Spontaneous emergence of non-convertible cell states with CD24-High phenotype results in phenotypic heterogeneity that associates with poor prognosis in oral cancer

Kavya Vipparthi[1], Kishore Hari[*2], Priyanka Chakraborty[*2], Subhashis Ghosh[1], Ankit Kumar Patel[1], Arnab Ghosh[1], Nidhan Kumar Biswas[1], Rajeev Sharan[3], Pattatheyil Arun[3], Mohit Kumar Jolly[2], Sandeep Singh[#1]

(*) These authors contributed equally

Affiliation:

[1] National Institute of Biomedical Genomics, Kalyani, India

[2] India Institute of Science, Bangalore, India

[3] Tata Medical Center, Kolkata, India

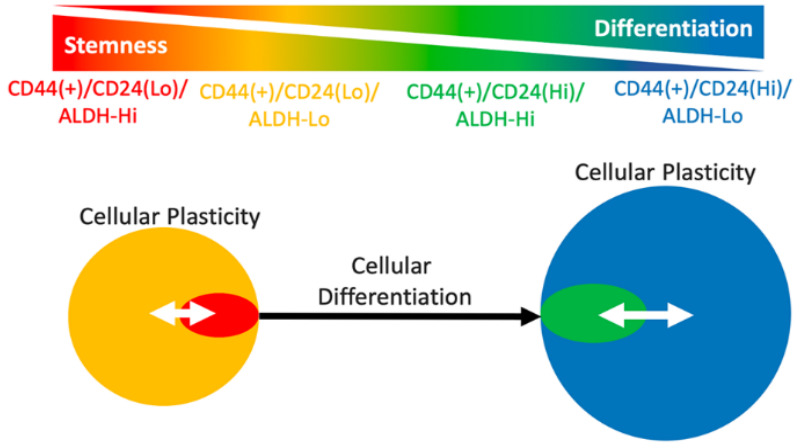
 # Correspondence Address:

Dr. Sandeep Singh, PhD
Associate Professor
National Institute of Biomedical Genomics
PO: NSS
Kalyani, WB-741251 India
Email: ss5@nibmg.ac.in
Phone: +91-89104264

Acknowledgement:
This work is supported by the grant (IA/I/13/1/500908) received from Wellcome trust-DBT India Alliance. KV acknowledges DST-INSPIRE, SG acknowledges DBT and AKP acknowledges ICMR for fellowship support. We thank Professors Saumitra Das and Partha P Majumder for scientific discussion and Dr. Arindam Maitra for assistance with transcriptomics platform.

Abstract

Cancer cells within individual tumors often exist in distinct phenotypic states contributing to intratumoral heterogeneity (ITH). However, studies on cell state dynamics among oral cancer cells are largely missing. Here, we have multiplexed phenotypic markers of putative oral-stem-like cancer cells (SLCCs) and characterized diversity among CD44-positive oral cancer cell subpopulations with respect to distinct expression of CD24 and aldehyde dehydrogenase (ALDH)-activity. Our *in vitro* experimental observations were explained by a Markov model where subpopulations followed two distinct patterns of spontaneous repopulation dynamics. Cells showed stochastic inter-conversions on ALDH-axis, harnessed by cancer cells to enrich $ALDH^{High}$ subpopulations in response to Cisplatin treatment. However, these cells followed a strict non-interconvertible transition of $CD24^{Low}$ to $CD24^{High}$ subpopulations, even in response to chemotherapy-induced stress. Using phenotype-specific RNAseq, we suggest the organization of subpopulations into hierarchical structure with possible maintenance of intermediate alternate states of stemness within the differentiating oral cancer cells. We also show that the population dynamics described here may influence tumor behaviour by increasing ITH in aggressive oral tumors. Overall, the described phenotypic subgroups not only reliably exhibited spontaneous or Cisplatin-driven cellular dynamics but also the distinct transcription states of oral cancer cells. Most importantly, our *in vitro* model system derived observations emphasized the prognostic power which may be translated for betterment of oral cancer patients.



Graphical Abstract: We have characterized diversity among CD44-positive oral cancer cells lines with respect to distinct expression of CD24 and ALDH-activity. Cells showed stochastic inter-conversions on ALDH-axis but a strict non-interconvertible transition of $CD24^{Low}$ to $CD24^{High}$ phenotype, even in response to chemotherapy-induced stress. RNAseq study suggested the organization of subpopulations into hierarchical structure with possible maintenance of intermediate alternate states of stemness within the differentiating oral cancer cells. The described population dynamics may influence tumor behaviour by increasing intratumoral heterogeneity in aggressive oral tumors.

Introduction:

Though genetic heterogeneity undeniably confers cell growth and survival advantages, it alone is not sufficient to explain several aspects of the cancer biology including the functional and phenotypic heterogeneity, emergence of drug resistance and metastasis phenotypes [1]. RNA sequencing studies from multiple tumor-types showed that individual tumors with similar genetic profile may exhibit transcriptional heterogeneity of tumors [2], that may be arising due to the presence of multiple phenotypic states along with diversity in interactions with tumor microenvironment [3-5]. Recent studies are uncovering the presence of extensive non-genetic heterogeneity at the phenotypic and functional levels within genetically identical sub-clones of cancer cells [6-8].

The ability of genetically identical cells to exist and switch to multiple phenotypic states is a long-standing notion of developmental biology [9]. However, cell state dynamics of generating phenotypic heterogeneity in the cancer cell populations independent of the genetic context has garnered attention after the pioneering work done by Dalerba *et. al.* [4]. Studies showed that the state transitions of cells can either occur spontaneously and stochastically without the need for any external signal or may occur due to perturbations driven by the external triggers [10-12]. Cells existing in each phenotypic state may exert distinct functions, such as therapy resistance, that overall benefits the tumor [13]. Two of the most common cellular processes associated with phenotypic state transitions are epithelial to mesenchymal transitions (EMT) and cancer stemness, that may also be closely related to each other [14]. The stem-like cancer cells (SLCCs) contribute to intratumoral heterogeneity by giving rise to different cellular phenotypes through self-renewal and differentiation [15].

Despite the intense debates, stemness in cancer cells continues to occupy a major portion of oncology research given their key attributes like tumor re-initiation, multipotency and therapy

resistance [16]. Multiple studies proved the existence of (SLCCs) in solid tumors [17]; however, phenotypic diversity in SLCCs has been the major challenge in defining the markers for SLCC-subpopulation in solid tumors [18, 19]. Adding further to this complexity, recent evidence suggests that rather than being fixed cellular entities, stem-like characteristics can be achieved over time, even by more differentiated cells [20, 21]. Hence, the unidirectional rigid hierarchy of CSC models is currently debated and challenged [22, 23]. Thus, stemness in cancer cells provides a conceptual framework for interrogating the aspects of complex cellular hierarchy, functional heterogeneity and plasticity in cancer.

Tumor cells with high CD44 expression were the first identified SLCCs in Head and Neck Squamous Cell Carcinoma (HNSCC) [24]. Later, along with CD44; cells with high levels of aldehyde dehydrogenase (ALDH) enzyme, i.e., $CD44^{+Ve}/ALDH^{High}$ cells were shown to be enriched with SLCCs properties [25]. Taking leads from breast cancers, cells with $CD24^{Low}/CD44^{High}$ cell surface marker profile were reported as representatives of SLCCs and EMT phenotypes in oral cancers [26]. The co-existence of heterogeneous oral-SLCCs with respect to EMT and their spontaneous interconversions has also been reported [27]. Despite these qualitative demonstrations, studies on population-dynamics in a quantitative and unbiased manner are largely missing in oral cancer. Questions like, how reproducibly these subpopulations arise and how these diverse subpopulations relate to each other with respect to molecular states has remained unanswered? Answers to these questions will help us understand the dynamic complexities of cellular and functional heterogeneity in oral cancer.

Here, we have attempted to investigate the cellular dynamics among oral cancer cell subpopulations derived by multiplexing phenotypic markers of putative oral-SLCCs (CD44, CD24 and ALDH-activity). We showed that cancer cells with distinct phenotypes co-exist in multiple cell lines. While subpopulations with differential ALDH-activity showed stochastic inter-conversions

among them, those with differential CD24 expression followed strict developmental hierarchy, even in response to chemotherapy-induced stress. We also identified transcriptional states of different subpopulations. Analysis demonstrated that the loss of transcriptional signature specific to the reported putative oral-SLCCs-state was significantly associated with poorer prognosis of head and neck cancer patients in TCGA cohort. Overall, our work has provided evidence of cellular dynamics leading to transcriptional heterogeneity relates to poor prognosis in oral cancer patient.

Results:

ALDH^{High} cells exhibited both CD24^{Low} and CD24^{High} phenotypes:

We first examined the co-expression of markers related to putative oral-SLCCs; CD44, CD24 and ALDH in five different oral cancer cell lines. All the tested cell lines were positive for CD44 marker expression. Three of five cell lines (GBC02, SCC-070 and SCC-029) showed co-existence of both CD24^{High} & CD24^{Low} sub-populations (Fig 1A, 1B and Supplementary Fig 1A, B) whereas GBC035 and SCC-032 were of CD24^{High} and CD24^{Low} phenotypes, respectively (Fig 1B and Supplementary Fig 1C, D). Analysis for ALDH-activity by ALDEFLUOR-assay showed that ALDH^{High} cells were not only present in previously reported CD24^{Low} subpopulation of CD44-positive cells, but also in relatively unexplored subpopulation with CD24^{High} phenotype (Fig 1A, B and Supplementary Fig 1). Hence, based on the differential status of these two markers, we identified four sub-populations of CD44-positive cells with marker profiles, CD24^{Low}/ALDH^{High}, CD24^{Low}/ALDH^{Low}, CD24^{High}/ALDH^{High}, CD24^{High}/ALDH^{Low} which were termed as 'Red', 'Orange', 'Green' and 'Blue' subpopulations, respectively. The frequencies of each of these subpopulations were analyzed in all tested cell lines. Importantly, ALDH^{High} phenotypes ('Red' and 'Green') were maintained at lower frequencies, while the ALDH^{Low} subpopulations ('Orange' and 'Blue') were the dominant ones across different cell lines (Fig 1B). Overall, these results emphasized the co-existence of ALDH^{High} cells overlapping with both CD24^{High} and CD24^{Low} phenotypes in CD44-positive oral cancer cells.

To explore if the phenotypes observed in cell lines were also present in primary oral tumors, we next tested for CD44, CD24 and ALDH status in seven freshly resected, treatment naïve, primary gingivobuccal tumor tissues. Enzymatically digested single cell suspension of tumor tissues was used for ALDEFLOUR assay and immunophenotyping using antibodies against

hematopoietic lineage markers and CD24, CD44. Cells negative for hematopoietic lineage markers were used for downstream analysis (Supplementary Fig 2). Entire tumor population was first gated into $ALDH^{High}$ and $ALDH^{Low}$ with the help of DEAB treated cells as control (Fig 1C). CD44 and CD24 expression were further assessed in gated populations. In accordance with the existing literature, we found $ALDH^{High}$ cells to be primarily enriched in $CD24^{Low}$ phenotype (Fig 1D, pie chart, Supplementary Fig 3). However, the $ALDH^{High}$ cells were also enriched in $CD24^{High}$ subpopulations in primary oral tumors at frequencies ranging from 0.3-11% (Fig 1D, Supplementary Fig 3). Collectively, results provided evidence for existence of heterogeneous $ALDH^{High}$ cells with diverse CD24 phenotypes in oral tumors like the established cell lines.

Subpopulations exhibited distinct repopulation dynamics:

Maintenance of subpopulations at different frequencies in a mixed population could be a result of differences in the proliferative potential of specific subpopulations. However, using dye dilution assay using 'Cell Trace-Violet' dye we did not observe a significant difference in proliferation index among the four subpopulations in both tested cell-lines, GBC02 and SCC-029B (Fig 1E). This observation prompted us to probe the population dynamics of these four subpopulations.

Isolated four subpopulations were allowed to propagate undisturbed for 10 days and reanalyzed for their CD24, CD44 and ALDH marker profiles. Results from GBC02, SCC-029B and SCC-070 cell lines showed that the 'Red' subpopulation was most 'potent' with efficiently reproduced all four subpopulations. Interestingly, 'Green' subpopulation could only give rise to itself and its $ALDH^{Low}$ counterpart, showing 'commitment' to generate only $CD24^{High}$ phenotype in all the tested cell lines (Fig 1F,G and Supplementary Fig 4A,B,C). While the $ALDH^{Low}$-'Orange' subpopulations efficiently repopulated the $ALDH^{Low}$ subpopulations of 'Orange' and 'Blue'

phenotypes; the 'Blue' subpopulation showed the phenomena of 'specification' by repopulating itself with higher efficiency (Fig 1F,G and Supplementary Figure 4A,B,C). Of note, the ALDH^{Low} ('Orange' or 'Blue') subpopulations showed higher plasticity and generated ALDH^{High} phenotypes at higher frequency in SCCC-070 cell line as compared to other cell lines (Fig 1G, Supplementary Fig 4B). Distinct from other cell lines, SCC-032 could not generate the CD24^{High} subpopulation ('Green' or 'Blue') from CD24^{Low} ('Red' or 'Orange') subpopulations. Also, similar to the 'Orange' subpopulation of SCC-070, the 'Orange' subpopulations of SCC-032 demonstrated the capacity to recapitulate both 'Red' and 'Orange' subpopulations (Supplementary Fig 4D).

Further, to confirm the property of CD24^{High} cells to be a committed subpopulation, we developed a separate sub-line termed as 'SCC-070 (CD24+)', by sorting and propagating CD24^{High} cells repopulated from 'Red' subpopulation of SCC-070 cells. Interestingly, though CD24^{High} cells were originated from CD24^{Low} subpopulation, these cells did not regenerate CD24^{Low} cells or parent phenotype (Fig 1H, Supplementary Fig 5A(i)), confirming the committed property of CD24^{High} subpopulation. Further, the specification of 'Blue' subpopulation could also be confirmed (Fig 1H, Supplementary Fig 5A(ii)). These results were further supported by an independent patient derived cell line, GBC035, which predominantly has phenotypically homogenous CD24^{High} phenotype. Similar to SCC-070 (CD24+) subline, sorted 'Green' and 'Blue' subpopulations of GBC035, also showed 'Green' subpopulation to be committed to generate only CD24^{High} subpopulations, whereas 'Blue' subpopulation to have specification for generating itself (Fig1, Supplementary Fig 5B).

Although with certain cell-line specific trends; overall, we uncovered two distinct patterns of repopulation dynamics; 1) Fixed CD24-Axis, which was more stringent and non-interconvertible and 2) Plastic ALDH-axis, where the transitions from ALDH^{High} to ALDH^{Low} and vice-versa were frequent.

Subpopulations exhibited the enrichment of cells with sphere forming ability and maintained repopulation dynamics:

Higher potency of 'Red' and 'Orange' as compared to 'Green' and 'Blue' subpopulations may indicate the possible hierarchical organization with the property of putative stem-like cancer cells (SLCCs) of 'Red' subpopulation, as reported earlier for oral-SLCCs [26]. Therefore, we next tested all four subpopulations for *in vitro* sphere formation assays in three cell lines with distinct phenotypes. Surprisingly, all the four subpopulations from GBC02 cell lines demonstrated sphere-forming potential in tested two generations of sphere propagation (Fig 2A-i). Next, a more critical evaluation of their sphere-forming efficiency was performed by limiting dilution sphere formation assays using 5-500 cells per well taken from the primary spheres of respective subpopulations. This revealed the highest enrichment of sphere-forming cells (1 in 3 cells) in the 'Red' subpopulations supporting its property of putative Oral-SLCCs, as reported earlier. However; interestingly, the 'Orange' (1 in 6 cells), 'Green' (1 in 11 cells) and 'Blue' (1 in 24 cells) subpopulations also demonstrated the enrichment of sphere-forming cells (Fig 2B-i; 2C-i). These differences in efficiency were significant as tested by pair-wise comparisons (Fig 2C-i, last column). Similar to these observations; SCC-070 (CD24+Ve) sub-line also demonstrated the sphere-forming efficiency of 'Green' (1 in 6 cells) to be higher than 'Blue' subpopulation (1 in 16 cells) (Fig 2A-ii 2B-ii, 2C-ii). Furthermore, the interconverting 'Red' and 'Orange' subpopulations of SCC032 showed similar (1 in 159 cells) sphere-forming efficiency (Fig 2A-iii and 2B-iii, 2C-iii).

As shown in Figure 1, the 2D cultures of each sorted subpopulations demonstrated specific dynamics of repopulation. Here, we tested if the clonal sphere-cultures of these subpopulations exhibit the similar pattern of repopulation. Thus, we picked single sphere from each subpopulation of GBC02 cell line after 15 days of sphere growth and plated in separate wells

of 96 well plate. Cells from each subpopulation-specific sphere generated colonies, which were passaged and expanded for 30 days (Supplementary Fig 6A). After 30 days of growth, these subpopulation-specific cultures were tested for CD24, CD44 and ALDH markers. Intriguingly, with close resemblance to the repopulation dynamics in 2D colonies, each clonal sphere-cultures of these subpopulations exhibited the similar pattern of repopulation (Fig 2D, Supplementary Fig 6B). The spheres generated from 'Red' subpopulation remained the most potent to generate all four subpopulations, whereas the 'Green' subpopulation spheres remained committed to generate $CD24^{High}$ subpopulations and 'Blue' demonstrated its specificity to maintain itself, efficiently (Fig 2D, Supplementary Fig 6B), thereby indicating principles underlying the emergence of phenotypic heterogeneity.

Quantitative model of phenotype-transition dynamics:

Next, to demonstrate all possible transition paths which these subpopulations may take in terms of self-maintenance or conversions resulting in the observed repopulation dynamics, we applied a Discrete Time Markov Chain (DTMC) model for a simple linear description of the evolution of phenotypic heterogeneity based on their mutual transitions. Assuming that the transition rates are constant over time, we used *CellTrans* [28] package to infer daily transition rates of all four subpopulation among each other in GBC02, SCC- 029B, SCC-070, GBC035 and SCC032 cell lines. Results were represented as transition graphs (Fig. 3A- i, ii, iii, iv), with arrows representing the direction of transition and the corresponding labels showing the calculated transition rates. All subpopulations showed high sustenance rates (rate with which the cells maintain their identity), with 'Blue' subpopulations having the maximum. Also, majority of transition paths led to the 'Blue' subpopulation with very low transition rate from 'Blue' subpopulation to other subpopulations (efflux rate). Thus, quantitative modelling explained the

experimental observation of low efficiency of transition to other subpopulations from sorted 'Blue' subpopulation.

