Stimulated Raman Scattering Micro-dissection Sequencing (SMD-Seq) for Morphology-specific Genomic Analysis of Oral Squamous Cell Carcinoma

Authors: Tao Chen[1,2†], Chen Cao[1†], Jianyun Zhang[3†], Aaron M. Streets[1,2], Yanyi Huang[1,2,4,5]* and Tiejun Li[3]*

Affiliations:

[1]Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Beijing 100871, China.

[2]College of Engineering, Peking University, Beijing 100871, China.

[3]Department of Oral Pathology, Peking University School and Hospital of Stomatology, Beijing 100081, China.

[4]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

[5]Beijing Advanced Innovation Center for Genomics (ICG), Peking University, Beijing 100871, China.

*To whom correspondence should be addressed: yanyi@pku.edu.cn (Y.H.) and litiejun22@vip.sina.com (T.L.).

†These authors contributed equally to this work.

Abstract: Morphologic and genetic alterations play crucial roles in tumorigenesis, and are the foundation of diagnosis in cancer research. However, visualization of tumor morphology often requires staining which compromises the genetic material, making the interpretation of genetic alterations a challenge. Histological staining is especially detrimental when measuring the transcriptomic changes that underlie variation of histological features at the microscopic level. Here we propose stimulated Raman scattering micro-dissection sequencing (SMD-Seq), which exploits the intrinsic vibrational signatures of chemicals to rapidly construct label-free histological images of cryo-sectioned tissues, and achieves in situ laser micro-dissection of small regions of interest for location-specific and simultaneous transcriptome and genome analysis. We applied SMD-Seq to unstained cryosections of human oral squamous cell carcinoma (OSCC) samples. SRS images proved to be comparable to H&E staining in revealing the morphological characteristics of tissues, and capable of differentiating the small cancer regions from normal epithelium of OSCC. With significantly reduced nucleic acid loss, accurate identification of copy number variations, gene expression levels, and gene-fusion events were obtained through genome and transcriptome analysis of SRS-guided high-purity micro-dissected regions. The high-resolution histological characteristics combined with preservation of high quality genetic material from specific regions of interest enabled the characterization of inter- and intra-tumor

heterogeneity using morphological and genetic analysis. Given histopathological features and matched biomolecular content, SMD-Seq provides complementary insights in the study of cancer, and opens a window for correlative analysis between morphology and genome and transcriptome sequencing in complex samples with intrinsic genetic mosaicism.

Introduction

Both morphological and genomic alterations are key factors of tumorigenesis. In clinical practice, tissue-based histopathologic observation, which assesses the cellularity, biochemical contents, and histoarchitecture of a biopsy, is the gold standard for cancer diagnosis. With the emergence of high-throughput sequencing technologies, the genomic analysis of tumor tissues have greatly extended our understanding of cancer in exquisite detail[1,2]. The combination of direct assessment of histological features of tissue and quantitative analysis of genomic variations should in principle enable better comprehension of the correlation between the phenotypes and the genotypes of cancer. Recent studies[3-5] have shown that the application of histology and genomic analysis provides complementary prognostic insight in cancer diagnosis and treatment. Whole genome sequencing has been extensively applied to quantitatively study the copy number variations (CNVs), single nucleotide variations (SNVs) and epigenetic modifications of cancer, and some studies have performed this analysis on single cancer cells[6]. RNA-Seq, on the other hand, provides gene expression profiling information that not only differentiates cell types and fates but also the transcriptional alternations such as alternative splicing, and its consequences[7]. Interestingly, for small amounts of cells whole genome amplification typically produces chimeras[8] that become false-positives of structural variation events such as gene fusions, one of the major driver mutations in neoplasia. Low-input RNA-Seq may become a solution to solve this problem by identifying the fused transcripts[9]. Particularly, RNA-Seq is capable of finding potential 'transcription-induced gene fusions (TIGFs)', which are the results of alternative splicing between genes[10]. However, complex spatial architecture[11,12] and the intra-tumor heterogeneity[13,14] of cancer create several technical challenges, one of which is relating morphology to genome and transcriptome variations. Many protocols for sample preparation, including DNA recovery and amplification, have been improved to handle various kinds of tissue sections. Although these techniques have been successfully developed to capture trace amounts of starting materials such as the genomic DNA or RNA of single cells[15,16], these methods have a few intrinsic limitations when handling histologically stained tissues. First, in many cases the admixture of cancer tissues contain a large amount of immune and stromal cells[17], which significantly contaminate and dilute the cancer signatures in the resulting sequencing data. Such ensemble sequencing is far from ideal to provide precise genomic alternations that should strongly correlate with the micro-scale histological cellular context, which is typically lost through lysing of complex tissues. Second, while laser capture micro-dissection is able to obtain morphologically pure tissue samples *in situ* at the cellular level[18], currently this technique still relies on

conventional staining and fixation protocols, such as hematoxylin and eosin (H&E) staining, a gold standard for morphological feature identification. During the H&E staining process the relatively stable DNA molecules may still be preserved but most RNA molecules are inevitably degraded, resulting in the loss of transcriptome information in the micro-dissected tissue specimens with high morphological purity[19].

Here we present a technology that enables simultaneous genome and transcriptome analyses of morphologically specific, small regions of tissue guided by label-free microscopy and coupled with *in situ* laser micro-dissection. This technology employs stimulated Raman scattering (SRS) microscopy[20], which exploits the intrinsic vibrational signatures of chemical species in order to rapidly construct histological images of cryo-sectioned tissues without staining or labeling, to preserve the RNA molecules in tissue samples. The specific small regions of interest (ROI) in the tissue section can be immediately dissected from the sample using the SRS microscope without sample transfer, enabling the capture of genetic material from a small number of morphologically specific cells. We applied our technology, named SRS micro-dissection sequencing (SMD-Seq), to characterize the heterogeneity of complex tissues of human oral squamous cell carcinoma (OSCC). The high-resolution and high-speed imaging preserves both intact histological features and vulnerable genetic materials in the frozen sections. We have differentiated the small cancer regions from normal epithelium, and obtained histology correlated genomic information to reflect the inter- and intra-tumor heterogeneity among patients in both genome and transcriptome profiles. In addition to improved CNV and SNV detection to unveil the genetic mosaicism in cancer, we have also shown that, with significantly reduced RNA loss, accurate identification of gene-fusion events can be obtained through transcriptome analysis of SRS-guided high-purity micro-dissected regions.

Results

Workflow design of SMD-Seq

SMD-Seq (**Fig. 1**) is an integration of microscopic imaging, *in situ* dissection, and low-input sequencing. In order to optimize both image quality and the sequencing data, we had to balance the tradeoff between imaging parameters including exposure time and laser power, and preservation of genomic material. Typically we used a 30-μm thick cryosection for the following analysis. To compose an image with histological features, we chose CH_2 (2850 cm^{-1}) and CH_3 (2950 cm^{-1}) stretching vibrations for contrast. These two bands were consistently present in the Raman spectra of both cancer and normal tissues, unlike the Raman bands within the fingerprint region which were not always detected in the samples(**Supplementary Fig. 1**, gray region). Owing to the spatial biochemical distribution, SRS images of these two bands exhibited a clear difference, particularly from lipid and protein ratio[21,22]. For each field of view, we applied a linear combination approach[23] to convert these dual-band imagesinto a reconstructed pseudo-color image, in which we represented protein- and lipid-rich regions with cyan and red, respectively. SRS microscopic images can be directly and rapidly obtained from tissue samples, generating high-quality and high-resolution histological images that resemble H&E staining.

To specifically harvest genetic material from certain tissue, we performed *in situ* laser micro-dissection on those cryosections placed on polyethylene naphthalate (PEN) membrane slides using the SRS microscope without sample transfer. We created a closed path for the regions of interest (ROI) after SRS histopathologic characterization and focused the scanning laser with elevated power (~180-190 mW) on the PEN membrane to cut the tissue slice. We then immediately recovered the dissected samples into a tube prepared with lysis buffer, and divided the lysate into two aliquots for DNA and RNA sequencing libraries preparation, separately.

