Facile semi-automated forensic body fluid identification by multiplex solution hybridization of NanoString® barcode probes to specific mRNA targets

Patrick Danaher, Ph.D.[1], Robin Lynn White, M.S.[1], Erin K. Hanson, Ph.D.[2] and Jack Ballantyne, Ph.D.[2,3]

[1]NanoString Technologies, 530 Fairview Avenue N, Suite 2000, Seattle, WA 98109
[2]National Center for Forensic Science, PO Box 162367, Orlando, FL 32816-2367, USA
[3]Department of Chemistry, University of Central Florida, PO Box 162366, Orlando, FL 32816-2366, USA

Short Title: Body fluid ID using NanoString® probes

Correspondence:

Jack Ballantyne, Ph.D.
Professor
Department of Chemistry
Associate Director (Research)
National Center for Forensic Science
University of Central Florida
PO Box 162367
Orlando, FL 32816-2367

Voice: +01 407 823-4041
Fax: +01 407 823-4042
E-mail: Jack.Ballantyne@ucf.edu

Keywords: Forensic Science; Body Fluid Identification; mRNA Profiling; Gene Expression; RNA; NanoString®; De-convolution of Body Fluid Mixtures

Abstract:

A DNA profile from the perpetrator does not reveal, *per se*, the circumstances by which it was transferred. Body fluid identification by mRNA profiling may allow extraction of contextual 'activity level' information from forensic samples. Here we describe the development of a prototype multiplex digital gene expression (DGE) method for forensic body fluid/tissue identification based upon solution hybridization of color-coded NanoString® probes to 23 mRNA targets. The method identifies peripheral blood, semen, saliva, vaginal secretions, menstrual blood and skin. We showed that a simple 5 minute room temperature cellular lysis protocol gave equivalent results to standard RNA isolation from the same source material, greatly enhancing the ease-of-use of this method in forensic sample processing.

We first describe a model for gene expression in a sample from a single body fluid and then extend that model to mixtures of body fluids. We then describe calculation of maximum likelihood estimates (MLEs) of body fluid quantities in a sample, and we describe the use of likelihood ratios to test for the presence of each body fluid in a sample. Known single source samples of blood, semen, vaginal secretions, menstrual blood and skin all demonstrated the expected tissue-specific gene expression for at least two of the chosen biomarkers. Saliva samples were more problematic, with their previously identified characteristic genes exhibiting poor specificity. Nonetheless the most specific saliva biomarker, HTN3, was expressed at a higher level in saliva than in any of the other tissues.

Crucially, our algorithm produced zero false positives across this study's 89 unique samples. As a preliminary indication of the ability of the method to discern

admixtures of body fluids, five mixtures were prepared. The identities of the component fluids were evident from the gene expression profiles of four of the five mixtures. Further optimization of the biomarker 'CodeSet' will be required before it can be used in casework, particularly with respect to increasing the signal-to-noise ratio of the saliva biomarkers. With suitable modifications, this simplified protocol with minimal hands on requirement should facilitate routine use of mRNA profiling in casework laboratories.

1. Introduction

Genetic identification of the donor of transferred biological traces deposited at the crime scene or on a person using STR analysis is now routine practice worldwide [1]. This represents potentially crucial 'source level' information for investigators [2]. A DNA profile from the perpetrator does not however reveal the circumstances by which it was transferred. This contextual information (sometimes known as the 'activity level' in Cook and Evett's classic 1998 paper [2]) is important for casework investigations because the deposition of the perpetrator's biological material requires some behavioral activity that results in its transfer from the body. The consequences of different modes of transfer of the DNA profile may dramatically affect the investigation and prosecution of the crime. For example a DNA profile from a victim that originates from skin versus the same DNA profile that originates from vaginal secretions may support social or sexual contact respectively. Thus tissue/body fluid sourcing of the DNA profile should be an important concern for, and service from, forensic genetics practitioners who are integral to the investigative team. The problem is that, up until the recent past, it was not possible to definitively identify many of the important body fluids of interest (e.g. vaginal secretions, saliva, and menstrual blood).

In order to overcome the limitations of currently used classical body fluid identification methods, the use of messenger RNA (mRNA) profiling, as described by Juusola & Ballantyne [3], was proposed to supplant conventional methods for body fluid identification. Terminally differentiated cells, whether they comprise blood monocytes or lymphocytes, ejaculated spermatozoa, epithelial cells lining the oral cavity or epidermal cells from the skin become such during a developmentally regulated program in which

certain genes are turned off (i.e. transcriptionally silent) and turned on (i.e. are actively transcribed and translated into protein) [4]. Thus, a pattern of gene expression is produced that is unique to each cell type in both the presence and the relative abundance of specific mRNAs [4]. The type and abundance of mRNAs, if determined, would then permit a definitive identification of the body fluid or tissue origin of forensic samples. This is the basis for mRNA profiling for body fluid identification. RNA profiling now offers the ability to identify all forensically relevant biological fluids using methods compatible with the current DNA analysis pipeline [5,6]. Despite the identification of numerous body fluid specific candidates there is some reluctance to utilize RNA profiling assays in the forensic community due to concerns over the perceived instability of RNA in biological samples. However, several studies have been conducted in order to assess the stability of RNA in dried forensic stains [7-10]. These have demonstrated that RNA of sufficient quantity and quality for analysis can be recovered from aged and environmentally compromised forensic samples [7-10]. The effective stability (i.e. 'recoverability') of mRNA in aged and compromised samples is not dissimilar to that of DNA and provides support to the use of mRNA profiling assays in forensic casework (Ballantyne, unpublished observations). The recently published EDNAP collaborative exercises on mRNA profiling for body fluid identification further demonstrate a significant interest in mRNA profiling by the forensic community in Europe and around the world as well as the ease in which this technology can be implemented into forensic casework laboratories [11-15]. Collectively, these studies demonstrate an interest in the use of mRNA profiling in forensic casework and its suitability of use with forensic samples and therefore warrant continued evaluation and development. Other classes of

RNA also exist in the cell and one in particular, microRNA (miRNA), has been investigated for potential forensic use since the short size of the molecule (~21-25 bases) makes it an attractive option for analyzing degraded specimens [16-22]. The field of forensic miRNA profiling, although promising, is less mature in terms of there being an international consensus on the identity and specificity of the best body fluid specific miRNA targets. Other non-RNA methods for body fluid identification have been recently investigated including the use of epigenetic [23-29] and proteomic [30-32] biomarkers. Although exhibiting some promise, epigenetic markers have not been identified for all of the important common body fluids and tissues such as vaginal secretions and skin. Proteomic markers suffer from a lack of demonstrated reproducibility studies among different laboratories, and paucity of peer reviewed reports demonstrating their forensic validity.

Gene expression differences are quantitative in nature meaning that a particular biomarker may be expressed in a particular cell type at low, intermediate or high levels. Even when it is not generally regarded as being expressed in a particular cell type it may exhibit basal level (or 'leaky') transcription with a few molecules present per cell. Thus far there have been three main methods developed for mRNA profiling of forensic samples: capillary electrophoresis (CE)-based analysis [5-7,33-36], quantitative RT-PCR (qRT-PCR) [7,37-39] and, more recently, high resolution melt (HRM) analysis [40]. Due to its facile multiplex capabilities and routine use in DNA profiling, CE-based analysis has been the platform of choice for casework mRNA assays [5,6]. However post–PCR CE peak heights/areas are, at best, semi-quantitative in nature with respect to biomarker expression levels. Similarly, HRM signal amplitude does not appear to correlate precisely

with RNA input [40]. Although qRT-PCR permits quantitation of biomarker targets, its low multiplex capability (typically 3-4 targets maximum compared to >20 for CE) appears to have limited its use.

In contrast to the aforementioned, digital gene expression (DGE) methods precisely count the number of individual transcripts in a sample [41] which facilitates the use of advanced statistical methods to better evaluate and interpret the experimental data. This facility would be expected to be of significant benefit when analyzing body fluid mixtures that are commonly encountered in forensic analysis. Deep sequencing of the transcriptome using next generation sequencing (NGS) technologies is capable of directly identifying and quantifying (by counting) all mRNA transcripts in a sample, a DGE technique known as RNA sequencing (RNA-Seq) [42]. RNA-Seq has been spectacularly successful in advancing our knowledge of cell-type-specific gene expression including transcript quantification and elucidation of their sequence diversity [42]. Although NGS heralds a new era of forensic genomics, impediments to its routine implementation in body fluid RNA analysis include its high cost of reagents and time-consuming, complex analysis. In this work we sought an alternative DGE method to NGS that is simpler and requires minimal hands-on experimentation. Here we describe the development of a prototype multiplex DGE method for forensic body fluid identification based upon solution hybridization of color-coded NanoString® probes [43] to 23 tissue/body fluid specific and 10 housekeeping gene mRNA targets present in forensic type samples. Concomitantly, to facilitate routine use, we also devised a simple 5 minute room temperature cellular lysis protocol as an alternative to standard RNA isolation for forensic sample processing.

2. Methods

2.1 Body fluid samples

Body fluids were collected from volunteers using procedures approved by the University's Institutional Review Board. Informed written consent was obtained from each donor. Blood samples were collected by venipuncture into vacutainers (K3-EDTA preservative) and 50 µl aliquots were placed onto cotton cloth and dried at room temperature. Freshly ejaculated semen was provided in sealed plastic tubes and stored frozen. After thawing, the semen was absorbed onto sterile cotton swabs and allowed to dry. Buccal samples (saliva) were collected from donors using sterile swabs by swabbing the inside of the donor's mouth. Semen-free vaginal secretions and menstrual blood were collected using sterile cotton swabs. Admixed body fluid samples were created by combining ½ of a 50 µl stain or single cotton swab from each body fluid. Environmental samples were prepared by exposing body fluid samples to the outside ambient heat, light and humidity protected ('covered') or non-protected ('uncovered') from precipitation for varying lengths of time (Supplementary Table1). Human skin total RNA was obtained from commercial sources: Stratagene/Agilent Technologies (Santa Clara, CA), Biochain® (Hayward, CA), Zenbio (Research Triangle Park, NC), and Zyagen (San Diego, CA). Human brain total RNA was obtained from a commercial source (Biochain®) (run as an internal positive control and not used in any data analysis). Cellular skin samples were collected by swabbing human skin or a touched object surface with a sterile water pre-moistened sterile swab. For all RNA isolations, ½ or a whole 50 µl stain or single cotton swab was used. All samples were stored at -20^{o}C until needed, except for the total RNA samples which were stored at -47^{o}C.

Suspected bio-particles from male shirt collar samples were collected as previously described [44]. Briefly, WF Gel-Film® x8 retention level (Gel-Pak®, Hayward, CA), was cut to a size appropriate for subsequent attachment to a glass microscope slide support (3" x 1" x 1mm, Fisher Scientific, Suwanee, GA). Using sterile tweezers, the back protective covering was removed to expose the adhesive back and the Gel-Film® was placed onto a clean glass microscope slide. The top protective plastic film layer was then removed using re-sterilized tweezers. The Gel-Film® surface was then repeatedly touched to the sample area (direct skin, clothing or object surface) several times to ensure sufficient transfer of biological material. Samples were stained with Trypan Blue (0.4%) (Sigma-Aldrich, St. Louis, MO) for 1 minute, then washed briefly by gentle flooding with sterile ultrapure water with a resistivity of 18.2MΩ at 25°C. Samples were then air-dried at room temperature prior to proceeding to sample collection. All samples were stored at room temperature in microscope slide boxes protected from light. Bio-particles were viewed, imaged and collected using a Leica M205C stereomicroscope (Micro Optics of FL, Inc, Davie, FL). Twenty-five, fifty and one hundred bio-particles (i.e. single cells or 'cellular agglomerates') were collected. Bio-particles were collected from Gel-Film® surface using 3M™ water-soluble wave solder tape (5414 transparent) on the end of a tungsten needle. The 3M™ water-soluble adhesive was adhered to a clean glass microscope slide using double sided tape and collected on the end of a tungsten needle under the stereomicroscope. The collected bio-particles were then transferred into a sterile 0.2ml PCR flat-cap tube (Phenix Research, Candler, NC)) containing lysis buffer: 100 bio-particle shirt collar sample - 10 µl of lysis buffer solution: 2.1X buffer-blue, 10% *forensic*GEM™ reagent (ZyGEM *forensic*GEM™ tissue kit, VWR, Suwanne,

GA), sterile water; 25 and 50 bio-particle shirt collar samples - 5 µl of lysis buffer solution: 1X buffer-silver, 5% *RNA*GEM™ reagent (ZyGEM *RNA*GEM™ tissue kit, VWR), sterile water.

2.2 *RNA Isolation*

Total RNA was extracted from blood, semen, saliva, vaginal secretions, menstrual blood and skin using a manual organic RNA extraction (guanidine isothiocyanate-phenol:chloroform) as previously described [33,45]. Briefly, 500 µl of pre-heated (56°C for 10 minutes) denaturing solution (4M guanidine isothiocyanate, 0.02M sodium citrate, 0.5% sarkosyl, 0.1M β-mercaptoethanol) was added to a 1.5mL Safe Lock extraction tube (Eppendorf, Westbury, NY) containing the stain or swab. The samples were incubated at 56°C for 30 minutes. The swab or stain pieces were then placed into a DNA IQ™ spin basket (Promega, Madison, WI), re-inserted back into the original extraction tube, and centrifuged at 14,000 rpm (16,000 x g) for 5 minutes. After centrifugation, the basket with swab/stain pieces was discarded. To each extract the following was added: 50 µl 2 M sodium acetate and 600 µl acid phenol:chloroform (5:1), pH 4.5 (Ambion by Life Technologies). The samples were then centrifuged for 20 minutes at 14,000 rpm (16,000 x g). The RNA-containing top aqueous layer was transferred to a new 1.5ml microcentrifuge tube, to which 2 µl of GlycoBlue™ glycogen carrier (Ambion by Life Technologies) and 500 µl of isopropanol were added. RNA was precipitated for 1 hour at -20°C. The extracts were then centrifuged for 20 minutes at 14,000 rpm (16,000 x g). The supernatant was removed and the pellet was washed with 900 µl of 75% ethanol/25% DEPC-treated water. Following a centrifugation for 10 minutes at 14,000 rpm (16,000 x

g), the supernatant was removed and the pellet dried using vacuum centrifugation for 3 minutes. Twenty microliters of pre-heated (60°C for 5 minutes) nuclease free water (Ambion by Life Technologies) was added to each sample followed by an incubation at 60°C for 10 minutes. All extracts were DNase treated to remove residual DNA using the Turbo DNA-*free*™ kit (Applied Biosystems (AB) by Life Technologies, Carlsbad, CA) according to the manufacturer's protocol. With each extraction, a negative control (extraction reagents without sample) was included.

Alternatively, total RNA was extracted from blood, semen, saliva, vaginal secretions, menstrual blood and skin using direct lysis without purification. One hundred microliters of Buffer RLT Plus (QIAGEN, Germantown, MD) with 1 µl β-mercaptoethanol was added to a 1.5mL Safe-Lock extraction tube (Eppendorf, Westbury, NY) containing the stain or swab. The samples were incubated at room temperature for 5 minutes with constant vortexing (20 second intervals). The swab or stain pieces were then placed into a DNA IQ™ spin basket (Promega, Madison, WI), re-inserted back into the original extraction tube, and centrifuged at 14,000 rpm (16,000 x g) for 5 minutes. After centrifugation, the basket with swab/stain pieces was discarded. All samples were stored at -20°C until needed.

Total RNA was extracted from bio-particles using the ZyGEM *forensic*GEM™ or *RNA*GEM™ tissue kits (VWR). For the *forensic*GEM™ kit, samples were lysed at 75°C for 15 minutes. For the *RNA*GEM™ kit, samples were lysed at 75°C for 5 minutes. All samples were stored at -20°C until needed.