Using these calculated transition rates, we next generated trajectories of the evolution of population heterogeneity starting from sorted homogeneous subpopulations using the calculated transition rates. As shown in Figure 3B for GBC02 cell line, when starting from a 'Red' sorted subpopulation, the reduction in the fraction of 'Red' subpopulation was associated with an initial rise and fall in the 'Orange' subpopulation and rise in 'Blue' subpopulation, indicating a transition from $CD24^{Low}$ to $CD24^{High}$ phenotype (Fig 3B- i). Furthermore, initial population of homogeneous $CD24^{High}$ phenotypes ('Blue' or 'Green' subpopulations) failed to give rise to $CD24^{Low}$ phenotypes, while the homogeneous 'Green' subpopulations $CD44^{+}/CD24^{High}/ALDH^{High}$ could give rise to $CD44^{+}/CD24^{High}/ALDH^{Low}$ 'Blue' subpopulation and vice-versa (Fig 3B-iii, iv). These trajectories revealed the direction(s) of population dynamics along the CD24- and ALDH-axes. Together, these results, obtained across different cell lines (Supplementary Fig 7A-D) reaffirm our experimental observations of uni-directional transition in CD24-axis and bi-directional plasticity in ALDH-axis.

Next, we performed sensitivity analysis to characterize the effects of cell-to- cell variability in transition rates on the emergence of observed phenotypic compositions. To do so, we shuffled the values in the transition matrices (that contain calculated transition rates), i.e. these *in silico* experiments created 'hypothetical' scenarios where, for instance, the influx and efflux rates were swapped and the effect on phenotypic distributions was calculated (see supplementary methods section for more details) (Supplementary Fig 7E-N). We found that as long as high influx and low efflux of the 'Blue' subpopulation rates were maintained, blue population retained its dominance. These observations identified the necessary and sufficient conditions enabling the 'Blue' subpopulation to be more dominant than the red, orange and green.

Cisplatin alters phenotype-transition dynamics due to plasticity in ALDH-Axis but not CD24-axis:

Previous studies showed that external cues such as chemotherapy-induced stress propels cells to transit into a more drug tolerant state [13, 29] often associated with stem-like phenotypes. Therefore, we explored the influence of chemotherapy on the transition dynamics of the subpopulations for SCC- 029B and SCC-070 cell lines. Sorted four subpopulations were plated separately and grown, as previously described. On the 6th day of cell plating, Cisplatin was added at 2μM dose for 48 hrs followed by recovery in drug free medium for next 2 days. Exposure of Cisplatin during the repopulation, the 'Red' subpopulation, retained higher frequency of itself and generated significantly higher frequency of 'Green' subpopulation in treated cells as compared to the 'Red' subpopulation unexposed to Cisplatin, for both cell lines (Supplementary Fig 8, 9). Similarly, isolated 'Green' subpopulation exposed to 2μM Cisplatin treatment during repopulation retained itself at significantly higher frequency under treated condition as compared to untreated in both cell lines (Supplementary Fig 8, 9). Moreover, the $ALDH^{Low}$ ('Orange' and 'Blue') subpopulations showed significant increase in $ALDH^{High}$ ('Red' and 'Green') subpopulations, respectively when exposed to Cisplatin compared to untreated cells (Supplementary Fig 8, 9). Therefore, our data suggested that the Cisplatin treatment enriched $ALDH^{High}$ subpopulations which may be a result of conversion from $ALDH^{Low}$ subpopulations, possibly due to their intrinsic plasticity. Further, it is also important to note that even under drug induced stress, $CD24^{High}$ subpopulation retained its commitment and did not generate $CD24^{Low}$ subpopulations.

Using the quantitative approach described earlier, compared the transition rates in response to Cisplatin treatment with those in repopulation of treatment *naive* subpopulations. Cisplatin treatment increased the sustenance rates for both $ALDH^{High}$ subpopulations ('Red' and 'Green') and *reduced the efflux rate* from 'Red' or 'Green' subpopulation to any other

subpopulations in both SCC-029B and SCC-070 cell lines (Fig 3C-i, ii). Further supporting the transition of $ALDH^{Low}$ subpopulations to $ALDH^{High}$ subpopulations upon Cisplatin treatment, repopulation from the 'Orange' and 'Blue' subpopulations in both SCC-029B and SCC-070 cell line showed stark contrast as compared to observations in untreated conditions. For both the cell lines, the quantitative model demonstrated a higher rate of transition from 'Orange' to 'Red' and 'Blue' to 'Green' subpopulation which was either absent or at a lower rate in untreated cells. The trajectories of the evolution of population heterogeneity in response to Cisplatin treatment is provided as Supplementary Fig 10.

To confirm that the bidirectional plasticity of ALDH-Axis is harnessed by $ALDH^{Low}$ cells to accumulate $ALDH^{High}$ subpopulation in response to Cisplatin treatment, we used SCC-070 (CD24+Ve) sub-line of cells. The 'Green' and 'Blue' subpopulations were separated by FACS and plated. On 6th day of cell plating, we added a sub-lethal-dose of 2μM Cisplatin for 48 hrs or kept the cells untreated as control. Drug treatment was followed by recovery in drug free medium for next 48 hrs. As anticipated, the 'Green' subpopulation maintained itself at higher frequency and 'Blue' subpopulation repopulated into 'Green' subpopulation. Importantly, both 'Blue' or 'Green' subpopulations failed to give rise to $CD24^{Low}$ phenotypes even in response to the Cisplatin treatment (Fig 3D, E). Assessment of this population dynamics revealed reduced sustenance rates of 'Blue' subpopulation and increased efflux rate but increased sustenance rate of the 'Green' subpopulations in Cisplatin induced condition (Fig 3F). Dynamic trajectories plotted to predict the evolution of subpopulations for untreated and Cisplatin treated conditions predicted that 'Green' subpopulation could maintain its higher fraction and homogeneous 'Blue' subpopulation resulted in generation of 'Green' subpopulation at higher rate (Fig 3G-ii, iv) as compared to the untreated homogeneous subpopulations (Fig 3G-i, iii). These results strongly demonstrated the plastic nature of ALDH-Axis to be responsible for accumulation of $ALDH^{High}$ subpopulations under

Cisplatin treatment condition; however unidirectional transitions ($CD24^{Low}$ to $CD24^{High}$) along the CD24-Axis was strictly maintained.

Transcriptome profiling revealed different levels of cell maturation among subpopulations:

The observed dynamic relationship among subpopulations prompted us to determine the molecular association between these subpopulations of cells. Towards this, we performed RNA sequencing of these four subpopulations isolated from 2D cultures and 3D- spheres generated from 'Red' and 'green' subpopulations separated from GBC02 in three biological repeats. Based on the overlap between sets of differentially expressed genes (DEGs) obtained from pair-wise comparisons we found varying degrees of overlap in DEGs between different comparisons (Fig 4A). For instance, 79% of overlap was seen for DEGs between 'Orange' vs. 'Blue' subpopulations and 'Orange' vs. 'Green' subpopulations. Similarly, 76% of DEGs for 'Red' vs. 'Green' subpopulations were found in 'Red' vs. 'Blue' DEG analysis. Conversely, 71% of DEGs noted in 'Red' and 'Blue' were witnessed in 'Red' vs. 'Green' DEGs-list. (Fig 4A). This pattern of overlap pointed out that 'Red' subpopulation may be approximately equidistant from 'Blue' and 'Green' subpopulations. Similarly, 'Orange' may also be approximately equidistant from 'Green' and 'Blue' subpopulations. Further, the DEGs between 'Blue' and 'Green' subpopulations had very few genes in common when overlapped with any of the five other DEG lists (see column 1 in Fig 4A), indicating that these subpopulations may be transcriptionally most similar to one another. These overlaps in DEGs, coupled with quantifying distances among these comparisons enabled us in arranging these subpopulations in a hierarchical structure with 'Red' subpopulations being the most upstream, followed by 'Orange', 'Green' and 'Blue' (Fig 4B).

Since the 'Red' subpopulation was the most distinct from other subpopulations; the pair-wise comparisons were made to tabulate DEGs between 'Orange', 'Green', 'Blue' and 'Red'

subpopulations. This showed differentially expression of 568 upregulated & 582 downregulated genes between 'Red' & 'Orange' subpopulations; 812 upregulated & 908 downregulated genes between 'Red' & 'Green' subpopulations and 959 upregulated & 968 downregulated genes between 'Red' & 'Blue' subpopulations with log fold change of more than 2 and adjusted *p* value of less than 0.05 (Fig 4C, Supplementary Table 1). Differentially expressed genes from pairwise comparisons are also depicted as volcano plots (Supplementary Fig 11). Interestingly, $ALDH^{High}$ ('Red' and 'Green') subpopulations showed significantly higher expression of ALDH1A3 as compared to their respective $ALDH^{Low}$ subpopulations, 'Orange' and 'Blue' (Supplementary Fig 11). To derive biological inferences of the differential expression of genes among these subpopulations, we performed gene set enrichment analysis (GSEA) from the larger list of DEGs ($q<0.1$, $\log_2 FC > 0.5$) between Red and Orange, Red and Green, Red and Blue subpopulations. We observed depletion for the GPCR-related gene expression signatures in all three subpopulations compared to the 'Red' subpopulations (Supplementary Fig 12A), indicating the loss of GPCR-signaling related to the change in transcription state from 'Red' to other subpopulations. Intriguingly, we observed enrichment of signatures related to keratinization and cornified envelope with the DEGs in 'Green' compared to the 'Red' subpopulations (Fig 4D). Similarly, the 3D-sphere cultures of the 'Green' subpopulation showed an enrichment of these signatures compared to the 3D-sphere cultures generated from 'Red' subpopulation (Fig 4E). This suggested the onset of differentiation in 'Green' subpopulation of cells, even when maintained in 3D sphere cultures.

A recent study by Jones *et. al.*, 2019 reported diversity within basal layer oral epithelial progenitor cells (OEPCs) in normal oral mucosa of mouse, using single cell RNA sequencing [30]. This study has demonstrated the maintenance of stemness with onset of differentiation in subsets of basal layer keratinocytes in mouse oral mucosa [30]. Therefore; to explore the similarities, we

overlapped our differentially expressed genes with the discovered gene sets of stem/progenitors and differentiating cells of basal layer keratinocytes. Interestingly, from the DEGs between 'Red' and 'Green' subpopulations; the upregulated genes in 'Red' subpopulation had significantly higher overlap with OEPC-stem cell cluster signature (Fig 4F); whereas, the upregulated genes in 'Green' subpopulation had almost equal number of genes overlapped to the gene-list of OEPC-stem cell cluster as well as differentiating cell cluster (Fig 4F). Very similar patterns were obtained even with DEGs between the 3D-spheres generated from 'Red' and 'Green' subpopulations (Fig 4G). Similar to the 'Green' subpopulation the upregulated genes in 'Orange' and 'Blue' subpopulations as compared to the 'Red' also had equal overlaps with the gene-list of OEPC-stem/progenitors cell cluster as well as differentiating cell cluster (Supplementary Fig 12B). Further, we explored the enrichment of transcription factors (TFs) specific to stem/progenitors and differentiating clusters from this study. Interestingly, all four subpopulations showed mosaic expression pattern of these TFs (Supplementary Fig 12C). This prompted us to explore if our four subpopulations exhibit the signature of stem cells. Hence, we performed single sample gene set enrichment analysis (ssGSEA) utilizing the previously reported gene-sets for Adult Tissue Stem Cells (ATSC) [31]. All the four subpopulations as well as the 3D-spheres of 'Red' and 'Green' subpopulations showed positive enrichment with no significant differences among them for the stemness signature (Supplementary Fig 12D). Encouraged by these results, we further evaluated the pattern of expression of 65 different genes which are highly studied for their roles in maintenance of stemness in various cancer types. Interestingly, we found specific subsets of genes expressed distinctly in 'Red' (Fig 4H(i)); or both 'Red' and 'Orange' (Fig 4H(ii)); or commonly in 'Orange', 'Green' and 'Blue' subpopulations (Fig 4H(iii), or among all the four subpopulations (Fig 4H(iv)). This suggests that the stemness may be maintained in these subpopulations and regulated by distinct gene expression networks.

To define the cell state specific gene expression signature of these four subpopulations, we generated a list of unique differentially expressed genes (unique-DEGs) specific to each subpopulation (Supplementary Table 2). Since, the 'Green' and 'Blue' subpopulations had very similar pattern of gene expression, the 'Blue' subpopulation was excluded from this analysis. Gene Ontology (GO) analysis for subpopulation specific unique-DEGs using the cytoscape tool- BiNGO, resulted in overrepresentation of specific pathways for these subpopulations (Supplementary Table 3,4,5). Interestingly, while the 'Red' subpopulation with $CD24^{Low/}ALDH^{High}$ phenotype overrepresented the process of organ development and lipid metabolism, the $CD24^{High/}ALDH^{High}$, 'Green' subpopulation had overrepresentation of organ development with differentiation processes.

Based on these results, we inferred that the spontaneous emergence of $CD24^{High/}ALDH^{High}$ subpopulation from the $CD24^{Low/}ALDH^{High}$ is a process of cellular differentiation with maintained transcriptional status of stem cells. We validated these results in spheres generated from four separated subpopulations of GBC02 by qRT-PCR for TFs responsible for maintenance of stemness (Fig 4J-i) as well as cytokeratin markers associated with basal and suprabasal layer of keratinocytes (Fig 4J-ii). While all four subpopulations-derived spheres maintained basal cytokeratin markers *CK14*, *CK5*, *CK19* at the similar levels; 'Green', 'Orange' and 'Blue' subpopulations expressed differentiation markers *CK13*, *CK1* and *CK10* at significantly higher levels compared to the 'Red' spheres. Interestingly, while *KLF4*, *CMYC*, *OCT4* and *NANOG* expression was variable among all subpopulations; *Sox2* was significantly downregulated while *SOX9* was expressed at higher level in 'Green', 'Blue' and 'Orange' spheres compared to the 'Red' spheres. siRNA mediated knockdown of *SOX9* expression in SCC-070 cells indeed displayed alteration in phenotype distribution (Fig 4K-i) and lower expression of *ALDH1A3* (Fig 4K-ii), suggesting the crucial role of *Sox9* in maintaining $ALDH^{High}$ phenotypes.

Evaluation of prognostic value of population-specific gene expression pattern:

To explore the functional relevance of transitions observed between cell states, we utilized the list of unique-DEGs to test its correlation with clinical outcome of HNSCC patients from TCGA cohort. Mean ssGSEA-score was calculated for subpopulation specific unique-DEGs, for the top most upregulated 50 to 100 genes. Similarly, with same subpopulation specific gene-set, mean ssGSEA score was calculated for each patient from their respective gene expression data and patients were classified in 'low' or 'high' groups. The statistical differences in the survival of these groups were estimated using Kaplan–Meier (KM) curves and Cox-regression analyses. Importantly, among all the comparisons, statistically significant difference in survival was obtained with the top 50 and 80 upregulated unique DEGs in 'Red' subpopulation (Fig 5A-i and Supplementary Fig 14A). Conversely, the patients with higher ssGSEA-score calculated from the most downregulated unique-DEGs showed a trend of poorer prognosis (Fig 5B-i and Supplementary Fig 14B). We verified this result by an alternate approach where survival analysis was performed using the most upregulated or downregulated unique-DEGs as query on cBioPortal web tool [32]. Similar to the results obtained with ssGSEA score; patients with alterations in upregulated unique-DEGs correlated significantly with better prognosis (Fig 5A-ii, iii); whereas, alterations in downregulated unique-DEGs significantly correlated with poorer progression free survival and a trend of poorer disease specific survival in patients (Fig 5B-ii, iii). Unique DEGs were highly expressed in the respective subpopulation and showed minimal expression and variance in other subpopulations. Therefore, individual patients with lower ssGSEA score for upregulated Unique-DEGs specific to 'Red' subpopulation or higher expression of genes which are uniquely downregulated in 'Red' subpopulation; represented the enrichment of other subpopulations, representing cellular heterogeneity and was associated with poor

prognosis in these patients. Therefore, one of the possible interpretations for this observation could be that patients with 'Red' subpopulation that lacked the ability to generate other subpopulations were likely to have better survival as compared to those where 'Red' subpopulation was able to generate heterogeneous subpopulations. For exploring this postulation, we calculated the IC_{50} value of Cisplatin among three different oral cancer cell lines with distinct 'Red' subpopulation driven heterogeneity pattern (Fig 5C). We found that the 3D-sphere cultures of heterogeneous cell lines GBC02 and SCC-029B (where the four subpopulations originates from 'Red' subpopulation) exhibited higher IC_{50} of Cisplatin compared to SCC-032 cell line (where 'Red' subpopulation fails to generate 'Green' subpopulation (Supplementary Fig 13B)), highlighting the spontaneous generation of $CD44^{+Ve}/CD24^{High}/ALDH^{High}$ 'Green' subpopulation from $CD44^{+Ve}/CD24^{Low}/ALDH^{High}$ 'Red' subpopulation (Fig 5C). Next, we tested the relative Cisplatin sensitivity of $CD24^{High}$ and $CD24^{Low}$ subtypes of cells isolated from GBC02 cell line. Interestingly, even at higher dose (10 μM) of Cisplatin treatment for 48 Hrs, $CD24^{Low}$ subtypes showed significantly higher Caspase 3/7 activity, indicative of higher induction of apoptosis compared to $CD24^{High}$ subtypes of cells (Fig 5D). RT-PCR analysis revealed highest expression of drug efflux gene ABCG2 in the 'Green' subpopulations among all four in both 2D and 3D-sphere culture of GBC02 in treatment naïve condition (Fig 5E), which may be responsible for Cisplatin refraction by $CD24^{High}$ subtype of cells. Further, the heterogeneous parent cultures in 2D as well as 3D-sphere conditions, significantly enriched for $CD44^{+Ve}/CD24^{High}/ALDH^{High}$ 'Green' subpopulation after 48 Hr of exposure to sub-lethal dose of Cisplatin (2 μM) (Fig 5F, Supplementary Fig 13A). Therefore, spontaneous transition of $CD24^{Low}$ to $CD24^{High}$ subtype and Cisplatin induced plasticity of $ALDH^{Low}$ to $ALDH^{High}$ state collectively result in emergence of $CD44^{+Ve}/CD24^{High}/ALDH^{High}$ 'Green' subpopulation as Cisplatin tolerant state of cells.

