Supplemental Material

Supplemental Materials and Methods

DNA Microarray. The RNA and DNA from each specimen were simultaneously extracted using the TRIzol method (Invitrogen, Carlsbad, CA). To increase DNA purity, we modified the DNA extraction protocol to include the use of a “back extraction buffer” (4 M guanidine thiocyanate, 50 mM sodium citrate, and 1 M Tris, pH 8.0). RNA was further purified with the use of an RNeasy mini kit (Qiagen, Valencia, CA) per Affymetrix (Santa Clara, CA) recommendations. For expression array analysis, 1.0 to 2.5 µg of total RNA was used to generate biotin-labeled cRNA using the GeneChip Expression 3’-Amplification Reagents Kit (Affymetrix) per manufacturer’s protocol. The cRNA was hybridized to an Affymetrix U133 2.0 Plus GeneChip arrays and scanned using an Affymetrix GeneChip arrays Scanner 3000 7G in the Fred Hutchinson Cancer Research Center’s Genomics Shared Resources per Affymetrix protocols. At least one clinically normal tissue sample from a control subject was processed in tandem with every seven to eight tumor tissue samples from OSCC cases.

Validation of Gene Expression of *LAMC2, COL4A1, COL1A1,* and *PADI1* by qRT-PCR. From the 167 OSCC cases, a subset of 30 was randomly chosen from amongst a group of 40 in which RNA were plentiful and dilutions were readily available. Among the 45 controls, 41 had plentiful RNA available, from which 30 samples were chosen at random. Each sample containing 7.5 ng purified total RNA was assayed in triplicate in 10 µl reaction volumes using the QuantiTect SYBR Green RT-PCR kit (Qiagen, Valencia, CA) and bioinformatically validated QuantiTect primers (Qiagen, Valencia, CA) on a 7900HT Sequence Detection System (ABI, Foster City, CA). The cycling conditions were as follows: 30 minutes at 50° C, 15 minutes at 95° C, and 40 cycles of 15 seconds at 94° C, 30 seconds at 55° C, and 30 seconds at

72° C. For *COL1A1* (NM_000088), a 118-bp amplicon spanning exons 1 and 2 was amplified. For *COL4A1* (NM_001845), a 119-bp amplicon spanning exons 6, 7, 8, and 9 was amplified. For *LAMC2* (NM_005562), a 74-bp amplicon spanning exons 18 and 19 was amplified. For *PADI1* (NM_013358), an 80-bp amplicon spanning exons 3, 4, and 5 was amplified. We used *ACTB* as the reference gene and amplified a 146-bp amplicon that spans exons 3 and 4. Ten point standard curves were generated using Universal Human Reference RNA (Stratagene, La Jolla, CA) for all genes, except *PADI1* in which Normal Adjacent Esophagus Total RNA (Ambion, Austin, TX) was used. The linear correlation coefficient (R^2) was 0.99 or greater for all runs. The mean threshold cycle (Ct) values were calculated from the triplicate Ct values. Mean Ct values were further normalized in relation to the mean Ct value of the *ACTB* gene.

