarigan,[2,4] Matthew R. Cooperberg,[4] Imelda Tenggara,[4] Peter R. Carroll,[4] June M. Chan,[2,4] John S. Witte,[1,2,3,4]* Jeffry P. Simko[4,5]*

[1] Program in Biological and Medical Informatics, University of California, San Francisco, San Francisco, California
[2] Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California
[3] Institute for Human Genetics, University of California, San Francisco, San Francisco, California
[4] Department of Urology, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California
[5] Department of Anatomic Pathology, University of California, San Francisco, San Francisco, California

*Correspondence to:

Jeffry P. Simko
1825 4th St., Room M2360, San Francisco, CA 94158
415-353-7171 (Phone), 415-353-7094 (Fax)
Email: Jeff.Simko@ucsf.edu

John S. Witte
1450 3rd St., San Francisco, CA 94158
415-502-6882 (Phone), 415-476-1356 (Fax)
Email: JWitte@ucsf.edu

Shortened Running Head Title

Comparison of Normal Tissue Genotypes

Grant Sponsorship

Grant sponsor: Department of Defense; Grant number: W81XWH-13-2-0074; Grant sponsor: National Institutes of Health; Grant number: CA088164; Grant sponsor: National Institutes of Health; Grant number: GM067547; Grant sponsor: National Institutes of Health; Grant number: K07CA197077; Grant sponsor: National Institutes of Health; Grant number: KL2TR000143; Grant sponsor: UCSF Discovery Fellows Program.

Conflict of Interest

The authors declare no potential conflicts of interest.

BACKGROUND

Benign tissue from a tumor-containing organ is commonly the only available source for obtaining a patient’s unmutated genome for use in cancer research. While it is critical to identify histologically normal tissue that is independent of the tumor lineage, few additional considerations are applied to the choice of this material for such measurements.

METHODS

Normal formalin-fixed, paraffin-embedded seminal vesicle and urethral tissues, in addition to whole blood, were collected from 31 prostate cancer patients having undergone radical prostatectomy. Genotype concordance was evaluated for DNA from each tissue source in relation to whole blood.

RESULTS

Overall, there was a greater genotype call rate for DNA derived from urethral tissue (97.0%) in comparison with patient-matched seminal vesicle tissues (95.9%, $p = 0.0015$). Furthermore, with reference to patient-matched peripheral blood, urethral samples exhibited higher genotype concordance (94.1%) than that of seminal vesicle samples (92.5%, $p = 0.035$).

CONCLUSIONS

These findings highlight the heterogeneity between diverse sources of DNA in genotype measurement and motivate consideration of normal tissue biases in tumor-normal analyses.

Introduction

Disease screening and risk-modeling involve the integration of increasingly diverse sources of biological information. Innovations in high-throughput assay technologies have enabled the acquisition of biological data at an unprecedented scale. Subsequently, the development of a clinically actionable model of disease risk now involves traversing multiple dimensions of biological measurements, including protein levels, gene expression levels, and germline DNA polymorphisms, in addition to clinical and sociodemographic variables. Recent studies have demonstrated that the predictive power of risk models that integrate diverse biomarkers may be greatly improved in comparison to traditional screening approaches based on clinical data and limited biomarkers [1,2]. Hence, the methods by which biological data are acquired deserve special attention, as they may influence downstream predictive performance.

One consideration is the choice of appropriate biospecimen from which biomarkers will be measured. In genetic association studies of complex disease, the DNA used for measuring germline variants is often purified from blood, oral scrapings, or saliva [3,4]. However, retrospective tumor-normal research analyzing mutations, copy number, and gene expression in tissue from biopsy or surgery often relies on tumor-adjacent normal tissue as the only possible source for germline DNA genotypes [5]. While previous studies have examined the genotyping performance of select normal tissues in comparison with blood [6–8], the issue of how different sources of normal tissue influence the result of germline DNA genotyping, and accordingly the validity of disease risk predictions that model such genotypes, has been generally overlooked.

In the development of an integrated risk prediction model to discern aggressive versus indolent prostate cancer, we hypothesized that distinct sources of normal tissue may perform differently in the context of high-throughput genetic analyses. Here we analyze surgically resected specimens from patients with prostate cancer to compare genotyping results for DNA samples derived from archival normal tissues of the prostatic urethra and seminal vesicle.

Materials and Methods

Tissue Preparation

We obtained 93 normal samples (patient-matched blood, urethral tissue, and seminal vesicle tissue) from 31 patients who had undergone radical prostatectomy. All tissue was obtained using a 2 mm dermal punch to biopsy archival formalin-fixed, paraffin-embedded (FFPE) tissue blocks. A new punch was used to collect each biopsy from each block, and a single punch was made each time and placed into an Eppendorf tube for DNA extraction. The region of interest from each block to be biopsied was marked for punching. For each prostatectomy, the slides and pathology report of each case were reviewed. Seminal vesicle tissue from the side opposite to that most involved by prostate cancer was used; the area marked included both the seminal vesicle epithelium and the muscle wall, and the punch was taken to include both. The area of the urethra to be punched was marked in an area at least 5 mm from any tumor foci and included both urothelium and underlying stromal tissue, in a manner to exclude prostate glandular tissue. Note that while all punches of seminal vesicle contained 100% tissue throughout, many punches of urethra were taken from the border of the tissue with surrounding FFPE such that the punch may not have been completely composed of tissue. Normal prostate tissue was excluded from consideration due to several known obstacles to the identification of histologically pure samples of normal prostate, including the presence of multiple, scattered heterogeneous tumor foci [9], prostatic intraepithelial neoplasia [10–12], and field effects due to the presence of nearby neoplasia(s) [13], all of which are known to induce genetic abnormalities.

DNA Purification and Genotyping

After the paraffin layer was removed, 1 mm diameter cores punched from FFPE tissue blocks were sectioned into 20–30 pieces using a sterile razor blade. Samples were then vortexed with 1 ml xylene, followed by 2min of centrifugation at room temperature. Next, samples were again vortexed with 1 ml of 100% ethanol and pelleted by centrifugation. The supernatant was discarded and residual solvent was evaporated at room temperature. Next, DNA was purified from blood samples (Promega Wizard Genomic DNA Purification Kit) and FFPE tissues (QIAamp DNA FFPE Tissue Kit). To boost DNA yields prior to genotyping, 200 ng of input DNA from each sample was amplified (Affymetrix Axiom 2.0 Reagent Kit) via isothermal incubation at 37°C for 48hr. The sample DNA was next fragmented into pieces ranging from 25 to 125 base pairs, followed by isopropanol precipitation. The Affymetrix GeneTitan Multi-Channel Instrument was used for sample genotyping.