Discussion:

Experimental cell cultures show the hallmark of maintaining diverse cancer cell subpopulations at a specific frequency during the extended period of passages and cell division. Studying such phenomenon in breast cancer; luminal cells ($CD44^{Lo}/CD24^{High}/EpCAM^{High}$) are shown to transition to stem ($CD44^{High}/CD24^{Negative}/EpCAM^{Low}$) and basal like states ($CD44^{High}/CD24^{Neative}/EPCAM^{Negative}$) and returning to a fixed equilibrium of cell-state proportions over time, in stochastic manner [20]. This phenomenon further proposed the *de novo* generation of SLCCs from non-SLCCs; having impact on the therapeutic strategies designed to exclusively target SLCC-subpopulations [20]. Similarly, in a different study where breast cancer cell populations demonstrated distinct dynamics and favored to drive the equilibrium to accumulate drug tolerant state of cells with $CD44^{High}/CD24^{High}$ phenotype in response to chemotherapeutic stress [33]. Thus, these phenotypic markers have successfully demonstrated the extent of intratumoral heterogeneity (ITH), biology of phenotype switching as well as the strategy of cancer cells to respond to the given therapy. Here, using diverse markers of putative oral-SLCCs, we have characterized diversity among CD44-positive oral cancer cells lines with respect to distinct expression of CD24 and ALDH-activity. Importantly, these characterized subpopulations served as experimental models and reliably demonstrated the population trajectories, their distinct transcription states as well as spontaneous or Cisplatin driven transitions. We also established the clinical relevance of phenotypic cell states driven ITH in oral cancer patients. Our study strongly suggested strict unidirectional conversion of $CD24^{Low}$ phenotype of cells to acquire $CD24^{High}$ phenotype; however, stochastic bidirectional plasticity was observed on ALDH- Axis in both $CD24^{Low}$ and $CD24^{High}$ phenotypes in oral cancer cells. Experimental observations of population trajectories were validated by mathematical modelling where $CD44^{+Ve}/CD24^{Low}/ALDH^{High}$ ('Red') subpopulation self-sustained itself and repopulated all other subpopulations to reconstitute the

phenotypic heterogeneity. Additionally, population trajectories also highlighted the sufficiency of $CD24^{High}$ subpopulation to maintain itself as long- term 2D and 3D cultures in 'Red' subpopulation depleted conditions, without being dependent on the 'Red' subpopulation or repopulating it. With this, $CD44^{+Ve}/CD24^{High}/ALDH^{Low}$ ('Blue') subpopulations showed highest self-sustenance with the majority of transition paths leading to this subpopulation. Collectively, these observations are in contrast to the stochastic interconversions of different phenotypic subpopulations as reported in breast cancer cells [20, 34] and suggest a hierarchical structure between these subpopulations.

Recently, Jones et al reported novel lineage relationships between diverse keratinocytes residing in the basal layer of mice oral mucosa [30]. In addition to the oral epithelial progenitor cells (OEPCs); mouse oral mucosal basal layer cells were found to harbour maturing keratinocytes; thus revealing a continuum of cell differentiation states. Interestingly, some of these differentiating keratinocytes continued to co-express high levels of cytokeratin 14 (CK14) and genes associated with both OEPCs and differentiation processes, representing the transitional intermediated state of cells [30]. Further, most of the basal layer cells were found to be cycling. Taking leads from this significant finding, our results demonstrated similarities between tissue hierarchy in normal mice oral mucosa and population trajectories in oral cancer cells. As anticipated, the 'Red' subpopulation with $CD44^{+Ve}/CD24^{Low}/ALDH^{High}$ phenotype showed similarity with the transcriptome state of OEPCs, supporting the reports of enrichment of putative Oral-SLCCs in $CD44^{+Ve}/ALDH^{High}$ cell population. Strikingly, other subpopulations emerging from the 'Red' subpopulation had transcriptome overlaps with both OEPCs and differentiating cells of oral mucosal basal layer cells representing the transitional intermediated state of cells. The GSEA result with the differentially expressed genes between 'Red' and 'Green' subpopulation clearly showed the onset of differentiation in 'Green' subpopulation. Thus, the nonreversible transition of $CD24^{Low}$ cells to $CD24^{High}$ phenotype is possibly due to the onset of differentiation in $CD24^{High}$

cells. Also, the DEGs between ALDH^{High} ('Red') and ALDH^{Low} ('Orange' and 'Blue') subpopulations did not result in any specific gene sets enrichment (data not shown). This could be due to the plastic nature of ALDH^{Low} subtype of cells which may have cells with mixed transcriptome states masked in the bulk RNA-sequence data. These are interesting propositions to be explored in future.

Further, Sox2 was highly expressed in 'Red' subpopulation and other subpopulations showed loss of Sox2 with gain of Sox9 expression. In concordance with the recent report where depletion of Sox9 had resulted in downregulation of ALDH1 in both Sox2 and Sox9 expressing oral cancer cells [35], we also found the downregulation of ALDH1A1 expression and loss of ALDH^{High} 'Red' and 'Green' subpopulation after silencing of Sox9 expression. Both Sox2 and Sox9 expression is shown to play important role in maintaining stemness in cancer [36, 37] and gain of Sox9 expression in cancer cells is linked with lineage infidelity in squamous cell carcinoma driving wound repair and cancer [38]. Therefore, we may suggest that the transitional intermediated state of 'Green' 'Orange' and 'Blue' subpopulations may enrich cancer cells with Sox9-positive alternate state of stemness in oral cancer along with the reported oral-SLCCs enriched in 'Red' subpopulation. Results obtained with multiple independent oral cancer cells lines for *in vitro* sphere formation assay strongly supported this possibility and must be explored in future. To the best of our knowledge, this is the first study in oral cancer where phenotypic states of cells could be mapped to the transcriptome states demonstrating the cellular hierarchy. Moreover, our study for the first time proposed the possibility of alternate states of putative stemness maintained within intermediate differentiating subpopulations in oral cancer cells irrespective of their genetic and epigenetic diversity.

The impact of ITH on clinical outcomes is a major focus of cancer research [39, 40]. Tumor samples are usually employed to estimate the extent of ITH using genomics approaches

[41, 42]. However, efforts on developing genomics-based data for investigating ITH and its clinical significance using isogenic subpopulations from patient derived cell cultures, have been limiting. This is worth to be mentioned because patient-derived cell cultures not only serve as essential tools for modelling disease heterogeneity but also is essential to perform functional validation of molecular mechanisms associated with observed complexities. Insights gained from such models can be cross-validated and translated using patient cohort. This is evident from our study where a set of subpopulation-specific unique-DEGs allowed us to identify subgroups of patients in HNSCC-TCGA cohort. With High- abundance of top-unique DEGs specific for 'Red' subpopulation was having better prognosis. In support of this finding, a recent study has similarly shown better prognosis in $Sox2^{High}/Sox9^{Low}$ expressing group as compared to $Sox2^{Low}/Sox9^{High}$ expressing HNSCC patients in TCGA cohort [35]. Since, we have found $Sox2^{High}/Sox9^{Low}$ gene expression pattern in 'Red' subpopulation with loss of Sox2 and gain of Sox9 in 'Green' 'orange' and 'Blue' subpopulation; this study along with our findings strongly suggest that patients having higher abundance of 'Red' subpopulation may have better prognosis than those where 'Red' subpopulation specific unique-DEGs were lost. One of the possible mechanisms for loss of 'Red' subpopulation specific unique-DEGs could be the population dynamics described by us, resulting in accumulation of heterogeneous subpopulations in aggressive oral tumors.

Tumor heterogeneity is a major contributor to therapy failure. We have only started to unravel the complexity of chemotherapy resistance. One of the emerging possibilities is the transition of cancer cells across a continuum of states in response to chemotherapy to develop drug tolerance [43]. In this study, we have demonstrated that the transient exposure of low doses of Cisplatin was sufficient to induce transition of $ALDH^{Low}$ subpopulations ('Orange' and 'Blue') into the $ALDH^{High}$ subpopulations ('Red' and 'Green'); whereas, the higher dose had resulted in increased induction of apoptosis in cells with $CD24^{Low}$ phenotype status. With the use of $CD24^{Low}$-

depleted cells, low-dose Cisplatin treatment and mathematical modelling, we validated that the accumulation of drug tolerant $CD44^{+Ve}/CD24^{High}/ALDH^{High}$ subpopulation was a result of the induction of phenotype switching rather than selection by Cisplatin, in multiple oral cancer cell lines. Indeed in earlier studies, cells with $CD24^{High}$-phenotype marked a transient chemo-resistant cell state in lung cancer [29] and breast cancer [33]. Also, increased CD24 expression with Oct4 and Nanog was recently correlated with poor response to chemo-radio therapy and unfavorable prognosis in oral cancer [44]. Interestingly, Cisplatin selected oral cancer cell line was shown to express higher levels of Sox9 [35]; which is expressed at higher levels in $CD44^{+Ve}/CD24^{High}/ALDH^{High}$ subpopulation in our study, too. Therefore, we support the notion that phenotype switching may act as a robust mechanism for cancer cells to rapidly acquire a drug tolerant state in changing environment from where these cells may eventually emerge as a drug-resistant population in relapsed tumor.

In summary, the current study identifies the maintenance of cell-state proportions by spontaneous transition due to the process of differentiation as well as interconversion of cells between specific cell states. Further, we unmasked the ability of oral cancer cells of gaining adaptive-resistance against Cisplatin by altering the inter-conversion rates of cell-states transitions *in vitro,* irrespective of their genetic and epigenetic diversity. This study is not only useful for our ability to relate cell-states for theoretical and experimental studies but also have translational implications. Our findings indicate that the co-existence of heterogeneous subpopulations may influence overall tumor behavior, which could be due to the cooperation between them. Therefore, the phenotypic subgroups described here may be optimum to estimate the functional heterogeneity among oral cancer patients. Further, our cellular models may serve as tool to define the mechanisms underlying cross-talk among phenotypes and identify targets against the population-dynamics, as novel therapeutic strategy.

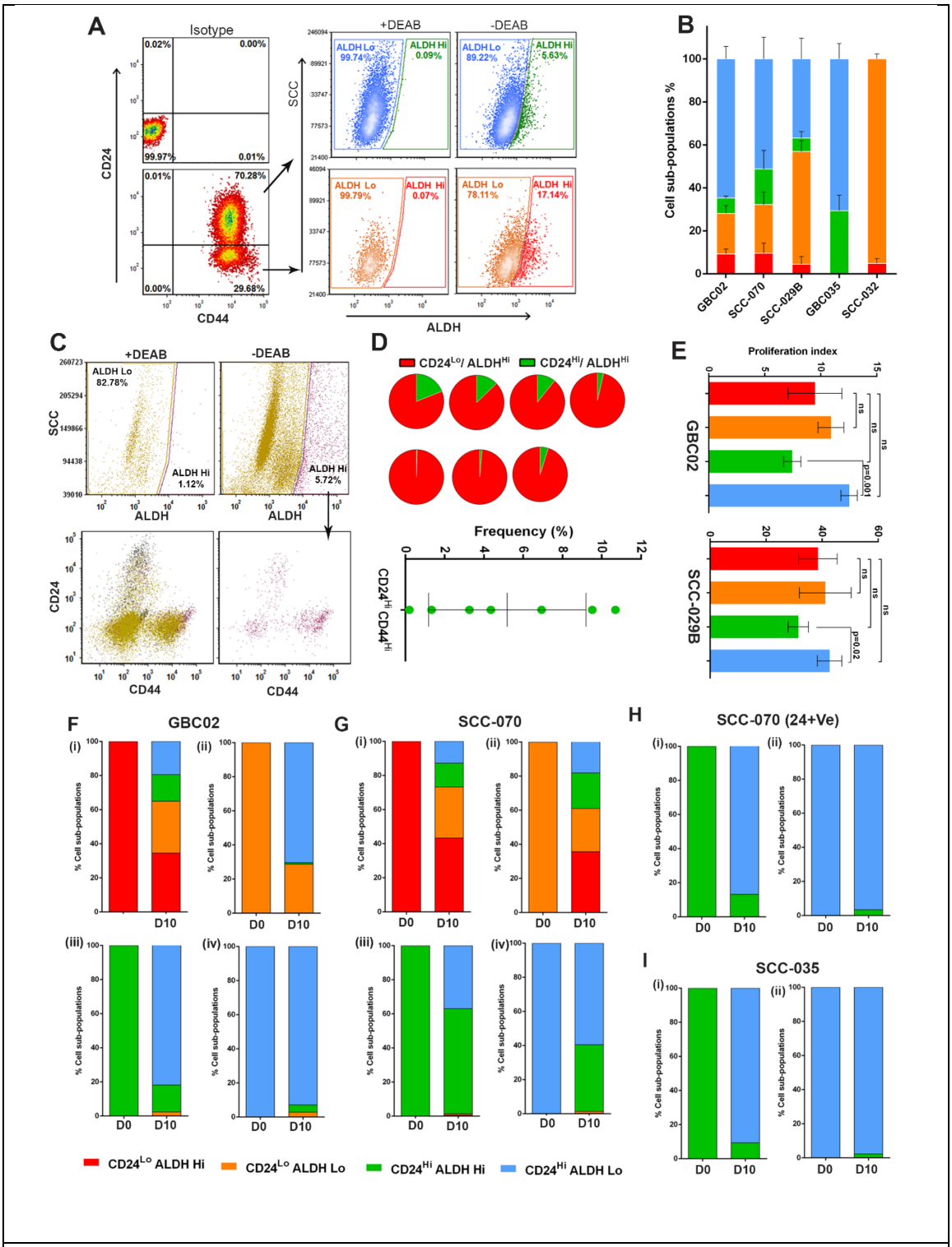
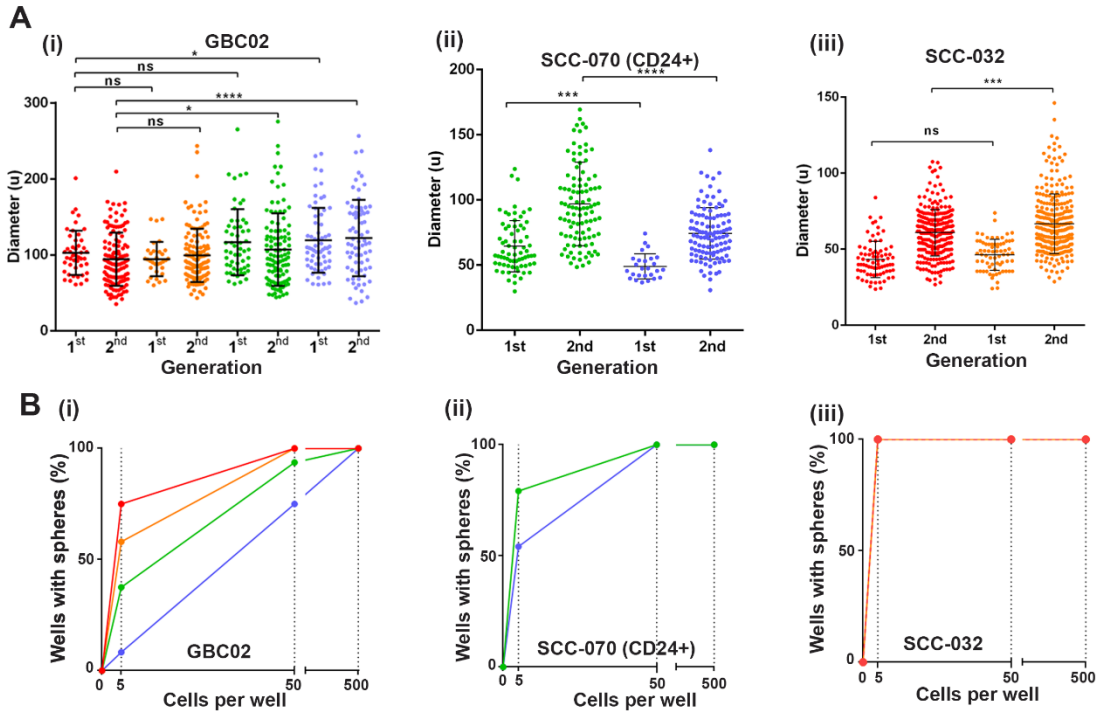


Figure 1: *Subpopulations exhibited distinct repopulation dynamics:* **(A)** Representative FACS dot plots of GBC02 cell line's CD24/CD44 staining (Left bottom) with respect to Isotype control (Left top). ALDEFLUOR (ALDH) phenotype of $CD24^{High}/CD44^{+Ve}$ sub-population (Right top) and $CD24^{Low}/CD44^{+Ve}$ sub-population (Right bottom) in presence or absence of DEAB. **(B)** Frequency distribution of the four sub-populations $CD24^{Low}/ALDH^{High}$ (Red), $CD24^{Low}/ALDH^{Low}$ (Orange), $CD24^{High}/ALDH^{High}$ (Green) and $CD24^{High}/ALDH^{Low}$ (Blue) in five genetically distinct oral cancer

cell lines. Error bars represent mean ± SEM from three biological repeats. **(C)** Representative FACS dot plots of a freshly resected and digested patient tumor. Hematopoietic lineage negative cells segregated into $ALDH^{High}$ and $ALDH^{Low}$ cells in presence or absence of DEAB (top). CD24/CD44 phenotype of all lineage negative cells (bottom left). CD24/CD44 phenotype of $ALDH^{High}$ cells only (bottom right). **(D)** Pie charts showing frequencies of $CD24^{High}/ALDH^{Hi}$ (Green) and $CD24^{Low}/ALDH^{High}$ (Red) subpopulations from 7 freshly resected human oral tumor samples and frequencies of $CD24^{High}/ALDH^{High}$ cells from all 7 oral tumor patient samples. **(E)** Graphs showing proliferation index of GBC02 and SCC029 cell lines based on CellTrace-Violet dye dilution assay. **(F, G, H, I)** Repopulation frequencies of each subpopulation in (i) Red (ii) Orange (iii) Green and (iv) Blue sorted cells on Day-0 and Day-10 of sorting in GBC02 (F), SCC-070 (G), GBC035 (H) and SCC-070 (CD24+) sub-line (I).