Supplemental Table 1. Selected Characteristics of Study Participants

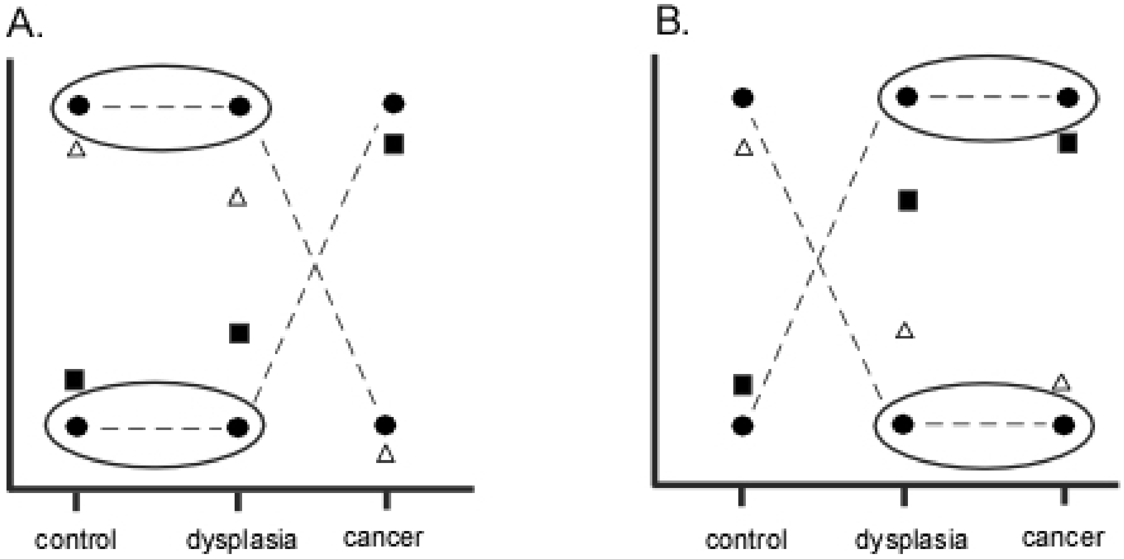
	Training Set				**Testing Set**			
	OSCC Case (n=119)		Control (n=35)		OSCC Case (n=48)		Control (n=10)	
	n	%	n	%	n	%	n	%
Age								
19-39	5	(4.2)	14	(40.0)	2	(4.2)	3	(30.0)
40-49	18	(15.1)	8	(22.9)	8	(16.7)	6	(60.0)
50-59	40	(33.6)	4	(11.4)	17	(35.4)	1	(10.0)
60-88	56	(47.1)	9	(25.7)	21	(43.8)	0	(0.0)
Sex								
Male	84	(70.6)	25	(71.4)	36	(75.0)	7	(70.0)
Female	35	(29.4)	10	(28.6)	12	(25.0)	3	(30.0)
Race								
White	106	(93.0)	24	(68.6)	40	(85.1)	6	66.7)
Non-white	8	(7.0)	11	(31.4)	7	(14.9)	3	(33.3)
Unknown	5		0		1		1	
Current smoking status*								
Never or Former	59	(49.6)	25	(71.4)	27	(56.3)	8	(80.0)
Current	60	(50.4)	10	(28.6)	21	(43.7)	2	(20.0)
Average alcoholic drinks per day in prior year*								
Never or Former	36	(30.8)	9	(25.7)	19	(40.4)	3	(30.0)
Current	81	(69.2)	26	(74.3)	28	(59.6)	7	(70.0)
Unknown	2		0		1		0	
AJCC Stage								
I/II	42	(35.3)	0		13	(27.1)		
III/IV	77	(64.7)	0		35	(72.9)		
Tissue Site - oral cavity vs. oropharynx								
Oral	88	(73.9)	1	(2.9)	29	(60.4)	0	(0.0)
Oropharynx	31	(26.1)	34	(97.1)	19	(39.6)	10	(100.0)

*As of the date of diagnosis (OSCC cases) or recruitment (controls)

Supplemental Table 2. Probe sets (n=67) potentially involved in the development of oral dysplasia (See attachment “Supplemental Table 2”)

Supplemental Table 2. Probe sets (n=67) potentially involved in the development of oral dysplasia

