Xi et al., Noninvasive genomic profiling of somatic mutations in oral cavity cancers

## SUPPLEMENTAL INFORMATION

## TABLE OF CONTENTS

Supplemental figures

Supplemental tables

## SUPPLEMENTAL METHODS

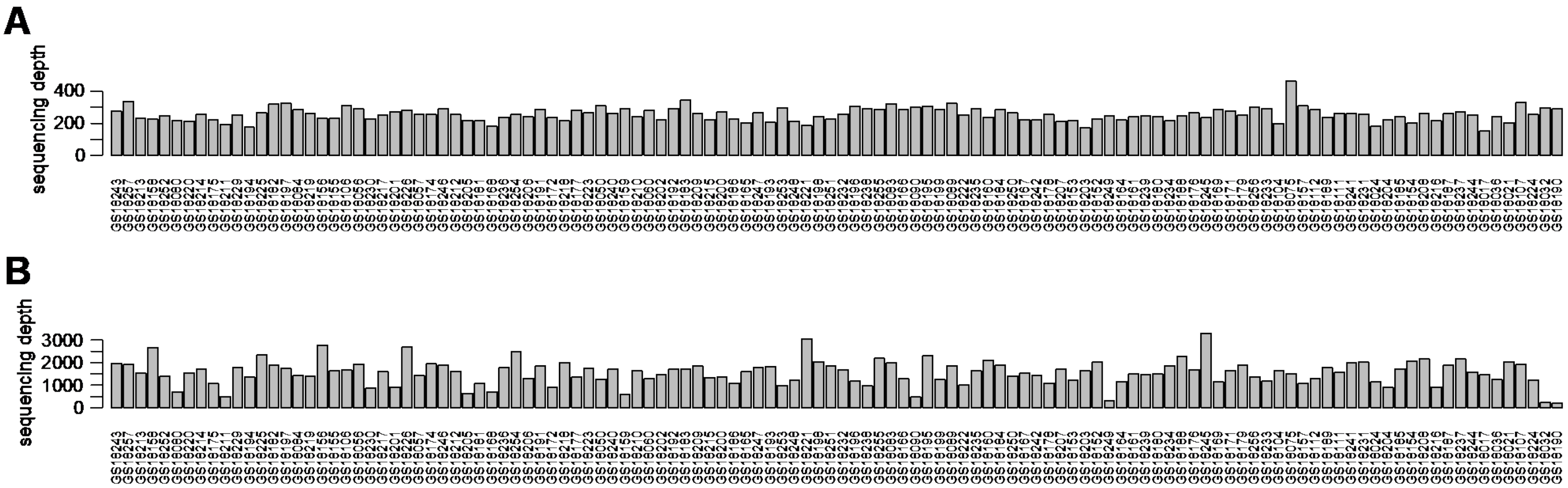
### *Additional approaches for clinical and genomic sensitivity analysis*

As described in the main text, we used two additional analytical approaches to detect somatic variant calls. In the first approach, we used downsampling to compensate for the absence of matched normal germline samples. Germline variants were called from downsampled ORS sequence data to facilitate identification of candidate somatic variants in ORS samples using VarScan 2 "somatic" functions (1). The ORS data were downsampled to the same sequencing depth as the matched fresh-frozen OSCC. The somatic calling parameters were set to "--min-var-freq 0.05". The called variants were first filtered by VarScan 2 "processSomatic" function with parameters "--p-value 0.01" to obtain high confidence somatic calls and then were annotated using ANNOVAR version 521 (2). Only somatic variants that have known COSMIC annotations were selected for downstream analysis. Synonymous variants and variants with known SNV annotations in the dbSNP138 database (3-5) were removed, unless they also have known COSMIC annotations (6). For ORS, variants with VAF >40% were also removed, and we applied the same filtering criterion for COSMIC and dbSNP annotations as for fresh frozen OSCC.