C

(i)

GBC02	Confidence intervals for SLCCs frequency 1/(Estimate)			Pairwise tests for differences in stem-like cell frequencies
Group	Lower	Estimate	Upper	Chisq P-value
$CD24^{Lo}/ALDH^{Hi}$	4.23	2.97	2.15	RO=0.00189, RG=7.26e-08, RB=6.57e-17
$CD24^{Lo}/ALDH^{Lo}$	8.64	6.24	4.55	OG=0.0172, OB=3.9e-08
$CD24^{Hi}/ALDH^{Hi}$	15.24	11.02	8.01	GB=0.00159
$CD24^{Hi}/ALDH^{Lo}$	33.1	23.88	17.27	

(ii)

SCC-070 (CD24+)	Confidence intervals for SLCCs frequency 1/(Estimate)			Pairwise tests for differences in stem-like cell frequencies
Group	Lower	Estimate	Upper	Chisq P-value
$CD24^{Hi}/ALDH^{Hi}$	9.92	5.98	3.71	GB=0.00525
$CD24^{Hi}/ALDH^{Lo}$	24.73	15.68	10.01	

(iii)

SCC-032	Confidence intervals for SLCCs frequency 1/(Estimate)			Pairwise tests for differences in stem-like cell frequencies
Group	Lower	Estimate	Upper	Chisq P-value
$CD24^{Lo}/ALDH^{Hi}$	317	159	79.3	RO=1
$CD24^{Lo}/ALDH^{Lo}$	317	159	79.3	

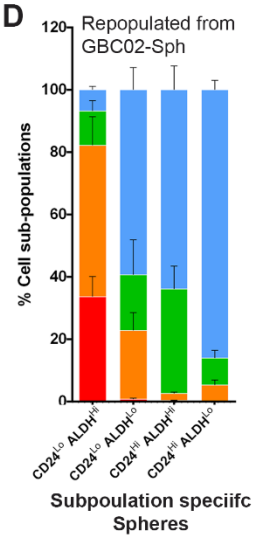


Figure 2: *Enrichment of cells with sphere forming ability by all four subpopulations:* **(A)** Size of the four subpopulations spheres from GBC02, SCC-070 CD24+ and SCC-032 cell lines. **(B)** Second generation spheroid formation frequencies in limiting dilution assay from GBC02, SCC-070 CD24+ and SCC-032 cell lines. **(C)** Estimation of the SLCC frequencies in each subpopulation of (i) GBC02, (ii) SCC-070 CD24+ and (iii) SCC-032 cell lines using ELDA software. **(D)** Frequencies of each subpopulation in colonies generated from Red, Orange, Green and Blue subpopulation specific single spheroids after 30 days of sorting. Error bars represent mean ± SEM from four biological repeats.

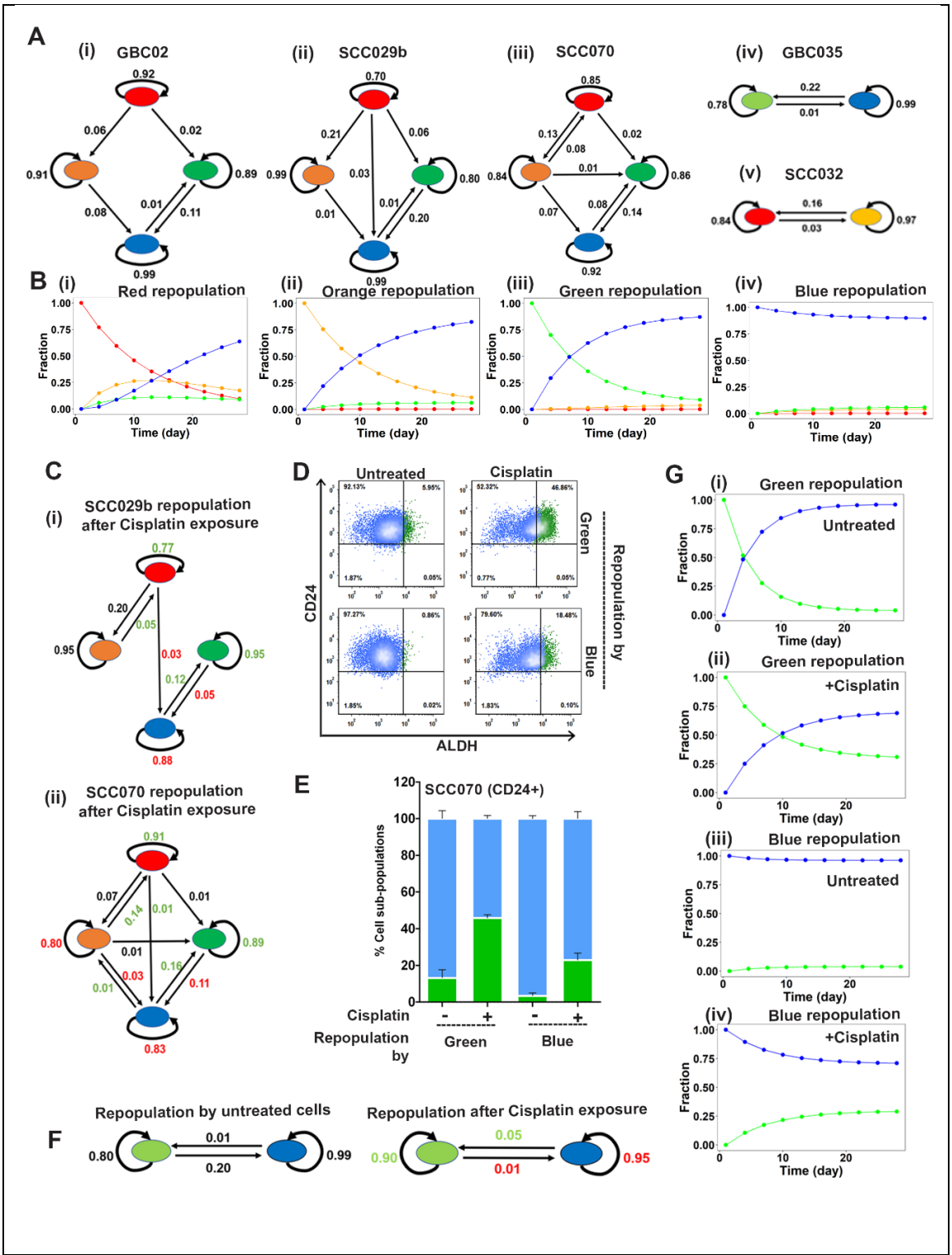
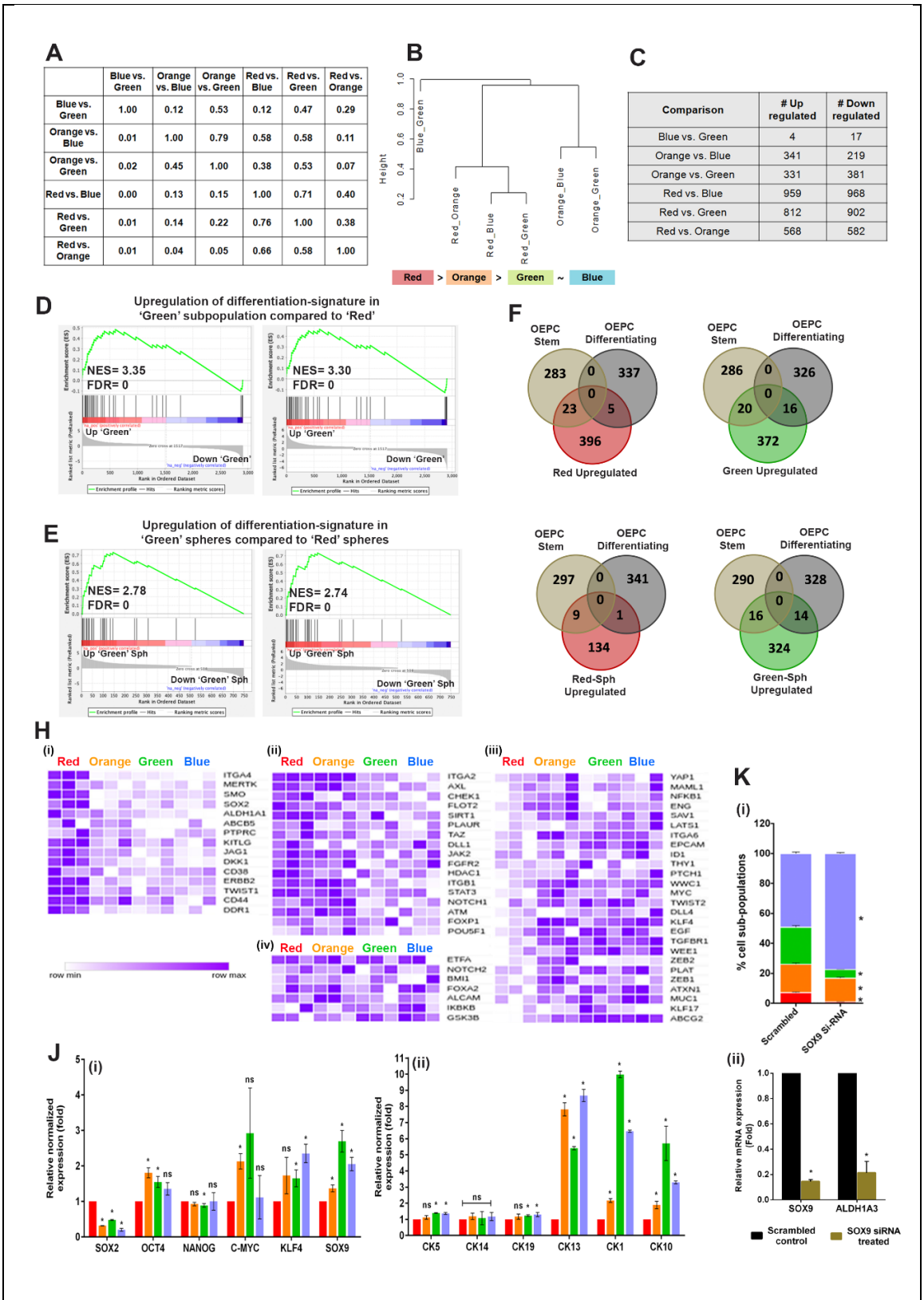


Figure 3: *Spontaneous and Cisplatin induced population dynamics:* **(A)** The phenotypic transition graphs of (i) GBC02, (ii) SCC029, (iii) SCC070, (iv) GBC035 and (v) SCC032. Lack of arrows between 2 subpopulations indicates that the transition rate in that direction is less than 0.01. **(B)** The phenotypic transition trajectories predicted based on the transition graph for GBC02 cell line's (i) Red (ii) Orange (iii) Green (iv) Blue subpopulations repopulations. **(C)** The

phenotypic transition graphs of SCC029 and SCC070 subpopulations treated with 2µM Cisplatin showing increased influx and sustenance rate of the green state and its reduced outflux, overall increasing the fraction of green subpopulation under Cisplatin treatment. **(D)** Representative dot plots of SCC070 CD24+ sub-line's repopulation showing increase in Green cells in sorted Green (top) and Blue (bottom) subpopulations in untreated and 2µm Cisplatin treated conditions. **(E)** Repopulation frequencies of SCC070 CD24+ sub-line's Green and Blue subpopulations in untreated vs. Cisplatin (2µm) treated conditions. Error bars represent mean ± SEM from two biological repeats. **(F)** The phenotypic transition graph of Green and Blue subpopulations of SCC070 (CD24+) sub-line in untreated and Cisplatin treated conditions. **(G)** The phenotypic transition trajectories of (i, iii) Untreated Green and Blue cells and (ii, iv) Cisplatin treated Green and Blue cells repopulations of SCC070 CD24+ sub-line showing increased fraction of Green cells in the population upon Cisplatin treatment.



	Blue vs. Green	Orange vs. Blue	Orange vs. Green	Red vs. Blue	Red vs. Green	Red vs. Orange
Blue vs. Green	1.00	0.12	0.53	0.12	0.47	0.29
Orange vs. Blue	0.01	1.00	0.79	0.58	0.58	0.11
Orange vs. Green	0.02	0.45	1.00	0.38	0.53	0.07
Red vs. Blue	0.00	0.13	0.15	1.00	0.71	0.40
Red vs. Green	0.01	0.14	0.22	0.76	1.00	0.38
Red vs. Orange	0.01	0.04	0.05	0.66	0.58	1.00

Comparison	# Up regulated	# Down regulated
Blue vs. Green	4	17
Orange vs. Blue	341	219
Orange vs. Green	331	381
Red vs. Blue	959	968
Red vs. Green	812	902
Red vs. Orange	568	582

Figure 4: *Subpopulation specific transcriptome analysis:* **(A, B)** Overlap in DEGs of different pairwise comparisons**.** Proportion of overlap between each pairwise comparison (A).

Hierarchical structure of different cell sub-populations (B). **(C)** Pairwise comparisons of differentially expressed genes of four subpopulations of GBC02 cell line. **(D)** Gene set enrichment analysis (GSEA) with genes up-regulated in 'Green' subpopulation as compared to 'Red' subpopulation from GBC02 monolayer cultures showing enrichment of genes involved in Keratinization and Cornified envelope formation. **(E)** Gene set enrichment analysis (GSEA) with genes up-regulated in Green subpopulation as compared to Red subpopulation from GBC02 3D spheroids showing enrichment of genes involved in keratinization and cornified envelope formation. **(F)** Venn diagram showing overlap of basal OEPC and basal differentiating gene sets with Red vs. Green comparison upregulated genes in Red & Green sorted subpopulations monolayer cultures and **(G)** Red & Green 3D spheroids from GBC02 cell line. **(H)** Heatmaps of mean normalized expression values from GBC02 monolayer cultures for genes expressed distinctly in (i) 'Red' (ii) both 'Red' and 'Orange' subpopulations (iii) commonly in 'Orange', 'Green' and 'Blue' subpopulations and (iv) among all the four subpopulations. **(I)** qRT-PCR of the four sorted subpopulations from GBC02 monolayer cultures for various (i) stemness genes and (ii) differentiating cytokeratins. **(J)** (i) Graph showing decreased frequency in $ALDH^{High}$ subpopulations in SCC-070 cell line upon siRNA mediated silencing of SOX9 and (ii) qRT-PCR after siRNA knockdown of SOX9 in SCC-070 cell line.

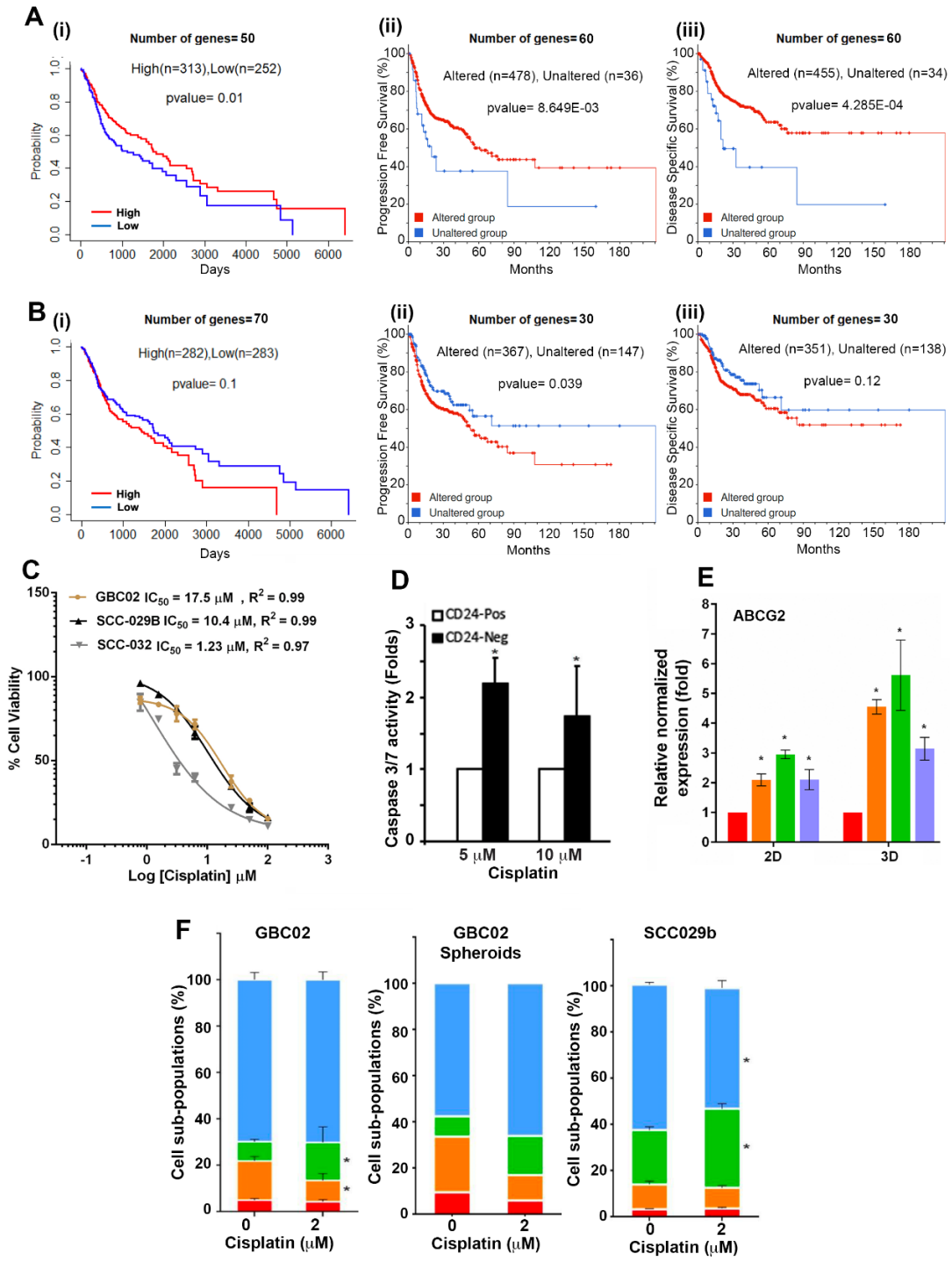


Figure 5: *Prognostic significance of emergence of CD24^{High} Cisplatin tolerant subpopulations:* **(A)** (i) Kaplan Meier curves of HNSCC patients segregated into High expression (Red line) and Low expression (Blue line) groups based on top 50 uniquely up-regulated genes of 'Red' subpopulation. (ii) Kaplan Meier curves for progression free survival and (iii) Disease specific

survival for HNSCC patients having altered (Red line) or unchanged (Blue line) expression for top 60 upregulated genes in 'Red' subpopulation. **(B)** (i) Kaplan Meier curves of HNSCC patients segregated into High expression ('Red') and Low expression ('Blue') groups based on top 70 uniquely down-regulated genes of 'Red' subpopulation. (ii) Kaplan Meier curves for progression free survival and (iii) Disease specific survival for HNSCC patients having altered (Red-line) or unchanged (Blue-line) expression for top 30 downregulated genes in 'Red' subpopulation. **(C)** Cell survival percentages and the IC50 values of GBC02, SCC-029 and SCC-032 cell lines 3D-sphere cultures with Cisplatin treatment. **(D)** Graph showing higher Caspase 3/7 activity in GBC02 cell line's $CD24^{Low}$ (CD24-Neg) cells compared to $CD24^{High}$ (CD24-Pos) cells upon Cisplatin treatment. **(E)** qRT-PCR result of ABCG2 expression in GBC02 2D sorted and 3D spheroid subpopulations. **(F)** (i) Graph showing significant increase in 'Green' subpopulation upon Cisplatin treatment for 48 hours in GBC02 cell line. Error bars represent Mean ± SEM from three biological repeats. Similar increase in 'Green' cells with 48 hours Cisplatin treatment in (ii) GBC02 3D spheroids (iii) SCC-029 parent cell line.