Custom Microarray Design

In collaboration with Affymetrix Inc., we designed a custom DNA microarray to assay functional and putative prostate cancer specific variation. While the array features many rare (< 1% minor allele frequency) and coding variants, its design was not limited to rare or exonic variation and broadly targeted genetic markers of interest genome-wide in a number of different functional categories.

The variant selection procedure was conducted as follows. First, a set of target markers, including both single nucleotide polymorphisms (SNPs) and insertion-deletion (indel) mutations, was constructed. The targets included previous GWAS findings (genome-wide significant and suggestive) in prostate cancer, associated traits (PSA level and prostate cancer gene-by-gene interactions), other correlated traits (breast cancer, height, BMI, obesity, diabetes), and uncorrelated traits (NHGRI GWAS catalog polymorphisms). Additionally, a list of pan-cancer candidate genes was compiled and rare variants in windows centered around these genes were included in the target set. Rare variants in frequently mutated genes from the somatic cancer database COSMIC were also included. Furthermore, rare variants from a series of in-house whole genome and whole exome sequence analyses (of African American prostate cancer patient normal genomes [14], normal prostate exomes from the TCGA and dbGaP [15,16], and prostate cell line DNAse I hypersensitive sites [17]) were added to the target set. Finally, variants from previous Affymetrix microarrays were also targeted. These included the Exome 319 chip and the UK Biobank [18] array (excluding the GWAS backbone), which covered a broad range of functional categories including missense mutations and putative deleterious variants from the Human Gene Mutation Database.

The next step was to select which probesets would be directly genotyped on the microarray. Probesets were selected from a pool of candidate markers by an iterative, greedy algorithm which prioritized candidates based on their coverage of the target set. In order to reduce redundancy with previous GWAS arrays, candidates were chosen with complementarity to GWAS arrays previously assayed in the Kaiser Permanente GERA cohort [3,19,20] by drawing from a candidate set disjoint from the GWAS array markers. This produced a set of markers optimized for coverage of the target set.

Sample and Variant Quality Control

We excluded samples from our analyses if there was insufficient resolution between marker probeset intensities (axiom_dishqc_DQC < 0.75) in any of the three tissue sources. This resulted in the exclusion of two samples and decreased the sample size from 31 initial subjects to 29 total. Out of the 29 subjects, 25 self-identified ethnically as Caucasian, one as African American, and three as “Other.” All subjects were designated as clinical T stage one or two and Gleason 6 (3 + 3) at diagnosis, although certain patients were upgraded and upstaged after surgery. These subject demographics and others are described in detail in Table I. Genotyping and sample quality control was performed using the Affymetrix Power Tools software suite.

To exclude variants susceptible to low-confidence genotype calls due to misclustering, variants with a minor allele frequency less than 5% are omitted from the reported concordance estimates. This minor allele frequency filter reduced the number of markers from 416,047 total variants to 127,847 common polymorphisms, from which call rates and concordance estimates were computed and summarized in Table I. However, for completeness, analyses where markers were stratified by minor allele frequency (in the main text and supplementary figures) include all 416,047 markers segregated into their respective minor allele frequency bins. These minor allele frequencies were based on the European (EUR) super population of the 1000 Genomes Project Phase 3 release [21].

To exclude variants susceptible to low-confidence genotype calls due to misclustering, variants with a minor allele frequency less than 5% are omitted from the reported concordance estimates. This minor allele frequency filter reduced the number of markers from 416,047 total SNPs to 127,847 common SNPs, from which call rates and concordance estimates were computed and summarized in Table I. However, for completeness, all minor allele frequency-stratified analyses in the main text and supplementary figures include all 416,047 markers segregated into their respective minor allele frequency bins. These minor allele frequencies were based on the European super population of the 1000 Genomes Project Phase 3 release [17].

Statistical Analyses

Tissue sources were compared using several sample statistics (DNA quantity, genotype call rate, and genotype concordance; Table I), as well as clinicopathologic factors (“subject-level” factors). For a given genetic variant, genotype concordance was defined as the agreement of both called alleles at a given marker (in samples from the same subject). Genotype pairs containing any no-calls were excluded from concordance calculations. Hypothesis testing for detecting statistically significant differences between tissue sources was conducted via paired-sample, two-tailed t-tests. Comparisons of genotype statistics (“variant-level” factors) between tissue sources (Figs. 1, 2, and S2) were likewise conducted using weighted, paired-sample, two-tailed t-tests, with the weight values equal to the number of markers in a given minor allele frequency bin. Linear regression model selection was conducted via stepwise bidirectional elimination using the Akaike Information Criterion. Concordance calculations and variant QC was conducted using PLINK [22], while all statistical analyses and figure generation were performed using the R statistical computing language [23,24].

Results

Sample Quality of Source DNA

We evaluated the concordance between genotypes calls in DNA samples isolated from patient-matched blood, prostatic urethra (UR), and seminal vesicle (SV) normal tissues for 31 men with prostate cancer. Quality control procedures are described in the Materials and Methods section, and yielded a dataset comprised of 127,847 common polymorphisms measured in 29 men across each of three DNA sources (blood, UR, SV).

As expected, we observed the superior performance of blood to both normal FFPE tissue sources with respect to several measures. Across all samples, post-amplification DNA yields (Table I) were significantly greater for blood than for UR ($p = 0.0091$) and for SV ($p = 0.0012$). In turn, genotype call rate was significantly greater in blood (98.3%) than in UR (97.0%; $p = 3.8 \times 10^{-6}$) and SV (95.9%; $p = 4.1 \times 10^{-10}$). This observation supported using blood genotypes as a gold-standard reference. Hence, in all subsequent comparisons, concordance estimates were computed with reference to blood genotypes.

Although the genotype call rate was higher overall for UR samples in comparison with SV (97.0% vs. 95.9%, $p = 0.0015$), DNA quantities did not differ significantly between UR and SV (p = 0.12), suggesting that the observed difference in call rate was not merely a consequence of DNA quality and may reflect physiological differences between normal tissue sources.

