Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2020 January 10.

Published in final edited form as:

Cell Rep. 2019 November 19; 29(8): 2164–2174.e5. doi:10.1016/j.celrep.2019.10.045.

Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung Tumors

Anchal Sharma[1], **Elise Merritt**[1], **Xiaoju Hu**[1], **Angelique Cruz**[2], **Chuan Jiang**[1], **Halle Sarkodie**[1], **Zhan Zhou**[3], **Jyoti Malhotra**[1], **Gregory M. Riedlinger**[1], **Subhajyoti De**[1,4,*]

[1]Rutgers Cancer Institute of New Jersey, Rutgers the State University of New Jersey, New Brunswick, NJ 08901, USA

[2]University of Miami, Coral Gables, FL 33124, USA

[3]Institute of Drug Metabolism and Pharmaceutical Analysis, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

[4]Lead Contact

SUMMARY

Impacts of genetic and non-genetic intra-tumor heterogeneity (ITH) on tumor phenotypes and evolvability remain debated. We analyze ITH in lung squamous cell carcinoma at the levels of genome, transcriptome, and tumor-immune interactions and histopathological characteristics by multi-region bulk and single-cell sequencing. Genomic heterogeneity alone is a weak indicator of intra-tumor non-genetic heterogeneity at immune and transcriptomic levels that impact multiple cancer-related pathways, including those related to proliferation and inflammation, which in turn contribute to intra-tumor regional differences in histopathology and subtype classification. Tumor subclones have substantial differences in proliferation score, suggestive of non-neutral clonal dynamics. Proliferation and other cancer-related pathways also show intra-tumor regional differences, sometimes even within the same subclones. Neo-epitope burden negatively correlates with immune infiltration, indicating immune-mediated purifying selection on somatic mutations. Taken together, our observations suggest that non-genetic heterogeneity is a major determinant of heterogeneity in histopathological characteristics and impacts evolutionary dynamics in lung cancer.

Graphical Abstract

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

*Correspondence: subhajyoti.de@rutgers.edu.

AUTHOR CONTRIBUTIONS

S.D. conceived the project. A.S. and S.D. designed the experiments. A.S. and A.C. performed the experiments. A.S., E.M., and X.H. analyzed the data. A.S., J.M., G.M.R., and S.D. interpreted the data. A.S. and S.D. wrote the manuscript with input from other authors.

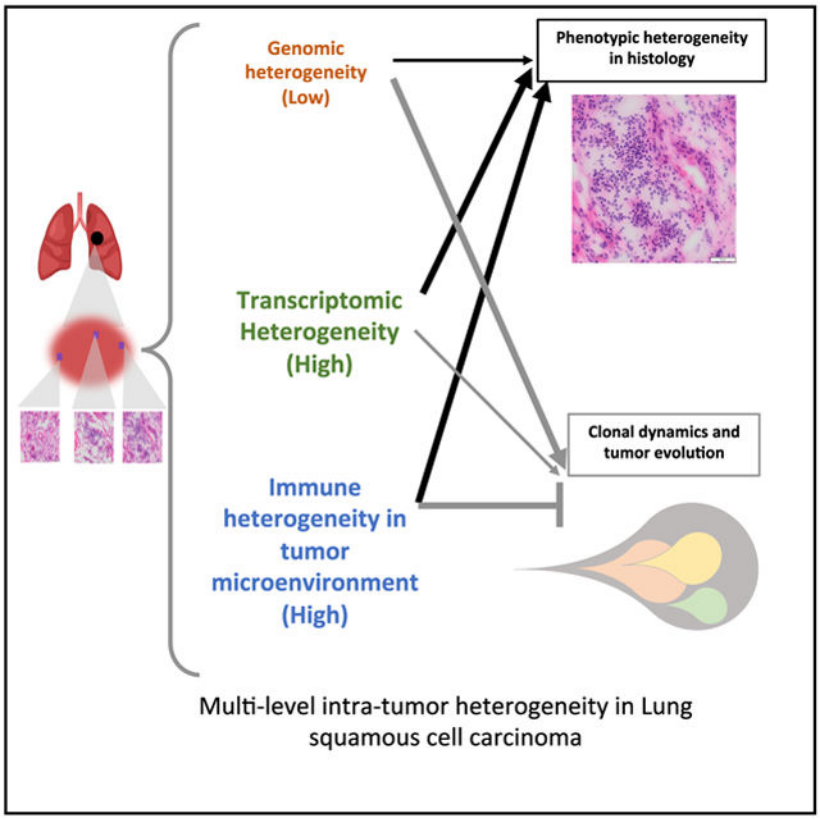
SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.10.045.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