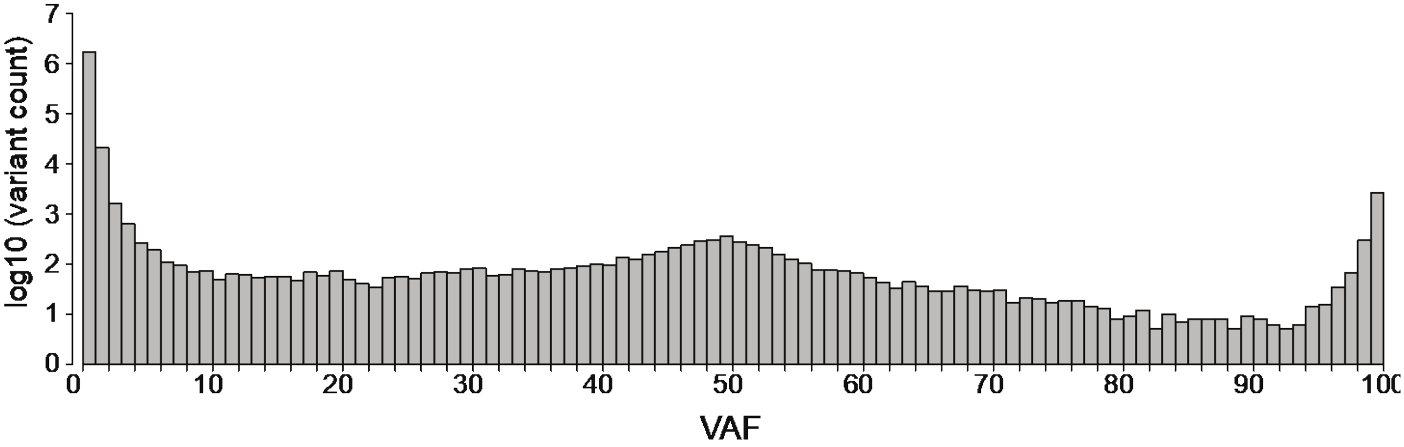
In the second additional approach, we used an alternative analytical pipeline to account for a possible high false-positive rate based upon a VAF cutoff alone. We applied a prediction model to distinguish true variants from possible sequencing errors. High confidence tumor variants were defined as variants that have known COSMIC annotations. We developed a naïve Bayesian classifier to model the probability of a specific variant being a true positive vs. a call due to sequencing error, using a combined set of features from behavioral and genomic characteristics. These included measures of tobacco and alcohol exposure (e.g., never, former, current use) as possible contributors to low VAF due to field cancerization, genomic site-specific nucleotide changes, functional amino acid changes, and the gene-specific characteristics for genes in the target panel. Variants in the ORS that were classified as false positives by this naïve Bayesian model were further excluded before applying the VAF cutoff to detect somatic variants in matched fresh-frozen OSCC. Assay performance (as measured by VAF cutpoint, sensitivity, specificity, AUC, precision) with and without application of the Bayesian model was compared.

## SUPPLEMENTAL REFERENCES

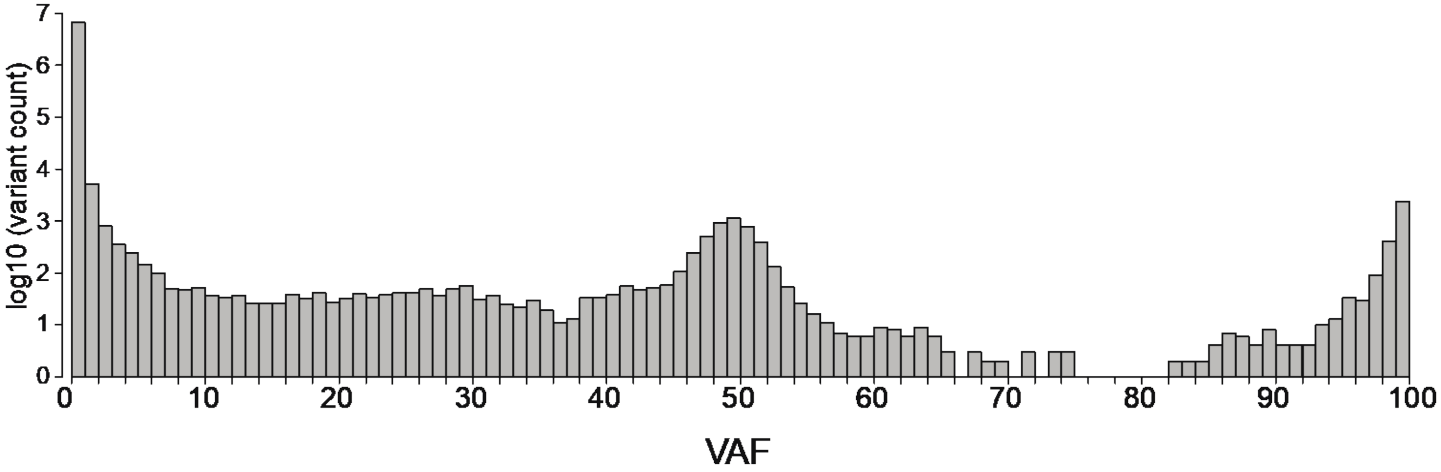
1. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22(3):568-76 doi 10.1101/gr.129684.111.
2. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164 doi 10.1093/nar/gkq603.
3. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 1999;9(8):677-9.
4. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29(1):308-11 doi 10.1093/nar/29.1.308.
5. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 2000;28(1):352-5 doi 10.1093/nar/28.1.352.
6. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 2019;47(D1):D941-d7 doi 10.1093/nar/gky1015.