2.3 RNA Quantitation

RNA extracts (manual organic RNA extraction only) were quantitated with Quant-iT™ RiboGreen® RNA Kit (Invitrogen by Life Technologies, Carlsbad, CA) as previously described [33,45]. Fluorescence was determined using a Synergy™ 2 Multi-Mode microplate reader (BioTek® Instruments, Inc., Winooski, VT).

2.4 *NanoString® Technology*

NanoString® standard gene expression chemistry utilizes two ~50 base probes, the reporter probe and the capture probe, for each mRNA target of interest [43]; when multiplexed, the probe pairs are referred to as a CodeSet. A multiplex CodeSet can be designed to have probe pairs targeting between 20 and 800 mRNAs. Each capture/reporter probe pair within the CodeSet is specifically designed to hybridize to an individual mRNA target. The reporter probe carries the signal and is comprised of a unique molecular fluorescent barcode binding to the 5' end of the mRNA target. The capture probe binds to the 3' end of the mRNA target and adheres the capture probe/barcode/target complex to the cartridge surface for data collection (see Figure 1). After overnight hybridization at 65°C in a thermal cycler (typical time of 12-24 hours), the complex is purified on the nCounter Prep Station with excess, unbound probes removed and intact complexes bound, stretched and immobilized on an nCounter Cartridge. Sample cartridges are then placed onto the nCounter Digital Analyzer for counting and data collection of each target complex. The number of times each barcode is counted is proportional to the abundance of that mRNA target in a given sample.

In this study, a NanoString® multiplex custom CodeSet was designed and created to target 23 genes known to be differentially expressed in forensically relevant body

fluids and tissues. As a reference, 10 ubiquitously expressed housekeeping genes were also included in the CodeSet, giving a 33-plex total. The body fluids and tissues targeted include: venous blood, menstrual blood, semen, saliva, vaginal secretions, and skin. The multiplex CodeSet consisted of 3 venous blood genes (ALAS2, ANK1, HBB) [11,13,34], 2 menstrual blood genes (LEFTY2, MMP10) [15,34,36], 3 saliva genes (HTN3, MUC7, STATH) [3,14,34], 3 semen genes (PRM2, SEMG1, TGM4) [14,34], 5 skin genes (CCL27, IL1F7, KRT9, LCE1C, LCE2D) [33,46], 7 vaginal secretion genes (CYP2A7, CYP2B7P1, DKK4, FUT6, IL19, MYOZ1, NOXO1) [45] and 10 reference (i.e. housekeeping) genes (B2M, COX1, HPRT1, PGK1, PPIH, S15, TCEA1, TFRC, UBC, UBE2D2) (Table 2). The CodeSet also included 6 positive control probes and 8 negative control probes. The 6 positive control probes are designed to assess overall assay performance and to normalize the data, accounting for any assay variability within the system. The 8 negative control probes have no corresponding targets within the sample and assess background noise in the system.

A total of 96 assays were included this study, involving 89 samples with technical replicates for 7 of the samples. A detailed summary of the 89 samples is provided in Table 1 and includes 14 blood, 17 semen, 17 saliva, 10 vaginal secretions, 10 menstrual blood, and 14 skin samples as well as 5 mixtures and 2 RNA-free controls. For each body fluid both standard and challenging or environmentally compromised samples were evaluated. Full sample descriptions, including number of donors, and the input (ng of total RNA or volume (μl) of extract) used for each of the 96 samples is provided in Supplementary Table 1.

Hybridization assays were performed according to the standard NanoString® gene expression assay protocol, as follows: Each individual assay consisted of 10µL Reporter Probe, 10µL Hybridization Buffer, 5µL Capture Probe and the specified RNA sample input (in most cases, 50ng of total RNA or 5µL crude lysate) for a total reaction volume of 30µL. Assays were placed into a thermal cycler at 65°C with a 70°C lid, and allowed to hybridize overnight for approximately 16 hours. Following this, assays were placed onto the nCounter Prep Station using the high-sensitivity protocol for purification and immobilization of the hybridized targets on the imaging cartridge. The cartridges were then scanned on the nCounter Digital Analyzer for counting of the hybridized targets, and data files were exported for analysis.

2.5 Statistical Methods

2.5.1 Overview of method

Our approach to the problem is motivated by three properties of *bona fide* casework samples: they often (i) comprise mixtures of two or more fluids, (ii) are limited in quantity and (iii) could be either partially or highly degraded. Our basic approach is as follows: First, we model the probability distribution of gene expression in body fluid samples. Next, we use this model to calculate the Maximum Likelihood Estimate (MLE) for the levels of each body fluid in a sample and to calculate the log-likelihood of a sample's profile given the estimated levels of each fluid. We then construct a likelihood ratio comparing the likelihood of a given sample's profile with and without the presence of a given fluid. If a sample's profile is far more likely when we include a specific fluid in the model, then we conclude the fluid is present in the sample.

2.5.1 Modeling gene expression in body fluids

Gene expression is best modeled on the log (multiplicative) scale: a doubling of a gene's expression level is generally considered a change comparable in magnitude to a halving of its expression level, and a gene increasing from 200 to 400 mRNA transcripts is as meaningful a difference in gene expression as a gene increasing from 2000 to 4000 counts. However, the mathematics of mixtures is additive: if a sample is half blood and half saliva, a gene's cumulative expression level will result from the summation of its expression levels in each tissue. We therefore model the contributions of each fluid to a mixture on the linear scale, but we measure discrepancies between observed and predicted expression on the log scale.

We develop the algorithm as follows: As a conceptual starting point, we first describe a model for gene expression in a sample from a single fluid. We then extend this model to mixtures of fluids. From there we describe calculation of maximum likelihood estimates (MLEs) of fluid quantities in a sample, and we describe the use of likelihood ratios to test for the presence of a fluid in a sample.

2.5.2 Model for gene expression in a sample from a single body fluid

On average, each gene represents a given proportion of total gene expression in each fluid. For example, in the average blood sample we might expect 15% of total RNA to be HBB, 1% to be ALAS2, etc. Call these expected proportions X_{HBB}, X_{ALAS2}, etc. Then in a given blood sample, the vector of expected gene expression is $\beta(X_{HBB}, X_{ALAS2}, \ldots)^T$, where β is the total amount of RNA in the sample.

Due to both biological and technical noise, actual expression will vary around its expectation. Per the multiplicative nature of gene expression, we model this variability as arising from a log-normal distribution, and we assume that each gene is equally variable. A single gene's expression in a sample can then be modeled:

$$\log(y_{HBB}) \sim N(\log(X_{HBB}\,\beta), \sigma^2),$$

where y_{HBB} is the expression of HBB in the sample, and σ^2 is the variance (on the log scale) of HBB's expression around its expectation.

2.5.3 Model for gene expression in mixtures of body fluids

The model for mixtures follows naturally from the model for single-fluid samples. First, let us define notation. We represent matrices with bold, uppercase letters, vectors with bold, lowercase letters, and scalars with lowercase letters. We index samples $i \in (1, \ldots, n)$, genes $j \in (1, \ldots, p)$, and tissues $k \in (1, \ldots, K)$. Call the gene expression profile for a given sample $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^T$, where y_{ij} is the expression of gene j in sample i. Call β_{ik} the amount of fluid k in sample i, and call $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iK})$ the vector of the amounts of all the fluids in sample i. Finally, define the matrix $\mathbf{X}$ to represent the expected proportion of each gene in each fluid type, with x_{jk}, the element in the j^{th} row and the k^{th} column of $\mathbf{X}$, representing the expected proportion of gene j in samples from fluid k.

Assuming the number of mRNA molecules in mixtures of fluids will be a sum of the number of mRNA molecules in each component of the mixture, we can write the expected counts of gene j in sample i:

$$E(y_{ij}) = \sum_{k=1}^{K} \beta_{ik} x_{jk},$$

and the expression for the sample's entire expected gene expression vector is simply

$$E(\mathbf{y}_i) = \mathbf{X}\boldsymbol{\beta}_\mathbf{i}.$$

Again assuming the variability of gene expression occurs on the log scale, we model gene expression in a sample as:

$$\log(\mathbf{y}_i) \sim N(\log(\mathbf{X}\boldsymbol{\beta}_\mathbf{i}), \sigma^2\mathbf{I}),$$

where $\mathbf{I}$ is the identity matrix and σ^2 is the common variance (on the log scale) of all genes. (Note that if $E(\mathbf{y}_i) = \mathbf{X}\boldsymbol{\beta}_\mathbf{i}$, then $E(\log(\mathbf{y}_i)) \neq \log(\mathbf{X}\boldsymbol{\beta}_\mathbf{i})$. However, under the values considered in this application, $E(\log(\mathbf{y}_i))$ very closely approximates $\log(\mathbf{X}\boldsymbol{\beta}_\mathbf{i})$.) As we lack the data to fully estimate the genes' covariance matrix, we approximate it with $\sigma^2\mathbf{I}$.

Before we can apply the above model for gene expression in body fluids, we must estimate two parameters: $\mathbf{X}$, the matrix of expected proportions of gene expression, and σ^2, the variance of gene expression. Estimation of the $\mathbf{X}$ matrix is described in Section 3.2. We estimated σ^2, the variance on the log scale common to all genes, as the average variance of each gene in each tissue or fluid.

2.5.4 Maximum likelihood estimation of the amounts of each tissue or fluid in a sample

Under the assumptions that log gene expression is normally distributed around the log of its expectation and that each gene is equally variable, the MLE for $\boldsymbol{\beta}_\mathbf{i}$ can be calculated as follows:

$$\widehat{\boldsymbol{\beta}}_i = argmin_{\boldsymbol{\beta}} \|\log(\boldsymbol{y}_i) - \log(\mathbf{X}\boldsymbol{\beta})\|_2^2 \text{ s.t. } \boldsymbol{\beta} \geq \mathbf{0},$$

i.e., $\widehat{\boldsymbol{\beta}}_i$ minimizes the sum of squared errors on the log scale between the observed gene expression $\mathbf{y}_i$ and the predicted gene expression $\mathbf{X\beta}$, subject to the constraint that all the elements of $\boldsymbol{\beta}$ are non-negative (a sample cannot have negative amounts of a fluid). As it is doubtful that a closed-form solution to this expression exists, we use numerical methods to optimize it [47]. The expression is not convex in $\boldsymbol{\beta}$; however, we find its estimates to be reasonably robust to differing initial conditions, returning similar estimates with very similar log-likelihoods.

To prevent the algorithm from overexerting itself trying to fit gene expression values in the background of the assay, we found it necessary to add one layer of complexity to the model: in addition to fitting β terms for each fluid, we added a β for background, with a corresponding column in the $\mathbf{X}$ matrix with equal weights on all genes. We further constrained this background β term to contribute no more than 15 counts to each gene. For the same reason, we truncated all gene expression values at 5 counts, a reasonable estimate of the average background counts.

2.5.5 Using likelihood ratios to test the presence of tissues or fluids

In any given sample $\mathbf{y}_i$, our goal is to determine which tissues or fluids are present. That is, we want to test whether each element of $\boldsymbol{\beta}_\mathbf{i}$ equals 0. A reasonable approach to this problem is to calculate the likelihood of the data under the MLE $\widehat{\boldsymbol{\beta}}_i$ and under a constrained MLE $\widehat{\boldsymbol{\beta}}_{i,-j}$ with the β_{ij} term corresponding to the tissue or fluid in question forced to 0. The likelihood ratio under the full and constrained MLEs will summarize the evidence for the presence of the tissue or fluid in question.

Calculation of a log likelihood for the data given a MLE is straightforward. Under our model, log gene expression is normally distributed around the log of the predicted gene expression. Then up to a constant, the log-likelihood of $\mathbf{y}_i$ given $\widehat{\boldsymbol{\beta}}_i$ is:

$$loglik\left(\boldsymbol{y_i}\middle|\widehat{\boldsymbol{\beta}}_i\right) =$$

$$-\frac{1}{2}log\left(det(\sigma^2\mathbf{I})\right) - \frac{1}{2}\left(\log(\boldsymbol{y_i}) - log(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_i)\right)^T \sigma^{-2}\mathbf{I}\left(\log(\boldsymbol{y_i}) - log(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_i)\right).$$

To test whether fluid j is present in sample i, we evaluate the above expression using $\mathbf{y}_i$ and $\widehat{\boldsymbol{\beta}}_i$ and again using $\mathbf{y}_i$ and the constrained MLE $\widehat{\boldsymbol{\beta}}_{i,-j}$, and we calculate a likelihood ratio.

3. Results

3.1 Selection of mRNA biomarkers

We designed a 'CodeSet' to probe 23 body fluid/tissue specific genes and 10 housekeeping genes (Table 2), which is well within the 800 target technological capability of the system. To take advantage of the high multiplex capability of the system, we deliberately included biomarkers that have been demonstrated to be highly specific to a particular body fluid (e.g. PRM2 and SEMG1 for semen) as well as some that have shown a lesser degree of tissue specificity (e.g. MYOZ1 for vaginal secretions and MUC7 for saliva).

3.2 Estimating expected body fluid profiles

Our algorithm requires accurate estimates of each fluid's average gene expression profile; below, we describe the results of this analysis.

Our dataset included samples of highly varying RNA concentration, and genes in the lower-concentration samples frequently dropped into the background noise of the assay. To ensure accurate estimates of each body fluid's average gene expression profile, we only used samples with high expression levels of housekeeping genes. As a set of 'training samples' we took the four highest-expressing samples from each fluid type with the exception of saliva, where a lack of high-expressing samples limited us to three training samples. Supplemental Figure 1 shows the overall housekeeping gene expression levels in the training samples and the remaining samples.

Per our model described in Section 2.5.3, we are interested in the relative expression levels of the genes within each body fluid; that is, in the proportion of total

signature gene expression expected from each gene in a given body fluid. (This is in contrast to most gene expression-based classifiers, which are more interested in each gene's absolute expression level. Since it is unrealistic to expect a housekeeping gene to be invariant across body fluid types, normalizing our data to attain "absolute" expression levels is impossible.) Therefore, we globally normalized each sample, rescaling them so the sum of all expression values was 1 and so that each gene's expression value was its proportion of the total signature gene expression. We then estimated each gene's expected proportion of expression in each fluid with its mean normalized expression value within each fluid.

The five body fluids and skin demonstrated highly distinct gene expression profiles, and although the signature genes varied between samples of the same fluid, their differences between fluids were much greater.

Figure 2 shows the expected proportion of total expression for each gene in each fluid. Supplemental Figure 2 shows the consistency of these profiles in the training data, and Supplemental Figure 3 organizes the information in Figure 2 by gene rather than by fluid. In all fluids the average expression profile exhibits elevated expression of the fluid's putative characteristic genes, although this trend was distinctly weaker in saliva samples.

HBB expression dominated the blood profiles, far exceeding the other blood markers ALAS2 and ANK1, although ALAS2 levels in blood greatly exceeded those of other genes. The putative blood marker ANK1 was not enriched in blood samples, surprisingly appearing most prominently in saliva samples instead. Expression in semen samples came almost entirely from the semen-specific genes PRM2, TGM4 and SEMG1,

although other genes, particularly HBB, were detectable. Saliva samples had the most diffuse profile, with the saliva-specific genes STATH, MUC7 and HTN3 contributing only 28% of total measured expression. Vaginal secretion samples had highly elevated levels of the vaginal markers DKK4, CYP2B7P1 and to a lesser extent FUT6. Menstrual blood samples alone showed elevated expression of their characteristic genes MMP10 and LEFTY2. Unsurprisingly, menstrual blood samples also contained blood (HBB, ALAS2) and vaginal secretion (CYP2B7P1) biomarkers. Skin samples showed elevated expression of the skin genes LCE1C, IL1F7 and CCL27, although these genes were also slightly elevated in vaginal secretions and menstrual blood. HBB was the most prevalent gene in the commercial skin preparation, probably due to the inevitable presence of contaminating endothelial tissue in such preparations.