Histopathologic characterization of tissue using label-free SRS images

We first examined if these SRS images of stain-free OSCC tissue cryosections were comparable to widely used H&E images for histopathologic applications. For each 30-μm thick cryosection imaged with SRS, we kept a 5-μm adjacent section for H&E staining as the histological reference for further comparison and validation (**Supplementary Fig. 2**). The nondestructive SRS imaging approach provided the intrinsic biochemical information from the molecular components in cells. Such microscale information could be combined with large-scale histoarchitectural features to better identify tissues at the single-cell level (**Fig. 2a**). For example, in both SRS and H&E images, the epithelium displayed gradually changing cell profiles, from prolate squamous cells to almost round basal cells (**Fig. 2a, Supplementary Fig. 3**). Cross-sections of bundles with characteristic high protein content represented muscle (**Fig. 2a**). Ducts, with a wall formed by a single layer of cells, could be found in crowed glands. Nerve tissue appeared as a large, lipid-rich fiber bundle with peripheral protein-rich fibrous connective tissue (peirneurium). We also identified features of a Watson tumor and a Mucoepidermoid Carcinoma sample (**Supplementary Fig. 4**), in addition to OSCC, to demonstrate the generalizability of SRS imaging for cancer histology.

We further evaluated the protein to lipid ratio (PLR) of different tissue types. Because of the unique chemical selectivity of SRS imaging (**Fig. 2a**), the PLR histogram of different tissue reflected a distinct biomolecular signature in their cellular contents. In addition to molecular characterization, we chose texture-based morphology identification to validate the applicability of SMD-Seq in histology. We then described the texture of SRS images with the histogram of orientation gradient (HOG)[24] features. The HOG features can characterize image texture, hence reflecting cellular packing from SRS image of tissue section. The HOG features of SRS images were compared to those of corresponding H&E staining images, followed by unsupervised clustering. The results revealed that each type of tissue harbored a unique pattern (**Fig. 2b**). It also implied that SRS images were comparable to H&E staining in revealing how cells were spatially organized, and were capable of supporting visual tissue identification based on histoarchitectural patterns. A non-trivial and more important demonstration was to separate normal epithelium from cancerous epithelium, which originated from the same tissue type. Their ROIs had similar PLRs, but harbored different histological features. We performed unsupervised clustering based on HOG features of 32 SRS images taken from the epithelium, including 16 cancer and 16 normal samples

(**Supplementary Fig. 5**). Two sets were clearly divided with only one exceptional cancer sample (**Fig. 2c**). This result showed that cancer and normal epithelium tissues displayed different histoarchitechtural patterns in SRS images. Furthermore, the correlation matrix of HOG features demonstrated a stronger correlation among normal epithelium compared with that of cancer samples, indicating that higher morphological heterogeneity might exist in cancer (**Supplementary Fig. 6**).

The large scale morphological features were described by a stitched SRS image (**Fig. 2d**). 80 fields of view (FOVs), with 0.65 × 0.65 mm^2 each, were used to construct this large field of view (9 × 4 mm^2). The cancerous part of the cryosection showed similar cellular mosaicism to the epithelium region. With these nondestructive stain-free SRS images, resembling H&E staining, tumor nests as small as 200 μm in diameter could be easily identified with conventional criteria such as hyperplasia, dysplasia, cancer nesting and keratin pearls (**Fig. 2d**). Higher resolution images of each FOV confirmed that SRS images revealed histological information almost identical with H&E staining, except that in some cases H&E stained samples were distorted due to the experimental process (**Fig. 2d**, the medium and small images of epithelium regions). It was noteworthy that keratin pearls, the landmark feature of highly differentiated OSCC, could be clearly discerned in SRS images as onion-like histological patterns (asterisks in **Fig. 2d**). Furthermore, the strong signal from the red channel of SRS images indicated that these keratin pearls are protein-rich, reflecting the molecular nature of their structure.

SRS-integrated in situ Laser Dissection

To illustrate the performance of SRS-image based laser dissection, we applied H&E staining on the entire cryosections after micro-dissection, and compared them with the adjacent H&E stained reference sections (**Supplementary Fig. 2**). With SRS images and H&E images of micro-dissected cryosections, we successfully pinpointed dissected sites on the reference H&E cryosection (**Fig 3a**). For each ROI, an SRS image was acquired, followed by determination of the dissection path (**Fig. 3b**, curves). Besides the dissected ROI in each FOV, the rest of the section was conserved in the SRS image, H&E stained micro-dissected section, and the adjacent H&E stained reference section. In some cases, the dissected regions marked in the reference H&E images exhibited different morphological features than in the SRS images because of the complexity of tumor spatial distribution (**Fig. 3b**). Our *in situ* micro-dissection approach, using the same instrument without sample replacement, did not rely on reference sections to restore the ROI position, and hence provided high precision seconds after acquisition of SRS images. We measured the width of laser incision to be 9 μm (**Fig. 3c**), and most cells in ROIs were intact. Although the size of cancer nests varied, a typical nest had an average diameter of 300 μm (**Fig. 3d**). We chose a 20×, NA 0.75 objective because (1) it offered a FOV of 635 μm wide, which was sufficient to fully record the majority (over 90%) of a single cancer nest in a single image; (2) we needed to avoid using immersion fluid, which may cause RNA degradation, section distortion, and counteract thermal effect of the laser that dissected the specimen. The dissected region of single cancer nest, covered a mean area of 0.16 mm^2 (**Fig. 3f**), and

contained about 200 cells in average (**Fig. 3e**), with variations among different tissues types. The SRS imaging process followed by morphology characterization and in situ micro-dissection, in total took no more than 2 min. The dissected samples were carefully collected into centrifuge tubes, where cells were lysed immediately. The released DNA and RNA were then prepared for genomic and transcriptome analysis, separately. Rapid imaging, dissection, and cell lysis are critical to preserve the quality of nucleic acid for sequencing in low quantity samples. This is especially important for the RNA molecules which can degrade rapidly after sectioning.

Transcriptome and Genome Sequencing and Analysis

To demonstrate the capability of SMD-Seq, we collected 28 *in situ* micro-dissection samples for sequencing. 27 samples passed the quality check for further analysis, including 12 normal and 9 tumorous samples by RNA-Seq, 13 normal and 8 tumorous samples by DNA-Seq from four patients (**Supplementary Table 1-3, Fig. 7**). We validated the reproducibility of SMD-Seq of these samples by checking the Spearman correlation coefficients (r) of expressed genes (average r = 0.7) and reads count (average r = 0.7) between biological replicates of the same patients (**Supplementary Fig. 8**). We estimated the cellular purities of micro-dissected cancer nests and found that they were significantly higher than the head and neck squamous cell carcinoma samples collected by TCGA (P value = 0.0098, **Supplementary Fig. 9**). We observed that one of the dissected cancer samples, P4S1C, displayed high purity, while its adjacent H&E staining reference section was proved to be infiltrated with stromal cells (**Supplementary Fig. 9**), indicating the highly diverse spatial structure of small cancer lesions and the importance of *in situ* dissection and analysis.

As H&E staining is the 'gold standard' of histological diagnosis, we compared RNA recovery between H&E stained and unstained cryosections by qPCR (**Supplementary Fig. 10**). For samples with the same thickness, the significant decrease of RNA quantity in H&E stained sections, as measured by Ct values, was not sufficient for sequencing library preparation. This proved the necessity of employing label-free histology to prevent RNA degradation, thus improving RNA-Seq measurements. We analyzed the whole transcriptome profiles of SMD-Seq samples, and ~9,000 genes were detected on average (FPKM > 0.1, Fig. 4A, Sup. Excel 1.1). Principal component analysis of transcriptome profiles showed that normal epithelium, gland, muscle and cancer samples clustered in distinct groups that corresponded with their identified morphology (**Fig. 4b, Supplementary Fig. 11**). Unsupervised hierarchical clustering was performed using 217 differently expressed genes (**Fig. 4c, Supplementary Excel 1.2**). Samples were classified by their morphologic characterization, with one exception of P4S2E. Clustering revealed four gene sets that distinguished the epithelium, cancer, gland, and muscle tissue. Gene ontology annotation terms of each set directly reflected the characteristics of the different tissue types and health conditions[25](**Supplementary Fig. 11**). The four sets which corresponded to epithelium, cancer, gland, and muscle were enriched for genes related to epidermal development, immune response regulation, digestion/secretion, and the motion of myofibril and actin. While showing relatively consistent expression patterns in normal samples, gene expression levels varied among