References:

1. Marine, J.-C., S.-J. Dawson, and M.A. Dawson, *Non-genetic mechanisms of therapeutic resistance in cancer.* Nature Reviews Cancer, 2020. **20**(12): p. 743-756.
2. Rambow, F., J.-C. Marine, and C.R. Goding, *Melanoma plasticity and phenotypic diversity: therapeutic barriers and opportunities.* Genes & development, 2019. **33**(19-20): p. 1295-1318.
3. Tirosh, I., et al., *Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma.* Nature, 2016. **539**(7628): p. 309-313.
4. Dalerba, P., et al., *Single-cell dissection of transcriptional heterogeneity in human colon tumors.* Nature Biotechnology, 2011. **29**(12): p. 1120-1127.
5. Puram, S.V., et al., *Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer.* Cell, 2017. **171**(7): p. 1611-1624. e24.
6. Inde, Z. and S.J. Dixon, *The impact of non-genetic heterogeneity on cancer cell death.* Critical reviews in biochemistry and molecular biology, 2018. **53**(1): p. 99-114.
7. Jolly, M.K., et al., *Phenotypic plasticity, bet-hedging, and androgen independence in prostate cancer: Role of non-genetic heterogeneity.* Frontiers in oncology, 2018. **8**: p. 50.
8. Shlyakhtina, Y., K.L. Moran, and M.M. Portal, *Genetic and Non-Genetic Mechanisms Underlying Cancer Evolution.* Cancers, 2021. **13**(6): p. 1380.
9. Huang, S., *Non-genetic heterogeneity of cells in development: more than just noise.* Development, 2009. **136**(23): p. 3853-3862.
10. Nguyen, L.V., et al., *Cancer stem cells: an evolving concept.* Nature Reviews Cancer, 2012. **12**(2): p. 133-143.
11. Beck, B. and C. Blanpain, *Unravelling cancer stem cell potential.* Nature Reviews Cancer, 2013. **13**(10): p. 727-738.
12. Marusyk, A., V. Almendro, and K. Polyak, *Intra-tumour heterogeneity: a looking glass for cancer?* Nature Reviews Cancer, 2012. **12**(5): p. 323-334.
13. Pisco, A.O., et al., *Non-Darwinian dynamics in therapy-induced cancer drug resistance.* Nature Communications, 2013. **4**(1): p. 2467.
14. Jia, D., et al., *Quantifying cancer epithelial-mesenchymal plasticity and its association with stemness and immune response.* Journal of clinical medicine, 2019. **8**(5): p. 725.
15. Prasetyanti, P.R. and J.P. Medema, *Intra-tumor heterogeneity from a cancer stem cell perspective.* Molecular cancer, 2017. **16**(1): p. 1-9.
16. Clarke, M.F., *Clinical and therapeutic implications of cancer stem cells.* New England Journal of Medicine, 2019. **380**(23): p. 2237-2245.
17. Prager, B.C., et al., *Cancer stem cells: the architects of the tumor ecosystem.* Cell Stem Cell, 2019. **24**(1): p. 41-53.
18. Tang, D.G., *Understanding cancer stem cell heterogeneity and plasticity.* Cell Research, 2012. **22**(3): p. 457-472.
19. Medema, J.P., *Cancer stem cells: the challenges ahead.* Nature cell biology, 2013. **15**(4): p. 338-344.
20. Gupta, P.B., et al., *Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells.* Cell, 2011. **146**(4): p. 633-644.
21. Dirkse, A., et al., *Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment.* Nature Communications, 2019. **10**(1): p. 1-16.
22. Meacham, C.E. and S.J. Morrison, *Tumour heterogeneity and cancer cell plasticity.* Nature, 2013. **501**(7467): p. 328-337.

23. Barkley, D. and I. Yanai, *Plasticity and clonality of cancer cell states.* Trends in cancer, 2019. **5**(11): p. 655-656.
24. Prince, M.E., et al., *Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma.* Proceedings of the National Academy of Sciences, 2007. **104**(3): p. 973-978.
25. Clay, M.R., et al., *Single marker identification of head and neck squamous cell carcinoma cancer stem cells with aldehyde dehydrogenase.* Head & neck, 2010. **32**(9): p. 1195-1201.
26. Ghuwalewala, S., et al., *CD44highCD24low molecular signature determines the cancer stem cell and EMT phenotype in oral squamous cell carcinoma.* Stem cell research, 2016. **16**(2): p. 405-417.
27. Biddle, A., et al., *Cancer Stem Cells in Squamous Cell Carcinoma Switch between Two Distinct Phenotypes That Are Preferentially Migratory or Proliferative.* Cancer Research, 2011. **71**(15): p. 5317-5326.
28. Buder, T., et al., *CellTrans: an R package to quantify stochastic cell state transitions.* Bioinformatics and biology insights, 2017. **11**: p. 1177932217712241.
29. Sharma, S.V., et al., *A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations.* Cell, 2010. **141**(1): p. 69-80.
30. Jones, K.B., et al., *Quantitative Clonal Analysis and Single-Cell Transcriptomics Reveal Division Kinetics, Hierarchy, and Fate of Oral Epithelial Progenitor Cells.* Cell Stem Cell, 2019. **24**(1): p. 183-192.e8.
31. Wong, D.J., et al., *Module Map of Stem Cell Genes Guides Creation of Epithelial Cancer Stem Cells.* Cell Stem Cell, 2008. **2**(4): p. 333-344.
32. Gao, J., et al., *Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal.* Science Signaling, 2013. **6**(269): p. pl1-pl1.
33. Goldman, A., et al., *Temporally sequenced anticancer drugs overcome adaptive resistance by targeting a vulnerable chemotherapy-induced phenotypic transition.* Nature Communications, 2015. **6**(1): p. 6139.
34. Chaffer, C.L., et al., *Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state.* Proceedings of the National Academy of Sciences, 2011. **108**(19): p. 7950-7955.
35. Sharma, A., et al., *Longitudinal single-cell RNA sequencing of patient-derivedprimary cells reveals drug-induced infidelity in stem cell hierarchy.* Nature Communications, 2018. **9**(1): p. 4931.
36. Zhang, S., X. Xiong, and Y. Sun, *Functional characterization of SOX2 as an anticancer target.* Signal Transduction and Targeted Therapy, 2020. **5**(1): p. 135.
37. Aguilar-Medina, M., et al., *SOX9 Stem-Cell Factor: Clinical and Functional Relevance in Cancer.* Journal of oncology, 2019. **2019**: p. 6754040-6754040.
38. Ge, Y., et al., *Stem Cell Lineage Infidelity Drives Wound Repair and Cancer.* Cell, 2017. **169**(4): p. 636-650.e14.
39. Morris, L.G.T., et al., *Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival.* Oncotarget, 2016. **7**(9): p. 10051-10063.
40. Stanta, G. and S. Bonin, *Overview on Clinical Relevance of Intra-Tumor Heterogeneity.* Frontiers in medicine, 2018. **5**: p. 85-85.
41. Sharma, A., et al., *Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung Tumors.* Cell reports, 2019. **29**(8): p. 2164-2174.e5.
42. Karaayvaz, M., et al., *Unravelling subclonal heterogeneity and aggressive disease states*

in TNBC through single-cell RNA-seq. Nature Communications, 2018. **9**(1): p. 3588.
43. Pisco, A.O. and S. Huang, *Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'.* British Journal of Cancer, 2015. **112**(11): p. 1725-1732.
44. Mishra, S., et al., *Increased Expression of Oct4, Nanog and CD24 Predicts Poor Response to Chemo-Radiotherapy and Unfavourable Prognosis in Locally Advanced Oral Squamous Cell Carcinoma.* Asian Pacific Journal of Cancer Prevention, 2020. **21**(9): p. 2539-2547.

Materials and Methods:

Cell culture:

Establishment of oral cancer cell lines GBC02 and GBC035 has been reported recently (KV et al. 2021). Other oral cancer cell lines SCC-029b, SCC-070, SCC-084 and SCC-032 were kindly provided by Dr. Susanne M. Gollin, University of Pittusburgh, USA (REF). These cell lines were maintained in MEM media with non-essential amino acids and 10% FBS.

Immunofluorescence staining:

Cells were seeded in 8-well chambered glass slides (Eppendorf) freshly coated with poly-D-Lysine. After 48 hours, cells were fixed with 10% buffered formalin (Sigma) for 30 minutes and washed with PBS contain glycine. They were further washed with ice-cold PBS containing glycine and permeabilized using 0.2% Tween-20 (Sigma) at room temperature. Blocking was performed with 10% normal goat serum (Cell Signalling). Primary antibody against Pancytokeratin (Cat No. ; Abcam), Cytokeratin-5 (Cat No. ; Abcam) and Cytokeratin-14 (Cat No. ; Abcam) were used. Antibody was diluted in 0.2% BSA in PBS and incubated at 4^0C, overnight. Next day, cells were washed and incubated with secondary antibody conjugated with Alexa Fluor 488 (Cat No. ; BD) or Alexa Fluor 568 (Cat No. ; Thermo), for 2 hrs at room temperature. Isotype-matched IgG was used to test antibody specificity. Cells were washed with PBS and mounted using DAPI containing mounting medium (ProLong® Diamond Antifade Mountant (Thermo). Stained

cells were examined and images were acquired using EVOS-FL (Thermo) microscope.

Cells were detached from the surface using Accutase cell dissociation reagent and re-suspended in HBSS buffer with 0.5 % FBS and 10mM HEPES. Cells were incubated with as per the experimental choice using CD44-FITC/BV786 conjugated (Cat No. ; BD), CD24-BV421/Alexafluor-647conjugated (BD), EpCAM FITC-conjugated (BD) for 30 minutes at 4^0C. After washing, Propidium Iodide (Sigma) was added to exclude dead cells. Isotype-matched IgG was tested for antibody specificity. ALDEFLUOR assay was performed (Cat. # 01700, Stem Cell Technologies) as per manufacturer's protocol. Briefly, $1x10^6$ cells/ml were resuspended in ALDEFLUOR buffer and incubated with 5ul/ml Aldefluor reagent BODIPY-aminoacetaldehyde (BAAA) for 60 minutes at 37^0C. Diethyl aminobenzaldehyde (DEAB), an inhibitor of Aldefluor reagent, was used as negative control. Propidium iodide was used to exclude dead cells. Flow cytometry data was acquired using BD FACS Aria Fusion cytometer and analysis was done with FCS Express 5 (DeNovo Software).

RNA preparation, cDNA preparation and real-time PCR:

RNeasy mini or micro kit (Qiagen) was used for RNA isolation as per manufacturer's instructions. Concentration of RNA was determined using Nanodrop (Thermo). For real-time qRT-PCR analysis, 450ng of RNA was transcribed into cDNA using Verso cDNA synthesis kit (Thermo) following manufacturer's instructions. Real-Time PCR was performed on ABI PRISM 7900HT Fast Real Time PCR System (Applied Biosystems) using KAPA SYBR FAST qPCR Kit (Kapa Biosystems). The expression levels of genes were

normalized with GAPDH and defined as $\Delta\Delta Ct = [\Delta Ct_{sample} - \Delta Ct_{GAPDH}]$. Fold change of the genes were calculated by $2^{-\Delta\Delta CT}$ method.

Sphere Formation Assay:

5-10 cells/ul were resupended in respective media used for cell lines growth without serum and with 20ng/ml EGF (Thermo), 20ng/ml bFGF (Thermo), 1x B-27 (Thermo) and 0.4μg/ml hydrocortisone (Merck, Sigma) along with 1.25% geltrex (Thermo) and plated in multiple wells of ultra-low attachment plate (Corning). Growth factors were supplemented every third day. Sphere formation was continued for 7-10 days. Spheres with size $\geq$60 μ were counted for calculating efficiency and size. Limiting dilution assays were conducted in the same manner with varying cell numbers ranging from 5 cells/100 μl to 500 cells/ 100 100 μl in each well of 96 well plate.

Colony formation:

1000 cells per 60mm dish were seeded and allowed to grow for 21 days. Media is changed every five days and after 21 days, crystal violet (Sigma) is used to stain the colonies. Upon washing the plates with water and drying, pictures were taken using chemiDoc imaging system (Biorad). Colonies were counted using ImageJ Fiji software.

Discrete time markov chain model construction:

A discrete time markov chain model was constructed to explain the repopulation patterns of the different phenotypes. The assumptions of the model are as follows:

1. Repopulation is due to transition of different phenotypes between each other. Other processes like cell division do not affect the phenotypic heterogeneity.
2. Transition rates among the phenotypes are independent of time and of each other.

To infer the transition rates, the package *CellTrans* [] (R 3.6.4) was applied on the repopulation data. Briefly, the package uses the repopulation data at time t_i obtained starting from a sorted population to construct a phenotypic proportion matrix (PPM_{t_i}). Using the theory of Markov chains, it calculates the transition matrix (TM) corresponding to the unit time by left-multiplying the proportion matrix with the identity matrix and taking a t_i^{th}root of the resultant matrix.

$$TM = PPM_{t_i}^{\frac{1}{t_i}}$$

The repopulation trajectories in Fig 4 were constructed using the transition matrix using the following formula:

$$p_t = p_0 * TM^t$$

Where p_0 is the population composition (a row vector) at time $t = 0$ and p_t is the population composition at time $t = t$.

For the sensitivity analysis of the estimated transition matrices, we generated random transition matrices by randomly shuffling the rows of the estimated TM and then randomly shuffling the columns.

Transition matrix randomization:

From the markovian model, we were able to obtain transition matrices and predicted possible trajectories leading to steady states. To understand our predictions further and to verify whether the predicted trajectories are unique to the transition matrix and hence to the experimental data, we randomized the transition matrices and compared the matrices and the steady states with those of "WT" matrix. To maintain the transition matrix property, i.e., all row-sums must be equal to one, the randomization was performed by shuffling the rates in a single row and shuffling entire rows (Supplementary Fig 7E)

We measured the distance between the steady states and the element-wise distance between the transition matrices using a standard distance metric: Euclidean distance. Based on these distances, we clustered the transition matrix-steady state pairs using the kmeans function of R, as demonstrated for the case of GBC02 in Supplementary Fig 7E. Further analysis of these clusters in terms of mean and standard deviation values of steady states (Supplementary Fig 7G-H) and transition rates (Supplementary Fig 7 I-J) suggested that in GBC02, the CellType $CD24^{High}/ALDH^{Low}$ remains the dominant fraction in the steady state population as long as the sustenance rate of that CellType (C4C4) is maintained high and transition from C4 to other cell types is maintained low. This is shown

by the fact that in cluster 1, the mean steady state is similar to that of WT (and SD is low), and only the variability of transition rates from C4 to other cell types and itself is low. We see a similar trend for SCC029B (Data not shown). For SCC070, C3 ($CD24^{High}/ALDH^{High}$) also has a significant fraction in the steady state population, and correspondingly the transition rates from C3 to C1 and C2 are maintained at low values with low variability (Supplementary Fig 7K-N).

RNA sequencing and differential gene expression analysis:

Quality of RNA was assessed on Agilent Bioanalyser 2100 (Agilent Technologies) by using Agilent RNA nano kit (Agilent Technologies). 400ng of RNA was used for library preparation. RNA sequencing was performed on illumina Hi-seq sequencer as per manufacturer's instructions. All the data analysis was performed using R 3.4.4 version and values were plotted using R package ggplot2. The raw read counts were normalized using the DESeq2Bioconductor package(Love et al., 2014). DEseq2 normalizes the counts by dividing sample-specific size factors determined by the median ratio of gene counts relative to geometric mean per gene. Post normalization genes with low expression values were excluded from the analysis. This was achieved by removing the genes that were not expressed in more than fifty percent of the total number of samples. After filtering the genes, a total of 22071 genes were used for differential expression analysis between different cell sub-populations (three biological replicates per condition). DESeq2 provided a statistical method to determine the differential expression using normalized gene expression data using a model based on the negative binomial distribution. The

resulting *P*-values were adjusted using the Benjamin and Hochberg's approach for controlling the false discovery rate. Significant DEGs were identified with *P*-values < 0.01 and fold-change ≥2.

Extraction of cell population-specific unique DEGs

Unique DEGs were obtained for each of the cell subpopulations based on all the pairwise comparisons. Genes were selected as unique DEGs for a cell population *i*as following - (1) by taking common set of n different set of DEGs (all pairwise comparison of cell population *i* with other cell population). (2) by excluding genes from common sets that are differentially expressed for any other pairwise comparison.

$A = ((i_1 \cap i_2) \cap \ldots i_n)$ … (1)

$B = ((j_1 \cup j_2) \cup \ldots . j_n)$ … (2)

Unique DEGs = A – B … (3)

Here, i_1 to i_n are set of DEGs of cell population i as compared to other n numbers of different cell populations and j_1 to j_n are set of DEGs for all other cell population pairwise comparisons.

Calculation of population-specific fold change

Cell population-specific fold change for each unique DEG for a given cell population was calculated as follows:

$$sFC_a = \frac{Xavg - Ymin}{Ymax - Ymin} \quad \ldots.(4)$$

sFCa is specific fold change of gene "a " in X cell population .Where,

X_{avg} – Mean expression of gene “a” in the X cell population

Y_{min} – Minimum of average expression of gene “a” in other cell populations

Y_{max} – Maximum of average expression of gene “a” in other cell populations

ssGSEA scores and survival analysis utilizing population-specific signature

The scores were calculated using GSVA R Bioconductor package and “ssGSEA” method was employed in the estimation of gene-set enrichment scores per sample.