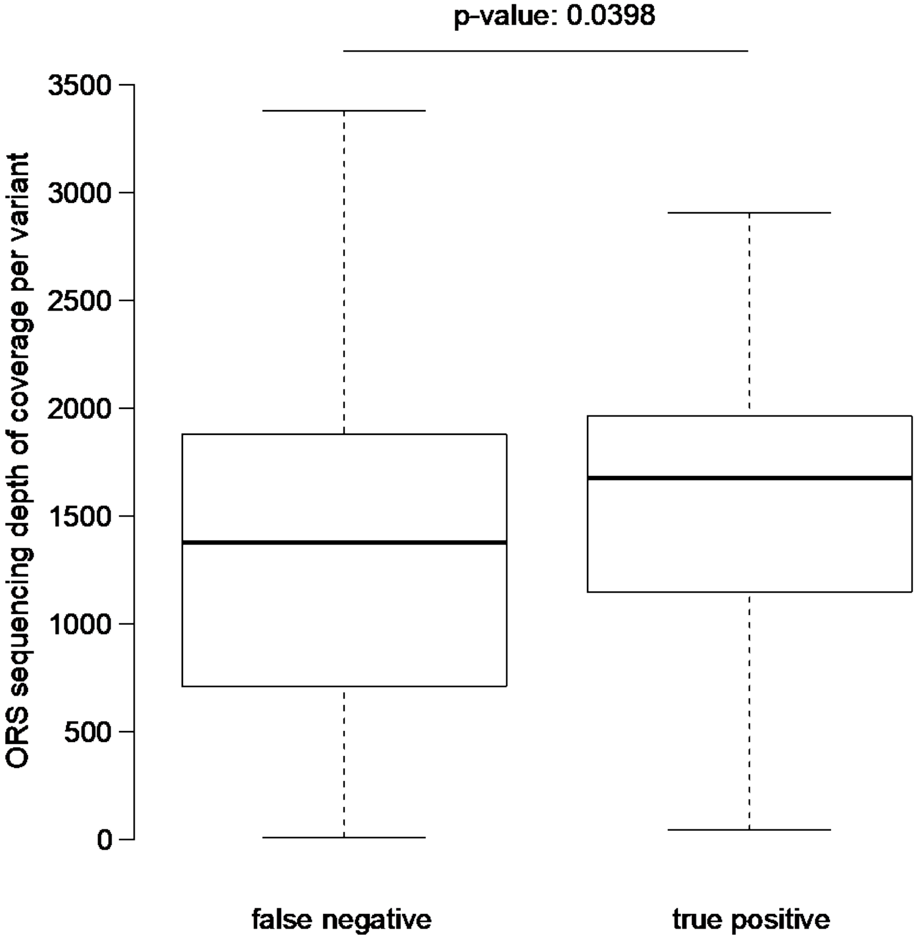
Supp. Fig. S1. **Sequencing coverage in samples.** Shown here are median depths of sequencing coverage (*y-axis*) across the targeted region for 120 patient sample pairs (*x-axis*): (A) Fresh-frozen OSCC samples; (B) ORS. Sample pairs from two patients with the lowest ORS sequencing coverage (i.e., GS18190 and GS18032) were excluded from further analysis, resulting in 118 OSCC sample pairs analyzed further.