cancer samples. For example, some genes related to the epithelial to mesenchymal transition (EMTs), KRT13, ELF3 and ESRP1[26,27](**Supplementary Fig. 12**), showed higher expression levels in all of the normal epithelium samples when compared to cancer samples. In contrast, SERPINE2[28] (**Fig. 4d**), AKR1B10*(29,30)* and UDP Glycosyltransferase 1A family (including UGT1A6, UGT1A7, UGT1A9 and UGT1A10, **Supplementary Fig. 12**) expressed at a lower level in all normal tissues but gained high expression levels in some cancer samples. This observation agreed with previously reported studies[28-30]. Specifically, KLK8, KRT1 and KRTDAP (**Fig. 4d**) only exhibited high expression levels in patient P3. KLK8 was found to be implicated in malignant progression of OSCC[31], and KRT1 and KRTDAP were genes strongly related to the differentiation and maintenance of stratified epithelium. Correspondingly, the "keratin pearl" structure of P3 in the SRS image (**Supplementary Fig. 12**) also indicated the high differentiation level of its cancer nests. Genes like GSTP1 and KRT13 (**Fig. 4d**), were found to be expressed similarly among patients and showed significant differences when compared with normal tissues. GSTP1 was previously studied in head and neck cancers[32,33], and its high expression level in tumor cells was reported to be associated with more aggressive cancer and poor patient survival. We performed immunofluorescence staining for GSTP1 to evaluate the level of protein expression (**Supplementary Fig. 13**) and confirmed the accumulation of GSTP1 in cancer nests. This observation correlates with the RNA-Seq results. Using SMD-Seq, we further found gene expression exhibited heterogeneity between different regions dissected from the same patient (**Fig. 4d**). The adjacent P4S1C and P4S2C exhibited similar gene expression feature, and at a distance from them, P4S3C highly expressed different groups of genes. The intra-tumor change of molecular evidence may indicate the progress of tumorigenesis.

We then exploited the RNA-Seq data to query the whole transcriptome for de novo identification of gene fusion events, which have been recognized as driver mutations in neoplasia[34]. The displacement and recombination of genes, especially oncogenes, had become the focus of many cancer studies as they may provide potential therapeutic targets[35,36]. Since gene fusions are dependent on cellular context, we have applied fusion transcripts analyses on RNA-Seq data of those morphology characterized and laser dissected OSCC samples. Though some genes, like KRT6A (P1S2C, P3S5C), FAM102A (P1S3C, P3S5C, P4S2C), etc., were involved in gene fusion events across different patients, heterogeneity of fusion events was discovered among patients (**Fig. 4e, Supplementary Excel 1.3-1.8**). A majority of these fusion events might be passenger events that came along with cancer development and thus their actual consequences remain unknown. However, we found that one of the fusion sets was recorded in TCGA (MYH9 and KRT14, from P2), and two of them included oncogenes AKT3 (AKT3 and LRRC45, from P3) and MAFB (MAFB and SAC3D1, from P3). To experimentally verify the fusion transcripts, fragments which harbored the joint junction of fusion genes were amplified and sequenced using Sanger sequencing (**Supplementary Fig. 14, Table 2, and File1.1**). The Sanger sequencing results confirmed the existence of the fusion junction between MYH9 (5' fusion partner: exon 20) and KRT14 (3' fusion partner: exon 8), AKT3 (5' fusion partner: UTR) and

LRRC45 (3' fusion partner: intron), showing that our approach can be applied to discover recurrent fusion events. The heterogeneity of fusion events was also found among ROIs of the same patient (**Supplementary Fig. 15**). For example, more fusion genes were found in P4S3C, compared with distantly dissected P4S1C and P4S2C, indicating the instability of the genome during tumorigenesis. Moreover, we also observed the existence of an oncogene-involved fusion (RAB3D and MTMR14) in the previously mentioned sample P4S2E (**Supplementary Fig. 14c**), which appeared to be normal according to histopathological characterization. The joint fragment between RAB3D (5' fusion partner: UTR) and MTMR14 (3' fusion partner: UTR) were amplified and validated by Sanger sequencing as well (**Supplementary Fig. 14d, File1.1**). The whole transcriptome analysis proved that the well preserved RNA from SMD-Seq provided genetic alteration evidence underlying morphological features among different cancer nests, increasing the information dimension of tumor heterogeneity studies. We performed whole genome amplification and sequencing through degenerate oligonucleotide primed PCR (DOP-PCR) using half of the lysate of micro-dissected samples, and further analyzed the genomic alterations and heterogeneity at the genome level. Copy number variation (CNV) analysis demonstrated that various patterns of CNVs existed between cancer and normal samples, and also in different patients (**Fig. 5a, Supplementary Fig. 16**), indicating that OSCC is a highly complex tumor with significant genetic mosaicism and heterogeneity. A few commonly shared large-size ploidy shifts, such as the losses of 3p and 8p and the gains of 3q and 8q, which have been observed in most tumors[37,38], were also observed in our OSCC samples. There were also various patient-specific CNVs, for example, chromosome 6 showed high instability in one patient's (P2) cancer sample but not found in others (**Supplementary Fig. 16, 17**). Unsupervised clustering of the CNVs also proved that each patient possessed a unique CNV pattern (**Fig. 5b, Supplementary Fig. 17**)[39]. For the same patient, samples dissected from different locations displayed subtle discrepancy in copy number alterations (**Fig. 5b**). The CNV pattern of chromosome 1 in P1S1C and P1S3C were different from that of P1S4C, with an obvious gain of 1q in the first two ROIs (**Supplementary Fig. 18**).

Among all the OSCC samples, 8 regions of recurrent copy number gain and 5 regions of recurrent copy number loss were identified ($q < 0.25$, **Supplementary Fig. 19, Excel 1.9-2.0**)[40]. Among these (copy number loss/gain or regions) 11q13.3, 8q24.3, 11p15.4 and 11q24.2 co-localized with differently expressed genes in cancer samples. GSTP1 located within the recurrent focal amplification of 11q13.3[37], which might imply that the high expression level of GSTP1 resulted from increased copy numbers. FAM83H was also co-localized with a focal amplification region, 8q24.3, and it specifically expressed at higher levels in patient P1. TP53AIP1 and PKP3 both expressed at a lower level in all the patients, and located in the regions of recurrent copy number loss 11q24.2 and 11p15.4, respectively. The expression level of GSTP1, FAM83H, TP53AIP1 and PKP3 were all reported to be involved in the development of cancer or had effects on patients' survival rate[41-44]. Additionally, we found some CNVs overlapped with gene fusion sites (**Fig. 5g**). For example, fusion genes CTSB and PPP1CA co-localized with focal amplification regions 8p23.1 and 11q13.3, separately.

CTSB was proved to be related to cancer progression and metastasis[45,46], and PPP1CA was reported to contribute to ras/p53-induced senescence[47]. Of 24 pairs of fusion genes we detected, 17 pairs (~71%) had at least one gene intersected with amplification and deletion regions (**Supplementary Excel 2.1**). The parallel observation of genomic rearrangement and gene expression fold change may illustrate that the instability of the cancer genome led to gene fusion events which were more likely to occur within amplification and deletion regions[48] (**Fig. 5g**).

Moreover, the mean expression level of genes within each genomic segment[49] was compared with the copy number in the same region (**Fig. 5c, d, e, Supplementary Fig. 18, 20**), and summarized across the whole genome (**Fig. 5f, Supplementary Fig. 21-22**). The average gene expression levels showed positive correlation with the copy number within the same segment (**Fig. 5f**), demonstrating that our technique is able to simultaneously detect and study copy number alterations and transcript abundance variation within the same section.

Discussion

Current approaches for *in situ* histology and transcriptome information acquisition are predominantly limited by RNA degradation during the staining process. Although it has been reported that shortening the H&E staining time or introducing RNase inhibitors may help preserve RNA with thousands of cells[50], obtaining both high quality of histological images and spatially correlated molecular analysis is still a challenge for the study of cancer heterogeneity at microscopic levels. SMD-Seq integrates stain-free nondestructive histological imaging with low-input genomic sequencing to provide a histoarchitecturally specific genome and transcriptome landscape in complex tissues. The intact cellular context was acquired through stimulated Raman scattering microscopy, yielding both histopathological features and biomolecular content for precise and accurate differentiation of small lesions from normal tissues. The nucleic acids, especially the RNA molecules, have been well preserved for further amplification and sequencing, opening a window to identification of various genomic alterations in complex samples with intrinsic genetic mosaicism. The correlative analysis between morphology and genome/transcriptome sequencing achieved in SMD-Seq is informative to tumorigenesis research. This analysis revealed a natural information flow: genomic alterations may be transcribed into expression alterations, and finally translated to observable functions that affect the phenotypes. With these complimentary data sets, characterization of cancer could be evaluated in a comprehensive manner, uncovering the hidden connections in tumorigenesis.