Upregulation in Dysplasia vs.Controls			Down-regulation in Dysplasia vs. Controls		
Gene Name	Gene Symbol	Z Score	Gene Name	Gene Symbol	Z Score
202311_s_at	COL1A1	26.04	241233_x_at	C21orf81	-21.57
211980_at	COL4A1	25.33	220149_at	FLJ22671	-18.67
202404_s_at	COL1A2	25.33	1569608_x_at		-18.16
202310_s_at	COL1A1	23.79	218885_s_at	GALNT12	-16.88
221729_at	COL5A2	21.66	225548_at	SHRM	-14.92
221730_at	COL5A2	20.65	205319_at	PSCA	-14.82
212488_at	COL5A1	20.65	218935_at	EHD3	-13.95
212489_at	COL5A1	19.35	230740_at		-13.91
225681_at	CTHRC1	18.82	242417_at	LOC283278	-13.86
217312_s_at	COL7A1	18.28	220962_s_at	PADI1	-13.69
212012_at	PXDN	18.06	218779_x_at	EPS8L1	-13.2
205157_s_at	KRT17	17.82	220016_at	AHNAK	-13.12
204415_at	G1P3	17.75	1553861_at	TCP11L2	-12.54
204715_at	PANX1	17.41	221665_s_at	EPS8L1	-12.53
217430_x_at	COL1A1	17.25	210868_s_at	ELOVL6	-12.33
1555778_a_at	POSTN	17.02	240000_at	LIPI	-11.57
225292_at	COL27A1	16.94	204378_at	BCAS1	-11.36
213869_x_at	THY1	16.92	1565661_x_at	FUT6	-11.23
204647_at	HOMER3	16.92	205730_s_at	ABLIM3	-11.19
229554_at	LUM	16.8	205428_s_at	CALB2	-9.5
203325_s_at	COL5A1	16.4	209975_at	CYP2E1	-8.52
209900_s_at	SLC16A1	15.83			
208851_s_at	THY1	15.72			
225288_at	COL27A1	15.55			
210809_s_at	POSTN	15.27			
205483_s_at	G1P2	14.85			
204114_at	NID2	14.66			
213668_s_at	SOX4	14.26			
226997_at		14.25			
41037_at	TEAD4	14.2			
214453_s_at	IFI44	14.12			
202235_at	SLC16A1	14.1			
217519_at	MACF1	13.96			
208156_x_at	EPPK1	13.53			
202238_s_at	NNMT	13.52			
223541_at	HAS3	13.43			
204619_s_at	CSPG2	13.33			
219863_at	HERC5	12.82			
218888_s_at	NETO2	12.5			
204972_at	OAS2	12.12			
222344_at	C5orf13	12.11			
210797_s_at	OASL	11.96			
209800_at	KRT16	11.91			
209969_s_at	STAT1	11.8			
203921_at	CHST2	11.75			
229055_at	GPR68	11.25			