In Brief

Sharma et al. show that in non-small-cell lung cancer, genetic heterogeneity is moderate and does not sufficiently reflect the extent of non-genetic heterogeneity at immune and transcriptomic levels, which contributes to regional differences in histopathological characteristics. Non-genetic heterogeneity also influences subclonal dynamics, shaping the trajectory of tumor evolution.

INTRODUCTION

Despite growing from a single, renegade somatic cell, by the time of detection, a tumor typically comprises of billions of cells that show considerable genetic and non-genetic differences among them, a phenomenon known as intra-tumor heterogeneity (ITH). Genetic and non-genetic ITH appear to be hallmarks of nearly all types of malignancies, providing substrates for evolvability and emergence of drug resistance and leading to unpredictable prognosis (Brock et al., 2009; Marusyk et al., 2012; McGranahan and Swanton, 2017). Emerging data show that certain patterns of genetic heterogeneity, as measured by the number of clones and presence of subclonal genetic alterations, predict poor survival in multiple cancer types (Andor et al., 2016; Jamal-Hanjani et al., 2017). Similarly, patterns of transcriptomic and immune heterogeneity appear to change with tumor subtyping, differentiation hierarchy, and response to treatment (Dalerba et al., 2011; Karaayvaz et al., 2018; Patel et al., 2014; Thorsson et al., 2018). Recent studies (Jia et al., 2018; Rosenthal et al., 2019) have suggested that immune editing influence both lung cancer evolution and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

survival, pointing toward complex interplay between the tumor and microenvironment. However, it is not well understood whether genetic and non-genetic ITH correlate and synergistically impact phenotypic variations and tumor evolution (De and Ganesan, 2017; Jia et al., 2018; Karaayvaz et al., 2018; Li et al., 2016; Patel et al., 2014; Suda et al., 2018; Zhang et al., 2018). In particular, there is a major gap in understanding the inter-relation between genetic, transcriptomic, and immunogenic heterogeneity and their contribution in tumor evolution and variability in clinically relevant phenotypic characteristics in lung squamous cell carcinomas.

RESULTS

We investigated patterns of intra-tumor spatial heterogeneity in 9 surgically resected stage I–IIIB lung squamous cell carcinoma specimens at genomic, transcriptomic, and tumor-immune cell interactions and histological characteristics by multi-region profiling (Figure 1A) and supplemented that with an analysis of single-cell sequencing data from an additional 5 samples, as described later. For each tumor, 3–6 geographically distant regions were profiled (pathological purity: 40%–80%). For each patient, tumor-proximal pathologically normal lung tissue was included as control. In total, 42 biopsies (10 normal and 32 tumor) were processed in this study. Clinical and pathological attributes of the samples are provided in STAR Methods section, and representative H&E staining images are shown in Figure S1.

Landscape of Genomic Changes in the Tumor Samples

We identified 120–1,606 somatic exonic single-nucleotide variatnts (SNVs) in the tumors (Figure 1B; Data S1). The frequency of detectable somatic mutation (2/Mb–26/Mb) was comparable to that reported elsewhere (Alexandrov et al., 2013) (Figure S1). A range of 34%–91% somatic SNVs were “ubiquitous” across all regions in a tumor, while 2%–55% SNVs were “shared” among multiple tumor regions, and a relatively small proportion of the somatic SNVs (0.8%–10%) was “unique” to any single region. The proportion of inferred genome-wide copy number alterations (CNAs; Data S1) that were ubiquitous was generally smaller than the corresponding proportion of SNVs. This is in line with observations that nearly half of the somatic mutations in tumor genomes likely arise in progenitor cells prior to tumor development (Tomasetti et al., 2013), while CNAs in pre-neoplastic tissues are rare (Aghili et al., 2014; De, 2011). Inferred telomere length varied across different regions within tumors, but overall, a majority of the tumor regions had shorter inferred telomere length than matched pathologically normal regions (Figure 1C).