To identify the role of different cell population gene signatures on patient survival, Head and Neck Cancer (HNSC) patients were classified into high and low enrichment groups, using the mean enrichment value as cut-off of single sample gene set enrichment analysis (ssGSEA) scores of a gene signature. To perform survival analysis R tool, *survival* was used. This analysis was performed as follows:

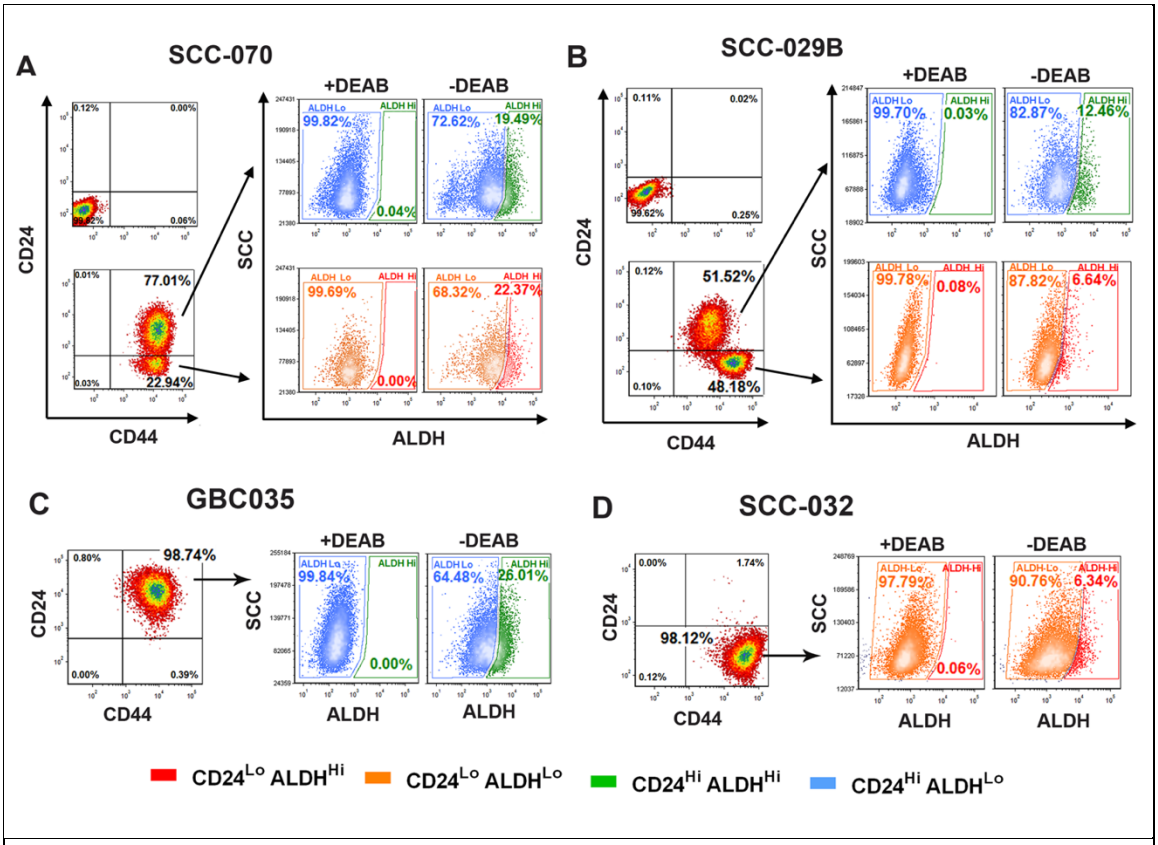
Select top n number of up regulated genes from unique DEG list sorted based on *sFC* value

(i) Calculate *ssGSEA* scores using the selected genes in the previous step

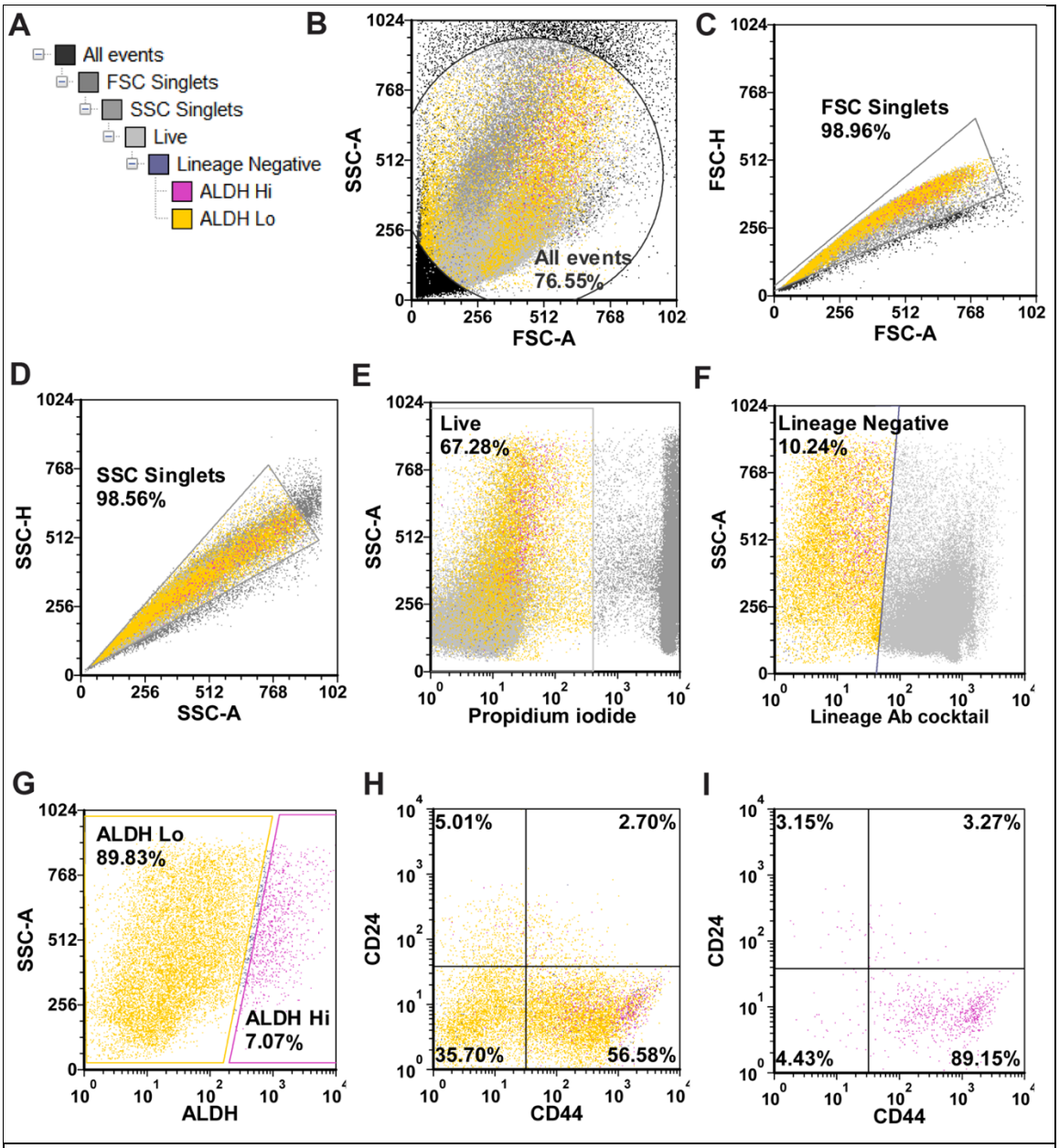
(ii) Segregate patients into “high” and “low” group. The HNSCC patients that showed *ssGSEA* scores above the mean *ssGSEA* value, were classed as the high expression group, and those below the mean value were classed as the low expression group

(iii) Perform survival analysis between the two patient groups

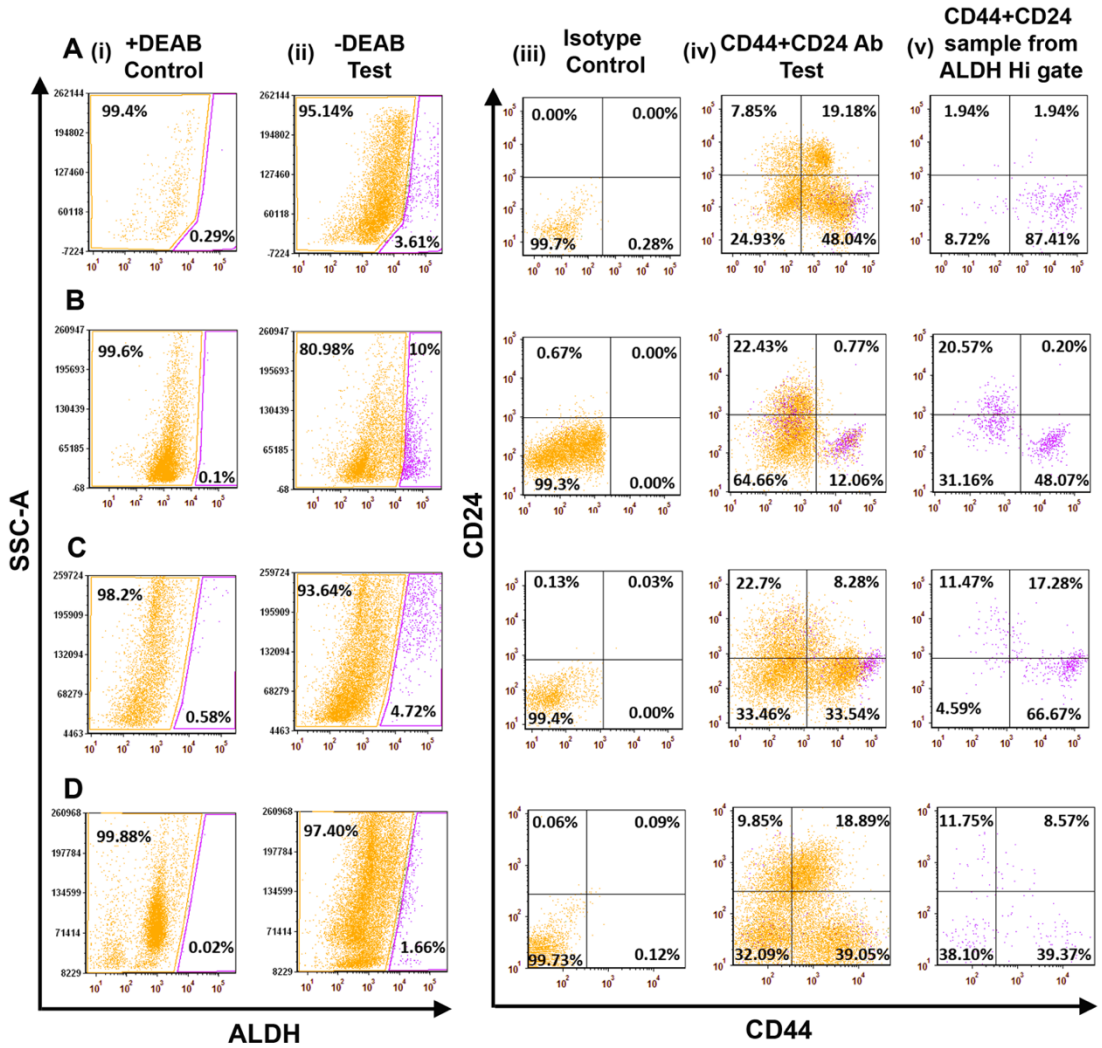
Supplementary Information



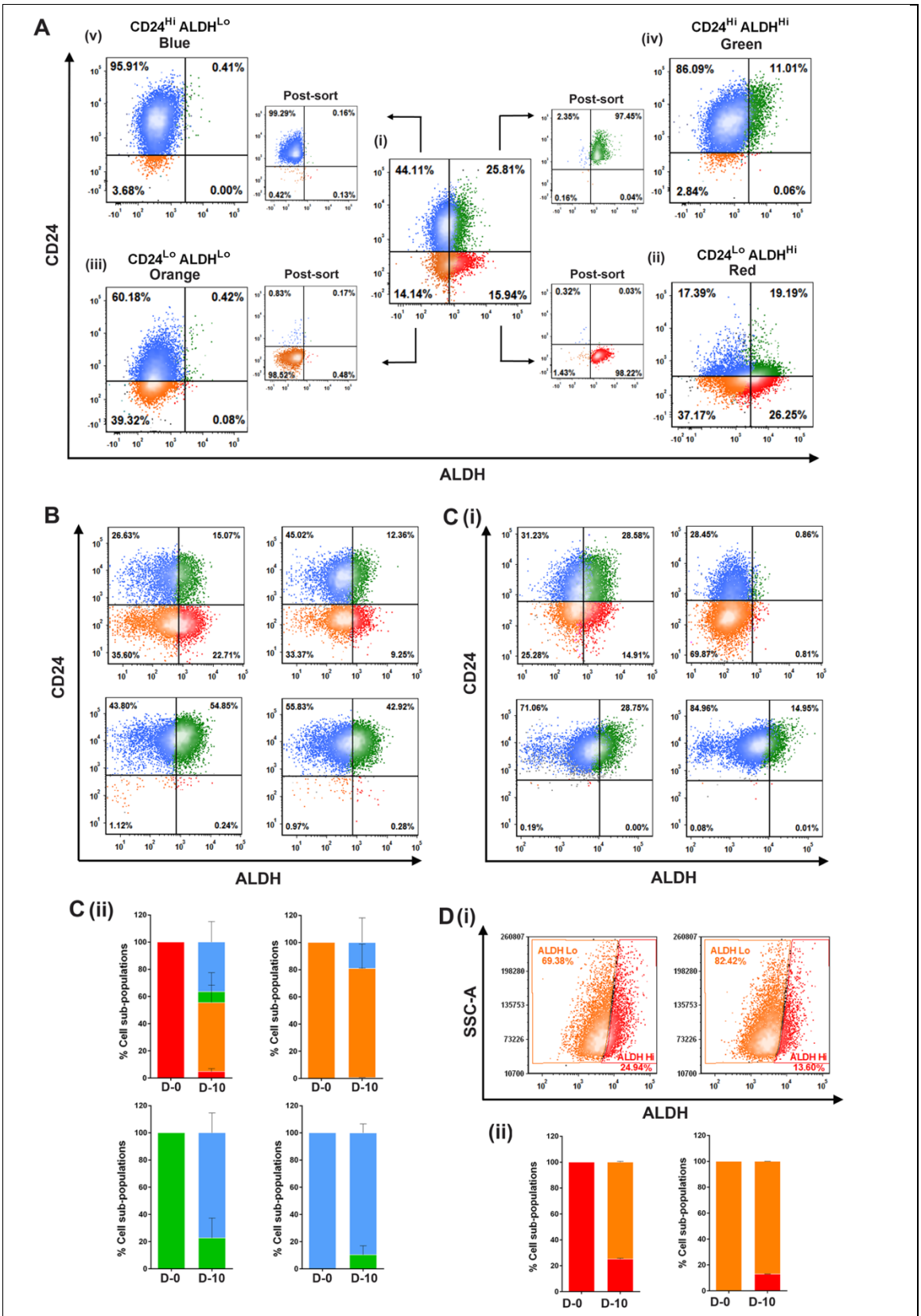
Supplementary Fig. 1: Representative FACS dot plots of **A)** SCC070 **B)** SCC029B **C)** GBC035 **D)** SCC032 cell lines CD24/CD44 staining (Left bottom) with respect to Isotype control (Left top). ALDEFLUOR (ALDH) phenotype of $CD24^{Hi}/CD44^{+}$ subpopulation (Right top) and $CD24^{Lo}/CD44^{+}$ subpopulation (Right bottom) in presence or absence of DEAB.



Supplementary Fig. 2: A) Gating strategy of ALDEFLUOR assay combined with CD24/CD44 staining in a freshly resected and dissociated patient tumor. **B)** Dot plot showing all events. **C, D)** Dot plots showing single cell selection upon doublets elimination from analysis. **E)** Selection of propidium iodide negative live cells. **F)** Selection of hematopoietic lineage negative cells using Lineage antibody cocktail. **G)** All the hematopoietic lineage negative cells segregated into $ALDH^{Hi}$ and $ALDH^{Lo}$ cells. **H)** CD24/CD44 phenotype of all the lineage negative cells. **I)** CD24/CD44 phenotype of $ALDH^{Hi}$ cells only.

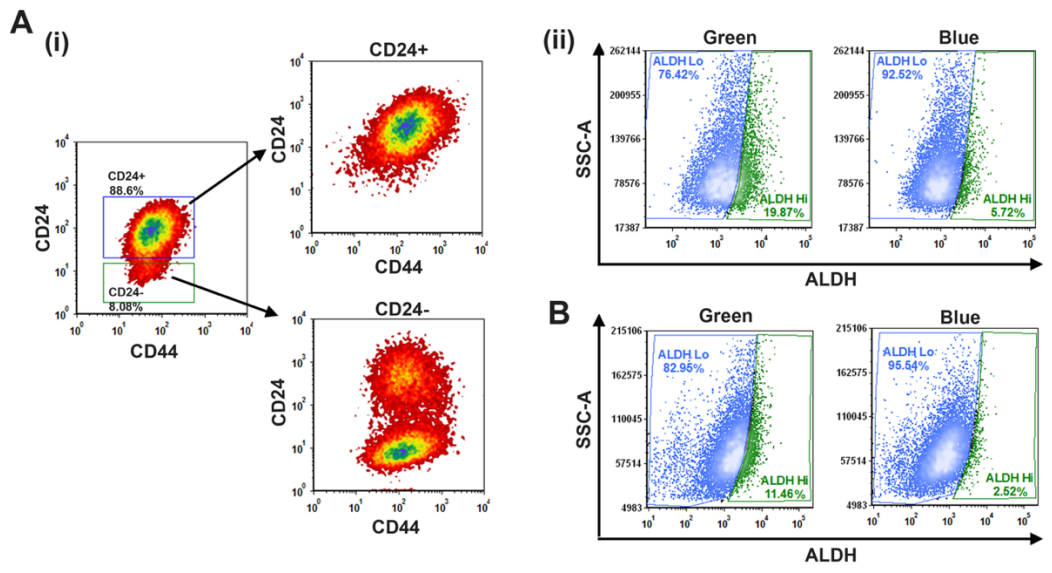


Supplementary Fig. 3: Representative dot plots of **A.** KV017, **B.** AP039, **C.** AP037, **D.** AP034 patient samples (row wise). **i)** With DEAB Control sample for setting background fluorescence for ALDH Hi cells. **ii)** Without DEAB Test sample showing $ALDH^{Hi}$ (yellow) and $ALDH^{Lo}$ (pink) cells. **iii)** Control sample with isotype antibodies. **iv)** Test sample with CD24, CD44 antibodies. **v)** CD24/CD44 phenotype OF $ALDH^{Hi}$ cells only.