Most genes were present at a non-negligible proportion of total expression in the saliva samples. This phenomenon results from this study's lack of a good saliva marker. If a gene highly expressed in saliva were measured, the relative expression of the other fluids' characteristic genes in saliva would shrink dramatically.

3.3 Using gene expression to predict the body fluid composition of samples

Our algorithm for body fluid detection is described in detail in the Methods section. Below, we summarize the performance of the method in predicting the body fluid composition of every sample in our study. Crucially for forensic applications, our test appears to have extremely high specificity; in fact, it returned zero false positives in this study's 89 samples.

We used a likelihood ratio cutoff of 100 to declare whether a body fluid was detected in a given sample, and found that 53/80 single-fluid, non-duplicate samples (66%) gave positive results. It is worth noting that our collection of samples was not necessarily representative of the real world population of forensic samples, as in many cases we intentionally chose degraded and miniscule samples to push the limits of the assay. Figure 3 shows the rate at which each body fluid was declared 'detected' in each actual fluid using an LR of 100. Supplemental Figures 4 and 5 indicate the performance of the algorithm in the training samples (abundant RNA) and in the remaining samples (low RNA quantity) respectively. The algorithm was successful in identifying the correct body fluid as long as the sample was abundant enough; in low input samples it detected blood, semen and vaginal secretions reliably while struggling to detect saliva, menstrual blood and skin. Across all samples, detection of blood, semen and vaginal secretions was nearly perfect. Menstrual blood was successfully detected 60% of the time. Blood and vaginal secretions were frequently detected in menstrual blood, though these cannot be considered false positives. Rather, it appears menstrual blood is best modeled as a variable mixture of blood, menstrual blood, and vaginal secretions. Saliva was successfully detected in only 25% of samples, likely due to fact that the characteristic saliva genes were not as informative as other fluids' characteristic signature genes and/or to the very low level of total RNA in most of the saliva samples. Skin also proved difficult to detect (31% success rate); however, the need to identify skin will probably be limited to specialized forensic cases. It is much more important to ensure that skin samples are not misclassified as other tissues.

The choice of a LR >100 cutoff for detecting fluids is arbitrary, and our algorithm could achieve better performance with a less strict cutoff. Figure 4 shows ROC curves for the True Positive Rate (TPR) and False Positive Rate (FPR) for detection of each fluid type in our data. As the LR threshold relaxes our test returns more of both false positives and false negatives. For the tissues with the worst performance in our data – menstrual blood, saliva and skin – the ROC curves reveal that a relaxation of the LR thresholds in some tissues would result in large increases in TPR without any increase in FPR.

3.4 Body fluid mixtures

As a preliminary indication of the ability of the method to discern admixtures of body fluids, five mixtures were prepared by combining ½ of a 50 µl stain or single cotton swab from each body fluid. The mixtures comprised four binary (2 x vaginal secretions/semen, 2 x blood/saliva) and one ternary mixture (semen/saliva/vaginal secretions). The blood/saliva and vaginal secretions/semen were biological, as opposed to technical, replicates since the donors were different. Using an LR of 100 as a decision threshold, two of the five mixtures were called perfectly, namely one of the vaginal secretions/semen and one of the blood/saliva samples (Figure 5). One of the component fluids was identified in each of the three 'false negative' mixtures: vaginal secretions (vaginal secretions/semen and semen/saliva/vaginal secretions) and saliva (blood/saliva). In the latter ternary mix the semen and saliva components were detected but with LRs of <100 (36.9 and 3.4 respectively). In the second blood-saliva sample, the LR for saliva was 95, falling just short of our strict bar for detection. In all but one of the mixture

samples, the component fluids are evident from their likelihood ratio profiles: using an LR cutoff of 5, four of the five mixtures were called perfectly. Significantly, no false positives were observed even under the very generous LR cutoff of 5.

3.5 Development of a routine-use 5 minute RNA direct lysis method

To facilitate routine analysis, we tested a simple 5 minute room temperature cellular lysis protocol as an alternative to standard RNA isolation for forensic sample processing using the NanoString® procedure (See Methods Section). The method is based upon the RLT buffer from QIAGEN which contains a high concentration of guanidine thiocyanate as well as a proprietary mix of detergents. β-mercaptoethanol (1% v/v) is also added before use to inactivate RNAses in the lysate. The NanoString assay involves direct hybridization to the RNA with no enzymatic steps, and neither the presence of the denaturing buffer nor the cellular debris in the lysate have a significant impact on the assay results.

We compared the reproducibility of the assay between standard RNA isolation/purification and direct lysis protocols from the same source material. Fourteen of the samples in our study were compared in this manner. Supplemental Figure 6 shows scatterplots comparing log expression values for each of these same source samples between the two protocols. In general we saw excellent concordance between the two protocols for all genes with a moderate to high degree of expression. The correlation between the protocols breaks down for very lowly-expressed genes, reflecting the greater noise in the assay when measuring vanishing target. The most dramatic differences between replicates (for example in the samples MB-2 and BD-5) are attributable to

expected variance in RNA input amounts between lysate and purified RNA since lysate concentration is not reliably measureable by current methods. The concordance observed between lysis and purified protocols suggest that the simpler, 5 minute lysis protocol would be an efficient option for routine forensic casework workflow.

4. Discussion

The results of this preliminary proof of principle study indicate that it is feasible to identify the common forensically relevant body fluids by multiplex solution hybridization of barcode probes to specific mRNA targets using a simple five minute direct lysis protocol. This simplified protocol with minimal hands-on requirement should facilitate routine use of mRNA profiling in casework laboratories. We first describe a model for gene expression in a sample from a single body fluid and then extend that model to mixtures of body fluids. From there we describe calculation of maximum likelihood estimates (MLEs) of body fluid quantities in a sample, and we describe the use of likelihood ratios (LR) to test for the presence of each body fluid in a sample. In contrast to most gene expression-based classifiers, we do not train a machine learning algorithm to optimize our ability to call samples correctly; rather, we define a biologically reasonable model of gene expression in body fluid samples and we use that model to evaluate the strength of evidence a sample provides for the presence of a particular fluid. This founding of our algorithm in sound statistical principles allows the calculation of log-likelihoods for detection of each fluid type, making the algorithm's results defensible in courtroom settings.

A further benefit of this principled approach is that it allows us to evaluate our algorithm on all samples, including those used in training: as our algorithm is based on an *a priori* model of gene expression in body fluid mixtures, and as we estimated its parameters without regard to model performance, the algorithm can only minimally overfit the training data. Our algorithm's performance in the training samples may therefore slightly overestimate its performance in future samples, while its performance

in the other, low-RNA samples will considerably underestimate future performance in high-quality samples. Although we initially used an LR of 100 as the decision threshold for all body fluid types, we subsequently demonstrated that it may be possible to use a less restrictive threshold to improve the positive call rate without generating false positives. Alternative approaches using body fluid-specific thresholds should be investigated.

While the prototype biomarker 'CodeSet' performed remarkably well in the work described herein, further optimization of the biomarkers will be required before the method can be used in casework. The HBB blood biomarker is approximately 1000-fold more highly expressed than ALAS2, the second-most prevalent blood marker in our data. This means that HBB's limit of detection (LOD) is so low that the possibility of false positives with non-blood body fluids increases due to possible low level contamination with vascular tissue products. This potentially confounding issue can be addressed by attenuating the HBB signal with the addition of precisely defined quantities of specifically designed unlabeled oligonucleotides complementary to the HBB RNA prior to hybridization with the full CodeSet. These competitively inhibit the hybridization reaction with the labeled probes.

In contrast to the need to attenuate one of the blood biomarkers, the signal for the saliva biomarkers needs to be enhanced. The most specific and highly expressed saliva biomarker is HTN3. Signal intensification could be accomplished by designing multiple probes that bind along a single HTN3 mRNA. In addition the current probes could be designed to hybridize to both HTN3 and HTN1, the latter of which is also saliva specific.

Alternative novel biomarkers identified by RNA-Seq studies could also be employed if the HTN3 intensification strategies fall short of expectations.

Some of the selected biomarkers did not perform as expected. For example, the ANK1 blood biomarker did not demonstrate blood specificity in the NanoString® assay with this sample set since the expression level was low in all tissues. Re-design of some probe sequences may be worthwhile, but it is likely that assay performance would be most significantly improved by the incorporation of additional body fluid specific biomarkers (e.g. commensal bacteria from the vagina, such as *Lactobacillus sp.*). Future iterations of the CodeSet will evaluate the performance of additional genes.

As a preliminary indication of the ability of the method to discern admixtures of body fluids, one ternary and four binary mixtures were prepared. The true fluid composition in four of the five mixtures was clear from their likelihood ratio profiles, and at least one fluid was correctly detected in all mixtures. Although these results were encouraging, a thorough investigation of the performance of a more optimized NanoString® assay with a variety of different mixtures will be necessary.

There needs to be a note of caution with respect to the skin assay results. The chosen skin biomarkers were selected using total skin RNA from commercial sources due to the difficulties in isolating sufficient quantities of total RNA from touch samples to perform the hundreds of assays required for the biomarker screening and confirmation process. It is likely that the highly purified commercial skin samples will contain mRNAs that originate from multiple layers of skin including both dermal and epidermal tissue as well as contaminating endothelial tissue and its contents (i.e. blood), and it is likely that *bona fide* touch samples, which presumably mainly consist of cortical cells from the

epidermis, will possess a different gene expression profile than that obtained from the commercial product. Some of the putative skin biomarkers were found in some of the other tissues, especially saliva (CCL27, LCE2D, IL1F7, KRT9), a finding perhaps due to common biomarker functions in skin and the alimentary tract or to the presence of skin cells in saliva. The highly expressed blood marker HBB was present in the commercial skin RNA preparations at comparable or higher levels than the highly expressed skin biomarker LCE1C, confirming the presence of contaminating endothelial tissue. In light of the extremely low abundance of tissue in most touch skin samples, it remains to be seen the degree to which skin biomarkers prove generally useful in forensic investigations. We suspect the inclusion of skin-specific genes will at a minimum help forensic assays avoid misclassification of skin samples as other tissues.

Housekeeping genes are typically added to gene expression assays to indicate that RNA of sufficient quality and quantity for analysis is present, and for normalization purposes [6,15,38]. Due to non-uniform expression of housekeeping genes their value as normalizers is questionable [48,49]. Here we show that the developed algorithm does not require normalization with housekeeping genes. However their presence indicates the recovery of suitable RNA for analysis and therefore still has a certain utility in the assay.

Acknowledgements

The authors would like to acknowledge all of the anonymous donors who provided samples for this study. Support for portions of this work was provided by the State of Florida through the National Center for Forensic Science at the University of Central Florida. The opinions, findings and conclusions or recommendations expressed in

this publication are those of the authors and do not necessarily reflect those of the State of Florida.

References

[1] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology, Elsevier/Academic Press, San Diego, CA, 2012.

[2] R. Cook, I. Evett, G. Jackson, P. Jone, A. Lambert, A hierarchy of propositions: deciding which level to address in casework, Science & Justice. 38 (1998) 231-239.

[3] J. Juusola, J. Ballantyne, Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification, Forensic Sci Int. 135 (2003) 85-96.

[4] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, Molecular Biology of the Cell,2nd, Garland Publishing, New York, NY, 1994.

[5] C. Haas, E. Hanson, J. Ballantyne, Capillary electrophoresis of a multiplex reverse transcription-polymerase chain reaction to target messenger RNA markers for body fluid identification, Methods Mol.Biol. 830 (2012) 169-183.

[6] E. Hanson, J. Ballantyne, RNA Profiling for the Identification of the Tissue Origin of Dried Stains in Forenic Biology, Forensic Sci Rev. 22 (2010) 145-157.

[7] C. Haas, B. Klesser, C. Maake, W. Bar, A. Kratzer, mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR, Forensic Sci Int Genet. 3 (2009) 80-88.

[8] M. Setzer, J. Juusola, J. Ballantyne, Recovery and stability of RNA in vaginal swabs and blood, semen, and saliva stains, J Forensic Sci. 53 (2008) 296-305.

[9] D. Zubakov, E. Hanekamp, M. Kokshoorn, I.W. van, M. Kayser, Stable RNA markers for identification of blood and saliva stains revealed from whole genome expression analysis of time-wise degraded samples, Int.J.Legal Med. 122 (2008) 135-142.

[10] D. Zubakov, M. Kokshoorn, A. Kloosterman, M. Kayser, New markers for old stains: stable mRNA markers for blood and saliva identification from up to 16-year-old stains, Int J.Legal Med. 123 (2009) 71-74.

[11] C. Haas, E. Hanson, W. Bar, R. Banemann, A.M. Bento, A. Berti, E. Borges, C. Bouakaze, A. Carracedo, M. Carvalho, A. Choma, M. Dotsch, M. Duriancikova, P. Hoff-Olsen, C. Hohoff, P. Johansen, P.A. Lindenbergh, B. Loddenkotter, B. Ludes, O. Maronas, N. Morling, H. Niederstatter, W. Parson, G. Patel, C. Popielarz, E. Salata, P.M. Schneider, T. Sijen, B. Sviezena, L. Zatkalikova, J. Ballantyne, mRNA profiling for the identification of blood--results of a collaborative EDNAP exercise, Forensic Sci Int Genet. 5 (2011) 21-26.

[12] C. Haas, E. Hanson, N. Morling, J. Ballantyne, Collaborative EDNAP exercises on messenger RNA/DNA co-analyis for body fluid identification (blood, saliva, semen) and STR profiling, Forensic Sci.Int.Genet.Supp.Ser. 3 (2011) e5-e6.

[13] C. Haas, E. Hanson, M.J. Anjos, W. Bar, R. Banemann, A. Berti, E. Borges, C. Bouakaze, A. Carracedo, M. Carvalho, V. Castella, A. Choma, C.G. De, M. Dotsch, P. Hoff-Olsen, P. Johansen, F. Kohlmeier, P.A. Lindenbergh, B. Ludes, O. Maronas, D. Moore, M.L. Morerod, N. Morling, H. Niederstatter, F. Noel, W. Parson, G. Patel, C. Popielarz, E. Salata, P.M. Schneider, T. Sijen, B. Sviezena, M. Turanska, L. Zatkalikova, J. Ballantyne, RNA/DNA co-analysis from blood stains--results of a second collaborative EDNAP exercise, Forensic Sci Int Genet. 6 (2012) 70-80.

[14] C. Haas, E. Hanson, M.J. Anjos, R. Banemann, A. Berti, E. Borges, A. Carracedo, M. Carvalho, C. Courts, C.G. De, M. Dotsch, S. Flynn, I. Gomes, C. Hollard, B. Hjort, P. Hoff-Olsen, K. Hribikova, A. Lindenbergh, B. Ludes, O. Maronas, N. McCallum, D. Moore, N. Morling, H. Niederstatter, F. Noel, W. Parson, C. Popielarz, C. Rapone, A.D. Roeder, Y. Ruiz, E. Sauer, P.M. Schneider, T. Sijen, Court DS, B. Sviezena, M. Turanska, A. Vidaki, L. Zatkalikova, J. Ballantyne, RNA/DNA co-analysis from human saliva and semen stains--results of a third collaborative EDNAP exercise, Forensic Sci Int Genet. 7 (2013) 230-239.