Cancer genes *PTPRS, KMT2D, FT4, ALK, ASXL2, ZNF521*, and *ARID1A* had ubiquitous, potentially pathogenic somatic SNVs in P2, P3, P5, and P6 (Data S1). *SOX2, HOXC13, PAX3, TERT*, and *TP63* were amplified in multiple samples, while tumor suppressor genes, such as *POU5F1, NTRK3*, and *GRIN2A*, were deleted in either all or some regions in different tumor samples (Data S1). A majority of the oncogenic events were ubiquitous and, therefore, probably arose reasonably early during tumor development. However, there are exceptions, e.g., in P2 potentially pathogenic mutation in *KMT2D*, a histone methyl-transferase implicated in non-small-cell lung cancer (Ardeshir-Larijani et al., 2018) was only

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

found in regions R2 and R3. Inferred CNA estimates were consistent with gene expression changes. For instance, amplification of *SOX2* was accompanied by higher *SOX2* expression in tumors relative to matched non-malignant tissue (Figure S1; Data S2). Deep targeted sequencing covering 257 cancer-related genes from multiple regions from two samples (P8 and P9) did not change the assignment of the existing catalog of somatic mutations in P8, and no additional oncogenic drivers were identified.

Patterns of Genetic Heterogeneity

To estimate the extent of intra-tumor genetic heterogeneity, we constructed dendrograms (Figure 1B; also see STAR Methods) for every patient based on variant allele frequencies for somatic SNVs across different regions, such that branch lengths in the dendrograms indicate the extent of genetic divergence among the tumor and normal regions in a patient. Overall, intra-tumor regional genetic heterogeneity was small compared to tumor-paired normal tissue differences. Patient samples P1, P3, P4, and P6 showed conservative patterns of genetic ITH, whereas patient samples P2, P5, P7, and P8 showed relatively high levels of ITH, with a subset of tumor regions harboring either substantially less tumor mutation burden (TMB) (R1 of P2) or an excess of low-frequency variations compared to other regions (R4 of P5 and R2 of P7).

Typically, 2–4 clonal clusters were identified for each tumor using Pyclone (Roth et al., 2014), and a majority of the clusters were present in all or most regions in a tumor (Figure S2). For instance, in P6, we identified 3 prominent mutation clusters; all 3 were present in all regions in the tumor but at varying proportions (Figure 1D). We also applied the method proposed by Williams et al. (2018) to assess clonal architecture and compare that with our estimates. Overall, the estimated clonal makeup of the tumors is similar to that inferred by PyClone, with 1–2 subclones per sample (Figure S2). We further computed Shannon's and Simpson's indices for each region to quantify intra-region species richness and diversity in terms of abundance of different mutation clusters (Figure S2); Shannon's and Simpson's indices were comparable across regions within a tumor.

Somatic mutations carried mutation signatures for smoking, consistent with the etiology of lung squamous cell carcinoma (LUSC) (Jamal-Hanjani et al., 2017), as well as signatures of defective DNA mismatch repair, APOBEC, and 5-methylcytosine deamination (Figure 1E). APOBEC mutation signature was proportionally higher in non-ubiquitous mutations, suggesting that it probably arises late during tumor development (Jamal-Hanjani et al., 2017). Overall, genetic heterogeneity manifested by regional differences in oncogenic mutations, genomic alterations, telomere length, and mutation signatures and was moderate in LUSC, consistent with that reported elsewhere (McGranahan and Swanton, 2017).

Landscape of Transcriptomic Heterogeneity in the Tumor Samples

An unsupervised principal-component analysis of normalized gene expression data across all samples showed that (1) non-malignant tissues cluster together, indicating a low level of between-sample variation, as expected in non-diseased tissue specimens; and (2) different regions from the same tumors typically cluster together, suggesting that intra-tumor transcriptomic variations were typically smaller than inter-individual differences (Figure

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2A). In P5, region 4 (P5-R4) clustered separately from rest of the three regions, which is consistent with the extent of genetic divergence this region had relative to non-malignant tissue and other regions in P5. P1-R1, P1-R3, and P2-R1 were closer to the non-malignant samples, which might be due to low tumor purity in these regions (Figure S1).

