Table of Contents

1. Materials and Methods
    a. Materials
    b. Animals and Breast Tumor Lysates
    c. SERS Substrate Preparation
    d. Liquid Chromatography
    e. SERS Detection
    f. LC-MS
    g. Data Analysis
        i. Background Removal
        ii. Signal Detection
        iii. Barcode Analysis

2. Supplementary Figures and Tables
    a. **Figure S1.** Parallel LC-MS and LC-SERS chromatograms of the same tumor sample
    b. **Table S1.** Selective metabolites identified by MS
    c. **Figure S2.** Continuous detection of two model metabolites using sheath-flow SERS detector
    d. **Figure S3.** Reference SERS spectrum of AHP
    e. **Figure S4.** Background spectra of 6-mercapto-1-hexanol monolayer covered SERS substrate after multiple LC-SERS runs
    f. **Figure S5**. SERS spectra of representative matched signals from 2 replicate LC-SERS experiments of the same tumor sample

3. References

Materials and Methods

Materials

Acetonitrile (HPLC grade, 99.9%), acetic acid (HPLC grade), water (HPLC grade), 6-mercapto-1-hexanol (97%), and 2-amino-3-hydroxypyridine (98%) were purchased from Sigma-Aldrich (St. Louis, MO). 1-(1,3-Benzodioxol-5-yl)ethanamine was purchased from ChemBridge (San Diego, CA). Bare fused silica capillary with 49 μm i.d. and 104 μm o.d. was purchased from Polymicro Technologies (Phoenix, AZ).

Animals and Breast Tumor Lysates

Mice used in this study were maintained under pathogen-free conditions in the University of Notre Dame Freimann Life Sciences animal facility. Animal experiments were conducted in accordance with the University of Notre Dame Institution Animal Care and Use Committee guidelines after IACUC approval (protocol # 15-10-2724 and 18-11-5000). Breast tumors derived from MMTV-Wnt1[1] mice and MMTV-Neu[2] and normal mammary glands from FVB/N mice were collected and used for this study.

The tumor was lysed by first grinding it with mortar and pestle in liquid nitrogen and then resuspending it into three times its volume of lysis buffer (10 mM Tris HCl, pH 7.6, 5 mM EDTA, and 120 mM NaCl). The sample then was lysed using a sonicator for 10 second lysis, pause, and repeat for one minute. The sample was then centrifuged at 14,000 *g* for 5 min, and the supernatant was collected. From the supernatant, a small sample was used to determine the protein concentration by Bradford assay using a standard curve of BSA. From the remaining supernatant, samples were diluted with methanol to the final protein concentration of 1 mg/mL, incubated at -80 C for 30 minutes to precipitate, and then centrifuged at 14,000 *g* for 20 minutes to remove proteins. The supernatant was then taken into a plastic vial, SpeedVac to dryness, and reconstituted in water with 0.1% acetic acid. The final metabolites solution was then used for SERS.

SERS Substrate Preparation

The SERS-active substrate was prepared by a thermal evaporation protocol as previously reported.[3] Briefly, silver shot was evaporated onto an aluminum anodized oxide (AAO) filter with 0.1 μm pores. The filter was dissolved with 0.1 M NaOH solution for 4 hours to reveal a thin layer of highly enhanced SERS substrate with Ag nanostructures on the surface. It is then affixed onto a standard microscope glass slide, predrilled with 2 holes (35 mm apart) and incorporated into a custom-built flow cell.[3]

For experiments reported in Figures 3 and 4, a self-assembled monolayer was attached to the bare substrate by immersing the substrate in a 10 mM ethanolic solution of 6-mercapto-1-hexanol for 24 h before the NaOH immersion.