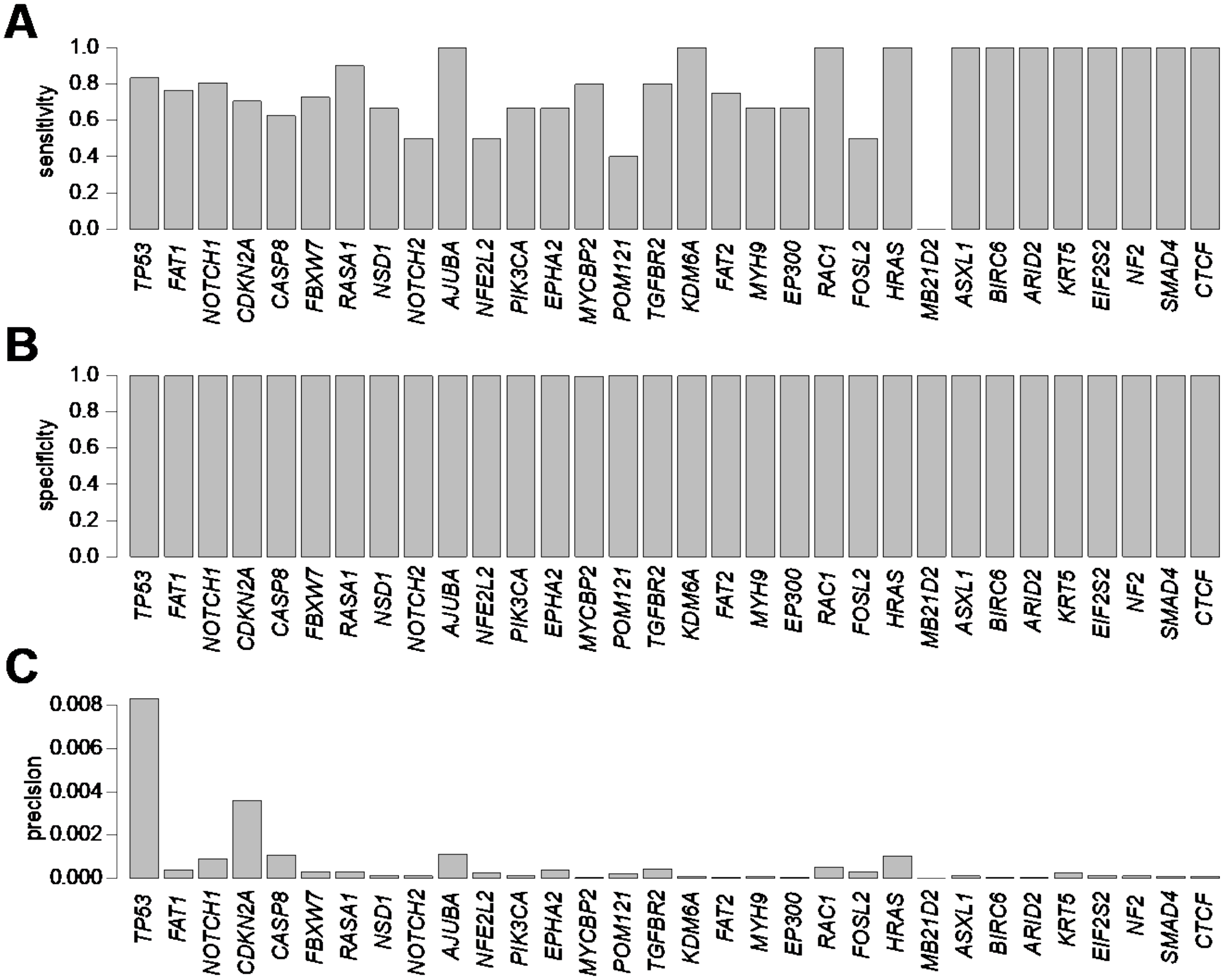
Supp. Fig. S2. **Distribution of VAFs in fresh-frozen OSCC.** Histogram displays the counts of all detected variants (log10 scale, *y-axis*) with indicated VAFs (*percentages, x-axis*). A total of 1,308,131 variants were identified across 118 OSCC samples. Most variants' VAFs were < 1%.



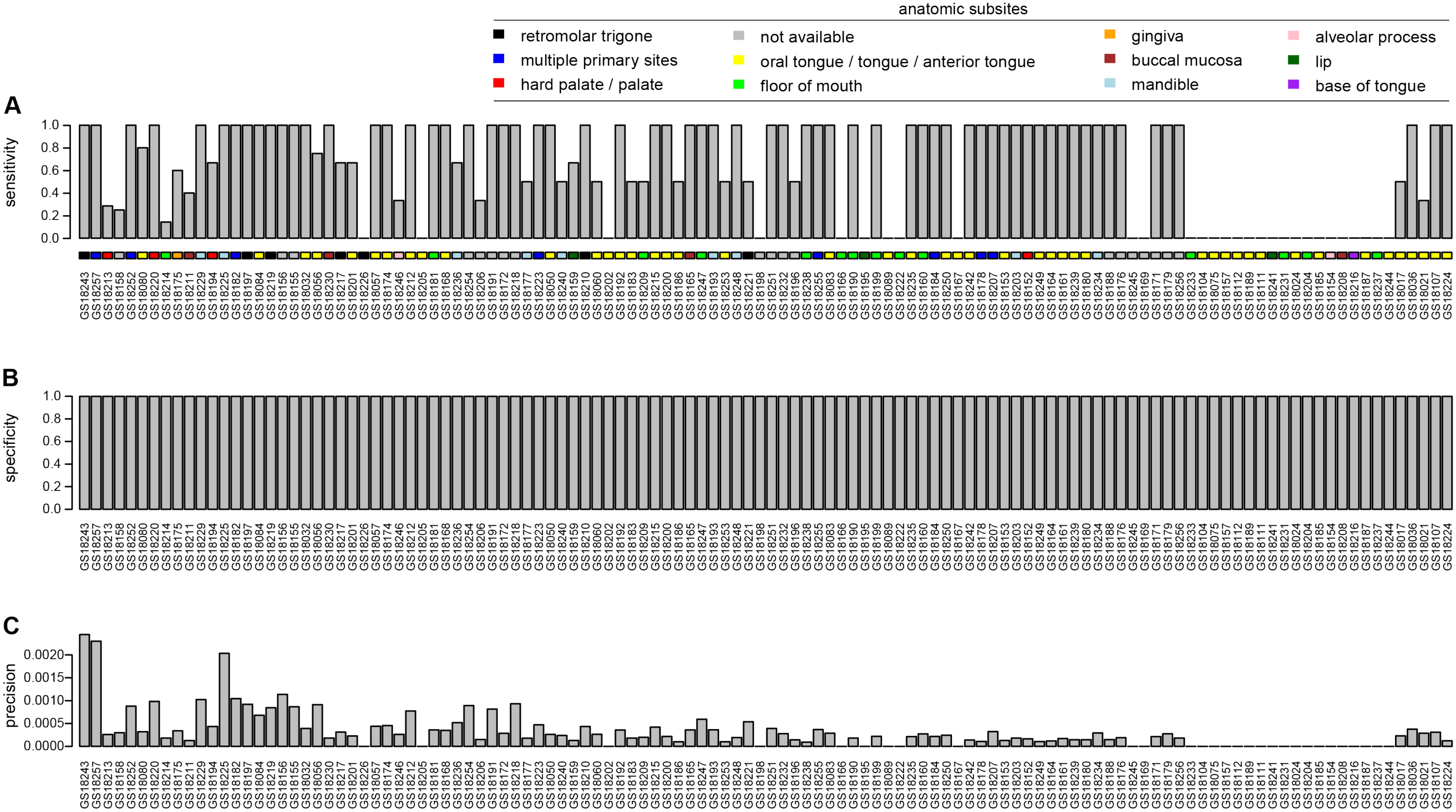
Supp. Fig. S3. **Distribution of VAFs in ORS.** Histogram displays the counts of all 5,046,099 variants (log10 scale, *y-axis*) identified across 118 OSCC samples, with indicated VAFs (*percentages, x-axis*). Most variants' VAFs were < 1%.



