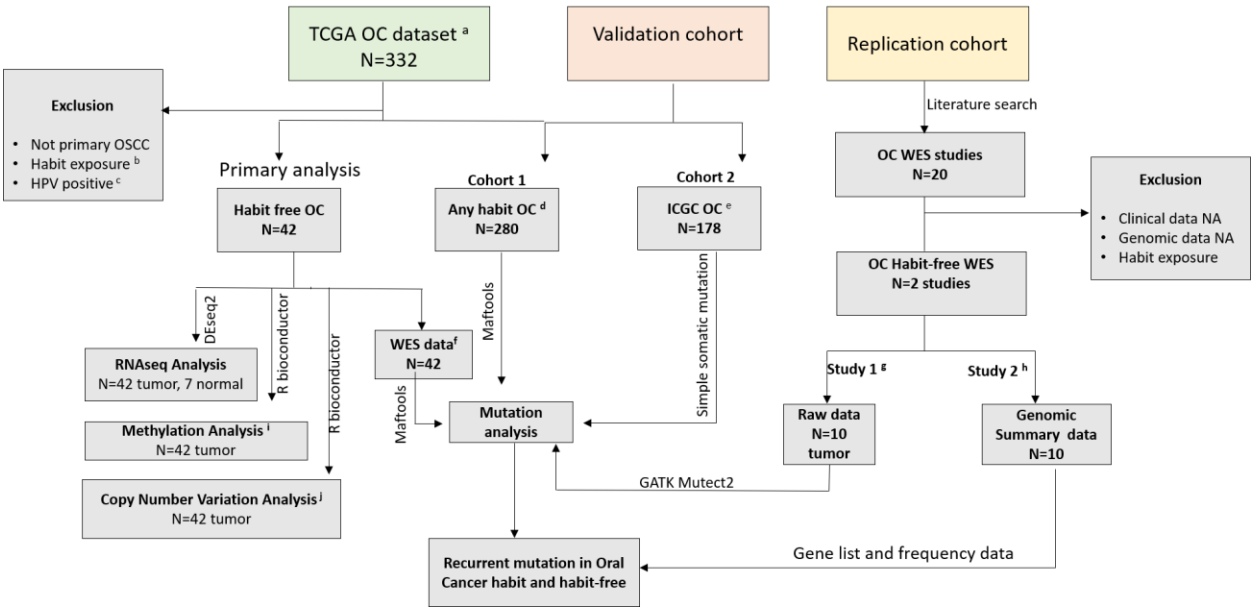

Figure S1: Flow chart showing the overall study design

Figure illustrates the datasets accessed and reasons for exclusion. The primary analysis included 42 habit-free OC subjects. Habit associated OC subjects from TCGA and ICGC were used as the validation set to compare the frequency of identified gene mutations. Two independent studies comprising habit-free OC subjects were used as the replication cohort.

[a] Oral cancer cases included in the cancer genome atlas (TCGA) data repository

[b] Smoker defined as having smoked ≥100 cigarettes in lifetime, and drinker defined as ever having consumed alcohol in their lifetime

[c] HPV positivity ascertained by immunohistochemistry for P16

[d] Tobacco and alcohol related oral cancers in the TCGA dataset

[e] International cancer genome consortium (ICGC) data repository including oral cancers

[f] Whole exome sequencing data were accessed

[g] Previously published whole exome data accessed (PRJNA740146)

[h] Previously published whole exome data accessed (PRJNA700466)