During the entire process, from sample collection to sequencing library preparation, a few critical issues need to be taken into consideration in order to eliminate the severe degradation of vulnerable RNA molecules. Though higher resolution of SRS images can be achieved in our system, current imaging parameters were carefully considered for balancing between the image quality and preservation of genomic material for sequencing. While conventional H&E staining requires the processing of thin slices to avoid the stacking of cells, SRS imaging, allows us to image relatively thicker slices, which benefits RNA recovery. Hence we typically used 30-μm thick slices, in which a

majority of the cells are protected from sectioning injury and thus kept intact. This keeps most of the RNA molecules away from environmental ribonucleases, preventing loss of RNAs from broken cells. Our practice also concluded that this thickness did not affect the quality or robustness of laser dissection. Another crucial factor is the maintenance of sample temperature throughout the entire procedure. The tissue samples were kept at low temperature to further eliminate the RNA degradation. With all of the above considerations, the imaging and micro-dissection (3 ROIs in each slice) must be completed within about 10 min in order to ensure high-quality RNA-Seq data.

In a few previously reported studies, phenotype-genotype integration has been shown to generate new insights in complex biosystems, using an algorithm developed to predict the spatial origin from transcriptome profile[11,12]. SMD-Seq, however, performed imaging and *in situ* micro-dissection, and thus directly matched the sequencing results with cells' position. Furthermore, SRS imaging not only resembles H&E-based histological images but also outperforms other alternative approaches. For example, one may use an adjacent H&E section to assist the stain-free cryosection micro-dissection and sequencing (assuming the nucleic acids are well preserved) without SRS imaging. However, histopathological pattern discordance between the neighboring sections occurred in our observation. It could result from natural difference of cellular histoartchitechtural pattern in microscopic scale, or a distortion due to staining process. Both could lead to a mismatch between histology and molecular signature. These challenges inevitably create difficulty when studying tissues with fine scale mosaicism such as OSCC, and signify the importance of *in situ* analysis. In our approach, correlative analysis of morphology and sequence information could unveil previously unobserved trace of tumorigenesis. For example, the P4S2E sample demonstrated inconsistency between morphology and sequence information. In both the SRS image and its H&E reference, the sample was identified as normal epithelium. It could be seen in PCA analysis, however, that P4S2E clustered more closely with the cancer samples along the PC1 and PC3 axes (**Fig. 4b**). In addition, the genetic and transcriptomic features of this sample reflected a cancer-like pattern. Furthermore, genes previously reported to be significantly mutated in OSCC, like FAT1, PPP2R1A, PTEN, HRAS, and CREBBP28, were also found in P4S2E (**Supplementary Fig. 23**). This histology-genomic inconsistency could imply that cancer-like gene expression profiles may arise before of morphological signatures.

Although SMD-Seq can preserve morphology and sequence information with high quality, this approach still bears limitations. First, SRS imaging in SMD-Seq is able to identify local tissue features, however, the resolution of air objective cannot guarantee visualization of subcellular structures, which hindered in depth image analysis. Increasing pixel density of image may reduce the noise level because of higher sampling rate (**Supplementary Fig. 24**), but prolonged exposure times lead to a higher chance of sample ablation during imaging (**Supplementary Fig. 24**). Second, the sequenced DNA and RNA molecules in this approach are from aliquots of one sample's lysate. In each aliquot, only DNA or RNA, can be sequenced. This small-bulk pool-and-split approach masks the single-cell-level heterogeneity within the sample and blurs the relationship between the genome and transcriptome. Additionally, the

simple operation of H&E staining is an advantage over SRS microscopy. The improvement and simplification of SRS imaging system should in principle broaden the clinical application range of SMD-Seq in the future.

In summary, in this pilot study we combined stimulated Raman scattering microscopy with transcriptome and genome sequencing through *in situ* micro-dissection. We have shown that SMD-Seq can readily detect morphology matched transcriptional variation and chromosomal alteration, and it is capable of discovering gene fusion events underlying different histological features. This approach has potential to extend to single cell imaging and micro-dissection, and the analysis of epigenetic information. In combination with deep genome sequencing, the detection of SNVs can also be incorporated into current methods. SMD-Seq offers a new possibility in cancer research, with the integrated analysis of histology, transcriptome and genome, it will enable a more comprehensive understanding of the tumorigenesis process and diagnosis basis.

References:

1. Mardis, E.R. & Wilson, R.K. Cancer genome sequencing: A review. *Hum. Mol. Genet.* **18**, 163–168 (2009).
2. Meldrum, C., Doyle, M. & Tothill, R.W. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin. Biochem. Rev.* **32**, 177–95 (2011).
3. Chowdhury, N. Histopathological and genomic grading provide complementary prognostic information in breast cancer: a study on publicly available datasets. *Patholog. Res. Int.* **2011**, 890938 (2011).
4. Kloth, M. & Buettner, R. Changing Histopathological Diagnostics by Genome-Based Tumor Classification. *Genes (Basel)*. **5**, 444–459 (2014).
5. Fan, Y.B., Ye, L., Wang, T.Y. & Wu, G.P. Correlation between morphology and human telomerase gene amplification in bronchial brushing cells for the diagnosis of lung cancer. *Diagn. Cytopathol.* **38**, 402–406 (2010).
6. Stephens, P.J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144**, 27-40(2011).
7. Patel, A.P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401(2014).
8. Lasken, R.S. & Stockwell, T.B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol*. **7**, 1(2007).
9. Maher, C.A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101(2009).
10. Kannana, K., *et al.* Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc. Natl. Acad. Sci.* **108**, 9172-9177(2011).

11. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

12. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–9 (2015).

13. Lee, M.C.W. *et al*. Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc. Natl. Acad. Sci.* **111**, E4726–E4735 (2014).

14. Navin, N. *et al.* Tumor evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

15. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Nature* **338**, 1622-1626(2012).

16. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777-782(2012).

17. Yuan, Y. *et al.* Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling. *Sci. Transl. Med.* **4**, 157ra143–157ra143 (2012).

18. Simone, N.L., Bonner, R.F., Gillespie, J.W., Emmert-Buck, M.R. & Liotta, L.A. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272-276 (1998).

19. Wang, H. *et al.* Histological staining methods preparatory to laser capture microdissection significantly affect the integrity of the cellular RNA. *BMC Genomics* **7**, 97 (2006).

20. Freudiger, C.W. *et al*. Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. *Science* **322**, 1857–1861 (2008).

21. Ji, M. *et al.* Detection of human brain tumor infiltration with quantitative stimulated Raman scattering microscopy. *Sci. Trans. Med.* **7**, 309ra163 (2015).

22. Ji, M. *et al*. Rapid, label-free detection of brain tumors with stimulated Raman scattering microscopy. *Sci. Trans. Med.* **5**, 201ra119 (2013).

23. Yu, Z. *et al*. Label-free chemical imaging in vivo: three-dimensional non-invasive microscopic observation of amphioxus notochord through stimulated Raman scattering (SRS). *Chemical Science* **3**, 2646 (2012).

24. Dalal, N. & Triggs, B. Histograms of Oriented Gradients for Human Detection. *CVPR '05 Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **1**, 886–893 (2005).

25. Huang, D.W., Sherman, B.T. & Lempicki, R.a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

26. De Craene, B. & Berx, G. Regulatory networks defining EMT during cancer initiation and progression. *Nat. Rev. Cancer* **13**, 97–110 (2013).

27. Worst, T.S. *et al.* IL1RN and KRT13 Expression in Bladder Cancer: Association with Pathologic Characteristics and Smoking Status. *Adv. Urol.* **2014**, 1–6 (2014).

28. Gao, S. *et al.* Overexpression of protease nexin-1 mRNA and protein in oral squamous cell carcinomas. *Oral Oncol.* **44**, 309–313 (2008).

29. Tang, X. *et al.* A mechanically-induced colon cancer cell population shows increased metastatic potential. *Mol. Cancer* **13**, 131 (2014).

30. Micalizzi, D.S. *et al.* The Six1 homeoprotein induces human mammary carcinoma cells to undergo epithelial-mesenchymal transition and metastasis in mice through increasing TGF-β signaling. *J. Clin. Invest.* **119**, 2678–2690 (2009).

31. Pettus, J.R. *et al.* Multiple kallikrein (KLK 5, 7, 8, and 10) expression in squamous cell carcinoma of the oral cavity. *Histology and histopathology* **24**, 197-207 (2009).

32. Troy, J.D. *et al.* Polymorphisms in NAT2 and GSTP1 are associated with survival in oral and oropharyngeal cancer. *Cancer Epidemiol.* **37**, 505–511 (2013).