[15] C. Haas, E. Hanson, M.J. Anjos, K.N. Ballantyne, R. Banemann, B. Bhoelai, E. Borges, M. Carvalho, C. Courts, C.G. De, K. Drobnic, M. Dotsch, R. Fleming, C. Franchi, I. Gomes, G. Hadzic, S.A. Harbison, J. Harteveld, B. Hjort, C. Hollard, P. Hoff-Olsen, C. Huls, C. Keyser, O. Maronas, N. McCallum, D. Moore, N. Morling, H. Niederstatter, F. Noel, W. Parson, C. Phillips, C. Popielarz, A.D. Roeder, L. Salvaderi, E. Sauer, P.M. Schneider, G. Shanthan, Court DS, M. Turanska, R.A. van Oorschot, M. Vennemann, A. Vidaki, L. Zatkalikova, J. Ballantyne, RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: results of a fourth and fifth collaborative EDNAP exercise, Forensic Sci Int Genet. 8 (2014) 203-212.

[16] C. Courts, B. Madea, Specific micro-RNA signatures for the detection of saliva and blood in forensic body-fluid identification, J.Forensic Sci. 56 (2011) 1464-1470.

[17] E. Hanson, K. Rekab, J. Ballantyne, Binary logistic regression models enable miRNA profiling to provide accurate identification of forensically relevant body fluids and tissues, For Sci Int Genet Supp Ser. 4 (2013) e127-e128.

[18] E. Hanson, H. Lubenow, J. Ballantyne, Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs, Forensic Sci.Int.Genet. Supplement Series 2 (2009) 503-504.

[19] E.K. Hanson, H. Lubenow, J. Ballantyne, Identification of Forensically Relevant Body Fluids Using a Panel of Differentially Expressed microRNAs, Anal.Biochem. 387 (2009) 303-314.

[20] Z. Wang, H. Luo, X. Pan, M. Liao, Y. Hou, A model for data analysis of microRNA expression in forensic body fluid identification, Forensic Sci.Int.Genet. 6 (2012) 419-423.

[21] Z. Wang, J. Zhang, H. Luo, Y. Ye, J. Yan, Y. Hou, Screening and confirmation of microRNA markers for forensic body fluid identification, Forensic Sci.Int.Genet. 7 (2013) 116-123.

[22] D. Zubakov, A.W. Boersma, Y. Choi, P.F. van Kuijk, E.A. Wiemer, M. Kayser, MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation, Int J.Legal Med. 124 (2010) 217-226.

[23] J.H. An, A. Choi, K.J. Shin, W.I. Yang, H.Y. Lee, DNA methylation-specific multiplex assays for body fluid identification, Int.J.Legal Med. 127 (2013) 35-43.

[24] A. Choi, K.J. Shin, W.I. Yang, H.Y. Lee, Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA, Int J.Legal Med. 128 (2014) 33-41.

[25] D. Frumkin, A. Wasserstrom, B. Budowle, A. Davidson, DNA methylation-based forensic tissue identification, Forensic Sci.Int.Genet. 5 (2011) 517-524.

[26] B.L. LaRue, J.L. King, B. Budowle, A validation study of the Nucleix DSI-Semen kit--a methylation-based assay for semen identification, Int.J.Legal Med. 127 (2013) 299-308.

[27] H.Y. Lee, M.J. Park, A. Choi, J.H. An, W.I. Yang, K.J. Shin, Potential forensic application of DNA methylation profiling to body fluid identification, Int.J.Legal Med. 126 (2012) 55-62.

[28] T. Madi, K. Balamurugan, R. Bombardi, G. Duncan, B. McCord, The determination of tissue-specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing, Electrophoresis. 33 (2012) 1736-1745.

[29] A. Wasserstrom, D. Frumkin, A. Davidson, M. Shpitzen, Y. Herman, R. Gafny, Demonstration of DSI-semen--A novel DNA methylation-based forensic semen identification assay, Forensic Sci.Int.Genet. 7 (2013) 136-142.

[30] J.L. Simons, S.K. Vintiner, Efficacy of several candidate protein biomarkers in the differentiation of vaginal from buccal epithelial cells, J.Forensic Sci. 57 (2012) 1585-1590.

[31] S.K. Van, C.M. De, M. Dhaenens, H.D. Van, D. Deforce, Mass spectrometry-based proteomics as a tool to identify biological matrices in forensic science, Int.J.Legal Med. 127 (2013) 287-298.

[32] H. Yang, B. Zhou, M. Prinz, D. Siegel, Proteomic analysis of menstrual blood, Mol.Cell Proteomics. 11 (2012) 1024-1035.

[33] E. Hanson, C. Haas, R. Jucker, J. Ballantyne, Specific and sensitive mRNA biomarkers for the identification of skin in 'touch DNA' evidence, Forensic Sci Int Genet. 6 (2012) 548-558.

[34] J. Juusola, J. Ballantyne, Multiplex mRNA profiling for the identification of body fluids, Forensic Sci Int. 152 (2005) 1-12.

[35] M.L. Richard, K.A. Harper, R.L. Craig, A.J. Onorato, J.M. Robertson, J. Donfack, Evaluation of mRNA marker specificity for the identification of five human body fluids by capillary electrophoresis, Forensic Sci Int Genet. 6 (2012) 452-460.

[36] A.D. Roeder, C. Haas, mRNA profiling using a minimum of five mRNA markers per body fluid and a novel scoring method for body fluid identification, Int J Legal Med. 127 (2013) 707-721.

[37] M. Bauer, D. Patzelt, Identification of menstrual blood by real time RT-PCR: technical improvements and the practical value of negative test results, Forensic Sci Int. 174 (2008) 55-59.

[38] J. Juusola, J. Ballantyne, mRNA profiling for body fluid identification by multiplex quantitative RT-PCR, J Forensic Sci. 52 (2007) 1252-1262.

[39] C. Nussbaumer, E. Gharehbaghi-Schnell, I. Korschineck, Messenger RNA profiling: a novel method for body fluid identification by real-time PCR, Forensic Sci Int. 157 (2006) 181-186.

[40] E.K. Hanson, J. Ballantyne, Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis, F1000Res. 2 (2013) 281.

[41] S. Audic, J.M. Claverie, The significance of digital gene expression profiles, Genome Res. 7 (1997) 986-995.

[42] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat.Rev.Genet. 10 (2009) 57-63.

[43] G.K. Geiss, R.E. Bumgarner, B. Birditt, T. Dahl, N. Dowidar, D.L. Dunaway, H.P. Fell, S. Ferree, R.D. George, T. Grogan, J.J. James, M. Maysuria, J.D. Mitton, P. Oliveri, J.L. Osborn, T. Peng, A.L. Ratcliffe, P.J. Webster, E.H. Davidson, L. Hood, K. Dimitrov, Direct multiplexed measurement of gene expression with color-coded probe pairs, Nat.Biotechnol. 26 (2008) 317-325.

[44] E.K. Hanson, J. Ballantyne, "Getting blood from a stone": ultrasensitive forensic DNA profiling of microscopic bio-particles recovered from "touch DNA" evidence, Methods Mol.Biol. 1039 (2013) 3-17.

[45] E.K. Hanson, J. Ballantyne, Highly specific mRNA biomarkers for the identification of vaginal secretions in sexual assault investigations, Sci Justice. 53 (2013) 14-22.

[46] E. Hanson, C. Haas, R. Jucker, J. Ballantyne, Identification of skin in touch/contact forensic samples by messenger RNA profiling, Forensic Sci Int Genet.Suppl Series. 3 (2011) e305-e306.

[47] R.H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM J.Scientific Computing.1995) 1190-1208.

[48] L.I. Moreno, C.M. Tate, E.L. Knott, J.E. McDaniel, S.S. Rogers, B.W. Koons, M.F. Kavlick, R.L. Craig, J.M. Robertson, Determination of an effective housekeeping gene for the quantification of mRNA for forensic applications, J.Forensic Sci. 57 (2012) 1051-1058.

[49] J. Vandesompele, P.K. De, F. Pattyn, B. Poppe, R.N. Van, P.A. De, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Genome Biol. 3 (2002) RESEARCH0034.

Table 1 List of Samples Tested

Sample Type	N	Description
Blood	***14***	
Organic Extraction	7	Blood stain on cotton cloth (-47°C storage after drying)
	1	Environmental (outside (FL) – heat, sunlight, humidity, rain (1 month)
	1	Environmental (outside (FL) – heat, sunlight, humidity, covered (3 days)
Direct Lysis (RLT)	5	Blood stain on cotton cloth (-47°C storage after drying)
Semen	***17***	
Organic Extraction	7	Dried on cotton swabs (-47°C storage after drying)
	2	Environmental (outside (FL) – heat, sunlight, humidity, covered (1 week)
	3	Sensitivity: 25ng, 12.5ng, 6.25ng (input achieved by use of 5µl of extract)
Direct Lysis (RLT)	5	Dried on cotton swabs (-47°C storage after drying)
Saliva	***17***	
Organic Extraction	7	Dried buccal sample on cotton swabs (-47°C storage after drying)
	1	Environmental (outside (FL) – heat, sunlight, humidity, rain (1 week)
	1	Environmental (outside (FL) – heat, sunlight, humidity, covered (1 month)
	3	Sensitivity: 25ng, 12.5ng, 6.25ng (input achieved by use of 5µl of extract)
Direct Lysis (RLT)	5	Dried buccal sample on cotton swabs (-47°C storage after drying)
Vaginal Secretions	***10***	
Organic Extraction	6	Dried sample on cotton swabs (-47°C storage after drying)
	1	Environmental (outside (FL) – heat, sunlight, humidity, rain (3 days)
Direct Lysis (RLT)	3	Dried sample on cotton swabs (-47°C storage after drying)
Menstrual Blood	***10***	
Organic Extraction	7	Dried sample on cotton swabs (-47°C storage after drying)
Direct Lysis (RLT)	3	Dried sample on cotton swabs (-47°C storage after drying)
Skin	***14***	
Organic Extraction	1	Swab of surface skin (male hand); swab moistened with sterile water
	1	Swab of coffee cup surface; swab moistened with sterile water
	1	Swab of computer mouse; swab moistened with sterile water
Direct Lysis (RLT)	1	Swab of surface skin (male hand); swab moistened with sterile water
	1	Swab of coffee cup surface; swab moistened with sterile water
	1	Swab of computer mouse; swab moistened with sterile water
Direct Lysis (*RNA*GEM)	1	25 bio-particles (clumps); shirt collar (male)
	1	50 bio-particles (clumps); shirt collar (male)
Direct Lysis (*forensic*GEM)	1	100 bio-particles (55 clumps/45 singles); shirt collar (male)
None	5	Skin total RNA (commercial source)
Mixtures	***5***	
Organic Extraction	2	Vaginal/semen (1/2 swab of each donor extracted in same tube)
	2	Blood/saliva (1/2 stain/swab of each donor extracted in same tube)
	1	Semen/saliva/vaginal (1/2 swab of each donor extracted in same tube)
Controls	***3***	
Organic Extraction	2	Clean sterile swab (negative control)
None	1	Brain total RNA* (commercial source)

Stain = 50 µl stain; Swab – saturated body fluid swab (sterile cotton)
Environmental samples (blood, semen, saliva) – on cotton cloth
Total RNA – commercial sources (see methods)
***** run as an internal positive control and not used in any data analysis

Table 2. Body Fluid Specific and Housekeeping Genes in the NanoString® Custom CodeSet

Gene	*Body Fluid Target*
ALAS2	Blood
ANK1	Blood
HBB	Blood
LEFTY2	Menstrual Blood
MMP10	Menstrual Blood
HTN3	Saliva
MUC7	Saliva
STATH	Saliva
PRM2	Semen
SEMG1	Semen
TGM4	Semen
CCL27	skin
IL1F7	skin
KRT9	skin
LCE1C	skin
LCE2D	skin
CYP2A7	vaginal
CYP2B7P1	vaginal
DKK4	vaginal
FUT6	vaginal
IL19	vaginal
MYOZ1	vaginal
NOXO1	vaginal
B2M	Housekeeping Gene
COX1	Housekeeping Gene
HPRT1	Housekeeping Gene
PGK1	Housekeeping Gene
PPIH	Housekeeping Gene
S15	Housekeeping Gene
TCEA1	Housekeeping Gene
TFRC	Housekeeping Gene
UBC	Housekeeping Gene
UBE2D2	Housekeeping Gene

Figure Legend

Figure 1. NanoString® digital gene expression technology

Figure 2. Average proportion of total expression for each gene in each fluid. The vertical axis shows the relative proportion of total gene expression attributable to each gene (on the log scale). For each fluid, each point shows a gene's relative expression in a single training sample, and each bar shows the average of the gene's relative expression across the fluid's training samples. Bar color indicates genes' putative tissues.

Figure 3. Performance of the algorithm on all single-source samples. Bars display the rate at which each fluid is called detected in each sample type. Fluids are called detected if their likelihood ratio exceeds 100.

Figure 4. ROC curves showing the algorithm's True Positive Rate (TPR) and False Positive Rate (FPR) for each tissue. Points indicate the performance achieved using a LR cutoff of 100. Relaxing this LR cutoff for detection of menstrual blood, saliva and skin could greatly increase the TPR without increasing the FPR. Line color indicates body fluid: blood – red, semen – blue, saliva – green, vaginal – orange, menstrual blood – pink, skin – purple.

Figure 5. Performance of the algorithm in five mixture samples. For each of five mixture samples, a bar plot shows the likelihood ratios for the presence of each fluid type**.** The dotted line indicates a LR of 100.

Figure 1.*http://www.nanostring.com/applications/technology*

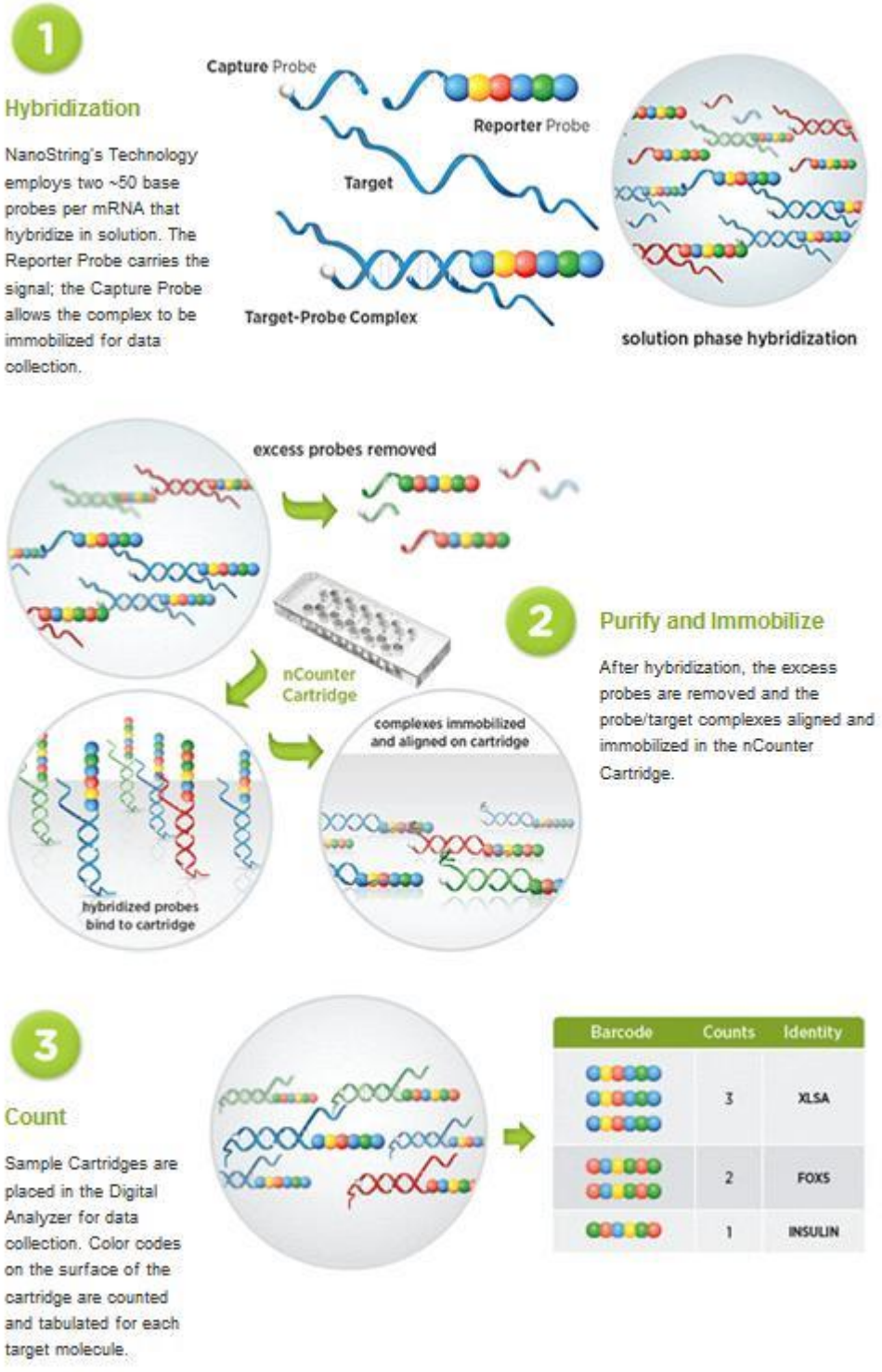


Figure 2.