Genotype Concordance Across Individual-Level Factors

Furthermore, UR genotypes were more concordant with blood than SV genotypes (94.1% vs. 92.5%, p = 0.035). To determine whether certain subject-level factors may explain this 1.6% concordance difference between UR and SV, we considered the potential confounding effect of specific variables on concordance. First, we stratified concordance estimates by subject age at diagnosis and found that the superiority of UR genotype concordance was consistent across age groups (Table I). Next, we stratified concordance with respect to two variables significantly associated (P < 0.05) with UR and SV concordance differences in a linear regression model: prostate specific antigen (PSA) level at diagnosis and the source DNA quantity difference (post-amplification) between UR and SV. Again, we found that concordance for UR was slightly better than for SV across all strata. This included the first and second quartiles of DNA quantity differences, where SV DNA was more abundant than UR DNA in all samples (Q1) or in the majority (Q2), although these subsets contained rather few counts and the differences therein were thus not individually significant. Finally, we stratified concordance with respect to two clinical variables of interest: pathologic Gleason score and pathologic T stage. These variables generally reflected the trend of higher concordance of UR with blood, with the exception of Gleason 7 (4 + 3), which was comprised of a small sample size of only two subjects. These observations support the notion that true differences between UR and SV tissue, rather than confounding by other factors, underlie the observed differences in genotype concordance with blood.

We also examined whether cigarette smoking status at diagnosis may have impacted our results. Smoking was categorized into three levels: never (18 subjects), past (8), and current (3). One current smoker at diagnosis had 17.1% higher genotype concordance between UR and blood than between SV and blood, by far the greatest concordance difference among all studied subjects. When this subject was removed from our analysis, the pairwise difference in concordance among the remaining 28 subjects weakened but remained statistically significant (94.0% vs. 92.9%; p = 0.04), and the concordance of UR with blood still exceeded that of SV concordance across all rows in Table I from which the subject was omitted.

We additionally identified another potential outlier subject for whom concordance between UR and blood was 57.4%, concordance between SV and blood was 56.9%, and concordance between UR and SV genotypes was 97.7%. Removal of this subject did not impact the statistical significance of UR and SV concordance differences (p = 0.037). However, it did increase the average concordance levels for UR and SV to 95.4% and 93.7%, respectively. Core punch slides for these two subjects were reviewed and revealed no tumor contamination, dysplasia, or general explanation for why these samples would have such poor concordance with blood.

Genotype Concordance Across SNP-Level Factors

We examined the concordance levels in different minor allele frequency (MAF) bins across all genotyped markers (total of 416,047 probesets, including previously filtered rare variants with MAF < 5%). We found that genotype concordance for UR samples exceeded that of SV samples across the MAF spectrum ($p = 8.2 \times 10^{-14}$; Fig. 1). In most cases, the margin of concordance differences within a given bin approached or exceeded one percent, reflecting the 1.6% difference observed over all common polymorphisms. However, while the trend of superior concordance of UR was maintained over all MAF bins, the margin narrowed substantially in two bins: MAF < 1% (+0.38%) and 1% ≤ MAF < 2% (+0.66%). One explanation for the observation of decreased concordance with blood and smaller differences in concordance between UR and SV in rare variants is simply a lack of variation, and hence potential differences, at such low MAFs. Another possible explanation is genotype misclustering: as the minor allele count at a given marker approaches zero, genotype clustering algorithms face the substantial difficulty of distinguishing heterozygotes from major allele homozygotes. This in turn contributes to errant clustering, whereby major allele homozygotes are incorrectly classified as heterozygotes and minor allele homozygotes, increasing the rate of heterozygosity. Accordingly, we observed a significant excess of heterozygosity in UR ($p = 8.8 \times 10^{-13}$) and SV ($p = 6.3 \times 10^{-18}$) in comparison with blood (Fig. 2) as well as an increasing proportion of samples with discordant genotypes in markers of decreased MAF (Figs. S1A and B). Moreover, SV heterozygosity significantly exceeded that of UR ($p = 4.7 \times 10^{-19}$) across the MAF spectrum and, as the difference between UR and SV heterozygosity narrowed in bins of increasing MAF, the difference in their concordance with blood simultaneously increased (Pearson's r = -0.67, 95% CI [0.80, 0.49], $p = 8.5 \times 10^{-8}$), suggesting that genotype misclustering may explain the narrower margins of concordance between UR and SV in rare variants.

To control for the effect of poor genotype clustering in rare variants, variant quality control was performed by Hardy–Weinberg equilibrium (HWE) filtering. When variants violating HWE were removed ($\alpha = 5 \times 10^{-5}$), heterozygosity for all tissue sources decreased significantly ($P < 5 \times 10^{-19}$) towards expected levels. However, heterozygosity of SV genotypes remained elevated in comparison with UR ($p = 7.2 \times 10^{-21}$) and blood ($p = 1.4 \times 10^{-13}$), suggesting that the superior concordance of UR and blood is not simply an artifact of poor genotype clustering (Fig. 2). This conclusion was further supported upon reexamination of concordance after HWE filtering, with UR concordance more clearly separated from SV concordance across all MAF categories ($p = 1.2 \times 10^{-36}$; Fig. S2A). Finally, while the differences between UR and SV call rate ($p = 0.0010$) and genotype concordance ($p = 0.037$) did not change substantially after HWE filtering, the overall genotype concordance with blood across the set of HWE filtered markers increased to 95.4% for UR and 94.0% for SV (Table I). These concordance figures for UR and SV increased to 96.6% and 95.4%, respectively, when increased stringency was applied to HWE filtering ($a = 0.05$; Fig. S2B). However, while more stringent variant quality control can increase the accuracy of FFPE tissue genotype calls in comparison with blood, there exists a tradeoff between increasing concordance and potentially eliminating large numbers of accurate genotype calls from the final dataset.

Finally, further examination of the classes of discordant genotype calls confirmed the trend of excess heterozygosity in the genotypes from UR and SV tissue. Among all genotyped variants (common and rare), the percentage of discordant genotypes switching from a homozygous call in blood to a heterozygous call was 69.8% for UR and 71.5% for SV. For both tissues, the next most frequent change among discordant genotypes was in the opposite direction, from heterozygous to homozygous (27.5% for UR, 26.1% for SV). To assess whether changes in copy number or loss of heterozygosity may contribute to the observed genotype discordances, we used the Affymetrix CNV Summary Tools Software package to examine copy number in each sample set (blood, SV, and UR) and calculate B allele frequencies for each sample. When considering the deviation of the B allele frequencies from expected diploid allelic ratios (1.0, 0.5, 0.0), we found that the genome-wide variance of this deviation was significantly greater for UR ($p = 7.1 \times 10^{-4}$) and SV ($p = 7.9 \times 10^{-8}$) in comparison with blood, and was also greater for SV than for UR ($p = 0.01$). Thus, significant differences were observed between DNA from FFPE tissue and blood DNA, and, more remarkably, between DNA from FFPE seminal vesicle and urethral tissue. Increased noise in the raw fluorescent intensities used to derive B allele frequencies (and genotype calls) may account for these increases in allelic fraction variance. However, it is also possible that there are true differences in copy number between these different DNA sources.