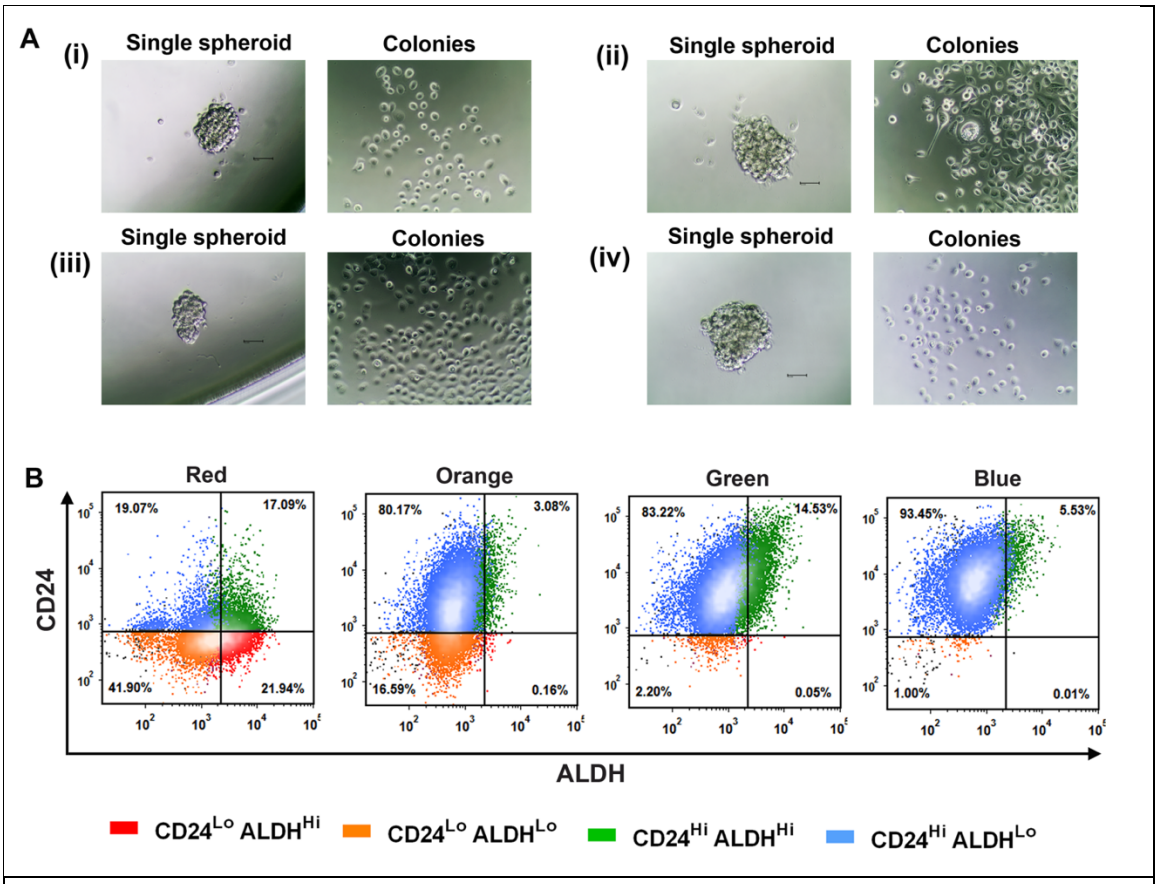


Supplementary Fig. 4: A) (i) Representative FACS dot plots of the GBC02 parent culture

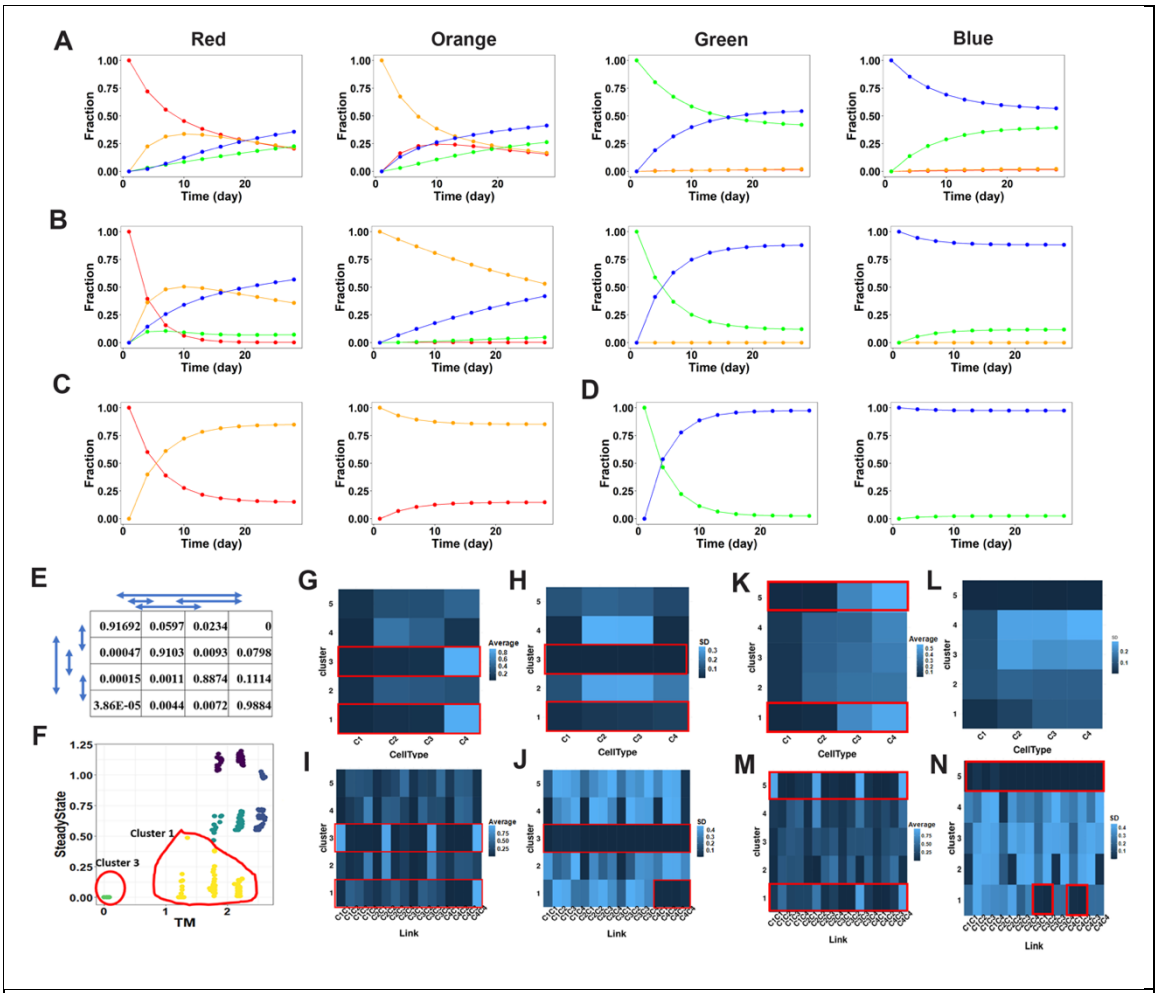
showing its CD24 and ALDH phenotype with existence of four populations and their post-sort data. Re-analysis for CD24 and ALDH showing repopulation of the **(ii)** Red **(iii)** Orange **(iv)** Green and **(v)** Blue phenotypes on Day -10 of sorting. **B)** Representative FACS dot plots of the SCC070 cell line's Red, Orange, Green and Blue subpopulations repopulation pattern on Day-10 of sorting. **C) (i)** Representative FACS dot plots of the SCC029b cell line's Red, Orange, Green and Blue subpopulations repopulation pattern on Day-10 of sorting. **(ii)** Repopulation frequencies of each subpopulation on Day-0 and Day-10 of sorting. Error bars represent mean ± SEM from four biological repeats. **D) (i)** Representative FACS dot plots of the SCC032 cell line's Red and Orange subpopulations repopulation pattern on Day-10 of sorting. **(ii)** Repopulation frequencies of each subpopulation on Day-0 and Day-10 of sorting. Error bars represent mean ± SEM from two biological repeats.



Supplementary Fig. 5: A) (i) Sorting of CD24+ and CD24- cells from SCC070 cell line. Established subline of CD24+ cells did not repopulate CD24- cells whereas CD24- cells repopulated both the subtypes. **(ii)** ALDH High and Low repopulation of SCC070 CD24+ subline's sorted cells.
B) ALDH High and Low repopulation of GBC035 sorted cells.

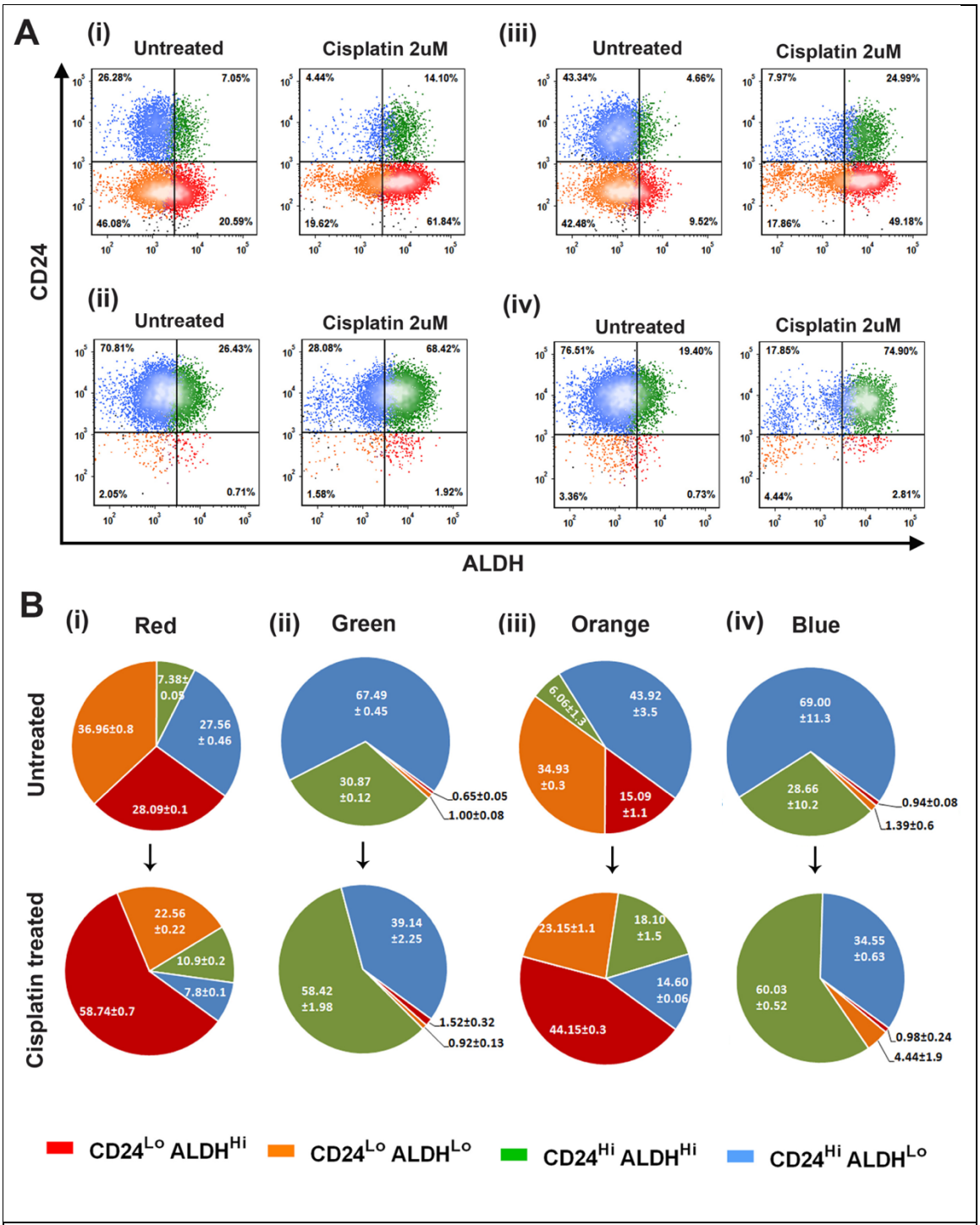


Supplementary Fig. 6: A) Representative pictures of single spheroids and colonies generated from them in (i) Red (ii) Orange (iii) Green and (iv) Blue subpopulations of GBC02 cell line. **B)** Representative FACS dot plots of repopulation of single spheroids of Red, Orange, Green and Blue subpopulations.

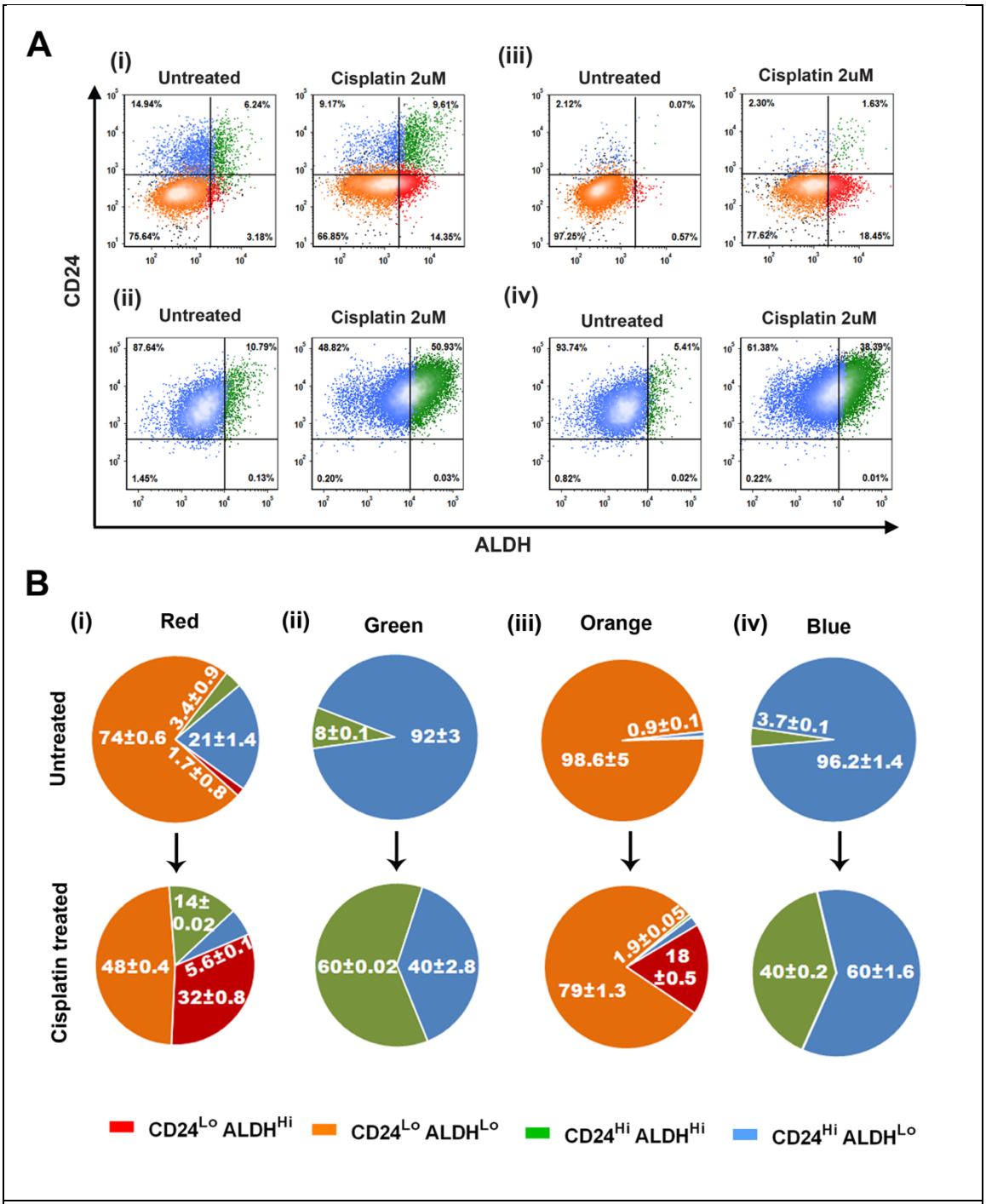


Supplementary Fig. 7: The phenotypic transition trajectories of Red, Orange, Green and Blue subpopulations of **A)** SCC070 **B)** SCC029B **C)** SCC032 and **D)** GBC035 cell lines. **E)** Demonstration of randomization of transition matrix for GBC02. **F)** Clustered scatterplot of distance of randomized transition matrices (x-axis) and the corresponding steady states (y-axis) from their wild-type (WT) counterparts, for GBC02. **G)** Cluster-wise heatmap for GBC02 of mean steady state values corresponding to the randomized transition matrices (TMs). The CellTypes are labelled as follows: C1 – $CD24^{Low}/ALDH^{High}$, C2 - $CD24^{Low}/ALDH^{Low}$, C3 – $CD24^{High}/ALDH^{High}$, C4 – $CD24^{Low}/ALDH^{Low}$. The highlighted rows indicate the cluster of TMs with similar steady state as WT. **H)** Same as G, but for Standard deviation instead of mean. **I)** Same as G, but for the elements of TM. The x-axis labels are to be read as follows: CiCj = Transition rate of CellType Ci to CellType Cj, i,j = {1,2,3,4}. **J)** Same as I, but for standard deviation instead of mean. Low SD regions are highlighted in red.

K) Cluster-wise heatmaps for SCC070 cell line showing the mean and **L)** standard deviation of the steady state fractions of the CellTypes (x-axis). Labelling of the cell types is the same as in GBC02 (G-H). **M, N)** Same as (K-L) but for transition matrices of SCC070.