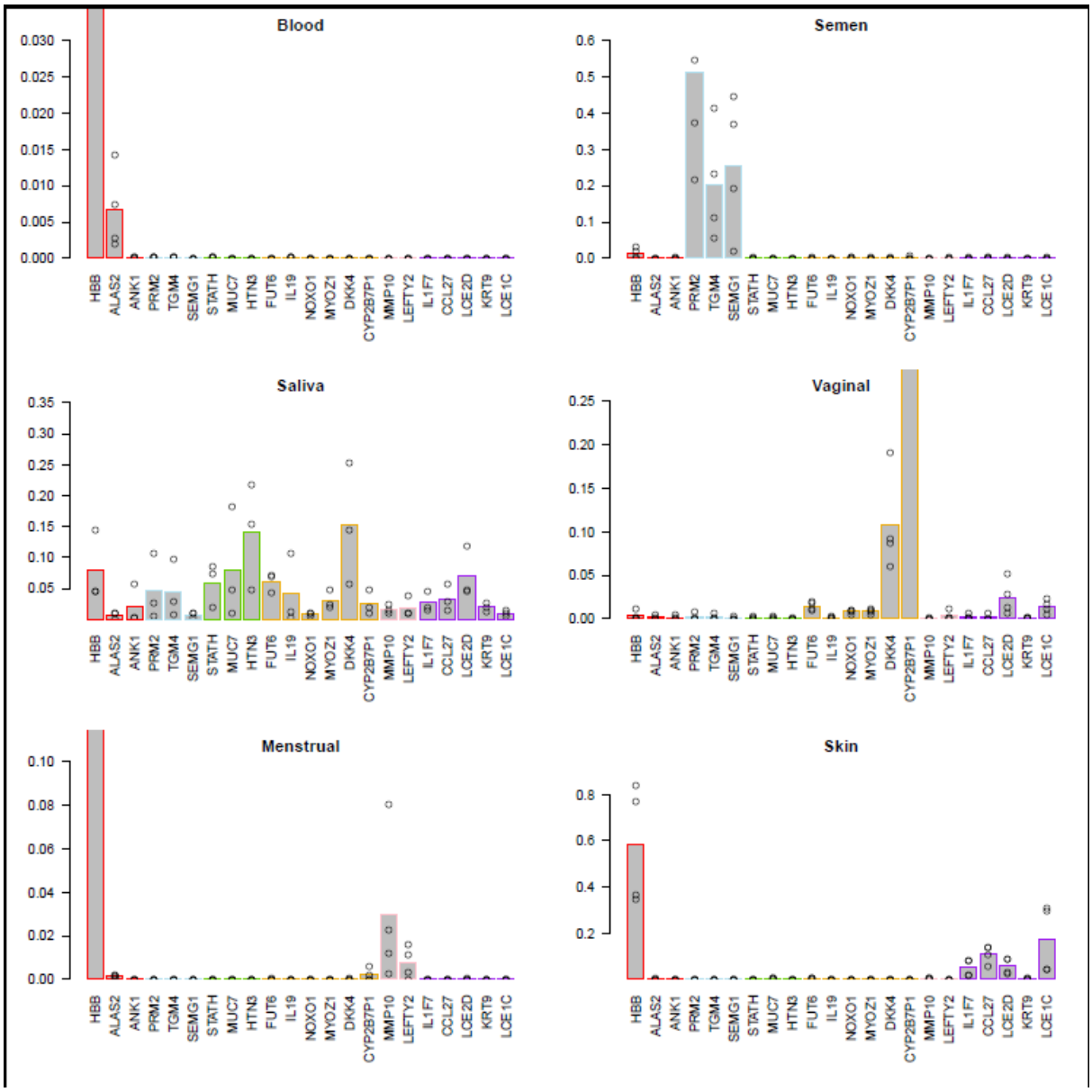


Figure 3.

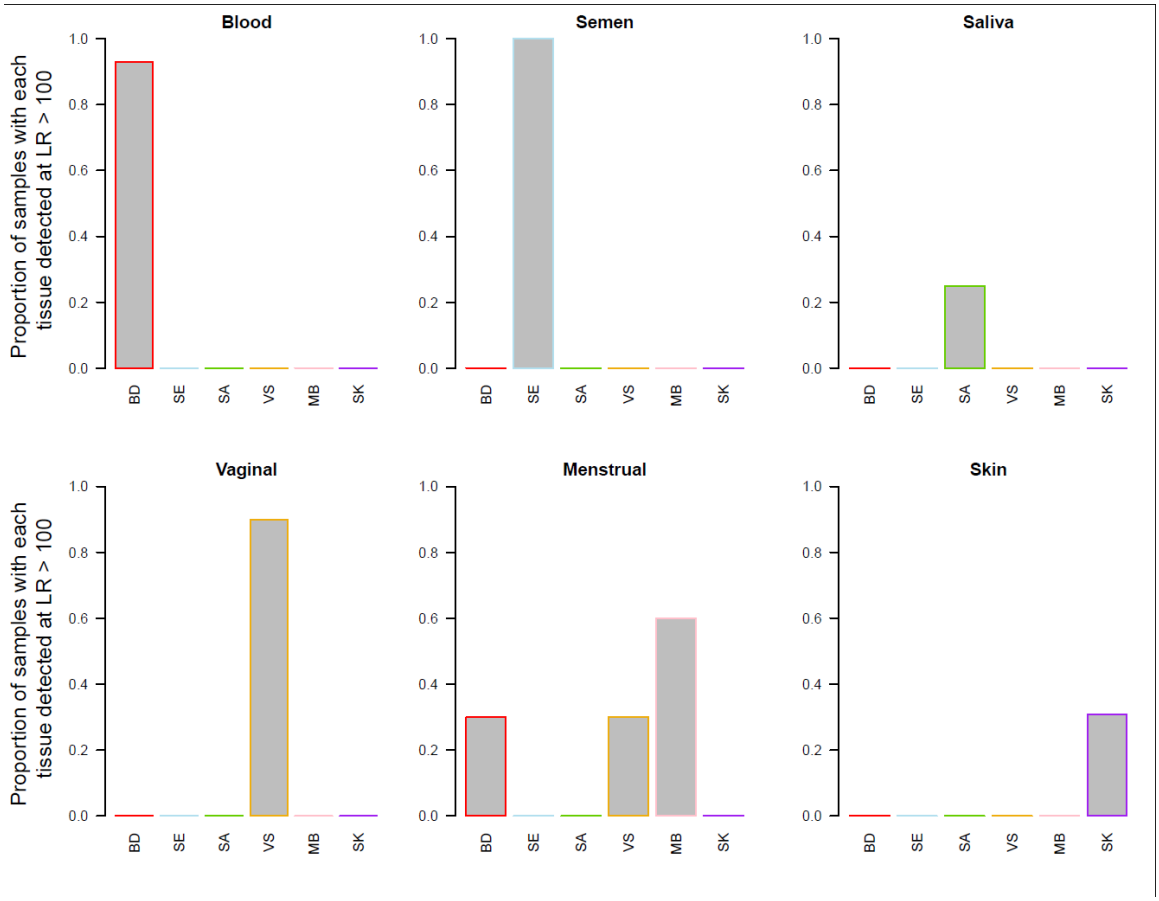


Figure 4.

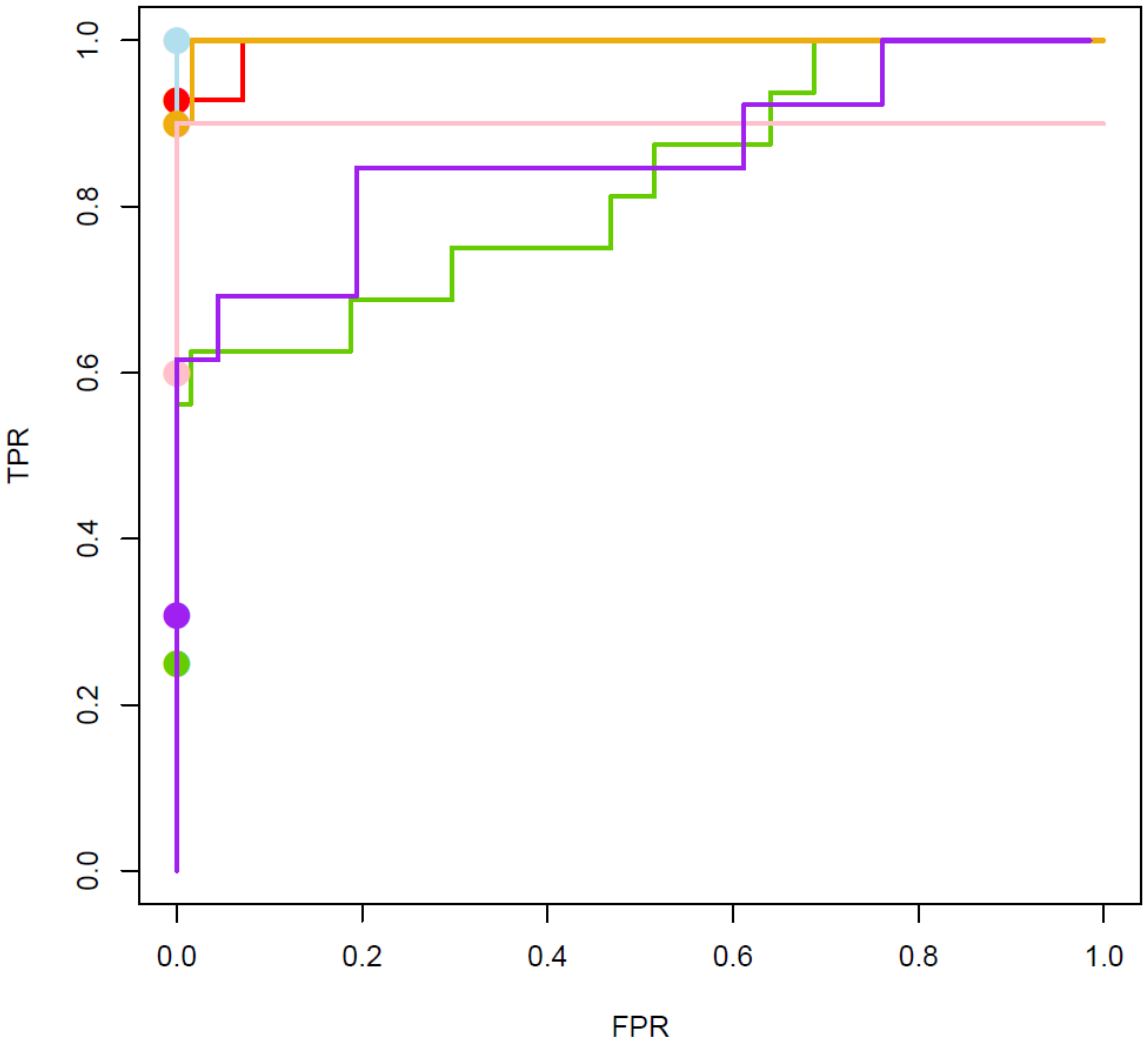
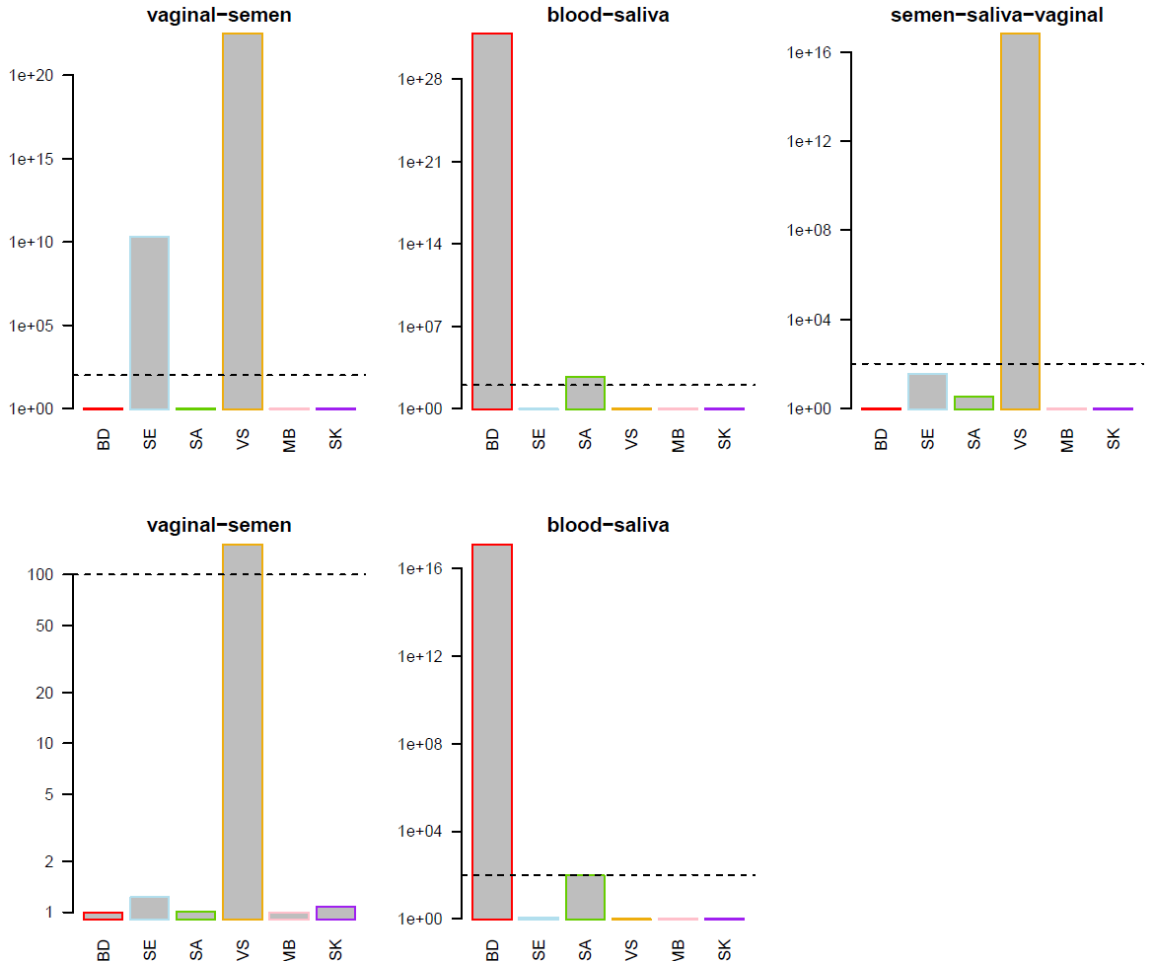


Figure 5.