[i] Methylation analysis were done for MUC16 and MUC5B only

[j] Copy number analysis were done for MUC16 and MUC5B only

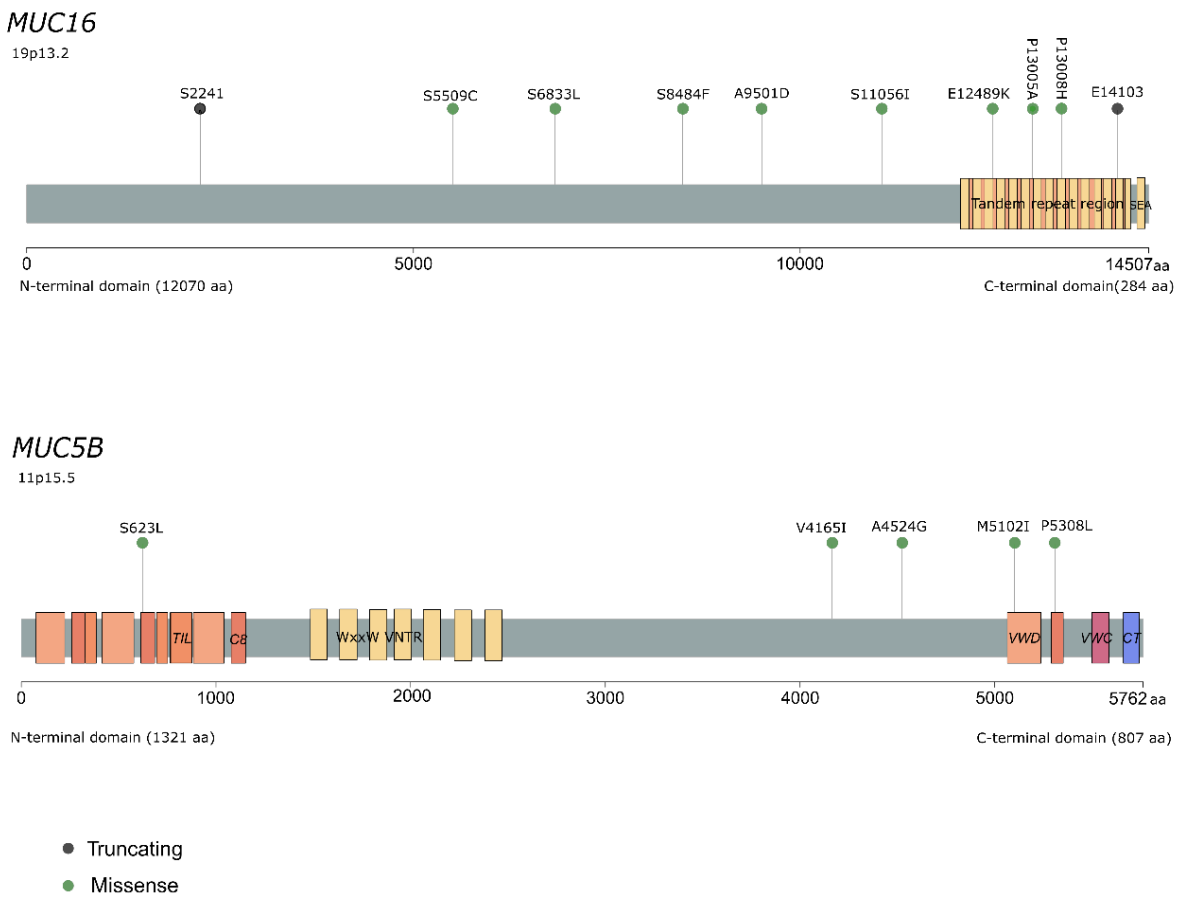

Figure S2: Lollipop plot of single nucleotide variant distribution on protein domain in *MUC16* (19p13.2) and *MUC5B* (11p15.5)

Each protein is shown to scale, with missense mutations indicated in green and truncating mutations in black. Domain annotations were derived from UniProt (*MUC16*: Q8WXI7; *MUC5B*: Q9HC84). Sea urchin sperm protein, Enterokinase, and Agrin (SEA), von Willebrand factor type D domain (VWD), von Willebrand factor type C domain (VWC), Trypsin Inhibitor like cysteine rich domain (TIL).