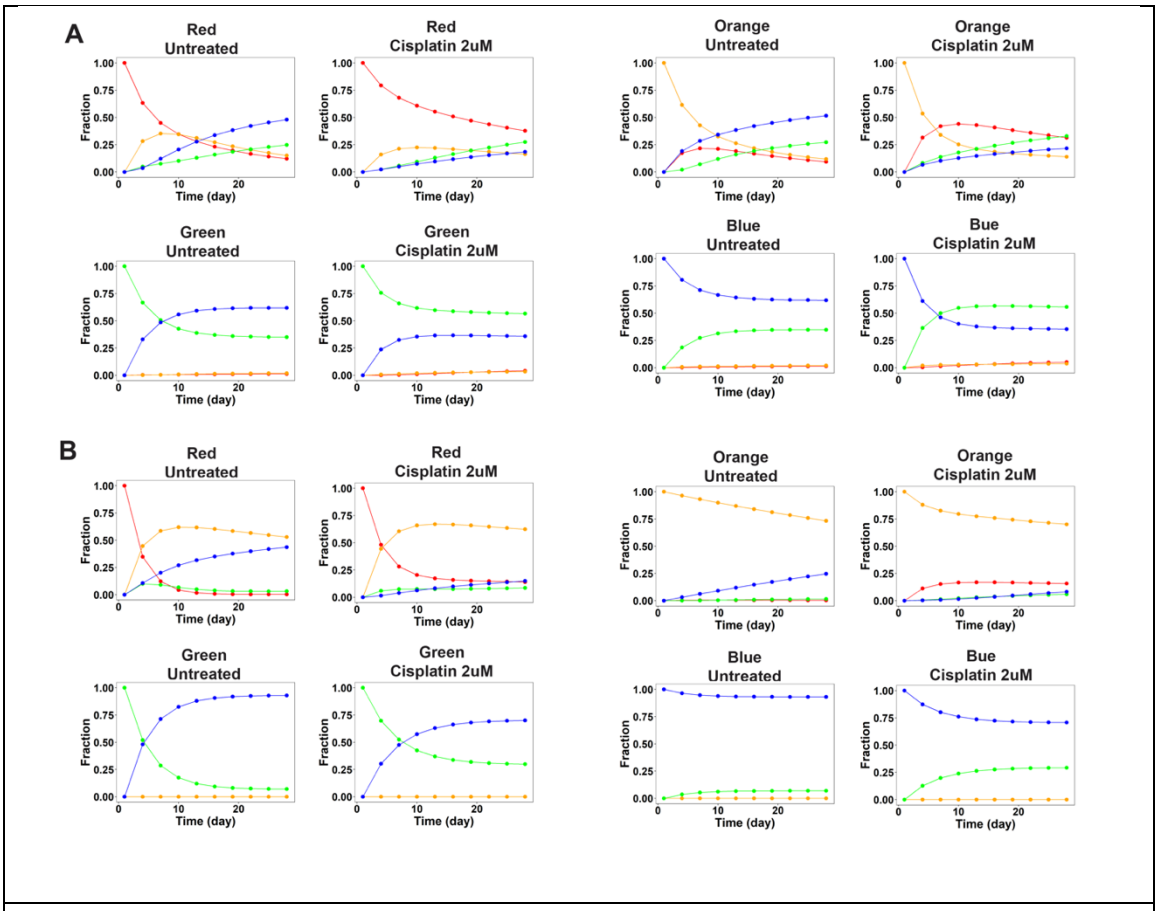


Supplementary Fig. 8: A) Representative dot plots of repopulation dynamics of SCC070 four subpopulations; (i) Red (ii) Green (iii) Orange and (iv) Blue in Untreated and Cisplatin (2µM) treated conditions. **B)** Repopulation frequencies of the four subpopulations (i) Red (ii) Green (iii) Orange and (iv) Blue in Untreated vs. Cisplatin (2µM) treated conditions from two biological repeats.

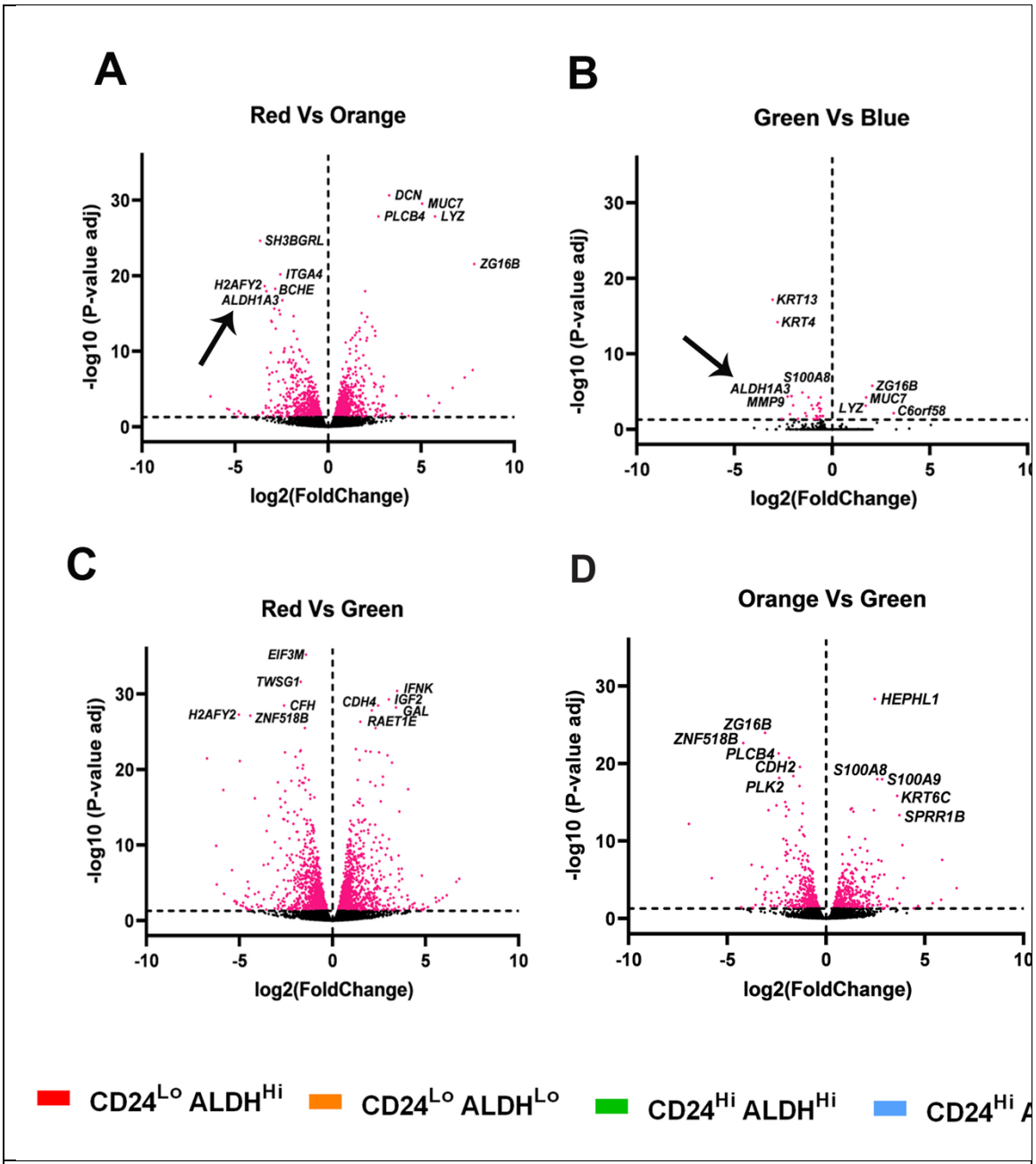


Supplementary Fig. 9: A) Representative dot plots of repopulation dynamics of SCC029 four subpopulations; (i) Red (ii) Green (iii) Orange and (iv) Blue in Untreated and Cisplatin (2µM) treated conditions.

B) Repopulation frequencies of the four subpopulations (i) Red (ii) Green (iii) Orange and (iv) Blue in Untreated vs. Cisplatin (2µM) treated conditions from two biological repeats.



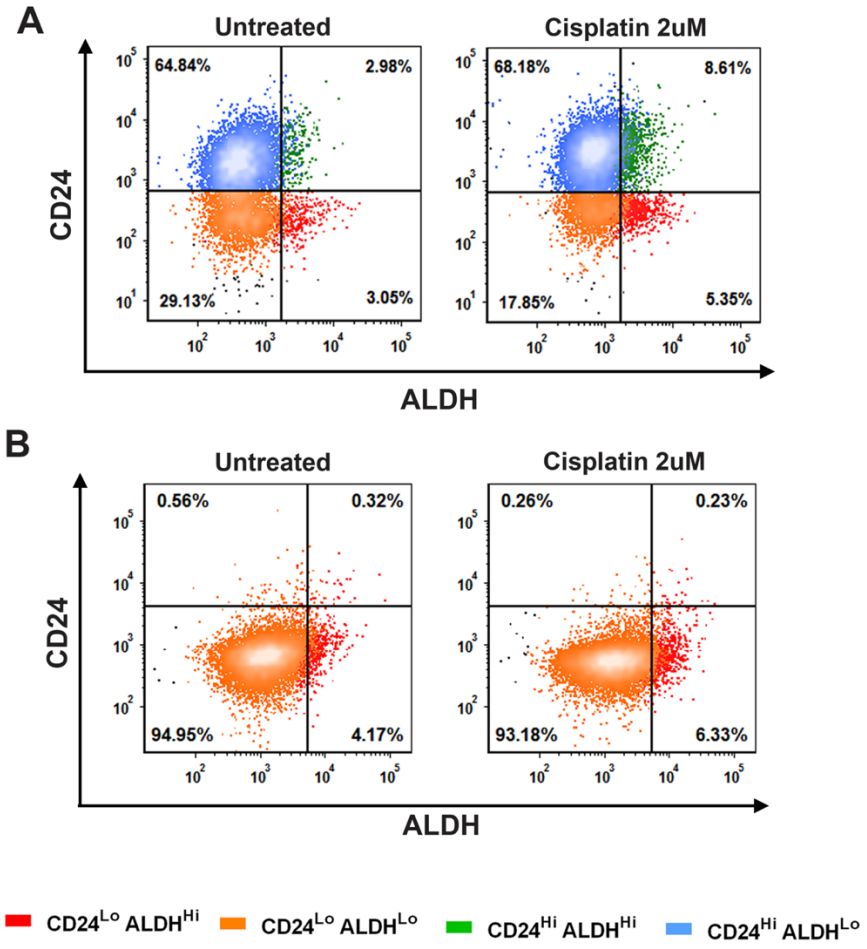
Supplementary Fig. 10: The phenotypic transition trajectories of the four subpopulations in Untreated vs. Cisplatin (2uM) treated conditions in **A)** SCC070 and **B)** SCC029B cell lines.



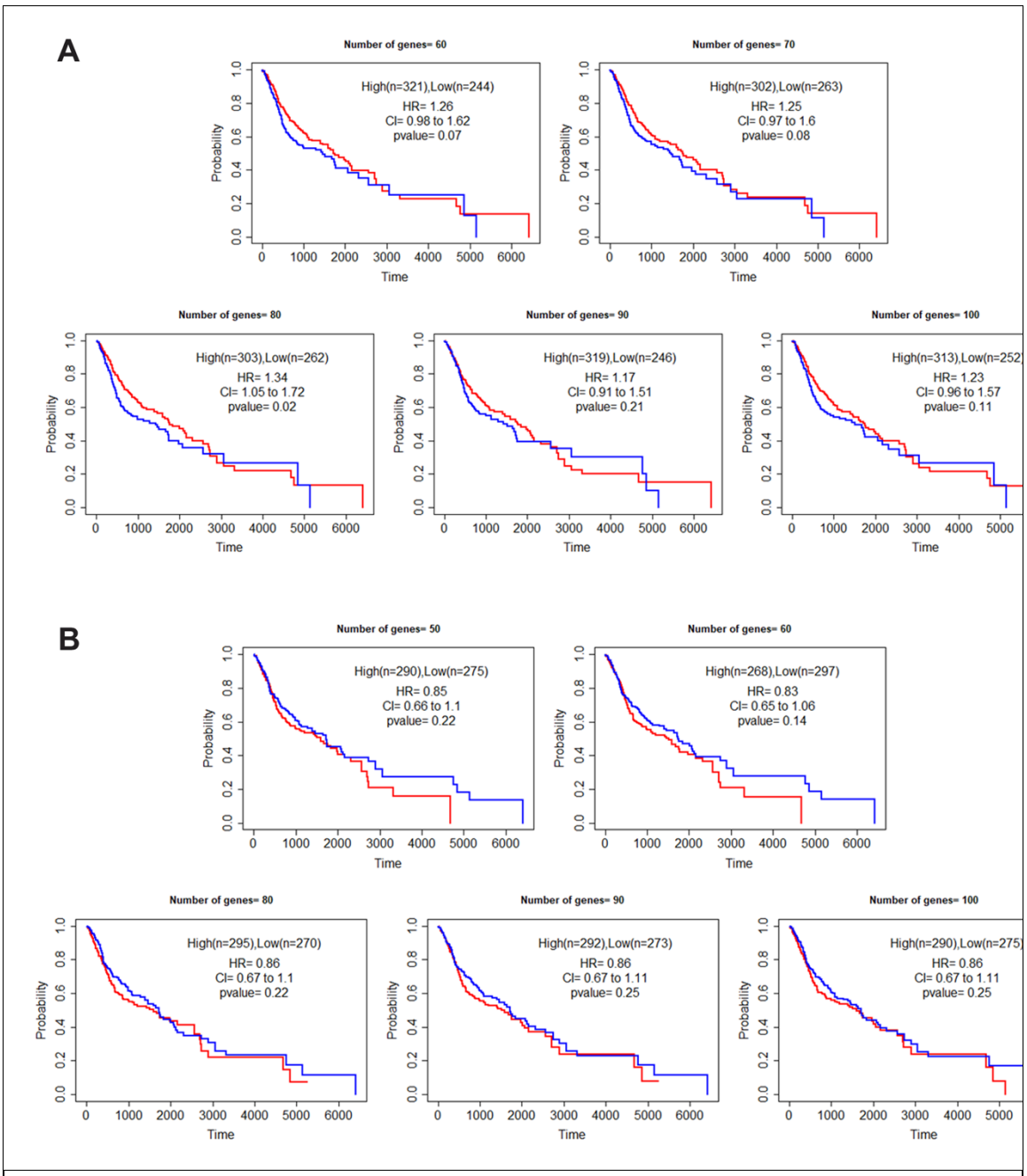
Supplementary Fig. 11: Volcano plots of **A)** Red vs. Orange, **B)** Green vs. Blue, **C)** Red vs. Green and **D)** Orange vs. Green subpopulations pair-wise comparisons. Significantly differentially expressed genes are represented in pink color. A horizontal line is used to mark the adjusted P-value threshold $\geq$ 0.05. The arrows point to *ALDH1A3* gene down regulated in $ALDH^{Lo}$ subpopulations compared to their respective $ALDH^{Hi}$ subpopulations in figures A and B. Five up regulated (right) and down regulated (left) genes are shown in each plot.

Downregulation of GPCR-signature in 'Orange' subpopulation compared to 'Red'
Enrichment plot: REACTOME_GPCR_LIGAND_BINDING
NES=-2.92
Up in Orange
Down in Orange
Enrichment plot: GO_G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY
NES=-2.30
Up in Orange
Down in Orange
DEGs (Red and Green)
Red Upregulated
Green Upregulated
OEPC Stem
OEPC Differentiating
396
23
0
5
283
0
337
372
20
0
16
286
0
326
Downregulation of GPCR-signature in 'Blue' subpopulation compared to 'Red'
Enrichment plot: REACTOME_GPCR_LIGAND_BINDING
NES=-2.53
Up in Blue
Down in Blue
Enrichment plot: GO_G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY
NES=-2.15
Up in Blue
Down in Blue
DEGs (Red and Green)
Red Upregulated
Orange Upregulated
261
19
0
9
287
0
333
218
9
0
6
297
0
336
Downregulation of GPCR-signature in 'Green' subpopulation compared to 'Red'
Enrichment plot: REACTOME_GPCR_LIGAND_BINDING
NES=-2.75
Up in Green
Down in Green
Enrichment plot: GO_G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY
NES=-2.18
Up in Green
Down in Green
DEGs (Red and Blue)
Red Upregulated
Blue Upregulated
499
22
0
14
284
0
328
369
13
0
12
293
0
330
Basal OEPC TF
id
red
orange
green
blue
SOX2
ASCL2
KLF15
TFCP2L1
NFATC1
POU3F1
GLI1
TSHZ2
ZFP64
CREB5
LEF1
CREM
MEIS2
BBX
PATZ1
ARID5B
ARID4A
Basal Differentiating TF
SP100
IRF7
STAT1
STAT2
SP110
HOPX
GRHL3
FOS
GRHL1
CERS3
OVOL1
FOXE1
KLF4
ELF5
MAF
HES1
BHLHE4
XBP1
KDM5B
MYCL
CEBPB
row max
row min
ATSC -ssGSEA
Blue
Green
Green_3DSph
Orange
RED
Red
Cell_population

Supplementary Fig. 12: A) Gene set enrichment analysis (GSEA) in (i) Orange (ii) Blue and (iii) Green subpopulations compared to Red subpopulation from GBC02 monolayer cultures showing depletion of GPCR ligand binding and receptor activity. **B)** Venn diagrams showing overlap of basal OEPC and basal differentiating gene sets (Jones *et al*, 2019) with upregulated genes from **(i)** Red vs. Orange comparisons in Red & Orange sorted subpopulations, and **(ii)** Red vs. Blue comparisons in Red & Blue sorted subpopulations, from GBC02 monolayer cultures. **C)** Heatmaps of mean normalized expression values from GBC02 monolayer cultures for (i) Basal OEPC TFs and (ii) Basal differentiating TFs adapted from Jones *et al*, 2019. **D)** ssGSEA scores of adult tissue stem cell (ATSC) gene signatures in the four subpopulations of GBC02 2D monolayers and 3D spheroids.



Supplementary Fig. 13: A) Representative dot plots of CD24 (y-axis) and ALDH (x-axis) phenotyping of GBC02 parent cell line in Untreated vs. Cisplatin (2uM) treated conditions. **B)** Representative dot plots of CD24 (y-axis) and ALDH (x-axis) phenotyping SCC032 parent cell line in Untreated vs. Cisplatin (2uM) treated conditions.



Supplementary Fig. 14: A) Kaplan Meier curves of HNSC patients segregated into High expression (Red) and Low expression (Blue) groups based on top 60-100 uniquely up-regulated genes of Red subpopulation. **B)** Kaplan Meier curves of HNSC patients segregated into High expression (Red) and Low expression (Blue) groups based on top 50-100 uniquely down-regulated genes of Red subpopulation.