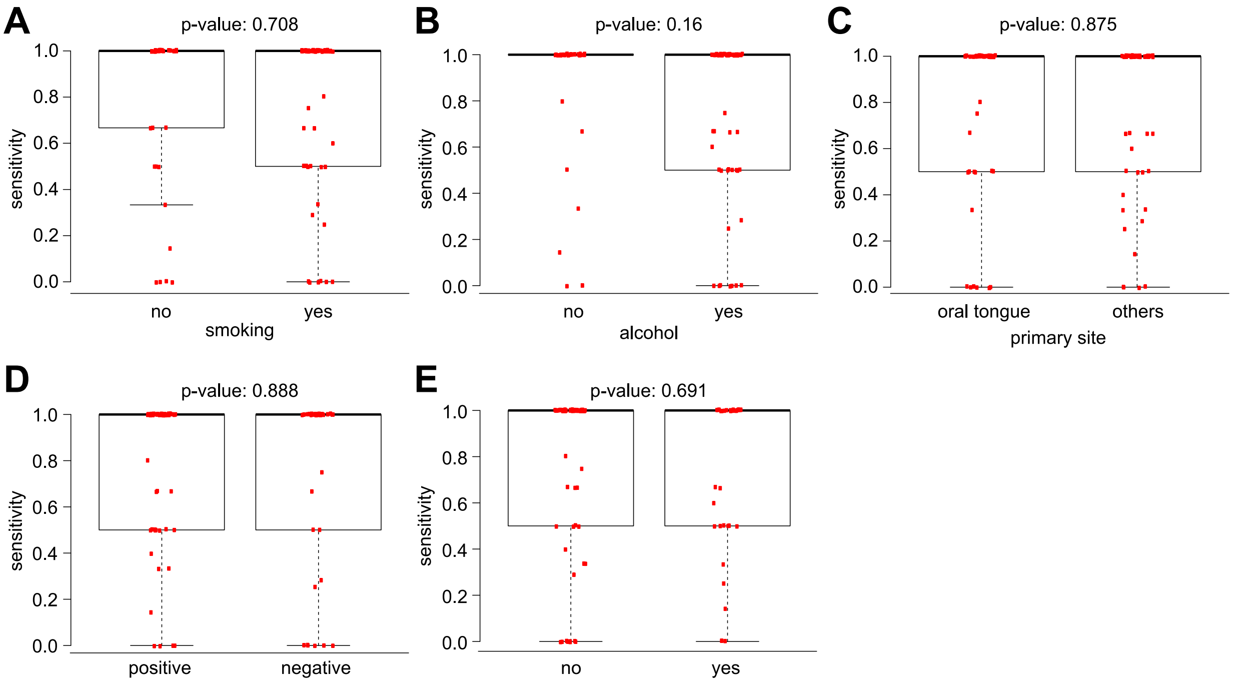
Supp. Fig. 4. **Association between sequencing depth of coverage and detection of true-positive variant calls in ORS.** Box-and-whiskers plot shows that the distribution of sequencing depths of coverage per variant in ORS samples for true-positive variants (*right*), as detected with VAF > 0.10%, is significantly higher overall than that for false-negative calls (*left*; p = 0.040). *Thick horizontal line*, median; *box*, first and third quartiles; *whiskers*, minimum and maximum.