Discussion

In this study, we evaluated the differences between sources of FFPE normal tissue from prostate cancer patients in assaying germline genetics and found that urethral tissue performs more favorably than seminal vesicle tissue in relation to patient-matched whole blood. While germline DNA from normal seminal vesicle tissue may serve as an adequately concordant proxy for blood DNA in the absence of alternatives, genotype measurements derived from urethral tissue DNA exhibited significantly higher call rate, lower heterozygosity, and greater concordance with blood in comparison with seminal vesicle derived genotypes. Although blood remains the ideal biospecimen for genomic analysis, normal tissue may serve as a suitable replacement, in particular for retrospective and tumor-normal studies when a blood specimen can no longer be obtained.

Although studies have revealed substantial technical reproducibility (generally exceeding 99%) among DNA biospecimens (including blood, FFPE tissue, saliva, and fresh frozen vs. FFPE tissue) genotyped in replicate [25–28], our findings suggest that significant heterogeneity may exist between genotype calls derived from different tissues. In general, special attention should be placed on the choice of normal tissue for germline genotyping, as distinct normal tissues may yield substantially different results. This insight may have particular relevance to tumornormal analyses such as whole genome and exome sequencing, array comparative genomic hybridization (aCGH), and RNA-seq, where the discovery of somatic aberrations in tumors is often predicated on the comparison to FFPE normal tissue as a reference [15,29–31]. Consequently, inaccuracies in germline measurements may lead to miscalled somatic mutations. While our results are based on data from a microarray genotyping platform, further study may determine that systematic differences among normal tissue sources potentially influence the results of next generation sequencing analyses.

There are several explanations for why genotype calls may vary significantly between normal tissue sources. One potential source of heterogeneity is somatic mosaicism, whereby mutations arising during development and aging propagate into specific tissues. Although the variability of somatic mutation rates among normal tissues is supported by observed differences in somatic mutation frequencies across tumor types [32], the expected number of somatic mutations is relatively modest when considering the generational human germline mutation rate [33]. Additionally, while studies of genome-wide somatic copy number mosaicism have discovered heterogeneity in several tissues, the size and number of validated somatic copy number variants suggests that structural variation may play only a minor role in germline genotype discordance across tissues [34]. Another potential source of heterogeneity is the differential invasion of the glands and ducts peripheral to the prostate: if one tissue is particularly susceptible to prostate tumor cell invasion, the purity of the DNA extracted from that tissue may be compromised and impact genotype call rate and concordance. While prostate cancer can metastasize to the urethra in rare cases, roughly 10–18% of patients having undergone radical prostatectomy are estimated to have pathological seminal vesicle invasion [35]. In our study, however, the majority of subjects were designated as pathologic T-stage 2 (Table I), and thus tumor cell invasion would not be expected to influence peripheral tissues. Furthermore, while field effects are known to influence many different classes of aberrations in tumor-adjacent, normal tissue, including epigenetic, genotypic, cytogenetic, and morphological changes [13,36,37], the extent to which field effects differ between different tumor-adjacent tissues has not been well characterized. The contributions of each of these determinants of heterogeneity and mosaicism to genotype discordance among normal radical prostatectomy tissues are subject to future research.

Finally, this work represents a novel application of the Affymetrix Axiom microarray technology to FFPE urethra and seminal vesicle tissue genotyping. Despite documented issues with purification of DNA fragments longer than 100–200 base pairs from formalin cross-linked tissue, researchers have been able to successfully profile FFPE samples that are up to 30 years old [38]. Furthermore, a recent study found expression profiles from paired fresh frozen and FFPE samples to be highly correlated, both between those newly collected and others archived 14 years earlier [39]. Although there is a tendency for sample degradation to increase with storage time, DNA isolated from FFPE tissue remains relatively intact, further demonstrating the potential to study the large numbers of samples stored in hospitals and tissue banks worldwide. Still, not all samples are equal, and for the purposes of obtaining the best quality DNA for germline genotyping from radical prostatectomy tissues, our findings suggest that urethral tissue DNA is preferential to that of the seminal vesicle.

Conclusions

By comparing germline genotype concordance between different sources of tissue from radical prostatectomy specimens, we found that various normal tissue sources may in fact have different levels of concordance with blood. Urethral tissue genotypes exhibited not only increased genotype call rate, but also increased genotype concordance with blood in comparison with seminal vesicle derived genotypes when controlling for subject-level and variant-level factors. These findings motivate characterization of different sources of genetic heterogeneity and mosaicism in radical prostatectomy normal tissues and highlight the importance of identifying the source of normal tissue that produces the greatest validity for any given biomarker assay, including microarray genotyping and tumor-normal sequencing.

Acknowledgements

This work was supported by the Department of Defense Transformative Impact Award (W81XWH-132-0074 to P.R.C.), the National Institutes of Health (GM067547; CA088164 to J.S.W.; K07CA197077 and KL2TR000143 to E.L.V.B.), and the UCSF Discovery Fellows Program. We thank Jeremy Gollub and Andrea Finn of Affymetrix Inc. for their assistance with custom microarray design.