Liquid Chromatography

For SERS detection, chromatographic separation was achieved using an Ultimate 3000 RSLCnano HPLC system (Thermo Fisher) with two C18 columns—one trap column (Thermo

Fisher, 0.075 x 20 mm, 3 μm) and one separation column (Thermo Fisher, 0.075 x 150 mm, 2 μm). Mobile phases are A: water (0.1 % acetic acid) and B: acetonitrile (0.1 % acetic acid). Flow rate was 300 nL/min. Two different separation methods were used. Method I: t = 0-3 min, 20% B; t = 8 min, 65% B; t = 10-14 min, 85% B; t= 16 min, 35% B; t = 18-20 min, 20% B. Method II: 20% B isocratic for 20 min.

SERS Detection

The sheath-flow SERS cell was connected online to the outlet of the HPLC system with a bare fused silica capillary (~90 cm). The flow cell consists a plastic base with an inlet and outlet for sheath flow, a silicone gasket (with a 2 mm slit defining the sheath-flow channel), a cover slip and stainless steel top plate. The end of the silica capillary is affixed onto the SERS substrate where analytes confinement occurs by hydrodynamic focusing. The sheath flow rate of water was 30 μL/min.

SERS spectral acquisitions were performed on a home-built setup. In general, a 632.8 nm HeNe laser was focused onto the SERS-active substrate in the flow cell through a 40× water immersion objective (Olympus, NA = 0.8). Raman scattering was collected through the same objective and directed to a Shamrock 303i spectrograph (Andor) and EMCCD (Newton 970, Andor). Raman spectra were recorded in series with a 0.2 s acquisition time and 0.5 mW of laser power at the sample.

LC-MS

LC-MS experiments were performed at the CCIC Mass Spectrometry and Proteomics (MSP) Facility of The Ohio State University.

For MS detection, chromatographic separation was achieved using an Ultimate 3000 RSLC HPLC system (Thermo Fisher) with a C18 column (Agilent Zorbax SB-Aq, 3 x 150 mm, 3.5 um). Flow rate was 350 μL/min. Separation method II was used. Injection volume was 3 μL.
MS measurement was performed on Thermo LTQ Orbitrap XL mass spectrometer. Positive mode with data dependent analysis was performed Mass Range was 100 to 1200 m/z. Metabolite identification was achieved by matching detected MS spectra with publicly available databases.

Data Analysis

Background Removal

The background removal algorithm addresses the following observations and concerns. First, the shape of background along frequency channels can change steeply, and thus no smoothing should be applied along the frequency. Second, the shape of background changes over time; this change is typically slower but may trend differently on different frequency regions. The algorithm is described as follows.

First, the original matrix of spectra is divided into time-frequency blocks by taking fixed-size windows at both the time domain and the frequency domain. For instance, a dataset with 5,000 spectra and 1,600 frequency channels will be divided into 50×16 blocks when the window

sizes in time and in frequency are both set to 100, and the first block, for example, contains the frequency channels from 1 to 100 in the first 100 spectra.

Next, within each time-frequency block, the fragments of spectra at different time points are scaled by their average intensities. This scaling removes the difference in the overall intensities and keeps only the shape. This shape is then captured by taking a pointwise median within the block. This median is taken over the time domain for every individual frequency channel and will not result in any smoothness on the frequency domain. Since median is used instead of mean, the signals, if present, will have no virtually effect on the estimation of this shape, and this shape reflects the shape of the background. Finally, the background (shape) is projected on each spectrum, and this projection is removed to give the background-removed spectrum. Algorithm 1 gives the whole algorithm for background removal.

Algorithm 1 for background removal is described as follows:

- Input: original SERS data in *T* time points and *W* frequency channels, which is given as a matrix $X = \left(X_{ij}\right)_{T\times W}$, window size in the time domain w_T, and window size in the frequency domain w_F.
- Output: a matrix of background-removed spectra $\left(Y_{ij}\right)_{T\times W}$.