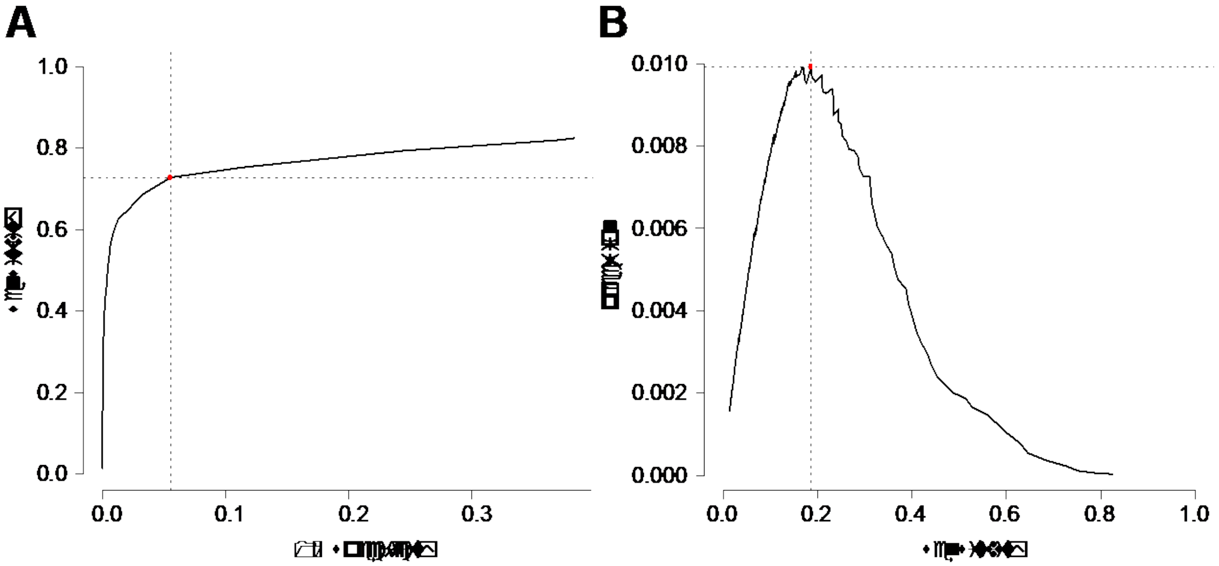
Supp. Fig. S5. **Gene-level sensitivity, specificity, and precision in ORS.** Bar graphs show (A) overall optimized sensitivity, (B) specificity, and (C) precision (*y-axis*) for each of the 32 targeted genes (*x-axis*) in which somatic variants were detected in any of the ORS. *X-axis, left to right*, genes ordered by frequency of detected somatic variants in OSCC.