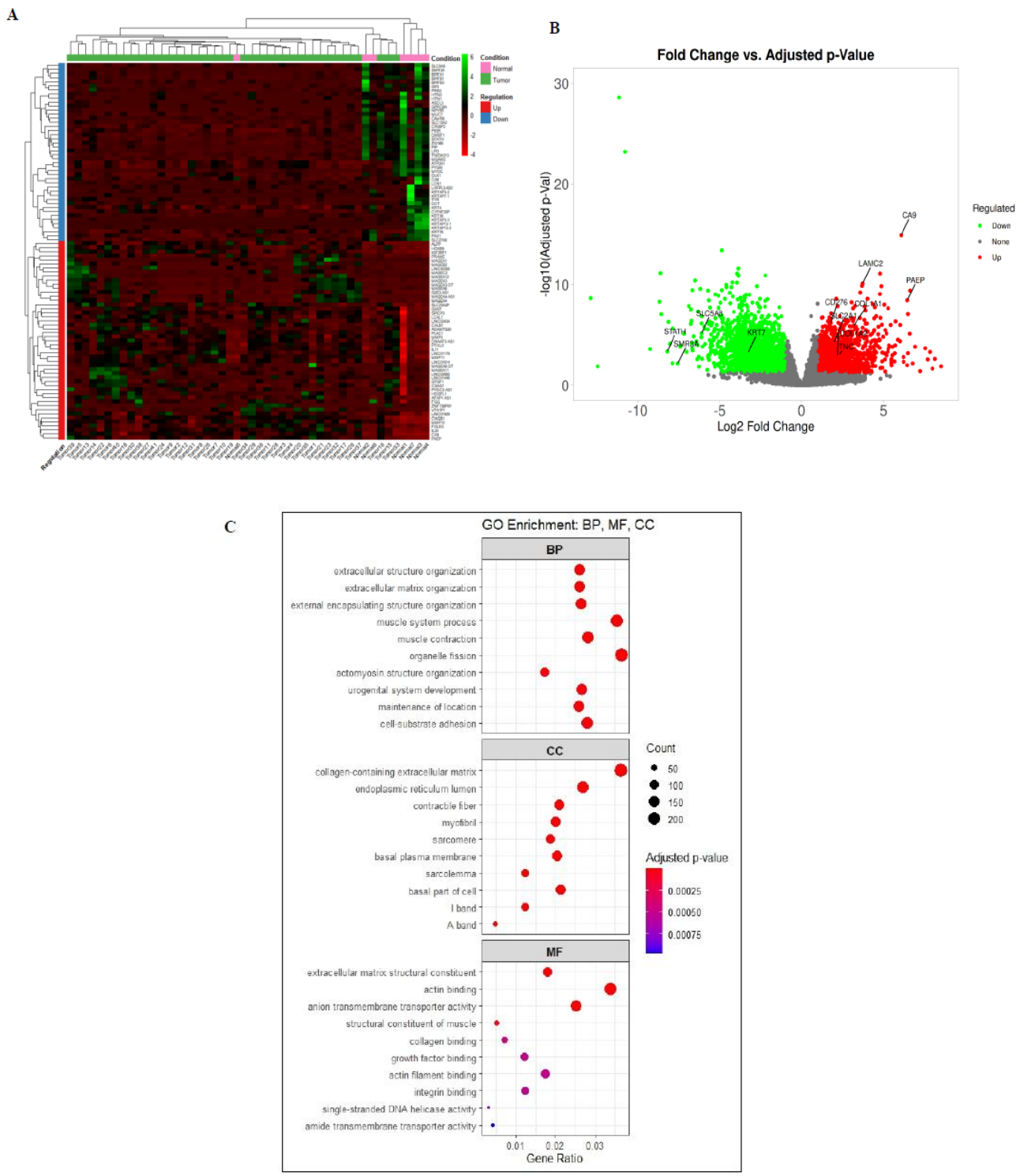

Figure S3: Differential gene expression analysis of habit-free TCGA dataset

(A) Heatmap showing the expression profiles of the top differentially expressed genes from RNA-seq analysis of OC samples in 42 habit-free cohort. The dataset includes 7 adjacent normal tissue samples and 42 tumor samples. Gene expression values are represented as

Z-scores of log-transformed normalized counts. Red indicates upregulation, green indicates downregulation.

(B) Volcano plot showing the distribution of differentially expressed genes based on $\log_2$ fold change and $-\log_{10}$ adjusted *p*-value. Significantly upregulated genes are shown in red, while downregulated genes are shown in green; grey dots indicate genes without significant differential expression. Thresholds for significance were set at $|\log_2 \text{ fold change}| \geq 2$ and adjusted $p < 0.05$.