Intra-tumor transcriptomic heterogeneity measured using dendrograms (see STAR Methods for details; Figure 2B) was generally lower than tumor-paired non-malignant tissue transcriptomic divergence. The extent of divergence of tumor regions from paired non-malignant regions both at the genomic and transcriptomic levels was largely similar (Figure 2C). The correlation between genomic and transcriptomic levels is not entirely due to regional differences in tumor purity because the correlation persisted even when adjusted for estimated tumor cell fraction in the cohort (Spearman partial correlation coefficient, 0.58; p = 4.36e-03). Overall, the topologies of the dendrograms at genomic and transcriptomic levels were qualitatively comparable, which is in line with convergent patterns of genetic and epigenetic evolution reported in other cancer types (Mazor et al., 2015).

Impact of Transcriptomic ITH on Cancer-Related Processes

Consistent with their copy number status, tumor suppressor genes like *ERBB4* and *HNF1A* had low expression in all regions in P4 and P6, whereas oncogenes such as *SOX2, ETV4, CCNE1*, and *TP63* had high expression in multiple tumors (Data S2). We measured dysregulation in key cancer-related pathways (Figure 2D; Figure S3) by using a network-based Jensen-Shannon Divergence (nJSD) score (Park et al., 2016). The nuclear factor κB (NFκB) activation pathway, which is associated with cell survival, was found to be highly deregulated in certain regions of tumors P3-R1 and P3-R3, P4-R1, and P5-R3. On the other hand, pathways like G2M DNA damage checkpoint, APC/CDC20-mediated degradation of cyclin B, and hedgehog pathways were moderately deregulated across all regions of all samples, except P2-R1 region.

Gene expression changes were consistent with their perceived impact on the tumor genomes. Apobec family genes (e.g., *APOBEC3B*) had high expression in P4 and P5, which had a high burden of APOBEC-related mutation signatures in somatic SNVs (Figure 1E). Likewise, in P5, altered mismatch repair activity was consistent with high mutation burden and mismatch repair mutation signatures. Telomere length had a negative correlation with *TERT* expression (Spearman correlation coefficient $r = -0.32$, $p = 3.9e-02$; Figure 2E) across the tumor regions, which is consistent with reports that *TERT* expression is negatively regulated by the telomere position effect over long distances (TPE-OLDs) mechanism (Kim et al., 2016).

Altered oncogenic pathway activities are expected to impact growth characteristics of tumors. We determined proliferation index (PI) and apoptotic index (AI) scores for all the samples (see STAR Methods for details). PI score was typically high in the tumor regions compared to adjacent normal tissues, but there were intra-tumor regional differences, e.g., P2-R2, P4-R3, P4-5, and P5-R1-3 showed a higher PI score than other regions in the same tumors (Figure 2D). These regions also clustered away from other regions in the same tumors in their overall transcriptomic profile (Figure 2A) and showed different mutational patterns at genomic levels (Figure 1B). In general, AI and PI correlated across tumor regions

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(Spearman correlation coefficient, 0.51), which is consistent with observations that *PCNA, CCNB1*, and *Caspase* 7 expression correlate significantly in the TCGA non-small-cell lung cancer cohorts (Figure S3). This is because replication stress is a major driver of apoptosis in rapidly proliferating tumors (Macheret and Halazonetis, 2015). P4 and P5 samples showed heterogeneity in epithelial-mesenchymal (EM) characteristic score across different regions (Figure 2D), with P4-R3 and P4-R4 depicting higher epithelial signature relative to other regions. Similarly P5-R1 and P5-R3 have a higher EM score than other regions. We also observed intra-tumor regional heterogeneity in gene expression signatures for multi-drug resistance (Wangari-Talbot and Hopper-Borge, 2013) (Figure 2D), including that for resistance to pemetrexed (Hou et al., 2012).