Figures

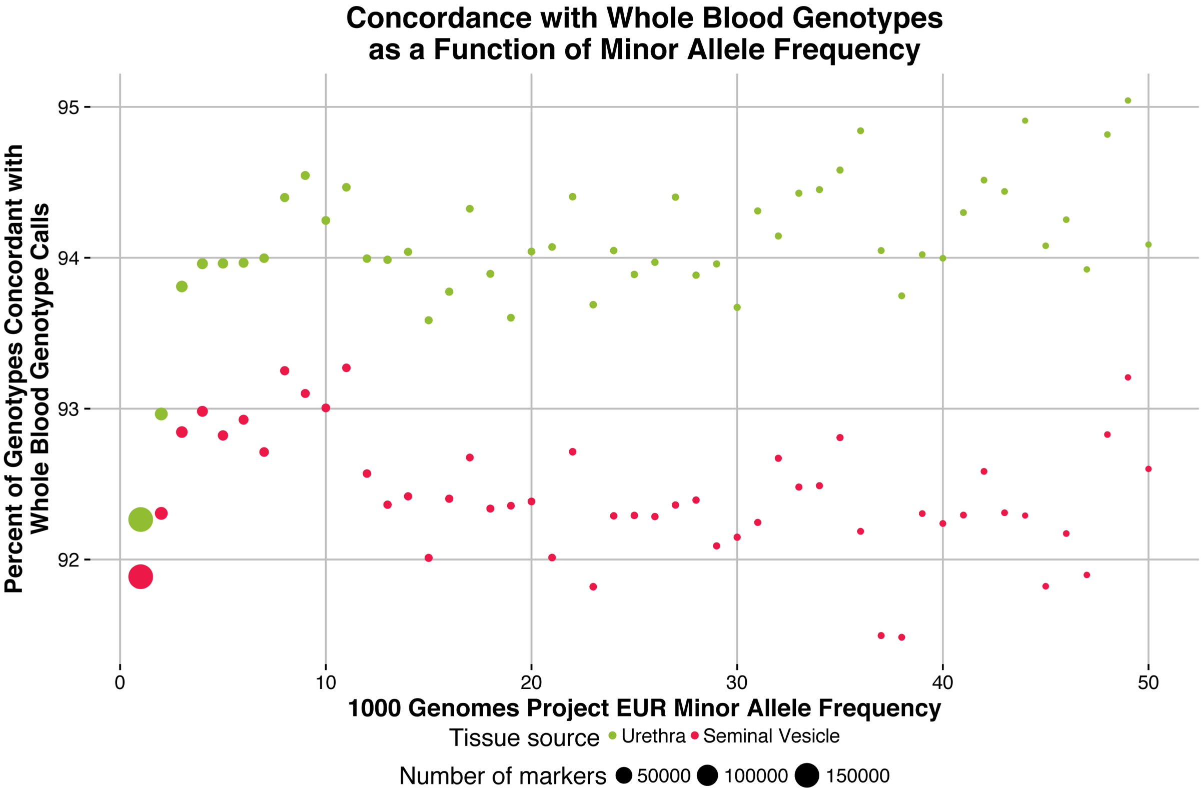


Fig. 1. Urethra-blood genotype concordance compared with seminal vesicle-blood over a range of variant minor allele frequency bins.

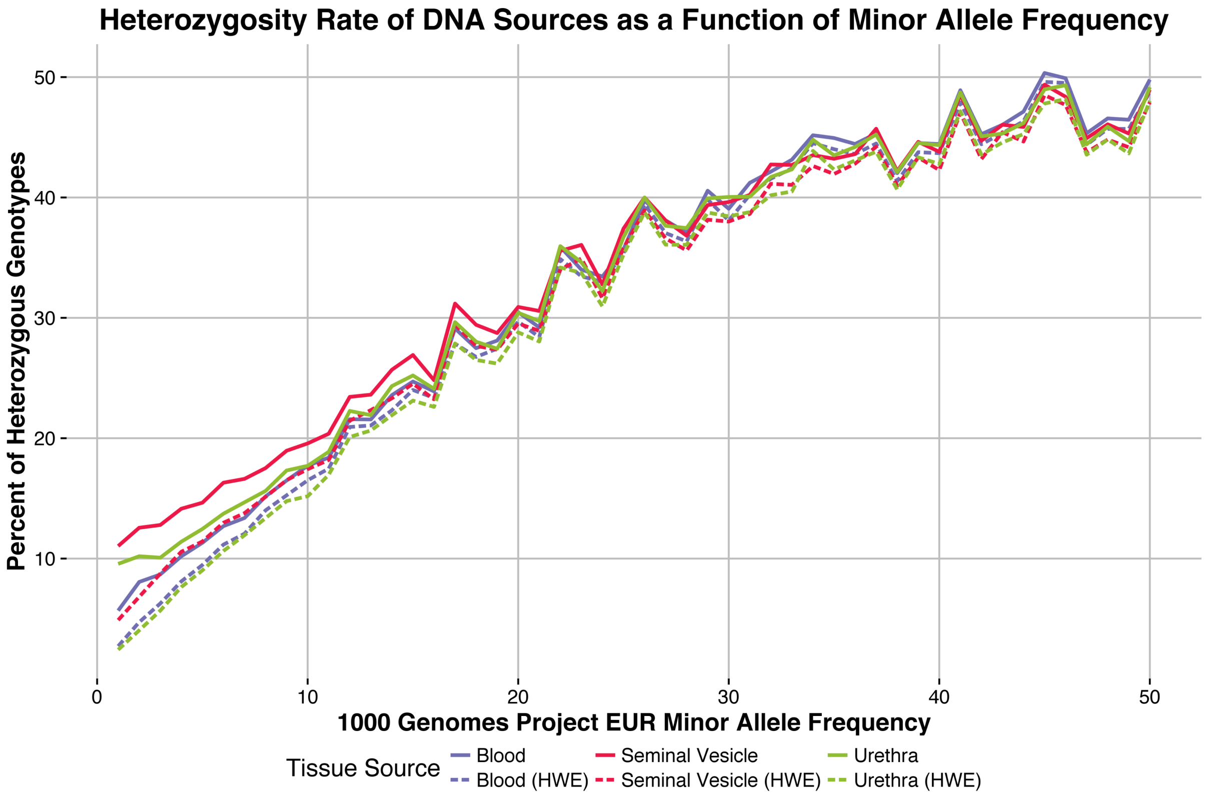
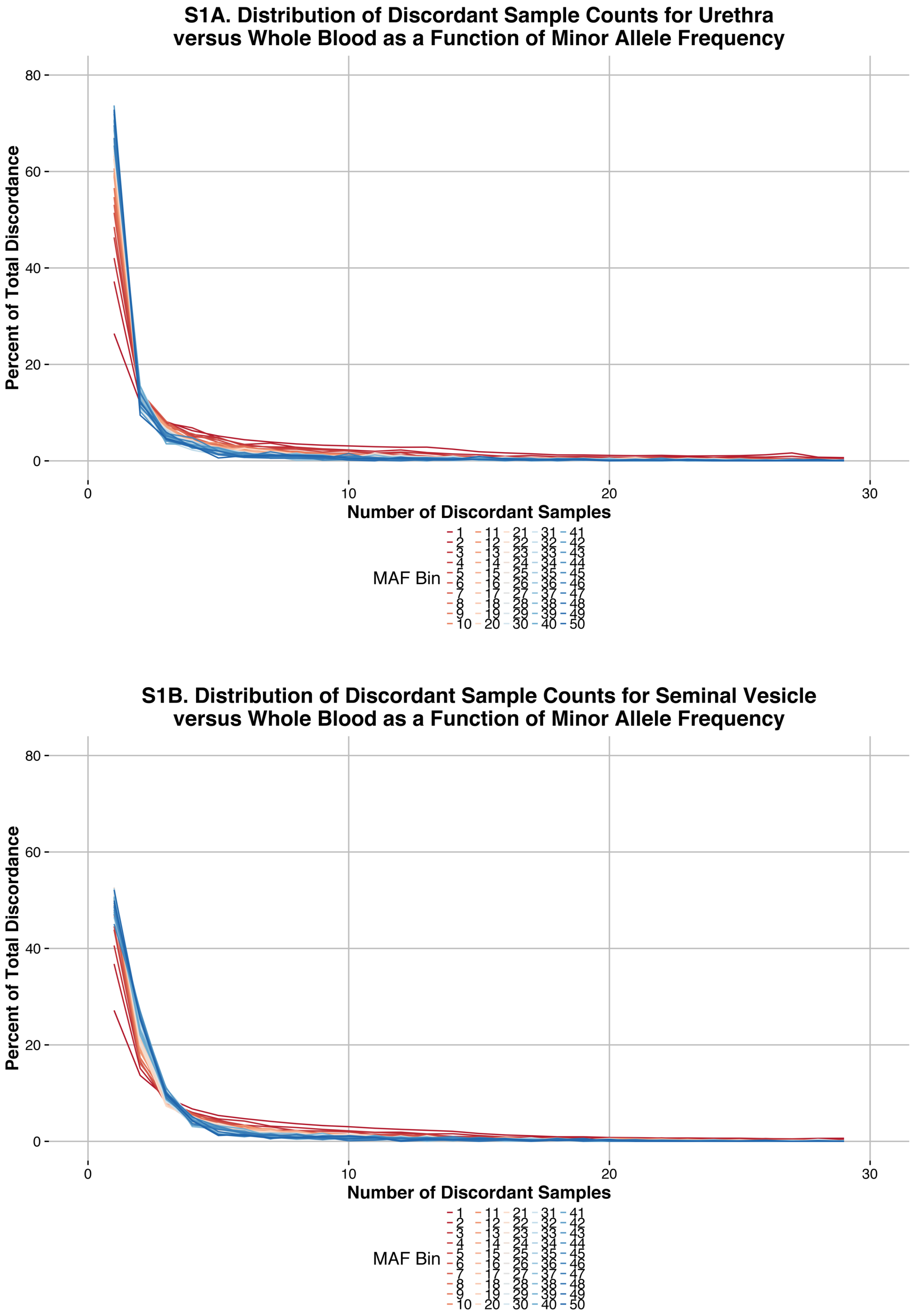