Supplementary Table 1. Sample Descriptions and Assay Input (Full Sample Set)

Sample	Description	Extraction Type	Input (µl)	Input (ng)
1	50µl bloodstain on cotton cloth; donor 1	Standard	5 µl	50 ng
2	50µl bloodstain on cotton cloth; donor 2	Standard	5 µl	50 ng
3	50µl bloodstain on cotton cloth; donor 3	Standard	5 µl	50 ng
4	50µl bloodstain on cotton cloth; donor 4	Standard	5 µl	50 ng
5	Env. Bloodstain: outside, covered 3 day (donor 5)	Standard	5 µl	NA
6	50µl bloodstain on cotton cloth; donor 4	Direct Lysis (RLT)	5 µl	NA
7	Sat. semen swab (cotton, dried); donor 1	Standard	5 µl	50 ng
8	Sat. semen swab (cotton, dried); donor 2	Standard	5 µl	50 ng
9	Sat. semen swab (cotton, dried); donor 3	Standard	5 µl	50 ng
10	Sat. semen swab (cotton, dried); donor 4	Standard	5 µl	50 ng
11	Env: 50µl semen on cotton cloth: outside, covered 1 week (donor 5)	Standard	5 µl	NA
12	½ Sat. semen swab (cotton, dried); donor 1	Direct Lysis (RLT)	5 µl	NA
13	Buccal swab (cotton, dried); donor 1	Standard	5 µl	50 ng
14	Buccal swab (cotton, dried); donor 2	Standard	5 µl	50 ng
15	Buccal swab (cotton, dried); donor 3	Standard	5 µl	50 ng
16	Buccal swab (cotton, dried); donor 4	Standard	5 µl	50 ng
17	Env: 50µl saliva on cotton cloth: outside, covered 1 month (donor 5)	Standard	5 µl	50 ng
18	½ buccal swab (cotton, dried); donor 6	Direct Lysis (RLT)	5 µl	NA
19	½ Vaginal swab (cotton, dried); donor 1	Standard	5 µl	50 ng
20	½ Vaginal swab (cotton, dried); donor 2	Standard	5 µl	50 ng
21	½ Vaginal swab (cotton, dried); donor 3	Standard	5 µl	50 ng
22	½ Vaginal swab (cotton, dried); donor 4	Standard	5 µl	50 ng
23	Env: ½ vaginal swab: outside, uncovered 3 days (donor 5)	Standard	5 µl	50 ng
24	½ Vaginal swab (cotton, dried); donor 2	Direct Lysis (RLT)	5 µl	NA
25	½ menstrual blood swab (cotton; dried) donor 1, day 2 of menstruation	Standard	5 µl	50 ng
26	½ menstrual blood swab (cotton; dried) donor 2	Standard	5 µl	50 ng
27	½ menstrual blood swab (cotton; dried) donor 3, day 1 of menstruation	Standard	5 µl	50 ng
28	½ menstrual blood swab (cotton; dried) donor 4, day 2 of menstruation	Standard	5 µl	50 ng
29	½ menstrual blood swab (cotton; dried) donor 5, Day 3 of menstruation	Standard	5 µl	50 ng
30	½ menstrual blood swab (cotton; dried) donor 1	Direct Lysis (RLT)	5 µl	NA
31	Skin – total RNA (commercial source)	None	5 µl	50 ng
32	Skin – total RNA (commercial source)	None	5 µl	50 ng
33	Skin – total RNA (commercial source)	None	5 µl	50 ng

34	Skin – total RNA (commercial source)	None	5 µl	50 ng
35	Surface swab (whole) of computer mouse	Standard	5 µl	NA
36	Surface swab (whole) of computer mouse	Direct Lysis (RLT)	5 µl	NA
37	Semen (donor 2) – dilution series	Standard	5 µl	25 ng
38	Semen (donor 2) – dilution series	Standard	5 µl	12.5 ng
39	Semen (donor 2) – dilution series	Standard	5 µl	6.25 ng
40	Saliva (donor 1) – dilution series	Standard	5 µl	25 ng
41	Saliva (donor 1) – dilution series	Standard	5 µl	12.5 ng
42	Saliva (donor 1) – dilution series	Standard	5 µl	6.25 ng
43	Human Brain – total RNA (commercial source)	None	5 µl	50 ng
44	Extraction blank (blank/clean swab)	Standard	5 µl	NA
45	100 bio-particles (55 clumps/45 singles); male shirt collar	Direct Lysis (FG)	5 µl	NA
46	Vaginal (donor3) -semen (donor 1) mixture (1/2 swab of each)	Standard	5 µl	50 ng
47	Blood (donor 1) -saliva (donor 2) mixture (1/2 swab of each)	Standard	5 µl	50 ng
48	Semen (donor 1)-saliva (donor 2)-vaginal (donor 3) (1/2 swab of each)	Standard	5 µl	50 ng
49	½ 50µl bloodstain on cotton cloth; donor 6	Standard	10 µl	60 ng
50	½ 50µl bloodstain on cotton cloth; donor 6	Direct Lysis (RLT)	5 µl	NA
51	Technical replicate of #50	Direct Lysis (RLT)	10 µl	NA
52	½ 50µl bloodstain on cotton cloth; donor 7	Standard	8 µl	104 ng
53	½ 50µl bloodstain on cotton cloth; donor 7	Direct Lysis (RLT)	5 µl	NA
54	½ 50µl bloodstain on cotton cloth; donor 8	Direct Lysis (RLT)	5 µl	NA
55	½ 50µl bloodstain on cotton cloth; donor 8	Direct Lysis (RLT)	10 µl	NA
56	½ Sat. semen swab (cotton, dried); donor 6	Standard	4 µl	108 ng
57	½ Sat. semen swab (cotton, dried); donor 6	Direct Lysis (RLT)	5 µl	NA
58	½ Sat. semen swab (cotton, dried); donor 7	Standard	5.3 µl	101 ng
59	½ Sat. semen swab (cotton, dried); donor 7	Direct Lysis (RLT)	5 µl	NA
60	Technical replicate of #59	Direct Lysis (RLT)	10 µl	NA
61	½ Sat. semen swab (cotton, dried); donor 8	Direct Lysis (RLT)	5 µl	NA
62	½ Sat. semen swab (cotton, dried); donor 8	Direct Lysis (RLT)	10 µl	NA
65	½ fresh buccal swab (cotton); donor 8	Standard	10 µl	470 ng
66	½ fresh buccal swab (cotton); donor 8	Direct Lysis (RLT)	5 µl	NA
67	Technical replicate of #66	Direct Lysis (RLT)	10 µl	NA
68	½ fresh buccal swab (cotton); donor 9	Direct Lysis (RLT)	5 µl	NA
69	½ fresh buccal swab (cotton); donor 9	Direct Lysis (RLT)	10 µl	NA
70	½ fresh buccal swab (cotton); donor 9	Direct Lysis (RLT)	5 µl	NA
71	½ fresh buccal swab (cotton); donor 9	Direct Lysis (RLT)	10 µl	NA
72	½ vaginal swab (cotton; dried); donor 6	Standard	1 µl	332 ng
73	½ vaginal swab (cotton; dried); donor 6	Direct Lysis (RLT)	5 µl	NA
74	½ vaginal swab (cotton; dried); donor 7	Standard	1 µl	255 ng
75	½ vaginal swab (cotton; dried); donor 7	Direct Lysis (RLT)	5 µl	NA
76	½ menstrual blood swab (cotton; dried); donor 6, day 2 of menstruation	Standard	1 µl	118 ng

77	½ menstrual blood swab (cotton; dried); donor 6, day 2 of menstruation	Direct Lysis (RLT)	5 µl	NA
78	½ menstrual blood swab (cotton; dried); donor 7	Standard	3.6 µl	101 ng
79	½ menstrual blood swab (cotton; dried); donor 7	Direct Lysis (RLT)	5 µl	NA
80	Technical replicate of #79	Direct Lysis (RLT)	10 µl	NA
81	Swab of human skin (male hand, left)	Standard	10 µl	80 ng
82	Swab of human skin (male hand, right)	Direct Lysis (RLT)	5 µl	NA
83	Technical replicate of #88	Direct Lysis (RLT)	10 µl	NA
84	Swab of metal coffee cup surface (side 1)	Standard	8.3 µl	100 ng
85	Swab of metal coffee cup surface (side 2)	Direct Lysis (RLT)	5 µl	NA
86	Technical replicate of #85	Direct Lysis (RLT)	10 µl	NA
87	25 bio-particles (clumps); male shirt collar	Direct Lysis (RG)	5 µl	NA
88	50 bio-particles (clumps); male shirt collar	Direct Lysis (RG)	5 µl	NA
89	Env: 50µl semen on cotton cloth: outside, covered 1 week (donor 9)	Standard	1.3 µl	100 ng
90	50µl bloodstain on cotton cloth; donor 9	Standard	7.1 µl	99 ng
91	Vaginal (donor 4)-semen (donor 9) mixture (1/2 swab of each)	Standard	1.0 µl	164 ng
92	Env: 50µl saliva on cotton cloth: outside, covered 1 week (donor 10)	Standard	7.7 µl	100 ng
93	½ Sat. semen swab (cotton, dried); donor 10	Standard	4.3 µl	99 ng
94	blood (donor 10)-saliva (donor 7) mixture (1/2 swab of each)	Standard	2.0 µl	98 ng
95	Extraction blank (blank/clean swab)	Standard	5.0 µl	0 ng
96	dried buccal swab (cotton); donor 1	Standard	1.0 µl	133 ng
97	Env: 50µl blood on cotton cloth: outside, uncovered 1 month (donor 11)	Standard	2.0 µl	106 ng
98	Skin – total RNA (commercial source)	Standard	2.0 µl	100 ng

Env = environmental; direct lysis (FG) = *forensic*GEM™; direct Lysis (RG) = *RNA*GEM™

Supplementary Figure Legend

Supplementary Figure 1. Housekeeping gene expression in training and test samples

Supplementary Figure 2. Profiles of the training samples from each fluid are plotted against each other.

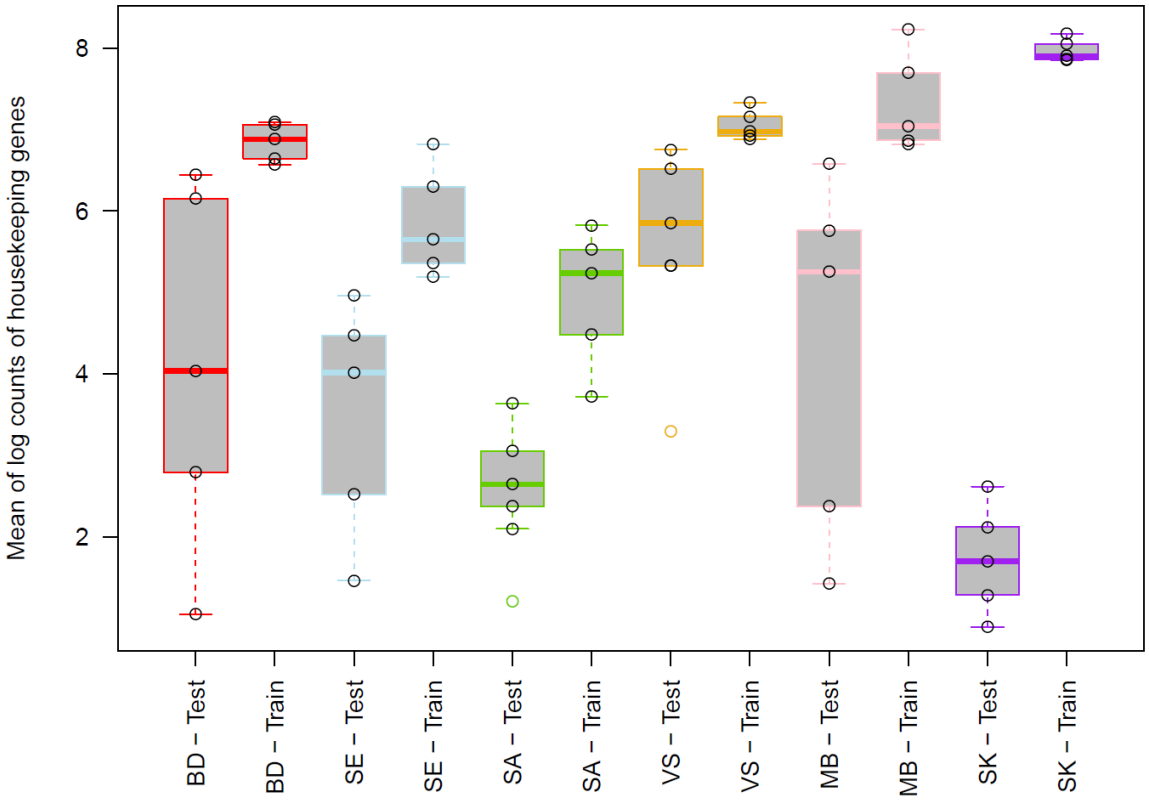
Supplemental Figure 3. Boxplots for individual genes' proportion of total expression in the different sample types.BD = blood, SE =semen, SA = saliva, MB = menstrual blood, VS = vaginal secretions, SK = skin.

Supplementary Figure 4. Performance of the algorithm in the training set. Bars display the rate at which each fluid is called detected in each sample type. Fluids are called detected if their likelihood ratio exceeds 100.

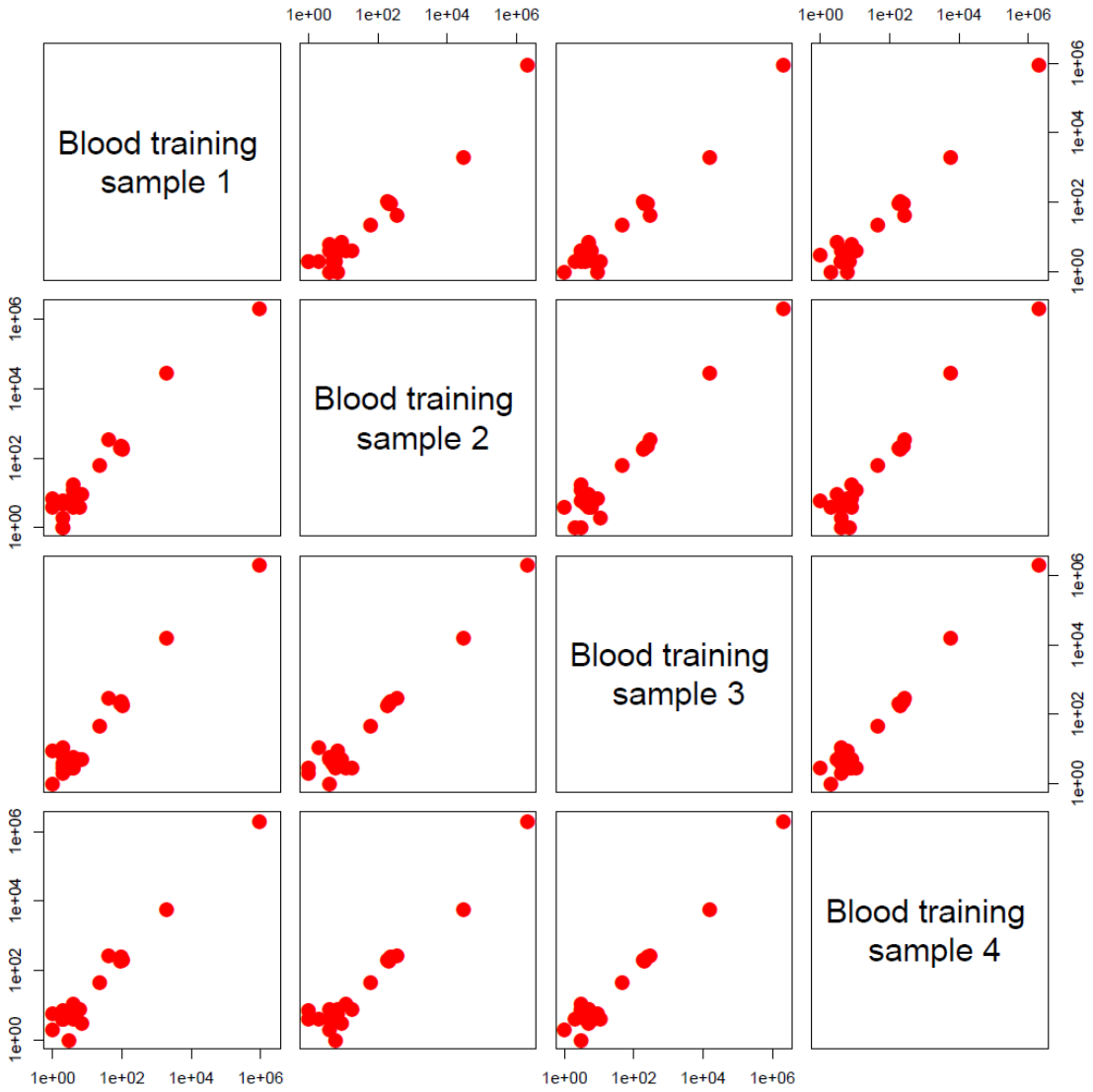
Supplementary Figure 5. Performance of the algorithm in the test set. Bars display the rate at which each fluid is called detected in each sample type. Fluids are called detected if their likelihood ratio exceeds 100.