(C) The top enriched GO terms are shown across three categories: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). The x-axis represents the gene ratio (number of input genes associated with a given GO term divided by the total number of input genes), while the size of each dot indicates the number of genes annotated to the term. The color scale corresponds to the adjusted p-value (Benjamini–Hochberg correction), with darker red indicating greater statistical significance. Enriched GO terms highlight extracellular matrix organization, muscle contraction, collagen binding, and actin filament dynamics, suggesting alterations in structural and motility-related pathways.

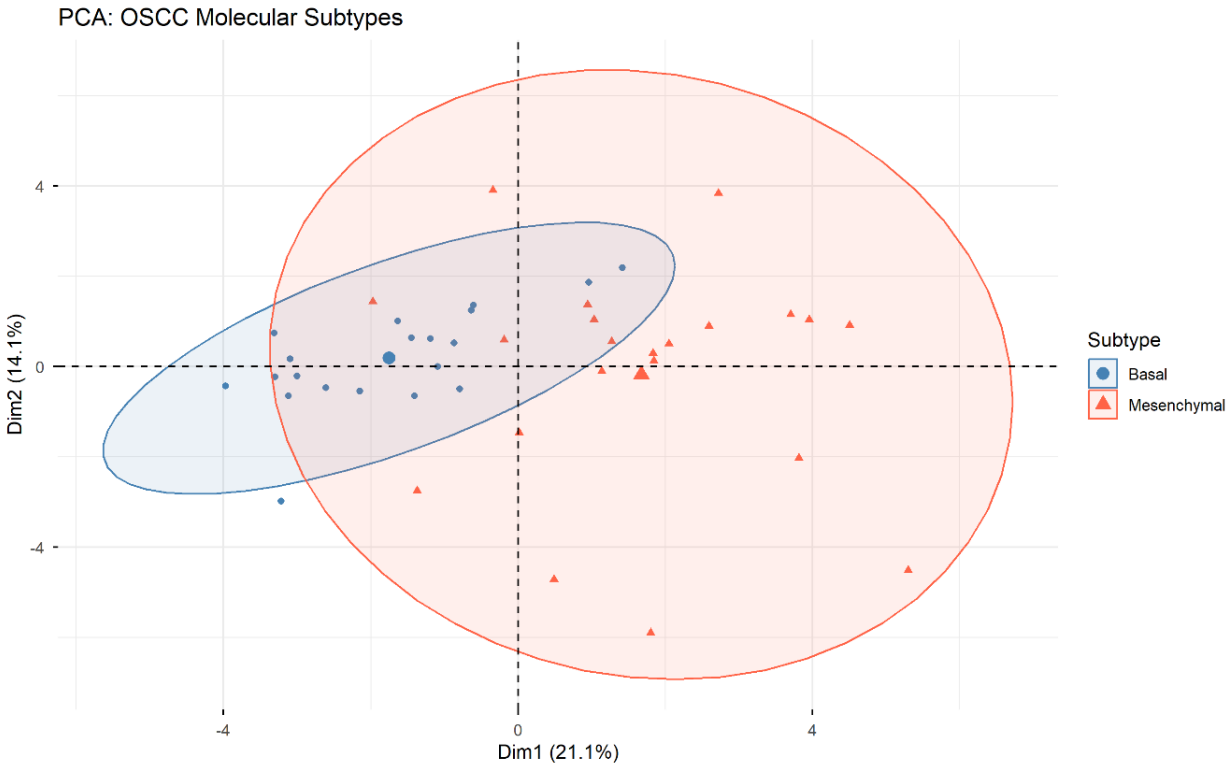

Figure S4: Principal component analysis of habit-free OC group

Habit-free OC samples (n=42) were classified into molecular subtypes based on gene expression profiles to visualize the clustering of samples across known OC subtypes: mesenchymal n=22 (red triangles) and basal n=20 (blue circles) using a centroid-based nearest template prediction method. PCA was then performed on the normalized gene expression data to visualize subtype separation. The distinct separation suggests transcriptional heterogeneity among the subtypes, with mesenchymal showing the greatest dispersion.