Fig. 2. Heterozygosity rate for blood, urethra, and seminal vesicle genotypes over a range of variant minor allele frequency bins, before (solid) and after (dashed) Hardy–Weinberg equilibrium filtering.



S1A. Distribution of Discordant Sample Counts for Urethra versus Whole Blood as a Function of Minor Allele Frequency

S1B. Distribution of Discordant Sample Counts for Seminal Vesicle versus Whole Blood as a Function of Minor Allele Frequency

Fig. S1. Proportions of discordant counts in Seminal Vesicle-Blood (S1A) and Urethra-Blood (S1B) comparisons over a range of 1000 Genomes Project (EUR) minor allele frequency bins

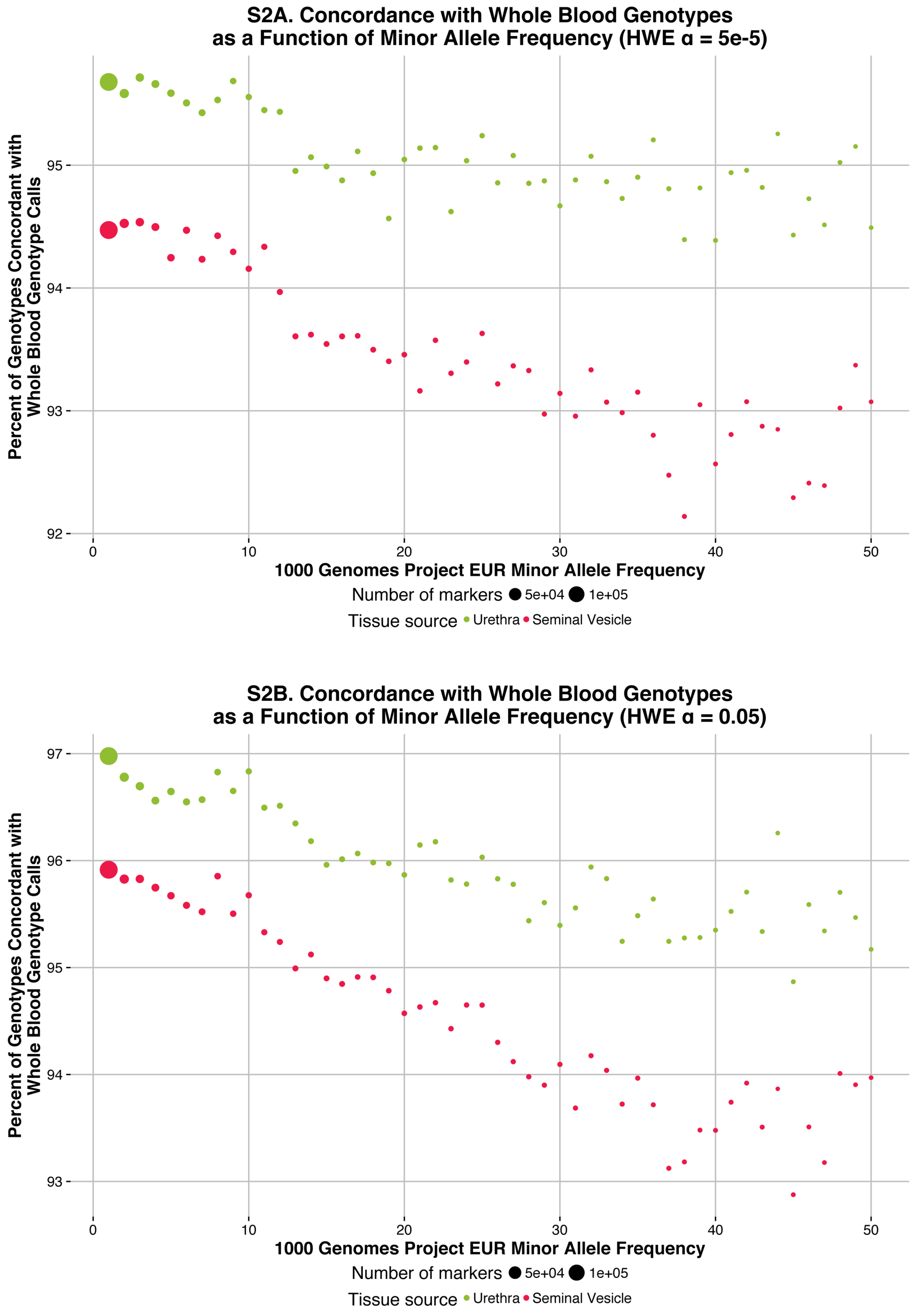


Fig. S2. Urethra-Blood genotype concordance compared to Seminal Vesicle-Blood concordance among markers in Hardy-Weinberg equilibrium with significance level α = 5 x 10^{-5} (S2A) and α = 0.05 (S2B) over a range of minor allele frequency bins