Lung tumors can be classified based on histological characteristics and associated expression signatures. The classical subtype is common in smokers and associated with xenobiotic metabolism, whereas the secretory subgroup has distinctive immune signatures, and the basal subgroup shows signatures of cell adhesion and epidermal development (Wilkerson et al., 2010). We found that two regions in P4 had basal-like characteristics, whereas the other three regions had more secretory type gene signatures (Figure 2D). Similarly, one region in P5 (current smoker) had signatures of secretory subgrouping, and others were classified as classical. P5 also showed amplification and high expression of the TP63 gene, which is another key characteristic of the classical subgroup. Samples such as P1 or P2, or regions therein, such as P4-R1, P4-R2, P4-R5, and P5-R4, which had enrichment for the secretory subgroup, also showed a high immune score. As we later show, these expression signatures correlated with their histopathological characteristics and phenotypic heterogeneity. Taken together, regional patterns of transcriptomic heterogeneity impacted molecular signatures of proliferation, EM characteristics, and ultimately clinically relevant subtype classification.

Comparative Assessment of Genomic and Transcriptomic ITH

Despite similarities between the topologies of dendrograms at genetic and transcriptomic levels (Figures 2B and 2C), there were important ITH differences at these two levels. We calculated Ω score, the ratio of ITH to tumor non-malignant tissue divergence for each tumor at genomic and transcriptomic levels (Ω_g and Ω_t, respectively), and compared them to study differences in ITH patterns between the two levels. Ω was generally higher at the transcriptomic level than that at the genetic level (Figure 2F). Interestingly, Ω_t/Ω_g was associated with histological subtyping and phenotypic characteristics; P1, P2, and P4 showed secretory characteristics, with low mean proliferative, apoptotic, and EM scores compared to P3, P5, and P6 (Figure 2G). Tumors with secretory subtype had modest proliferative characteristics and showed substantial transcriptional heterogeneity beyond that appreciated at the genomic level.

Heterogeneity in Immune Cell Infiltration and Associated Signatures

The extent of immune cell infiltration differed within and across tumors. P1, P2, and P4 samples had higher immune infiltration (Figure 3A), whereas P5 showed regional immune heterogeneity marked by higher immune infiltration in R4. CD4+ and M2 macrophages were present consistently across all tumor regions, whereas CD8+ T cells showed more

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

regional variations (Figure 3A). Dendrograms constructed using estimated proportions of different immune cell types indicated overall high intra-tumor immune heterogeneity (Figure 3B). Multiple potential T cell clonotypes were detected based on T cell receptor status (Shugay et al., 2015), but few of them were ubiquitous (Figure S4).

Next, we analyzed both histocompatibility leukocyte antigen (HLA) and somatic mutation status (see STAR Methods for details) to determine tumor neo-epitope burden (TNB) (Table S1) and found that TNB varied 8–341 per region (Figure 3C), of which ~10% were classified as high-affinity neo-epitopes (Figure S4). A total of 55%–100% of the neo-epitopes were present in only one tumor region and only up to 7% were ubiquitous (Figure 3D). Even though TMB and TNB correlate (Spearman correlation coefficient, 0.82; $p = 2.1e-06$; Figure S4), this observation indicates that TNB has a high level of intra-tumor regional heterogeneity, unlike that observed at the level of somatic mutations (Figures 1B and 3D). Reduction in the burden of ubiquitous neo-epitopes relative to that observed for somatic mutations is suggestive of immune-mediated negative selection purging tumor cells carrying mutations that could elicit immune response. In agreement with that reported earlier (Rosenthal et al., 2019), we observed that TNB negatively correlated with immune infiltration (Spearman correlation coefficient, −0.85; Figure S4) and was high in samples with low immune infiltration (P3, P5, and P6). This was not biased by mutation burden, and the negative correlation was significant even after adjusting for TMB (Spearman partial correlation coefficient, −0.57; $p = 5.1e-03$). The samples with low immune infiltration also showed low anti-PD1 favor score (Figure 3C; Figure S4). These observations together with the absence of ubiquitous, dominant T cell clonotypes suggest that these appear to be cold tumors in terms of their potential to elicit response to immunotherapy. Both TMB and T cell inflamed gene expression profile are needed to elicit a strong immune response (Cristescu et al., 2018), and LUSC is known to generally escape immune surveillance (Thorsson et al., 2018).