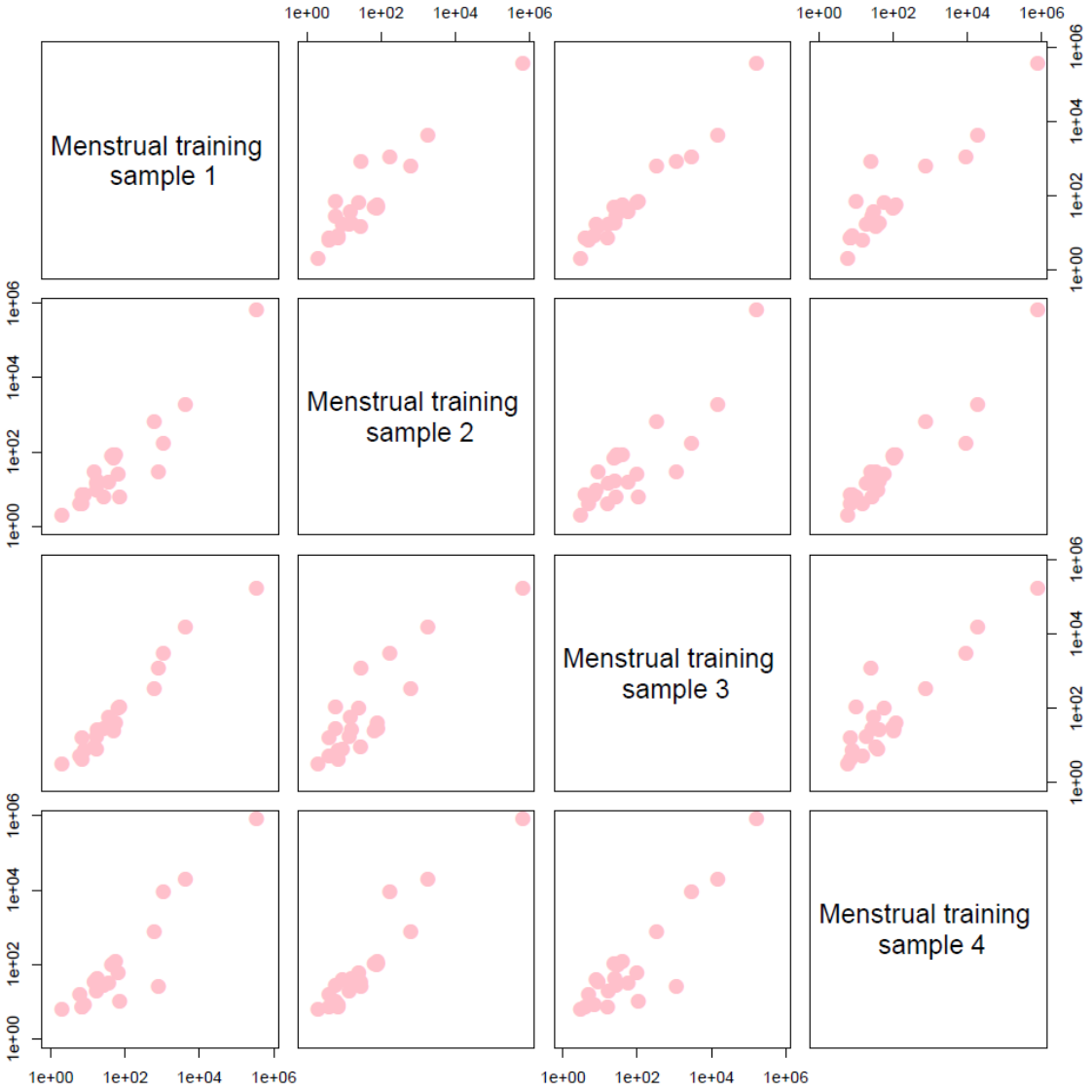
Supplementary Figure 6. Concordance of the assay between purification and lysis protocols. For the 14 samples with replicates run under each protocol, the natural log gene expression profile under the lysis protocol (vertical axis) is plotted against the profile under the purification protocol (horizontal axis).

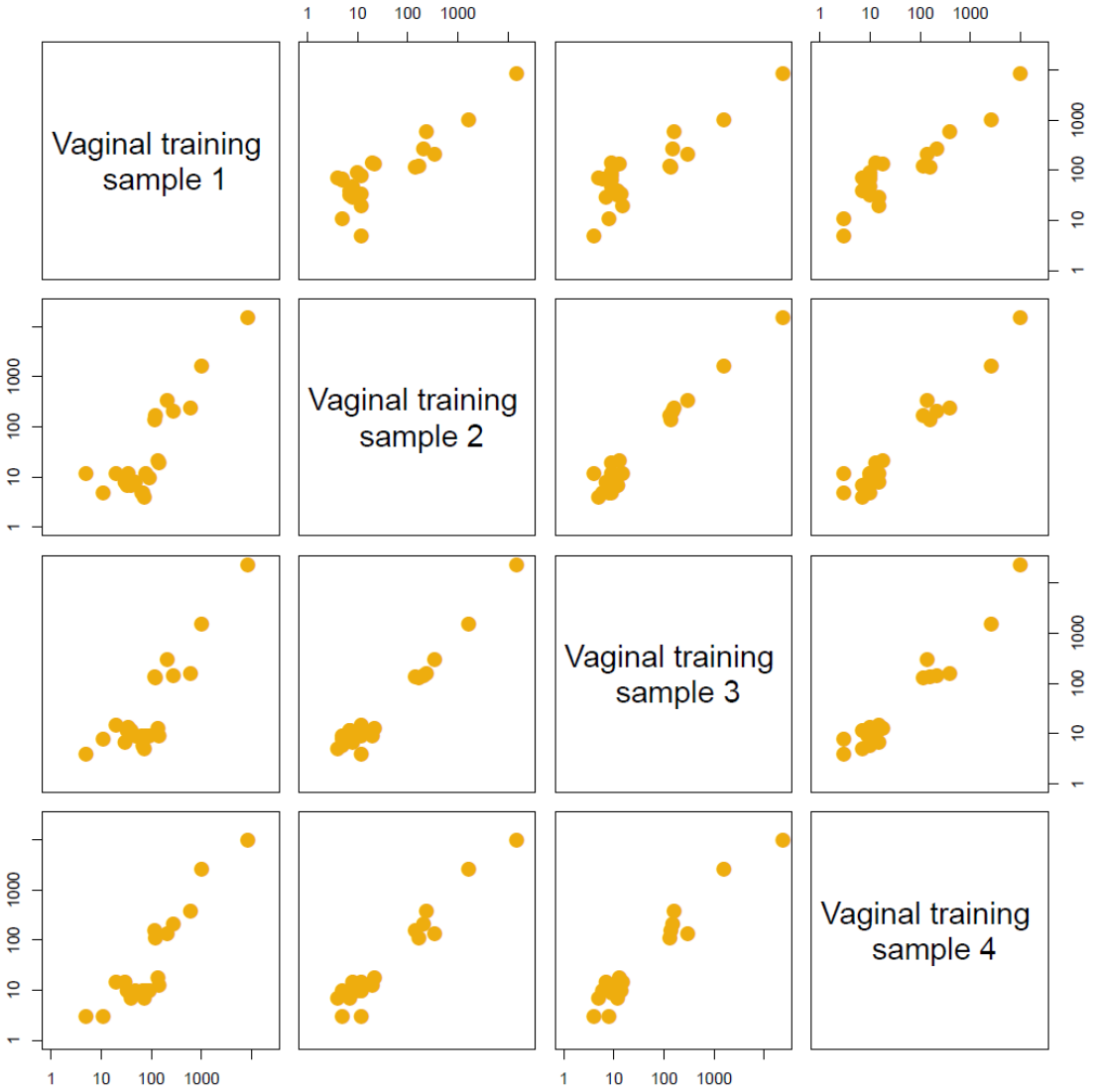
Supplementary Figure 1.

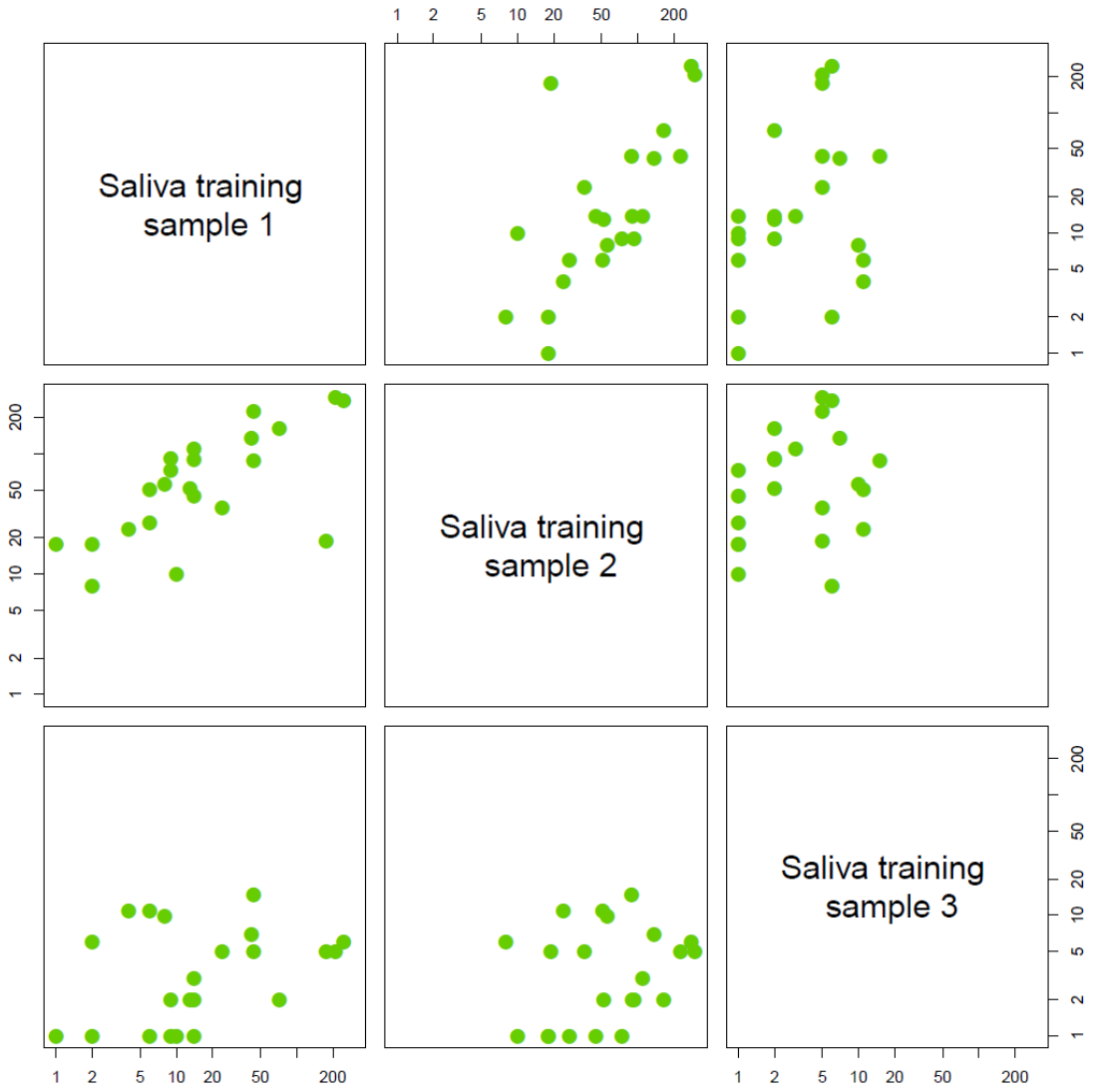


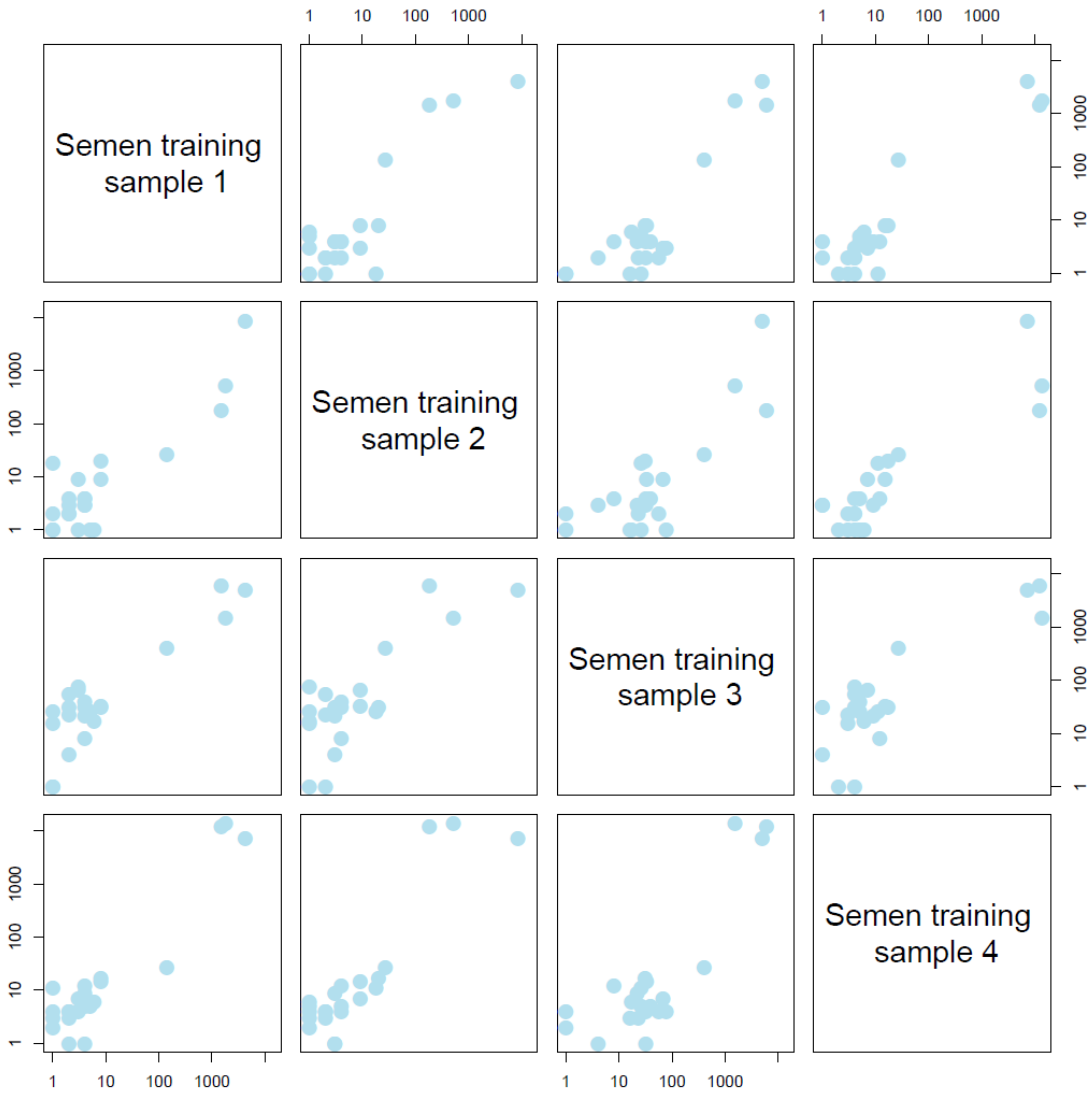
Supplementary Figure 2.