1. Segment time and frequency dimensions evenly by the corresponding window sizes to obtain $n_T \times n_F$, where $n_T = T/w_T$ and $n_F = W/w_F$. Denote the fragments of spectra in each block by a matrix $X^* = \left(X^*_{ij}\right)_{w_T\times w_F}$.
2. Scale each spectrum fragment at a time point, $X^*_{i\cdot}$ $(i \in \{1, 2, \dots, w_T\})$, by its average intensity $X'_{i\cdot} = X^*_{i\cdot}/mean(X^*_{i\cdot})$. Then estimate the background for this block $B = (B_1, \dots, B_{w_F})$ in a pointwise manner by $B_j = median\{X'_{ij}, 1 \le i \le w_T\}$. Note that median, instead of mean, is used to make the estimate robust to the possible presence of signals.
3. For each spectrum fragment in the block, calculate its pointwise projection vector onto the background $P_i = (P_{i1}, \dots, P_{iw_F})$ by $P_{ij} = X_{ij}/B_j$. Then take the *q*'th ($q = 40$ was usedby default) percentile of the values in P_i as an overall scaling factor and denote it by Q_i. Finally, remove the estimated background at the original intensity scale by $Y_{i\cdot} = X^*_{i\cdot} - Q_i \cdot B$.

Signal Detection

Background-removed spectra consist of signals of interest as well as random noises. Noises typically have relatively low magnitudes, and/or their values alter rapidly between positive and negative values. Signals, on the other hand, are usually positive and look like a set of bumps, which are defined as consecutive positive sections with relatively high magnitudes. Based on this, we have a mathematical definition (shown in Algorithm 2) that depends on three cutoffs: a cutoff for statistical significance that controls the false positive findings measured by false discovery rate (FDR), a cutoff for practical significance that controls the minimum magnitude of signals compared to the noise, and a cutoff of the length of the bump. The last cutoff is introduced based on the observation of presence, although rare, of sharp peaks with large magnitude but minimal length in frequency domain. These peaks are speculated to be due to

cosmic rays, and a length cutoff effectively rules them out. The whole algorithm is shown in Algorithm 2.

Algorithm 2 for signal detection is described as follows:

- Input: A matrix of background-removed spectra *Y* obtained from Algorithm 1. A cutoff α for the relative intensities of signals and an FDR cutoff β, a cutoff γ for bump length.
- A set of time indices of signals, denoted by *t*.

1. For a background-removed spectrum, $Y_{i\cdot}$ $(i = 1, \dots, T)$, estimate the standard deviation of noises σ_i by $\hat{\sigma}_i = k \cdot median\{|Y_{ij}|, j = 1, \dots, W\}$, where $k = 1/(\Phi^{-1}(0.75))$, and Φ^{-1} is the inverse cumulative function of the standard normal distribution. This estimate that uses MAD (median absolute deviation) is highly robust to the possible presence of signals.
2. Calculate the p-value for frequency channel *j* of $Y_{i\cdot}$ by
$$p_{ij} = \begin{cases} 2 \times \Phi\left(-\frac{|Y_{ij}|}{\hat{\sigma}_i}\right) & Y_{ij} > 0 \\ 1 & Y_{ij} \le 0 \end{cases}$$
where Φ is the cumulative function of the standard normal distribution. Then convert *p*-values $(p_{i1}, \dots, p_{iW})$ into $(F_{i1}, \dots, F_{iW})$, where F_{ij} is the FDR of the frequency channel *j* in spectrum *i*.
3. Find all bumps in Y_i, where a bump is defined as a consecutive region of frequencies on which the magnitude satisfies $Y_{ij} > \alpha$ and $F_{ij} < \beta$.
4. Let $L_{i\cdot}$ be a vector that records the length of bumps in $Y_{i\cdot}$. If $\max\{L_{ij}, j = 1, \dots, W\} \ge \gamma$, claim that spectrum $Y_{i\cdot}$ has at least one signal and add its time index *i* into set *t*. Otherwise, claim $Y_{i\cdot}$ as a spectrum without any signal. Repeat this procedure for all background-removed spectra. Finally, the time index set *t* contains all the time indices that have at least one signal.