33. Ma, H. *et al*. Decreased expression of glutathione S-transferase pi correlates with poorly differentiated grade in patients with oral squamous cell carcinoma. *J. Oral Pathol. Med.* **44**, 193–200 (2015).

34. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).

35. Ben-Neriah, Y., Daley, G.Q., Mes-Masson, A.M., Witte, O.N. & Baltimore, D. The chronic myelogenous leukemia-specific P210 protein is the product of the bcr/abl hybrid gene. *Science* **233**, 212–214 (1986).

36. Kwak, E.L. *et al.* Anaplastic Lymphoma Kinase Inhibition in Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **363**, 1693–1703 (2010).

37. Lawrence, M.S. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).

38. Hammerman, P.S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

39. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).

40. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

41. Troy, J.D. *et al.* Polymorphisms in NAT2 and GSTP1 are associated with survival in oral and oropharyngeal cancer. *Cancer epidemiology* **37**, 505-511 (2013).

42. Kuga, T. *et al.* A novel mechanism of keratin cytoskeleton organization through casein kinase I alpha and FAM83H in colorectal cancer. *J. Cell Sci.* **126**, 4721-4731 (2013).

43. Yamashita, S. I. *et al.* P53AIP1 expression can be a prognostic marker in non-small cell lung cancer. *Clin. Oncol-Uk.* **20**, 148-151 (2008).

44. Demirag, G.G., Sullu, Y., Gurgenyatagi, D., Okumus, N.O. & Yucel, I. Expression of Plakophilins (PKP1, PKP2, and PKP3) in Gastric Cancers. *Diagn. Pathol.* **6** (2011).

45. Sevenich, L. *et al.* Synergistic antitumor effects of combined cathepsin B and cathepsin Z deficiencies on breast cancer progression and metastasis in mice. *Proc. Natl. Acad. Sci.* **107**, 2497-2502 (2010).

46. Bengsch, F. *et al.* Cell type-dependent pathogenic functions of overexpressed human cathepsin B in murine breast cancer progression. *Oncogene* **33**, 4474-4484 (2014).

47. Castro, M.E. *et al.* PPP1CA contributes to the senescence program induced by oncogenic Ras. *Carcinogenesis* **29**, 491-499 (2008).

48. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371-381 (2015).

49. Olshen, A.B., Venkatraman, E.S. Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

50. Yee, J.Y. *et al.* Ensuring good quality rna for quantitative real-time pcr isolated from renal proximal tubular cells using laser capture microdissection. *BMC Res. Notes* **7**:62 (2014).

Materials and Methods

Study design

Our goal was to develop a new technique for obtaining high-quality histological images of OSCC while simultaneously profiling the transcriptomic and genomic variation that correlate with morphological features at the microscopic level. We first validated that simulated Raman scattering images of stain-free OSCC tissue cryosections were comparable to widely used H&E images for histopathologic applications. We compared SRS images with H&E stained images from 4 different tissue types (epithelium, gland, muscle and nerve), and verified that SRS images accurately revealed features of OSCC . The texture characteristics extracted from normal samples (n=16) and lesion samples (n=16) were evaluated in SRS images. At 3 different scales, the similarity between a stitched SRS image (FOVs, n=80) of one full slice and its corresponding H&E staining image was investigated. To prove the necessity of employing label-free histology to prevent RNA degradation, we evaluated the RNA recovery between H&E stained cryosections (n=24) and unstained cryosections (n=48). The SRS system was then optimized to perform fast, successive imaging and *in situ* micro-dissection (SMD-Seq). We applied SMD-Seq to 13 cryosections from four patients who suffered various stages of OSCC at different ages and with different genders. All the biopsies were collected by protocols reviewed by and approved by the Ethics Committee of Peking University School and the Hospital of Stomatology

(PKUSSIRB-201418116). We collected 28 *in situ* micro-dissection samples for sequencing, and 27 samples passed the quality check for further analysis, including 12 normal (5 muscle; 2 gland; 5 epithelium) and 9 cancer samples by RNA-Seq, 13 normal (muscle, n=5; gland, n=2; epithelium, n=6) and 8 cancer samples by DNA-Seq. Bioinformatics analysis of the transcriptome and genome demonstrated the intra- and inter- tumor heterogeneity, and the correlative analysis between morphology and genome/transcriptome sequencing in the samples indicated intrinsic genetic mosaicism.

Stimulated Raman scattering (SRS) microscope

The home-built SRS system used a pump laser integrated optical parametric oscillator (picoEmerald, APE, Germany). It provided two spatially and temporally overlapped pulse trains, with the synchronized repetition rate of 80 MHz. One beam, fixed at 1064 nm, was used as the Stokes light. The other beam, tunable from 780 to 990 nm, served as the pump light. The intensity of the Stokes beam was modulated at 20.2 MHz by a resonant electro-optical modulator (EOM). The overlapped lights were directed into an inverted multi-photon scanning microscope (FV1000, Olympus, Japan). The collinear laser beams were focused into the sample by a 20× objective (UPlanSAPO, NA 0.75, Olympus, Japan). Transmitted light was collected by a condenser (NA 0.9, Olympus, Japan). After filtering out the Stokes beam, the pump beam was directed onto a large area photo diode (FDS1010, Thorlabs, USA). The voltage from photo diode was sent into lock-in amplifier (HF2LI, Zurich Instruments, Switzerland) to extract the SRS signal. Image was reconstructed through software provided by manufacture (FV10ASW, Olympus, Japan).

SRS imaging and laser micro-dissection in situ

Each slide was surveyed with SRS microscopy in two channels immediately after sectioning. The two Raman bands are 2850 and 2950 cm^{-1}, representing CH_2 symmetric vibration and CH_3 vibration, respectively. *In situ* micro-dissection was performed right after image acquisition. Details and parameters were described and discussed in Supplementary Materials and Methods.

Image analysis

Dual-color SRS images were analyzed with Matlab (Mathwork, USA) and R, as describe in Supplementary Materials and Methods.

Transcriptome and genome sequencing

For each tissue section, we selected at least one dissected region of tumor, and two dissected areas of similar size from normal tissues, one of them from the epithelium (the origin of this tumorigenesis) and the other from gland or muscle. The dissected samples were put into lysis buffer[51] separately and immediately centrifuged at 13000 rpm for 30s. After lysis, each sample was equally divided into two aliquots for RNA-Seq and genomic DNA sequencing, respectively. The protocol of RNA-Seq was adapted from the pipeline of single-cell transcriptome analysis[51]. In brief, mRNA was reverse transcribed into first strand cDNA with polyT primer which has an anchor

sequence. After other used primers were digested, polyA was added to the 3' end of cDNA and second strand cDNA was formed and amplified with polyT primer with another anchor sequence by PCR. We employed degenerate oligonucleotide primed PCR (DOP-PCR) for amplifying the whole genome of each lysed tissue sample by the GenomePlex Single Cell Whole Genome Amplification Kit (WAG4-50RXN, Sigma-Aldrich, USA). For each sample, 50 ng of amplified genomic DNA and cDNA were used as the start amount of libraries preparation, separately. The pair-end sequencing libraries with ~300 bp insert size were constructed following the instructions of NEBNext Ultra DNA Library Prep Kit for Illumina (E7370, New England Biolabs, USA). Illumina HiSeq 2500 systems were used for sequencing.

Sequencing data analysis

Adaptor contamination and low-quality reads (phred quality< 20) were discarded from the raw data. Only samples with coefficient of variation (CV) of reads count per 1M bin<0.25 (genomic DNA) and gene number more than 6000 (FPKM>0.1, RNA) were kept for analysis. For RNA-Seq data, TopHat (v2.0.10) were used for sequencing alignment. Reference genome assembly hg19 and gene annotation files were downloaded from UCSC Genome Browser. FPKM values used for analyses were generated by Cufflinks (v2.1.1), and Cuffdiff (v2.2.1) was used for gene expression levels comparison. Significantly different expressed genes between muscle, gland, epithelium and cancer were selected under the criteria that p value < 0.05 and |log2 (fold change)| >1. Gene functional annotation was performed by The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.717[25]. The purity of tumor samples were estimated by ESTMATE[46] with gene expression data. Gene fusion analysis were carried out by FusionCatcher (v0.99.4a)[53] with four mapping tools (Bowtie, Bowtie2, BLAT, STAR). Matched normal samples were used for each patient to exclude the fusion genes that are also found in normal samples. Under following situations the fusion were discarded:(1) both fusion genes are mutual paralogs;(2) one or both of the fusion genes were pseudogene;(3) reported only by one mapping tool or reported by 2 mapping tools only once;(4) no known genes existed in between the fusion genes;(5) the distance between both genes were less than 100 kbp. Under this criteria, 24 fusion genes were discovered with more than 10 paired reads spanning two different genes sequences. The circular diagram of fusion gene was generated by CIRCOS (v0.67-7)[54]. RNA-Seq data was used for variant calling by GATK (v3.4-0) according to GATK Best Practices recommendations[55]. We performed duplicate removal, SplitNCigarReads, base quality score recalibration before SNP calling, and filtered out SNPs by Fisher Strand values (FS>30.0), Qual By Depth values (QD<2.0) and sequencing depth passing the quality filter (DP<10). Annotation of SNPs was performed by SnpEff (v4.0)[56]. Significantly mutated genes in HNSCC (Head and Neck Squamous Cell Carcinoma) were inferred from COSMIC (Catalogue of Somatic Mutations in Cancer) and a comprehensive previous study[57]. Spearman correlation coefficient was computed between tissue samples by function 'cor' in R. The unsupervised hierarchical clustering was performed by the function 'pheatmap' of package 'pheatmap' in R. and the method of measuring the distance in clustering columns was 'manhattan'. 3D PCA plot was generated by R package 'scatterplot3d'.