Table S1: Integrated analysis of mucin expression, genomic and epigenetic alterations

***MUC16* Status**	**N cases[a]**	**MUC16 Mean Expression (95% CI) [b]**	**P-value [c]**	***MUC5B* Status**	**N cases**	**MUC5B mean expression (95% CI) [a]**	**P-value [c]**
All Tumors[d]	41	2.02 (-0.78-4.84)	NA	All Tumors[d]	42	1.66 (-0.15-3.46)	NA
FC (FDR)	41	-1.34, (0.19)				-4.01 (0.03)	
***MUC16* Mutation status**			0.24	***MUC5B* Mutation status**			0.96
Wild type	32	2.55 (-1.08- 6.18)		Wild type	38	1.67 (-0.33- 3.67)	
Mutated	9	0.18 (0.03-0.32)		Mutated	4	1.55 (-1.26- 4.36)	
Mutation function prediction			0.31	**Mutation function prediction**			0.98
Deleterious	2	0.05 (-0.22- 0.31)		Deleterious	0	NA	
Moderate	7	0.21 (0.028- 0.39)		Moderate	4	1.61 (-1.12- 4.35)	
Not known	32	2.55 (-1.08- 6.18)		Not known	38	1.66 (0.34- 3.66)	
Copy number status			0.09	**Copy number status**			**0.02**
Loss	3	0.14 (-0.43- 0.71)		Loss	2	0.98 (0.01- 1.94)	
No change	37	2.22 (-0.89- 5.35)		No change	38	1.67 (0.01- 37.26)	
Gain	1	0.38 (NA)		Gain	2	2.14 (0.16- 4.12)	
Methylation status [f]			r^2= 0.15	**Methylation status**			**r^2= 0.04**

[a] One outlier sample for *MUC16* expression was excluded

[b] Mean expression and 95% confidence interval of means in each category

[c] P-value from comparison of means, t-test for two group comparisons and Anova for three groups

[d] Reflects overall *MUC16/ MUC5B* expression levels in 41 tumors, data expressed as expression counts

[e] Reflects overall *MUC16/ MUC5B* expression levels in 41 tumors compared to 7 normal tissues, data expressed as fold change and false discovery rate (FDR)

[f] Represents correlation between *MUC16/ MUC5B* gene expression counts and average beta value of CpGs assigned to corresponding gene body or promoter regions. Pearson's correlation coefficient represented as r-squared value. Bold represents statistically significant values.

Table S2: Association between *MUC16/MUC5B* mutation status, molecular subtype, TMB and overall survival

	***MUC16* Status**				***MUC5B* Status**			
Characteristic	Wild Type (N= 32)	Mutated (N= 10)			Wild Type (n= 38)	Mutated (n= 4)		
	n (%)[a]	n (%)[a]	OR (95% CI)[b]	P-value	n (%)[a]	n (%)[a]	OR (95% CI)[b]	P-value
Molecular subtype								
Basal	11	8	Reference		18	1	Reference	
Mesenchymal	21	2	**0.12 (0.02-0.73)**	**0.02**	20	3	2.70 (0.26-28.34)	0.41
OR (95% CI)[c]			**0.12 (0.19-0.72)**	**0.02**			5.30 (0.29-96.62)	0.26
Tumor Mutation Burden								
Nonsynonymous mutations/ MB	2.57	3.51	1.393(0.90-1.96)	0.16	2.58	4.83	**1.79 (1.02-3.11)**	**0.04**
OR (95% CI)[c]			1.31 (0.84-2.05)	0.23			2.94 (0.93-9.25)	0.07
Overall Survival								
Alive	18	5	Reference		21	2	Reference	
Dead	14	5	1.28 (0.46-3.56)[d]	0.64	17	2	0.88 (0.20-3.83)[d]	0.87
HR (95% CI)[e]			1.17 (0.41-3.3)	0.77			1.02 (0.23-4.50)	0.98

[a] Reflects the number of samples in that category and column percentages

[b] Shows the odds ratios and 95% CI from univariate logistic regression model with wild type *MUC16/MUC5B* as reference category

[c] Shows the odds ratios and 95% CI from logistic regression model adjusted for age, sex and stage (early/ late) at diagnosis with wild type *MUC16/MUC5B* as reference category

[d] Hazards ratio from univariate Cox proportional models

[e] Hazards ratio from multivariate Cox proportional models adjusted for age, sex and early/ late stage at diagnosis. Significant associations are shown in bold.