Tables

TABLE I. Genotype Summary Statistics and Concordance Stratified by Tissue Source and Clinicopathologic Factors

Summary statistics and variables for 29 research subjects and 127,847 common SNPs (MAF > 5%)			**Peripheral blood**	**Urethra (UR)**	**Seminal vesicle (SV)**	**Percent difference (UR - SV)**	**P-value**
Average DNA quantity (± std. dev.) post-amplification			1543 ng (± 163)	1472 ng (± 129)	1425 ng (± 192)	-	0.12
Genotype call rate (post-Hardy-Weinberg Equilibrium QC)			98.3% (97.9%)	97.0% (96.8%)	95.9% (95.8%)	+1.1% (+1.0%)	0.0015 (0.0010)
Concordance of genotype calls with peripheral blood genotypes (post-Hardy-Weinberg Equilibrium QC)			100% (100%)	94.1% (95.4%)	92.5% (94.0%)	+1.6% (+1.4%)	0.035 (0.037)
Age at Diagnosis (years)	< 55	n = 6	-	95.8%	94.2%	+1.6%	0.045
	55 - 60	n = 12	-	92.0%	91.1%	+0.9%	0.21
	> 60	n = 11	-	95.6%	93.2%	+2.4%	0.25
DNA quantity difference post-amplification, by quartile (UR - SV, ng)	Q1 [-178.4, -47.7]	n = 7	-	95.6%	95.4%	+0.2%	0.49
	Q2 [-47.6, 14.2]	n = 7	-	89.2%	88.0%	+1.2%	0.70
	Q3 [14.3, 90.4]	n = 8	-	95.8%	94.7%	+1.1%	0.33
	Q4 [90.5, 663.8]	n = 7	-	96.0%	91.6%	+4.4%	0.063
Pathologic Gleason Score	6	n = 20	-	93.5%	92.2%	+1.3%	0.042
	7 (3+4)	n = 7	-	95.7%	92.8%	+2.9%	0.26
	7 (4+3)	n = 2	-	93.8%	94.5%	-0.7%	0.79
Pathologic T-stage	T2a	n = 1	-	92.0%	91.6%	+0.4%	-
	T2c	n = 25	-	94.0%	92.3%	+1.7%	0.048
	T3a	n = 3	-	95.6%	94.8%	+0.8%	0.32
PSA at diagnosis (ng / mL)	< 4.5	n = 10	-	95.7%	92.9%	+2.8%	0.08
	4.5 - 6.5	n = 11	-	95.8%	94.9%	+0.9%	0.14
	> 6.5	n = 8	-	89.3%	88.5%	+0.8%	0.95

References

1. Grönberg H, Adolfsson J, Aly M, Nordström T, Wiklund P, Brandberg Y, Thompson J, Wiklund F, Lindberg J, Clements M, Egevad L, Eklund M. Prostate cancer screening in men aged 50-69 years (STHLM3): a prospective population-based diagnostic study. Lancet Oncol 2015;16(16):1667-1676.
2. Vachon CM, Pankratz VS, Scott CG, Haeberle L, Ziv E, Jensen MR, Brandt KR, Whaley DH, Olson JE, Heusinger K, Hack CC, Jud SM, Beckmann MW, Schulz-Wendtland R, Tice JA, Norman AD, Cunningham JM, Purrington KS, Easton DF, Sellers TA, Kerlikowske K, Fasching PA, Couch FJ. The contributions of breast density and common genetic variation to breast cancer risk. J Natl Cancer Inst 2015;107(5):dju397.
3. Kvale MN, Hesselson S, Hoffmann TJ, Cao Y, Chan D, Connell S, Croen LA, Dispensa BP, Eshragh J, Finn A, Gollub J, Iribarren C, Jorgenson E, Kushi LH, Lao R, Lu Y, Ludwig D, Mathauda GK, McGuire WB, Mei G, Miles S, Mittman M, Patil M, Quesenberry CP, Jr., Ranatunga D, Rowell S, Sadler M, Sakoda LC, Shapero M, Shen L, Shenoy T, Smethurst D, Somkin CP, Van Den Eeden SK, Walter L, Wan E, Webster T, Whitmer RA, Wong S, Zau C, Zhan Y, Schaefer C, Kwok PY, Risch N. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics 2015;200(4):1051-1060.
4. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84(3):362-369.
5. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455(7216):1061-1068.
6. Vos HI, van der Straaten T, Coenen MJ, Flucke U, te Loo DM, Guchelaar HJ. High-quality genotyping data from formalin-fixed, paraffin-embedded tissue on the drug metabolizing enzymes and transporters plus array. J Mol Diagn 2015;17(1):4-9.
7. Yu YP, Michalopoulos A, Ding Y, Tseng G, Luo JH. High fidelity copy number analysis of formalin-fixed and paraffin-embedded tissues using Affymetrix Cytoscan HD chip. PLoS One 2014;9(4):e92820.
8. Cannon-Albright LA, Cooper KG, Georgelas A, Bernard PS. High quality and quantity Genome-wide germline genotypes from FFPE normal tissue. BMC Res Notes 2011;4:159.
9. Cheng L, Song SY, Pretlow TG, Abdul-Karim FW, Kung HJ, Dawson DV, Park WS, Moon YW, Tsai ML, Linehan WM, Emmert-Buck MR, Liotta LA, Zhuang Z. Evidence of independent origin of multiple tumors from patients with prostate cancer. J Natl Cancer Inst 1998;90(3):233–237.
10. Bostwick DG, Qian J, Frankel K. The incidence of high grade prostatic intraepithelial neoplasia in needle biopsies. J Urol 1995;154(5):1791–1794.
11. Jung SH, Shin S, Kim MS, Baek IP, Lee JY, Lee SH, Kim TM, Lee SH, Chung YJ. Genetic progression of high grade prostatic intraepithelial neoplasia to prostate cancer. Eur Urol 2016; 69(5):823–830.