Barcode Analysis

Here we describe the algorithm to obtain the barcode plot, which exhibits the recurrent signals across technical replicates and the reproducibility of them. First, for each signal in a replicate, we match it with signals in other replicates. Signals are matched according to their Pearson's correlation and the difference between the time they appear, and Pearson's correlation represents the reproducibility of the signal. Second, we record how many times each signal reoccurs across replicates. If the times of recurrence is big enough, for example larger than three, then the corresponding signals will be exhibited in the barcode plot. Besides, we sum up the Pearson's correlation coefficients of each recurrent signal and turn it into color shade in the barcode plot. Hence, the darker color indicates the better reproducibility of a signal. The details of the algorithm are as follows.

Algorithm: barcode plot
Input: the set of signals from *N* replicates, tolerance of time difference denoted by *w,* the cutoff of Pearson's correlation denoted by α, and the cutoff of the recurrence time *K*. Output: a vector *B* for the barcode plot
For $i = 1,2, \dots N$: 1. $s_{i,t}$ is a signal in replicate *i* coming at time *t.* For $j \neq i$, find all signals s_{j,t^*}'s that satisfies:

$$cor(s_{j,t^*}, s_{i,t}) > \alpha$$
$$|t^* - t| < w$$

If there exists such a signal in replicate *j,* we say $s_{i,t}$ is recurrent in replicate *j.*

2. $R_{i,t}$ denotes the times that $s_{i,t}$ recurs in other replicates, and the reproducibility of $s_{i,t}$ is defined as $C_{i,t} = \sum_{j \neq i} \sum_{t^*} cor(s_{j,t^*}, s_{i,t})$.
3. Considering the variance of time, we modify $R_{i,t}$ by $R^*_{i,t} = \max_{\{t':|t'-t|<w\}} R^*_{i,t'}$ and $C_{i,t}$ by $C^*_{i,t} = \sum_{\{t':|t'-t|<w\}} C_{i,t}$

Across all replicates, R^*_t is the times of recurrence for the signal at time t, where $R^*_t = \sum_{i=1}^{N} R^*_{i,t}$, and C^*_t is the accumulative reproducibility, where $C^*_t = \sum_i^N C^*_{i,t}$. For the barcode plot, $B = (b_1, \ldots, b_T)$ is calculated as:

$b_t = C^*_t$ if $R^*_t \geq K$ and $b_t = 0$ otherwise.

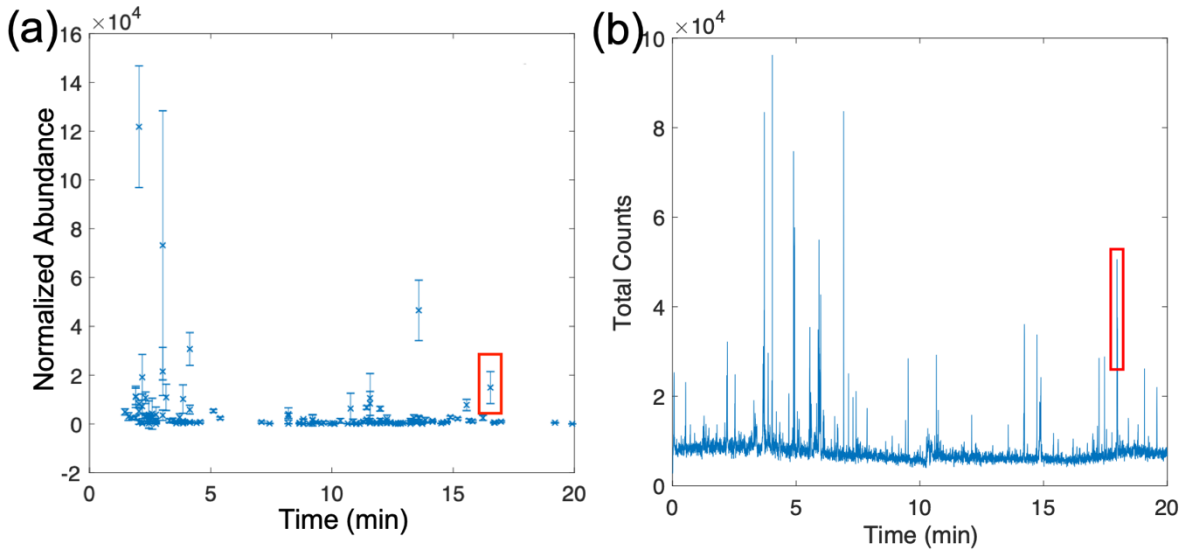
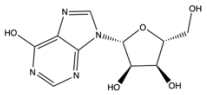
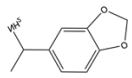
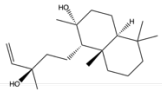
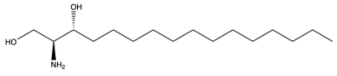
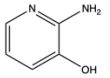