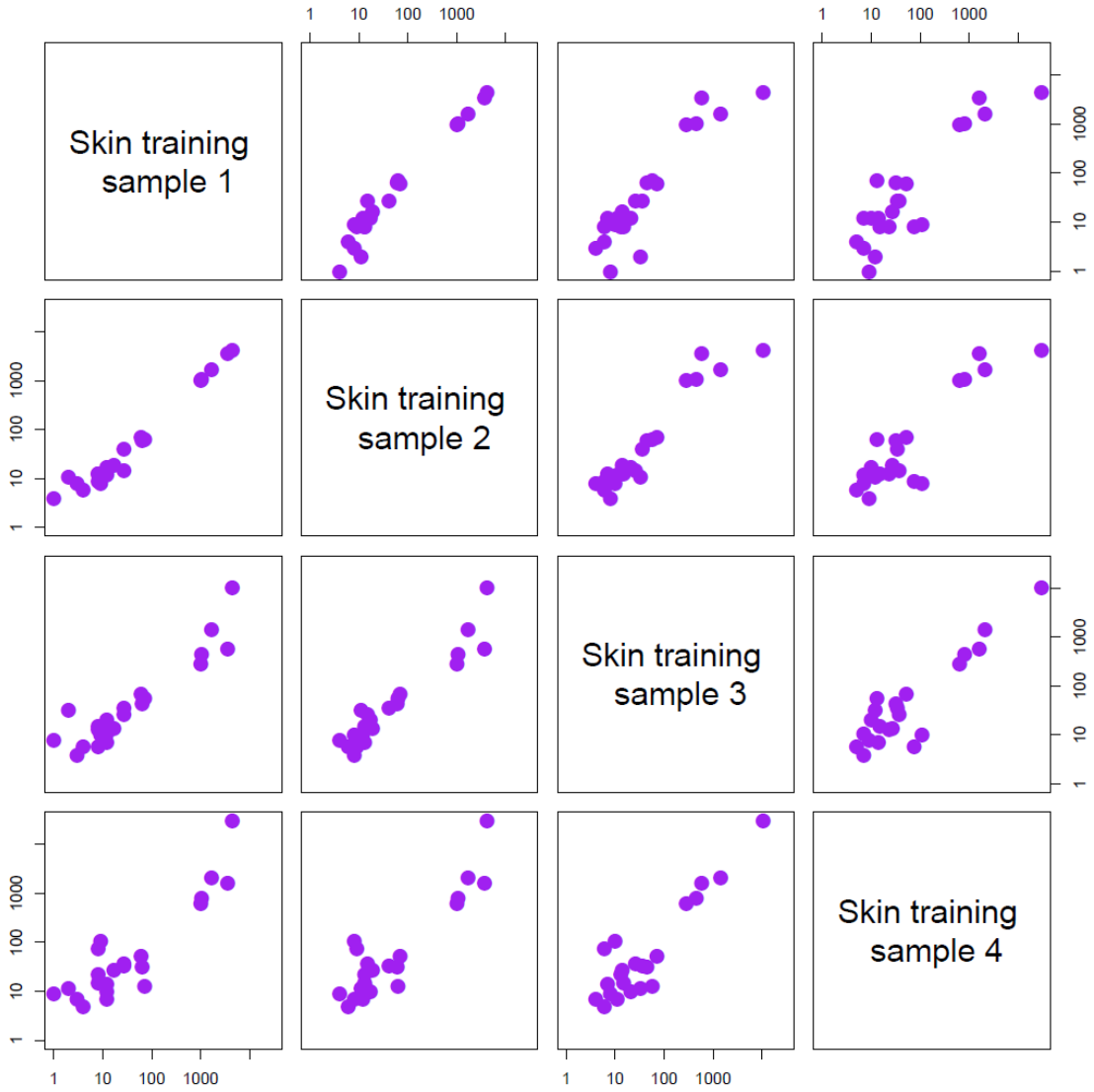




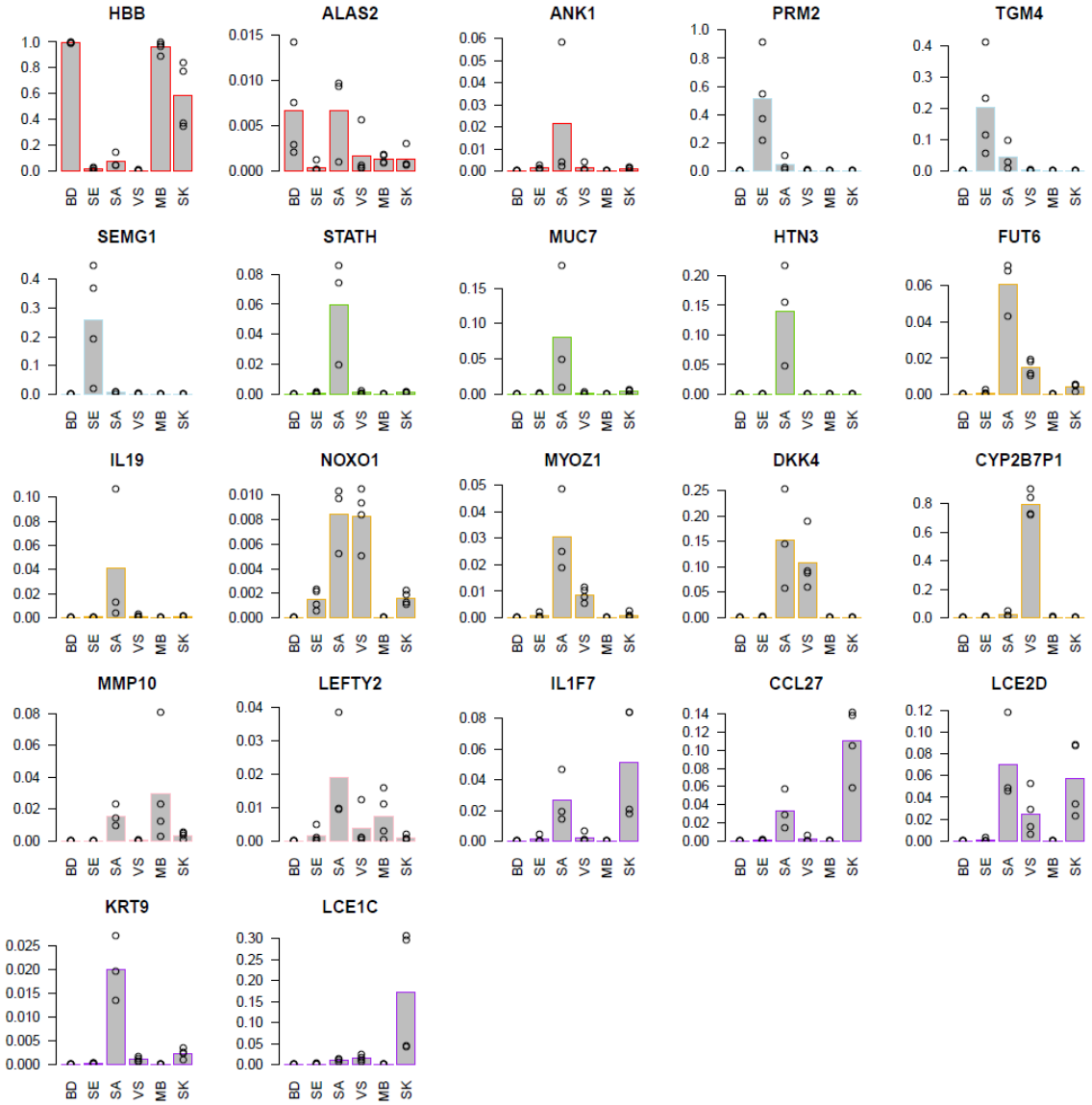




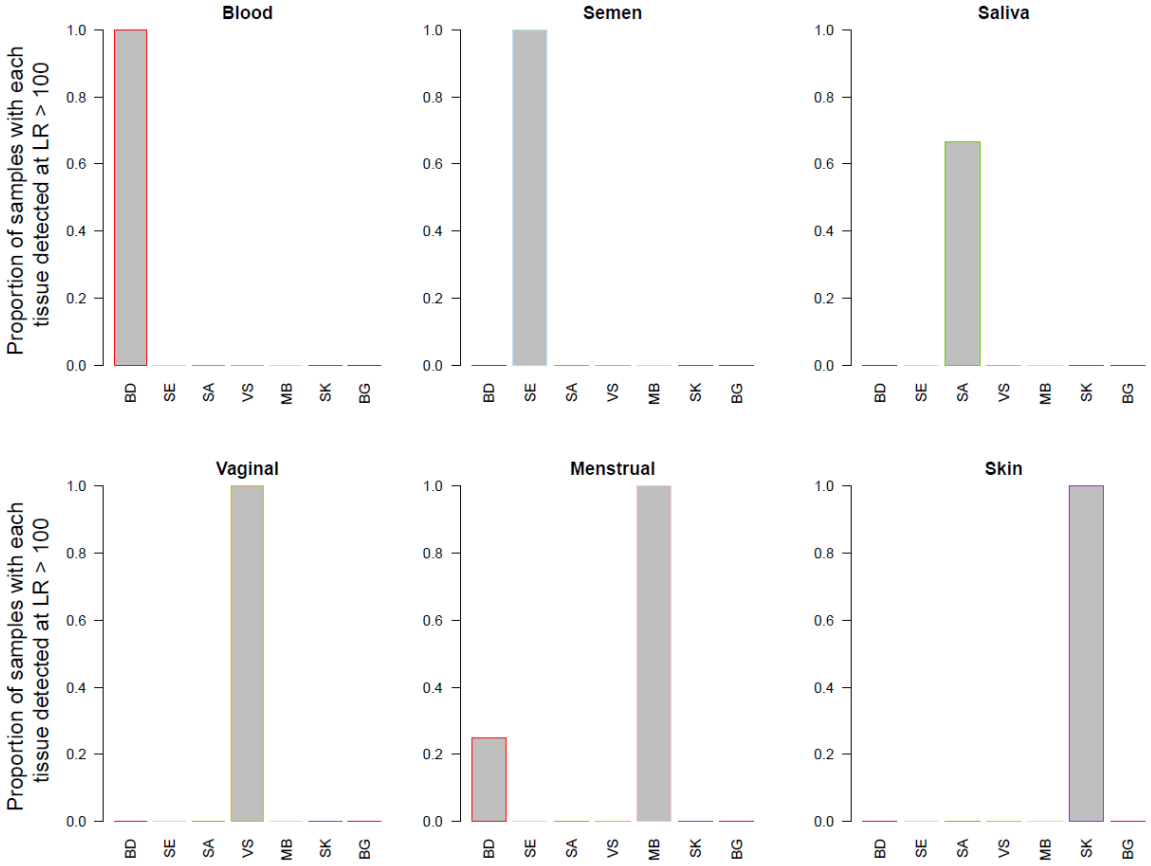




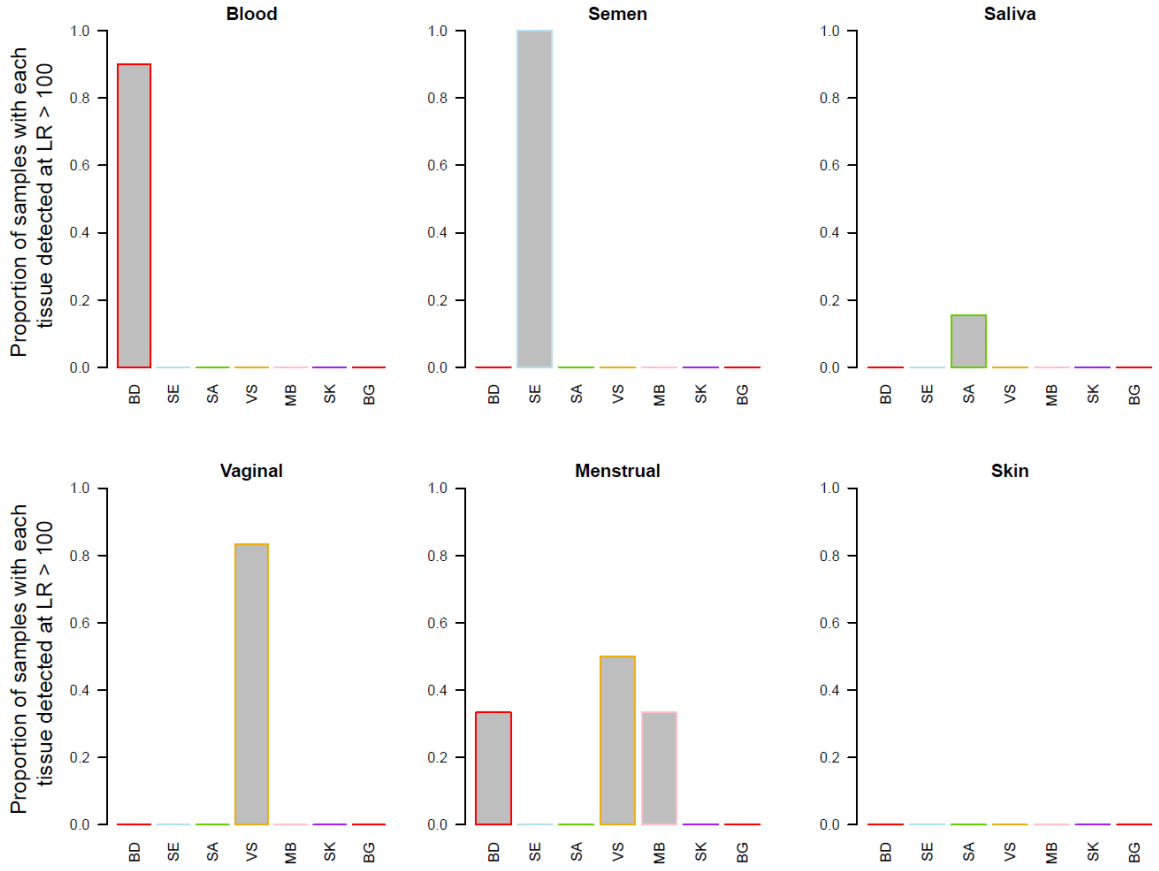
Supplemental Figure 3.



Supplementary Figure 4.



Supplementary Figure 5.



Supplementary Figure 6.