12. Epstein JI, Grignon DJ, Humphrey PA, McNeal JE, Sesterhenn IA, Troncoso P, Wheeler TM. Interobserver reproducibility in the diagnosis of prostatic intraepithelial neoplasia. Am J Surg Pathol 1995;19(8):873–886.
13. Chai H, Brown RE. Field effect in cancer-an update. Ann Clin Lab Sci 2009;39(4): 331-337.
14. Lindquist KJ, Paris PL, Hoffmann TJ, Cardin NJ, Kazma R, Mefford JA, Simko JP, Ngo V, Chen Y, Levin AM, Chitale D, Helfand BT, Catalona WJ, Rybicki BA, Witte JS. Mutational Landscape of Aggressive Prostate Tumors in African American Men. Cancer Res 2016;76(7):1860-1868.
15. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell 2015;163(4):1011-1025.
16. Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF, Coleman I, Ng SB, Salipante SJ, Rieder MJ, Nickerson DA, Corey E, Lange PH, Morrissey C, Vessella RL, Nelson PS, Shendure J. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci U S A 2011;108(41):17087-17092.
17. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57-74.
18. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779.
19. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, Cawley S, Chung E, Connell S, Eshragh J, Ewing M, Gollub J, Henderson M, Hubbell E, Iribarren C, Kaufman J, Lao RZ, Lu Y, Ludwig D, Mathauda GK, McGuire W, Mei G, Miles S, Purdy MM, Quesenberry C, Ranatunga D, Rowell S, Sadler M, Shapero MH, Shen L, Shenoy TR, Smethurst D, Van den Eeden SK, Walter L, Wan E, Wearley R, Webster T, Wen CC, Weng L, Whitmer RA, Williams A, Wong SC, Zau C, Finn A, Schaefer C, Kwok PY, Risch N. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 2011;98(2):79-89.
20. Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, Lu Y, Mei G, Purdy MM, Quesenberry C, Rowell S, Shapero MH, Smethurst D, Somkin CP, Van den Eeden SK, Walter L, Webster T, Whitmer RA, Finn A, Schaefer C, Kwok PY, Risch N. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics 2011;98(6):422-430.
21. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature 2015;526(7571):68-74.
22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(3):559-575.

23. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.
24. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2009.
25. Abraham JE, Maranian MJ, Spiteri I, Russell R, Ingle S, Luccarini C, Earl HM, Pharoah PP, Dunning AM, Caldas C. Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. BMC Med Genomics 2012;5:19.
26. Hong H, Xu L, Liu J, Jones WD, Su Z, Ning B, Perkins R, Ge W, Miclaus K, Zhang L, Park K, Green B, Han T, Fang H, Lambert CG, Vega SC, Lin SM, Jafari N, Czika W, Wolfinger RD, Goodsaid F, Tong W, Shi L. Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. PLoS ONE 2012;7(9):e44483.
27. Wang Y, Carlton VE, Karlin-Neumann G, Sapolsky R, Zhang L, Moorhead M, Wang ZC, Richardson AL, Warren R, Walther A, Bondy M, Sahin A, Krahe R, Tuna M, Thompson PA, Spellman PT, Gray JW, Mills GB, Faham M. High quality copy number and genotype data from FFPE samples using Molecular Inversion Probe (MIP) microarrays. BMC Med Genomics 2009;2:8.
28. Zhang S, Tan IB, Sapari NS, Grabsch HI, Okines A, Smyth EC, Aoyama T, Hewitt LC, Inam I, Bottomley D, Nankivell M, Stenning SP, Cunningham D, Wotherspoon A, Tsuburaya A, Yoshikawa T, Soong R, Tan P. Technical reproducibility of single-nucleotide and size-based DNA biomarker assessment using DNA extracted from formalin-fixed, paraffin-embedded tissues. J Mol Diagn 2015;17(3):242–250.
29. Wang M, Escudero-Ibarz L, Moody S, Zeng N, Clipson A, Huang Y, Xue X, Grigoropoulos NF, Barrans S, Worrillow L, Forshew T, Su J, Firth A, Martin H, Jack A, Brugger K, Du MQ. Somatic Mutation Screening Using Archival Formalin-Fixed, Paraffin-Embedded Tissues by Fluidigm Multiplex PCR and Illumina Sequencing. J Mol Diagn 2015;17(5):521-532.
30. Fisher KE, Zhang L, Wang J, Smith GH, Newman S, Schneider TM, Pillai RN, Kudchadkar RR, Owonikoko TK, Ramalingam SS, Lawson DH, Delman KA, El-Rayes BF, Wilson MM, Sullivan HC, Morrison AS, Balci S, Adsay NV, Gal AA, Sica GL, Saxe DF, Mann KP, Hill CE, Khuri FR, Rossi MR. Clinical Validation and Implementation of a Targeted Next-Generation Sequencing Assay to Detect Somatic Variants in Non-Small Cell Lung, Melanoma, and Gastrointestinal Malignancies. J Mol Diagn 2016;18(2):299-315.
31. Clynick B, Tabone T, Fuller K, Erber W, Meehan K, Millward M, Wood BA, Harvey NT. Mutational Analysis of BRAF Inhibitor-Associated Squamoproliferative Lesions. J Mol Diagn 2015;17(6):644-651.
32. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjo€rd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, J€ager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otın C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P,

Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van ’t Veer L, Vincent-Salomon A, Waddell N, Yates LR. Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. Nature 2013;500(7463):415–421.

33. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, Dominiczak A, Morris A, Porteous D, Smith B, Stratton MR, UK10K Consortium, Hurles ME. Timing, rates and spectra of human germline mutation. Nat Genet 2016; 48(2):126–133.
34. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP. Extensive genetic variation in somatic human tissues. Proc Natl Acad Sci U S A 2012;109(44):18018-18023.
35. Lee HM, Solan MJ, Lupinacci P, Gomella LG, Valicenti RK. Long-term outcome of patients with prostate cancer and pathologic seminal vesicle invasion (pT3b): effect of adjuvant radiotherapy. Urology 2004;64(1):84-89.
36. Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, Fasching PA, Widschwendter M. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun 2016;7:10478.
37. Troester MA, Hoadley KA, D’Arcy M, Cherniack AD, Stewart C, Koboldt DC, Robertson AG, Mahurkar S, Shen H, Wilkerson MD, Sandhu R, Johnson NB, Allison KH, Beck AH, Yau C, Bowen J, Sheth M, Hwang ES, Perou CM, Laird PW, Ding L, Benz CC. DNA defects, epigenetics, and gene expression in cancer-adjacent breast: a study from The Cancer Genome Atlas. Npj Breast Cancer 2016;2:16007.
38. Blow N. Tissue preparation: Tissue issues. Nature 2007;448(7156):959-963.
39. Hedegaard J, Thorsen K, Lund MK, Hein AM, Hamilton-Dutoit SJ, Vang S, Nordentoft I, Birkenkamp-Demtroder K, Kruhoffer M, Hager H, Knudsen B, Andersen CL, Sorensen KD, Pedersen JS, Orntoft TF, Dyrskjot L. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS One 2014;9(5):e98187.