Figure S1. Parallel LC-MS (a) and LC-SERS (b) chromatograms of the same MMTV-Wnt1 tumor sample. Red box: 2-amino-3-hydroxypyridine (AHP). The retention times for AHP in LC-MS and LC-SERS are 16.537 min and 17.977 min, respectively. The absolute difference in retention time is expected due to the differences in the LC instruments used for MS and SERS experiments. The similar relative retention time supports the detection of AHP.

Table S1. Selective metabolites identified by MS.

Retention Time (min)	Peak Width (min)	Abundance	ID	Structure
3.022	0.992	21501	Inosine	
3.162	0.835	10852	1-(1,3-Benzodioxol-5-yl)ethanamine	
9.664	0.618	4110	Sclareol	
10.770	1.009	5287	Hexadecasphinganine	
16.537	**2.357**	**14869**	**2-amino-3-hydroxypyridine**	

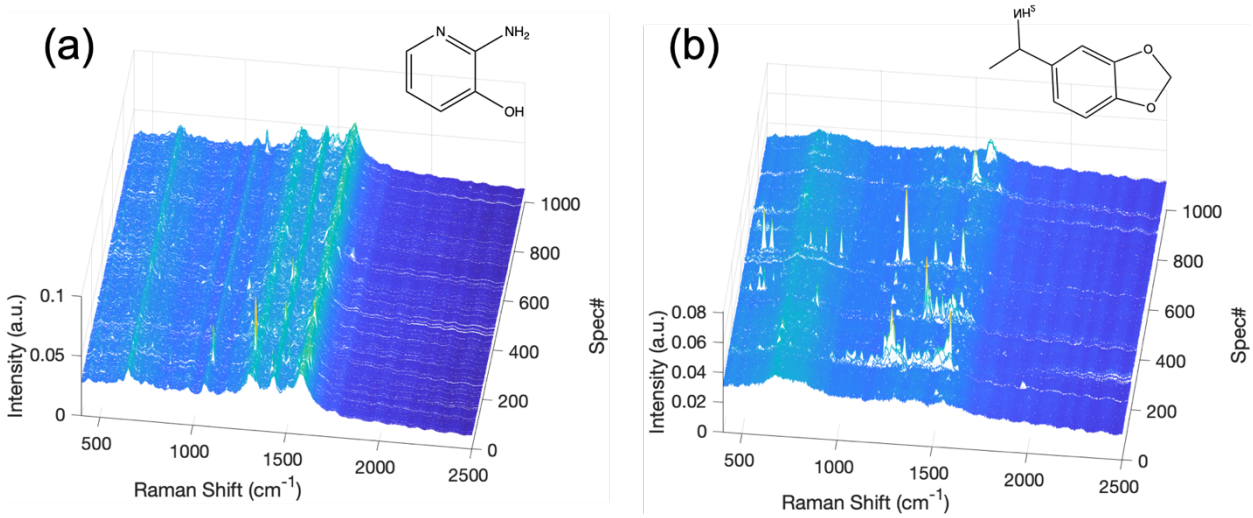


Figure S2. Continuous detection of metabolites (a) 2-amino-3-hydroxypyridine (AHP, 287 µM) and (b) 1-(1,3-Benzodioxol-5-yl)ethanamine (142 µM) using sheath-flow SERS detector. Sample solutions were injected continuously at a flow rate of 1.5 µL/min, while sheath flow was kept at 60 µL/min.