The sequencing depth of DNA-Seq was ~0.1×. Genomic DNA sequencing reads were mapped to reference genome by bowtie2 (v2.2.3)[58]. After duplication removal of mappable reads, the counts of aligned reads were calculated in each 1M bin along the genome (Figure 5A, C, D, E). For each bin, the read count of each tumor sample was normalized by sequencing depth and the median read count of all normal tissue samples, and the generated copy number went through segmentation by Circular Binary Segmentation[59] (the significance level was set as 0.05). The mean gene expression values of cancer samples within each segment were also calculated and normalized by mean expression values of normal samples which had corresponding qualified (CV<0.25) gDNA reads (Figure 5C, D, E). Function "pheatmap" in R was adopted for CNVs clustering (Figure 5B). GISTIC 2.030 was adopted to analyze the significantly reoccurring focal alterations for the gDNA segmented data.

51. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5**, 516–535 (2010).
52. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
53. Nicorici, D. *et al.* FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv* (2014). doi:10.1101/011650
54. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
55. Van der Auwera, G.A. *et al.* in Current Protocols in Bioinformatics 11, 11.10.1–11.10.33 (John Wiley & Sons, Inc., 2013).
56. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*. **6**, 80–92 (2012).
57. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* **91**, 355-359 (2004).
58. Langmead, B. & Salzberg, S.L., Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
59. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

Acknowledgments: We thank BIOPIC and Peking University Sequencing Center for the experimental help and sequencing service.

Funding: Supported by Ministry of Science and Technology of China (2015AA0200601), National Natural Science Foundation of China (21327808 and 21525521), and the stimulation grant for the collaboration between fundamental sciences and clinical researches.

Author contributions: Y. H., T. L. conceived the project; Y. H., T. L., T. C., C. C., and J. Z. designed all experiments; T. C. performed SRS imaging, *in situ*

microdissection and image analysis; C. C. performed molecular biology experiments and bioinformatics analysis. J. Z. performed histology identification, H&E staining and immunofluorescent staining. T. C., C. C., J. Z., Y. H. and A. M. S wrote the manuscript.

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: The sequence of SMD-Seq samples have been deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/) under accession number SRP075236.

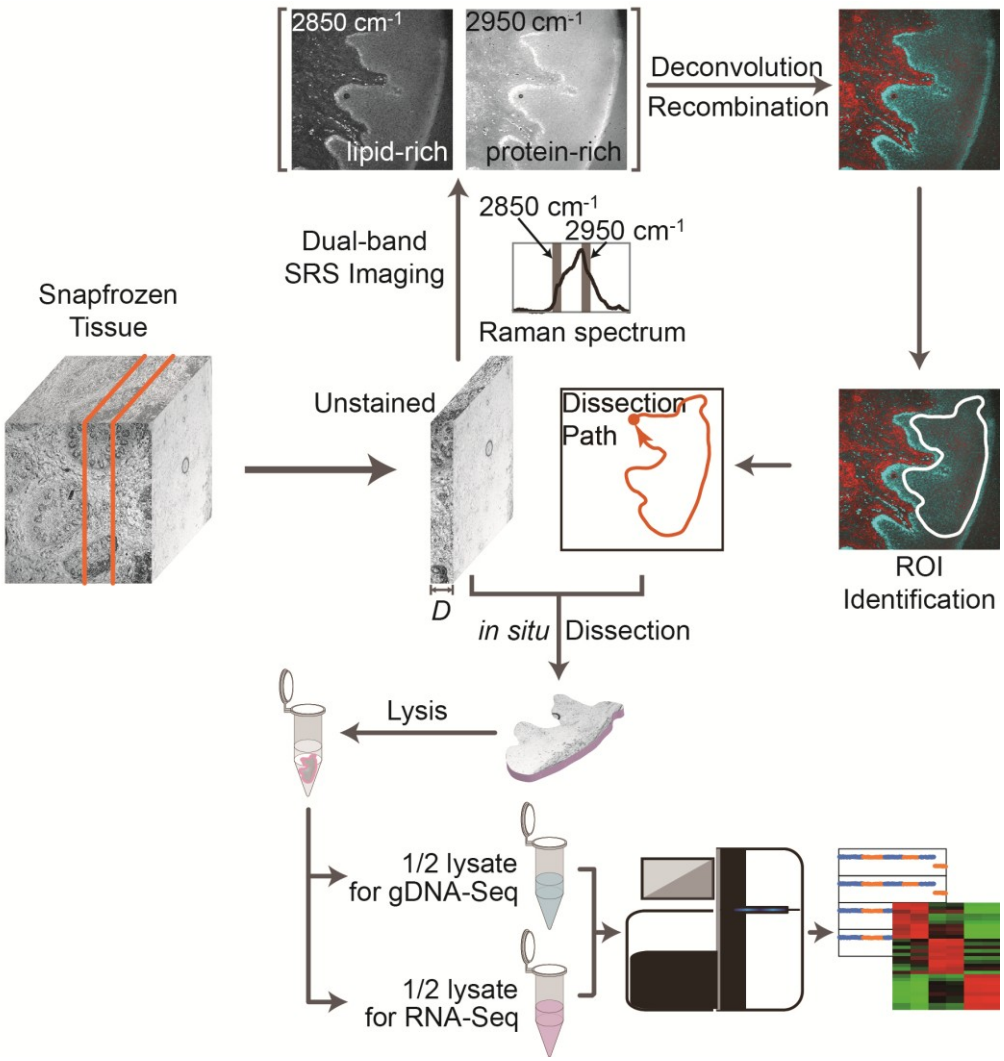

Fig. 1. Illustration of SMD-Seq. Snap-frozen tissues were dissected into 30 μm sections and immediately imaged with SRS microscopy. The SRS images constructed by deconvolution and recombination processes reflected the features of imaged tissue. After regions of interest were identified on the SRS images, the dissection path was determined and samples were dissected *in situ* by laser. Dissected samples were then lysed and aliquoted for following DNA and RNA sequencing.