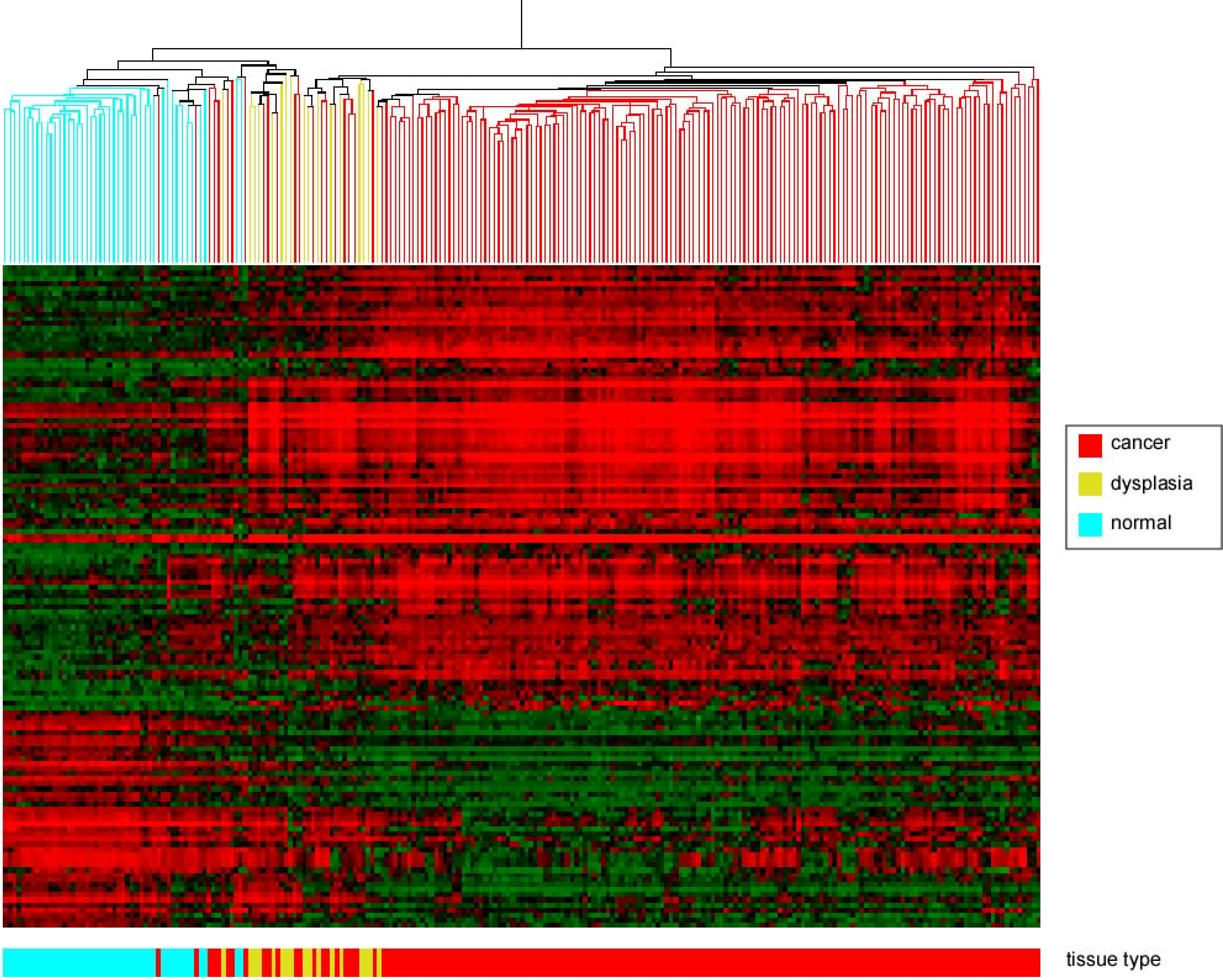


Supplemental Figure 1: Schematic representation of the method for selecting the differentially expressed genes specific to oral cancer. A) To obtain the list of differentially expressed genes only between cancer and control that were not also differentially expressed in dysplastic lesions, the dysplastic lesions and controls were grouped together and the gene expression was compared to that of the invasive tumors. Using this list of genes, we then excluded those that were differentially expressed between control and dysplasia (i.e. Δ and ■) using NFD=1 criterion. The remaining genes were those whose expression levels remained the same between control and dysplasia (i.e. ●), but were up- or down-regulated in cancer. B) Conversely, we sought to determine those genes whose differential expression between cancer and controls appears to occur early in the process of carcinogenesis. For this analysis, we grouped the dysplastic lesions with the cancer samples and the gene expression was compared to that of the controls. Using this list of genes, we then excluded those that were differentially expressed between dysplasia and cancer (i.e. Δ and ■) using NFD=1 criterion. The remaining genes were those whose expression levels remained the same between dysplasia and cancer (i.e. ●), but were up- or down-regulated compared to controls.

Supplemental Figure 2

Legend for Supplemental Figure 2:

Supplemental Figure 2. Hierarchical clustering of the gene expression data using the top 131 probe sets differentially expressed in OSCC when compared to normal controls. The dendogram at the top lists all of the samples tested and measures their degree of relatedness in gene expression. Each column represents the expression levels for all the probe sets in a particular sample, whereas each row represents the relative expression of a particular probe set across all samples. The expression level of any probe set in any given sample (relative to the mean expression level of that probe set across all samples) is reported along a color scale in which red represents transcription up-regulation, green represents down-regulation, and the color intensity indicates the magnitude of deviation from the mean. The color bar underneath the heat map color codes the tissue type of each sample in the heat map as normal control (aqua), OSCC (red), or dysplasia (yellow). These colors are also used in the dendogram at the top of the heat map.


cancer
dysplasia
normal
tissue type