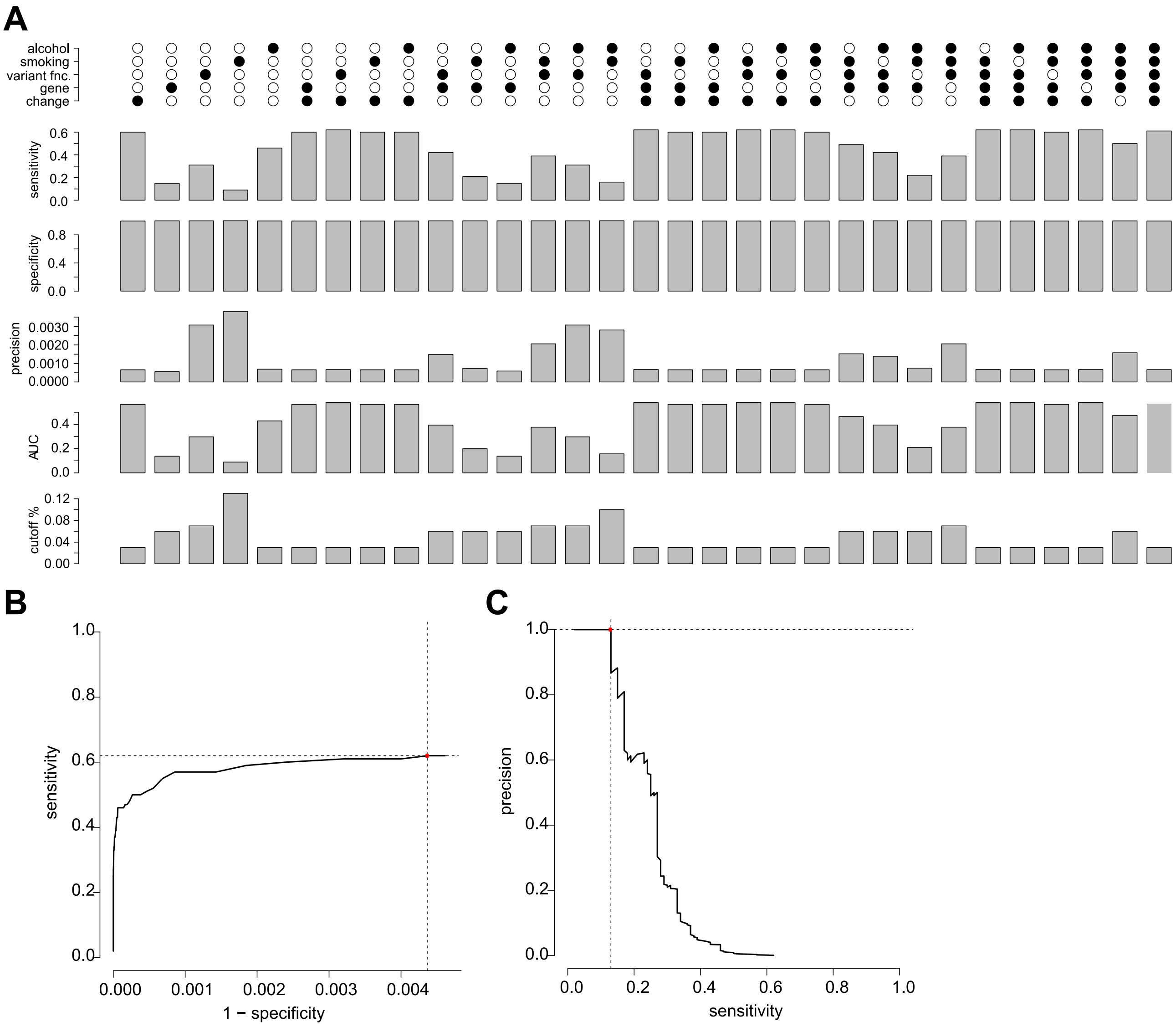
Supp. Fig. S6. **Optimized ORS sample-level sensitivity, specificity, and precision per ORS sample annotated by anatomic subsite.** (A-C) Bar graphs show (A) sensitivity, (B) specificity, and (C) precision for each of the 118 paired ORS samples. *X-axis, left to right*, individual ORS samples, annotated by (*key,* panel A) anatomic subsites. See Fig. 3B.



Supp. Fig. S7. **Lack of impact of clinical factors in optimization of ORS mutation detection.** Box and whiskers plots display a lack of significant associations between sensitivities of ORS variant detection and the presence or absence of various patient characteristics including (A) tobacco exposure; (B) alcohol consumption; (C) primary tumor site involving the oral tongue vs. other sites; (D) nodal status; and (E) status of disease recurrence. *Red dots,* mean of sensitivity of detection for all variants in individual patients.



Supp. Fig. S8. **Downsampling of ORS sequence reads in assay optimization.** (A) ROC and (B) PR (precision-recall) curves for assay optimization based on downsampling of ORS sequence reads to assess impacts on sensitivity, specificity, and precision of somatic variant detection. From the ROC curve, with VAF cutoff 0.1%, the optimal sensitivity was 0.73, the specificity was 0.94, and the AUC was 0.76. From the PR curve, with cutoff = 1.40, the optimal sensitivity was only 0.19 and precision was 0.0099.



Supp. Fig. S9. **Bayesian modeling for optimization of mutation detection in ORS**. (**A**) Results of Bayesian models including (*top*) different subset combinations of variables: alcohol status, smoking status, functional change, nucleotide change, and host genes. *Solid dots*: inclusion of variable(s) in models. *Bottom*, bar graphs (from *top* to *bottom*): sensitivity, specificity, precision, AUC, and optimal cutoffs. (**B**) ROC curve for Bayesian model involving optimization of functional change, nucleotide change, host genes, and smoking status, for which AUC was 0.58. At a VAF cutoff of 0.03%, sensitivity was only 0.62, specificity was ~0.996, and precision was 0.00068. (**C**) Optimization of assay precision in a PR curve based on Bayesian model at cutoff 2.57: sensitivity was only 0.13 and precision was 1.0.