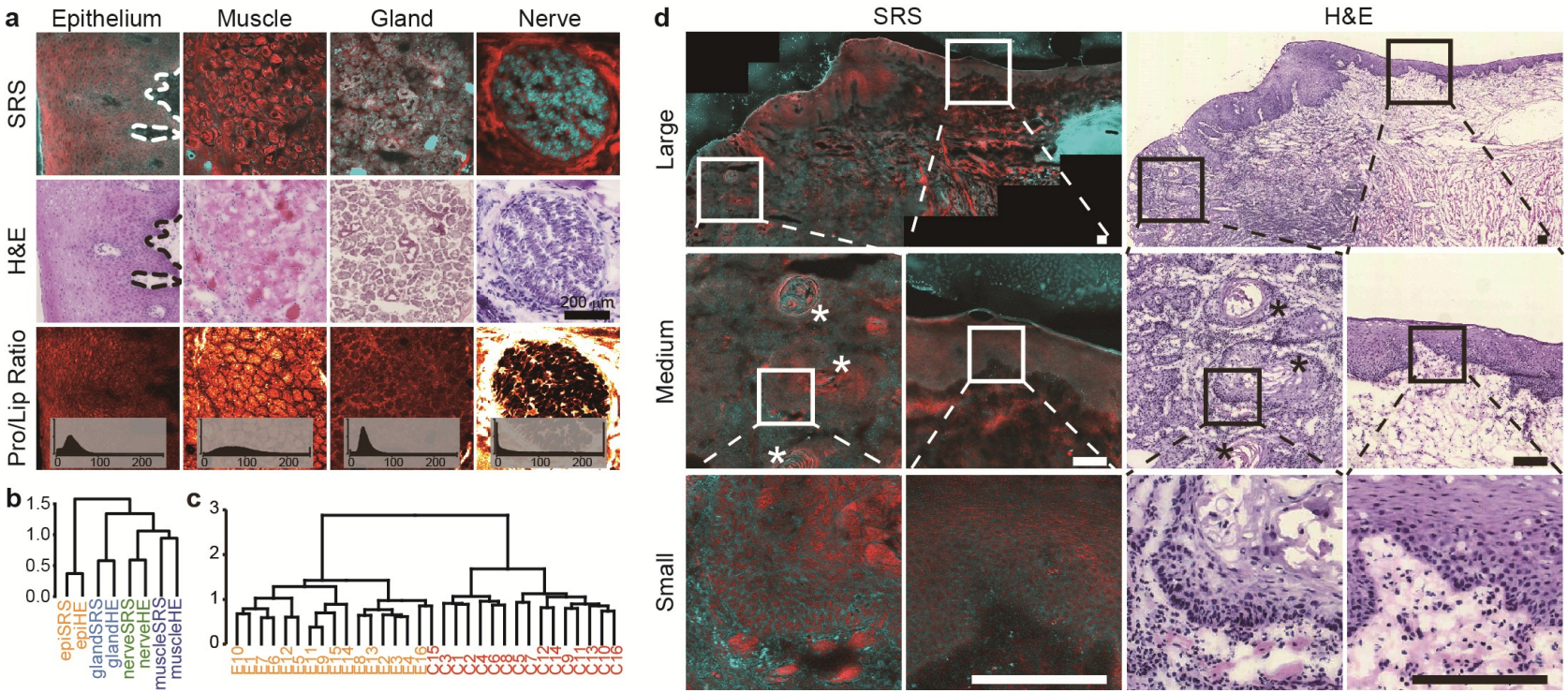

Fig. 2. SRS images of different tissue regions on-slide. (**A**) SRS images (top row) of different oral tissues, with corresponding H&E images (middle row), and protein to lipid content ratio (PLR, bottom row). The histograms showed the distribution of PLR. (**B**) Unsupervised clustering of different tissue types' HOG features in both SRS and H&E images. (**C**) Unsupervised clustering of HOG features extracted from 16 oral cancer (C1-C16) and 16 epithelium samples (E1-E16). (**D**) Stitched SRS image of one full slice and its corresponding H&E staining image. SRS and H&E images appeared in high similarity at different scales. In middle and bottom rows, images in the left column were oral cancer, the right column represented epithelium. Asterisks were labeled on 'keratinized pearls', a specific structure in OSCC. Scale bars are 200 μm.

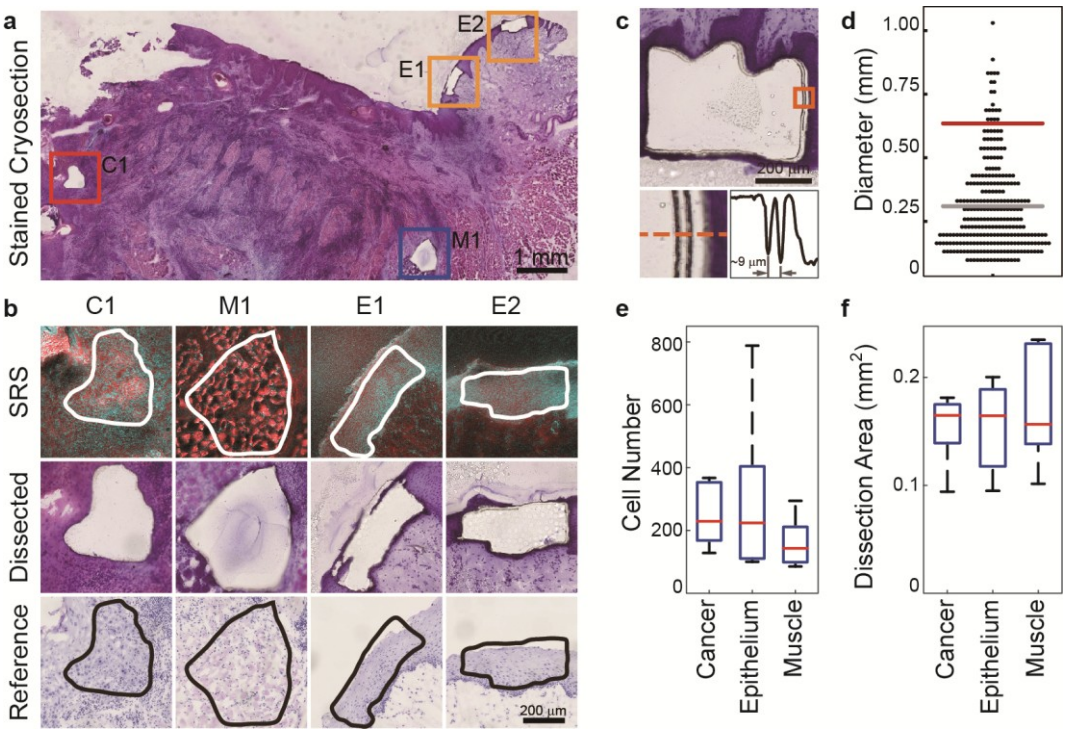

Fig. 3. Micro-dissections of tissue slices. (**A**) A single 30 µm cryo-section stained with H&E after SRS imaging and micro-dissection. Red, orange and white squares highlighted the dissected cancer (C1), epithelium (E1 and E2) and muscle (M1) regions. (**B**) SRS images before dissection (top row), H&E staining images after dissection (middle row), and H&E references (bottom row). White and black curves in SRS and reference images highlighted the dissected regions from the cryo-section shown in (A). The dissection regions were identified by a pathologist based on SRS images. (**C**) The diameter distribution of cancer nests. Gray bar showed the average size, red bar showed the width of 20× objective's field of view. (**D**) The linewidth of dissection path. Part of the path (orange box) is selected to measure linewidth of incision. (**E**) Cell number and (**F**) area in each dissected tissue.