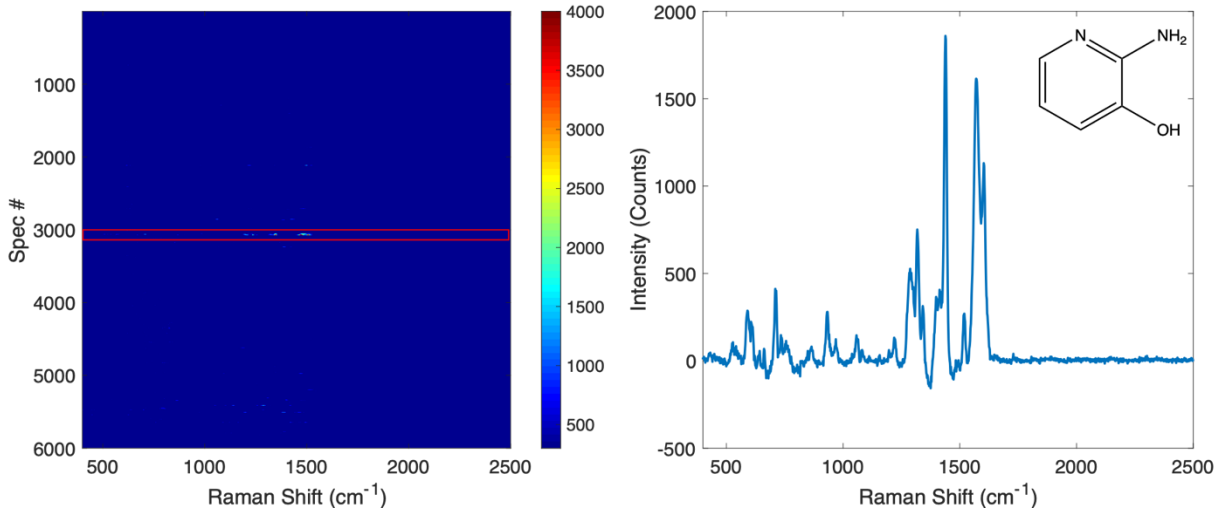


Figure S3. Reference SERS spectrum of AHP. AHP solution (1 µL, 14.4 µM) was injected and run through a 20-min LC separation using Method II. Spectrum on the right is the average of 24 spectra in the SERS chromatogram (highlighted in the red box).

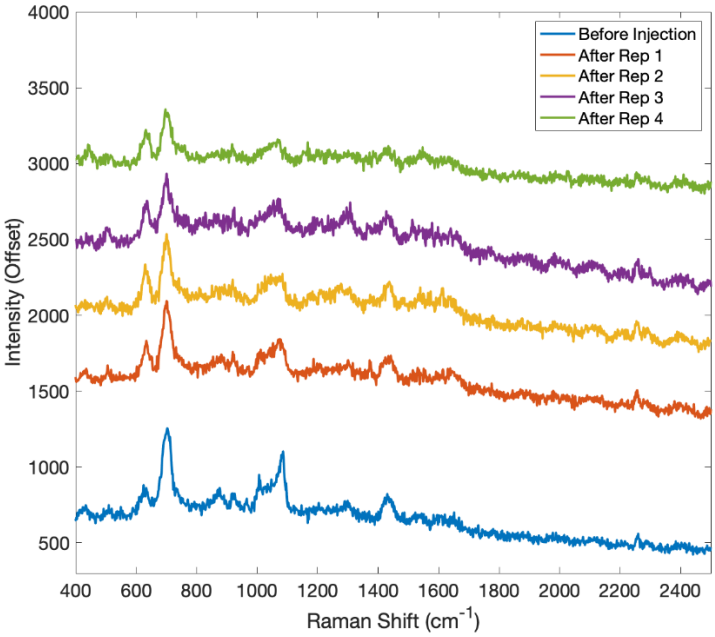


Figure S4. Background spectra of 6-mercapto-1-hexanol monolayer covered SERS substrate before and after 4 consecutive LC-SERS runs.