| gene | MutSig q-value | oncodrive adj p-value | oncoclust adj p-value | drgap adj p-value |
|---|---|---|---|---|
| *AJUBA* | 0.00E+00 | 5.45E-07 | 1.15E-01 | 9.72E-20 |
| *ARID2* | 1.00E+00 | 2.84E-06 | NA | 5.33E-05 |
| *ASXL1* | 1.00E+00 | 5.47E-02 | 8.11E-02 | 5.11E-05 |
| *BIRC6* | 1.00E+00 | 1.72E-01 | NA | 5.16E-01 |
| *CASP8* | 0.00E+00 | 1.09E-10 | 3.53E-02 | 2.29E-67 |
| *CDKN2A* | 0.00E+00 | 1.09E-10 | 3.65E-04 | 5.13E-190 |
| *CHEK2* | 1.00E+00 | 9.08E-01 | 3.55E-06 | 7.91E-08 |
| *CTCF* | 1.49E-02 | 3.89E-05 | NA | 2.69E-06 |
| *EIF2S2* | 9.69E-02 | 9.08E-01 | NA | 2.04E-05 |
| *EP300* | 5.76E-01 | 1.44E-06 | NA | 3.10E-04 |
| *EPHA2* | 1.26E-06 | 9.13E-04 | 1.15E-01 | 1.80E-13 |
| *FAT1* | 0.00E+00 | 1.09E-10 | 2.85E-01 | 2.23E-125 |
| *FAT2* | 1.00E+00 | 1.82E-05 | NA | 3.96E-07 |
| *FBXW7* | 1.25E-05 | 1.09E-10 | 1.93E-01 | 8.55E-22 |
| *FN1* | 1.00E+00 | 1.08E-01 | NA | 3.14E-01 |
| *FOSL2* | 8.66E-03 | 3.43E-01 | NA | 9.90E-07 |
| *HERC1* | 1.00E+00 | 1.60E-01 | NA | 4.36E-02 |
| *HLA-A* | 1.00E+00 | 3.51E-04 | 7.23E-03 | 3.92E-28 |
| *HLA-B* | 0.00E+00 | 1.58E-04 | 2.64E-02 | 2.55E-15 |
| *HRAS* | 0.00E+00 | 7.36E-01 | 6.96E-04 | 3.67E-36 |
| *KDM6A* | 1.00E+00 | 8.50E-02 | NA | 5.05E-02 |
| *KMT2C* | 1.00E+00 | 1.72E-01 | NA | NA |
| *KMT2D* | 1.35E-03 | 9.08E-01 | NA | NA |
| *KRT5* | 8.90E-01 | 8.69E-01 | NA | 3.48E-06 |
| *MB21D2* | 1.00E+00 | 9.08E-01 | 4.58E-02 | 2.99E-01 |
| *MYCBP2* | 1.00E+00 | 8.30E-02 | NA | 2.20E-01 |
| *MYH9* | 1.00E+00 | 4.21E-06 | 1.14E-01 | 2.48E-01 |
| *NF2* | 5.23E-03 | 1.53E-01 | NA | 6.01E-03 |
| *NFE2L2* | 5.69E-02 | 1.44E-06 | 1.39E-02 | 1.35E-09 |
| *NOTCH1* | 5.88E-10 | 1.09E-10 | 4.01E-01 | 7.98E-46 |
| *NOTCH2* | 9.69E-02 | 1.25E-02 | NA | 3.73E-06 |
| *NSD1* | 3.03E-05 | 1.09E-10 | 4.35E-01 | 1.47E-13 |
| *PARG* | 1.00E+00 | 9.08E-01 | 2.02E-06 | 3.65E-06 |
| *PIK3CA* | 2.66E-08 | 5.01E-03 | 1.35E-02 | 3.68E-51 |
| *POM121* | 1.00E+00 | 9.18E-01 | 1.73E-03 | 1.31E-01 |
| *PSIP1* | 5.82E-02 | 1.70E-01 | NA | 5.79E-06 |
| *RAC1* | 7.24E-02 | 8.18E-02 | 7.23E-03 | 2.16E-06 |
| *RASA1* | 5.76E-04 | 2.84E-06 | 3.31E-01 | 1.33E-09 |
| *SMAD4* | 5.69E-02 | 5.86E-03 | NA | 4.44E-08 |
| *SMARCA4* | 1.00E+00 | 1.49E-01 | 3.60E-01 | 4.43E-01 |
| *TGFBR2* | 0.00E+00 | 9.23E-07 | 1.15E-01 | 6.16E-14 |
| *TP53* | 0.00E+00 | 1.09E-10 | 4.85E-02 | 0.00E+00 |

Supplemental Table S1. **Highly mutated genes in OSCC probed by custom hybrid capture baits.** As described in the text, sequencing data from 4 genes (*HLA-A*, *HLA-B*, *KMT2C* and *KMT2D*) were not reliable and therefore were excluded from further analysis. Also shown are adjusted p-values (q-values) for enrichment of mutations in each of the 42 genes as called by MutSig, OncoDrive, OncoClust and DrGap (cf. Methods).

(Table included as separate Excel file.)

Supplemental Table S2. **List of somatic variants in fresh-frozen OSCC.** Shown for each variant (rows) is sample ID name, gene name, chromosomal coordinates, variant allele fraction in ORS, VAF in OSCC, functional impact of variant, presence in COSMIC database, known status as a germline variant, and population allele frequency in GnomAD database.