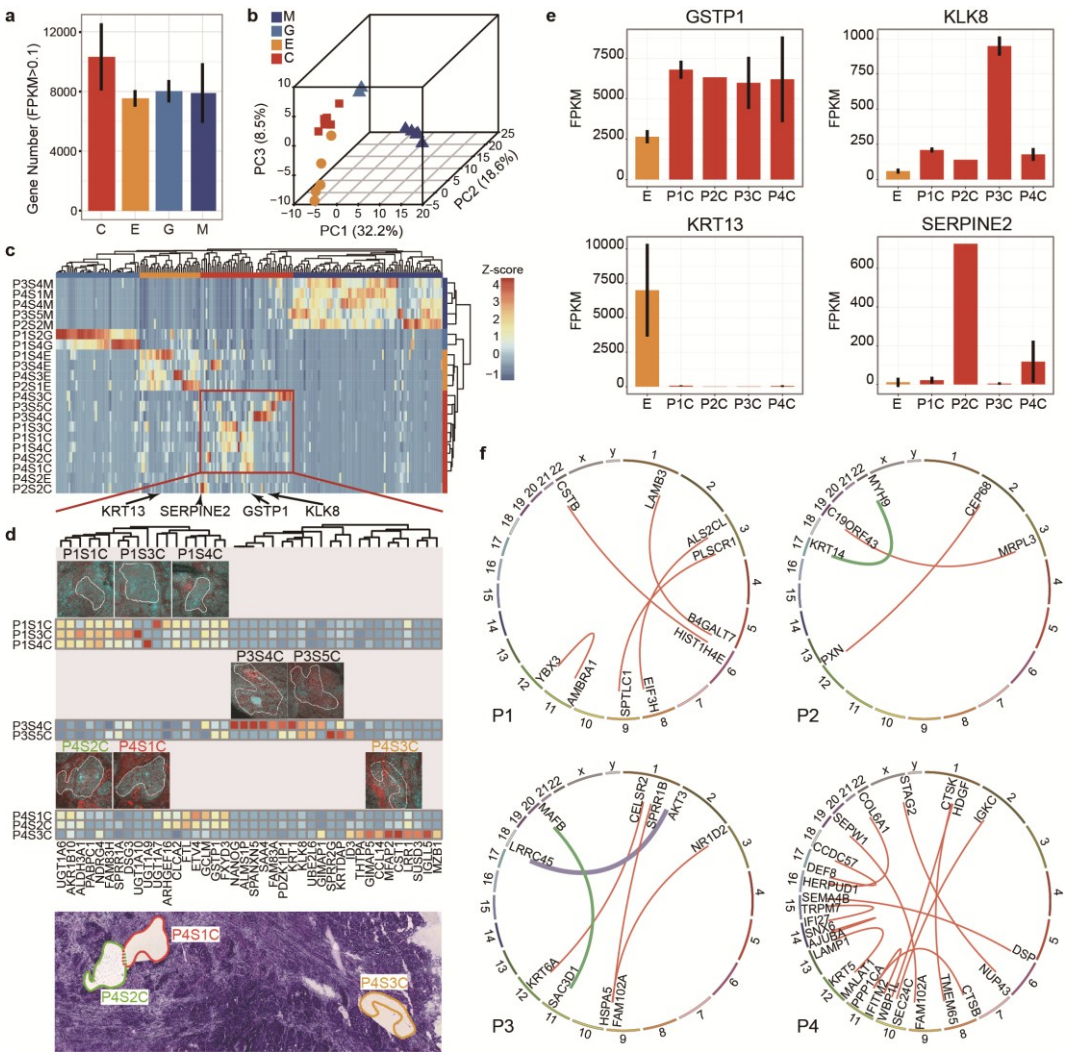

Fig. 4. Transcriptome analysis of laser dissected gland (G, light blue), muscle (M, blue), epithelium (E, orange) and cancer tissues (C, red). (**A**) Detected gene number (FPKM > 0.1) of 21 samples dissected from four patients' slices. (**B**) Principal Component Analysis (PCA) results of expressed genes. (**C**) Unsupervised hierarchical clustering by 217 differently expressed genes. Both samples and genes were clustered into four groups. (**D**) Details of unsupervised hierarchical clustering of differently expressed genes in cancer samples with at least two dissected slices. Highly expressed genes of patient P1 (top), P3 (middle) and P4 (bottom) and their corresponded SRS morphology images were shown. The H&E stained cryo-section of P4 was demonstrated, and the locations and dissection paths of P4S1C (red), P4S2C (green) and P4S3C (orange) were labelled. (**E**) Gene expression levels of GSTP1, KRT13, KLK8 and SERPINE2 between normal epithelium (E, orange) and cancerous epithelium (C, red). (**F**) Detected gene fusion events of four patients. Orange lines indicated fusion genes with at least 10 span pair reads (Online Methods). The green and purple lines represented oncogenes involved gene fusions.

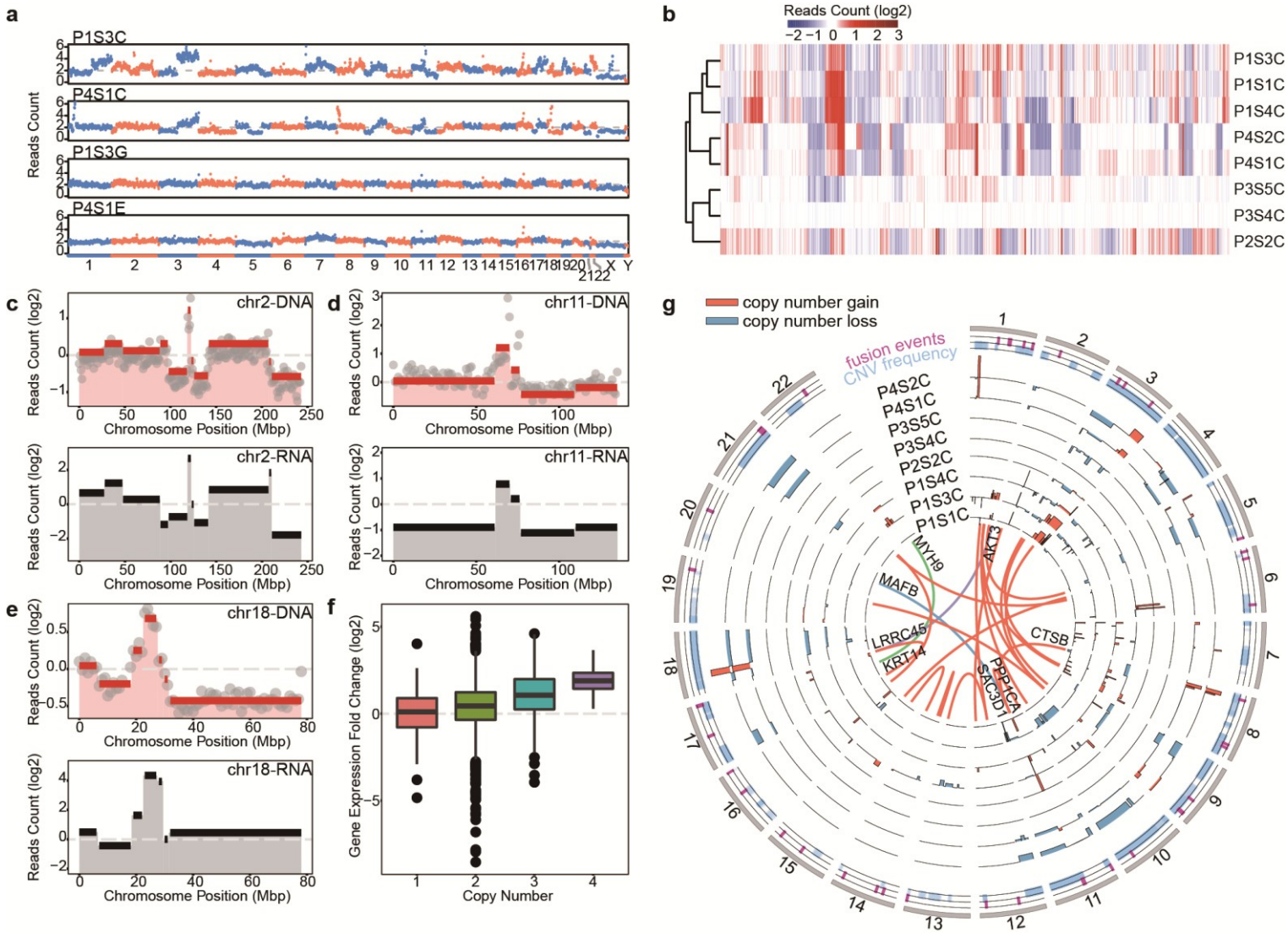
a
P1S3C
P4S1C
P1S3G
P4S1E
Reads Count
b
Reads Count (log2)
P1S3C
P1S1C
P1S4C
P4S2C
P4S1C
P3S5C
P3S4C
P2S2C
c
chr2-DNA
chr2-RNA
d
chr11-DNA
chr11-RNA
e
chr18-DNA
chr18-RNA
Reads Count (log2)
Chromosome Position (Mbp)
f
Gene Expression Fold Change (log2)
Copy Number
g
copy number gain
copy number loss
fusion events
CNV frequency
P4S2C
P4S1C
P3S5C
P3S4C
P2S2C
P1S4C
P1S3C
P1S1C
MYH9
MAFB
LRRC45
KRT14
AKT3
CTSB
PPP1CA
SAC3D1

Fig. 5. Genomic and transcriptomic analysis of laser dissected tissue samples. (**A**) Raw reads count across the whole genome of two paired cancer-normal tissues dissected from the same slice. Chromosome numbers were labeled at the bottom. (**B**) Unsupervised clustering of normalized reads count of all the dissected cancer samples. (**C**, **D**) Top panels demonstrated the normalized read count (grey dots) and copy numbers (red lines) identified by CBS algorithm in chromosome 2 and 11 of sample P1S1C and P1S3C, respectively. Mean expression levels within the same segment were shown in black lines in the bottom panel. (**E**) Averaged read counts of all the cancer samples (top) and their corresponding gene expression values (bottom) of chromosome 18. Genome and transcriptome variations of other chromosomes and samples were shown in Fig. S16, 18-19. (**F**) Averaged gene expression fold changes were computed per 1M bin across the whole genome, and plotted against copy numbers. (**G**) The distribution of CNV and fusion genes across the genome of all the cancer samples. Orange and blue bars indicated the copy number gains and loss, respectively. Orange lines in the inner circle indicated fusion genes with at least 10 span pair reads, and the green and purple lines represented oncogenes involved gene fusions. The position of fusion events (magenta) and CNVs (light blue) of all the cancer samples were shown in the 2 outmost circles.