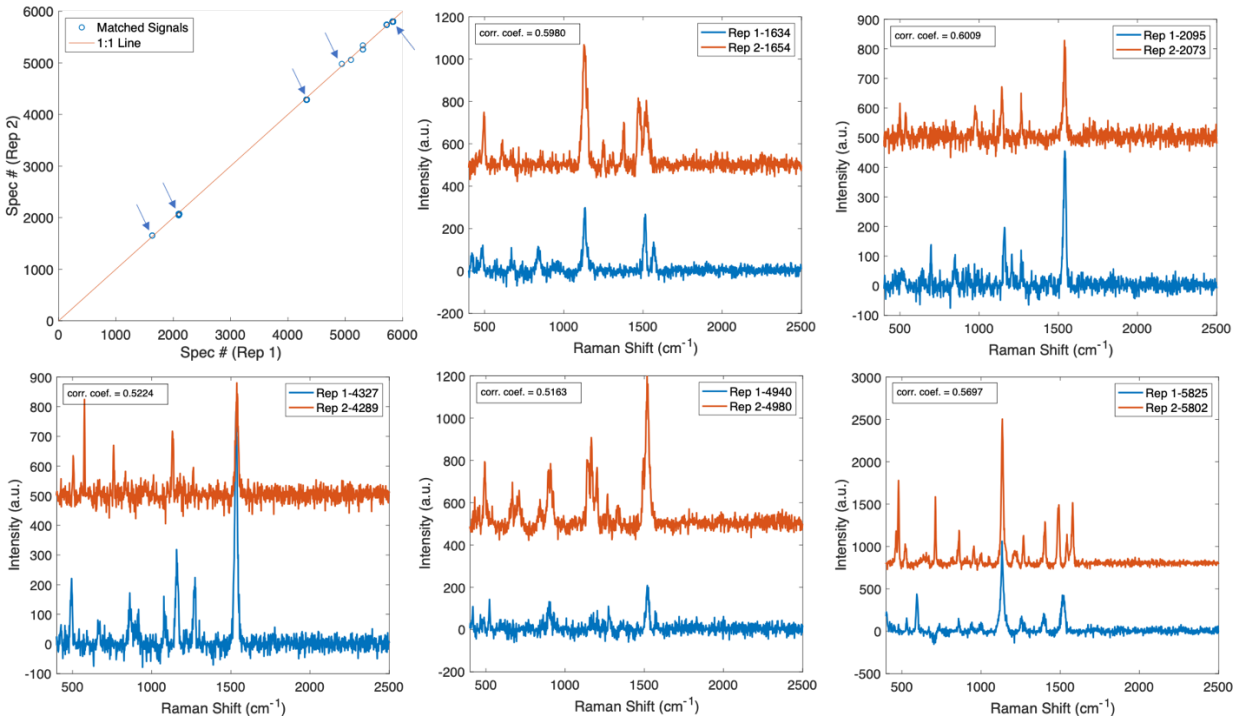


Figure S5. SERS spectra of representative matched signals from 2 replicate LC-SERS experiments of the MMTV-Wnt1 sample. Figure legends denoted replicate#-spectrum#.

References:

[1] A. S. Tsukamoto, R. Grosschedl, R. C. Guzman, T. Parslow, H. E. Varmus, *Cell* **1988**, *55*, 619-625.

[2] C. T. Guy, M. A. Webster, M. Schaller, T. J. Parsons, R. D. Cardiff, W. J. Muller, *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10578-10582.

[3] P. Negri, K. T. Jacobs, O. O. Dada, Z. D. Schultz, *Anal. Chem.* **2013**, *85*, 10159-10166.