We computed Shannon's and Gini-Simpson's indices for each region to quantify intra-region species richness and diversity in terms of abundance of different immune cell populations (Figure S4). These intra-region diversity and richness measures at the level of immune features have higher values and showed greater regional variation than that observed at the genetic level (Figure S2), which is consistent with recent reports (Jia et al., 2018). In general, ITH patterns at the levels of immune cell abundance were high and less similar to the genomic and transcriptomic ITH patterns than the latter were to one another. We compared the extent of divergence of tumor regions from paired normal regions and found that immune divergence weakly correlates with genomic divergence (Spearman correlation coefficient, 0.5; $p = 0.01$; Figure 3E). In a similar note, by comparing Ω between genomic, transcriptomic, and immune levels, we found no strong correlation between heterogeneity patterns at immune and other levels (Figure S4).

In general, the abundance of immune cell types showed ITH higher than that at other levels, and joint patterns of regional variations in immune cell infiltration and neo-epitopes suggest that immunological pruning of tumor cell populations by neo-epitope depletion impose negative selection during tumor evolution, an emerging theme observed elsewhere as well (Jia et al., 2018; Zhang et al., 2018).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Meta-Heterogeneity and Regional Difference in Histopathological Characteristics

Overall, regional differences at genomic, transcriptomic, and immune levels appear to be broadly consistent and collectively impact clinically relevant histopathological characteristics. As an example, we showcase patient P4, a male, 56-year-old exsmoker with a stage IIIA malignancy (Figure 4A). The five geographically distant regions profiled from the patient had regional differences in histological characteristics and immune infiltration, which were reflected in gene-expression-based classification, with R1, R2, and R5 designated as secretory and R3 and R4 designated as basal subtype. R3 and R4 also had a higher proliferative potential than the other regions and more epithelial-like features, which corroborated with their basal subtype classification. This observation for P4 is in line with the comparative analyses of ITH at multiple levels in Figures 2 and 3, where we observe that transcriptomic and immune-level heterogeneity rather than genetic heterogeneity correlate with intra-tumor regional differences in histopathological characteristics. Furthermore, immune infiltration negatively correlates with mutation burden at the genetic level. These observations show that ITH at different levels are related and impact phenotypic characteristics (Figure 4B; Figure S1); yet, their complex inter-relation may not be sufficiently captured by assessment at any one level.

Heterogeneity Analysis at Single-Cell Resolution

We further analyzed single-cell RNA sequencing (RNA-seq) data from tumors of 5 non-small-cell lung cancer patients (Figure 5A). A total of 1,472–18,496 single cells were profiled from the core, middle, and edge regions per tumor by using the 10X Genomics platform (Lambrechts et al., 2018) (Figure 5B). Somatic CNA status and tumor clonal architecture inferred from RNA-seq data (see STAR Methods for details) indicated that the tumors had a small number of distinct subclonal clusters, with minor differences among the cells within those subclones, as evident from the branch lengths of the inferred tumor phylogenetic tree (Figure 5C). Significant differences in proliferation rate among the subclones in a tumor would be suggestive of non-neutral evolution. We calculated the pathway-level score for proliferation-related genes, as before, and then compared the distributions of the score among sets of tumor cells grouped by their subclonal membership and/or tumor location. Most subclones were ubiquitous, i.e., had presence in tumor core, middle, and margin regions, but in a majority of the tumors there were significant differences between the subclones in their proliferation scores, and these differences were generally consistent across the tumor regions (Figures 5D and 5E; nested ANOVA p values are listed in Table S2). This observation provides support for the difference in growth rates between subclones in these tumors. Proliferation scores for the subclones were typically lower in the core relative to that in the tumor periphery (Figure 5F), which is consistent with the fact that much of the tumor growth occurs near the tumor margin. Similar regional differences were observed for apoptosis (Figure S5) and other pathways as well. This suggests that regional contexts can affect proliferative characteristics and growth dynamics within a tumor, beyond that attributed to subclonal differences. Although the tumors were genetically well-mixed, i.e., most major subclonal clusters had sufficient presence in all regions in a tumor (Figure 5F), we found large intra-tumor regional differences in the overall burden of infiltrating immune cells (Figure 5G), and the abundance of different immune cell types therein (Figure 5H). This corroborates the observation made using bulk sequencing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

data that genetic heterogeneity does not sufficiently capture the extent of intra-region non-genetic heterogeneity.

DISCUSSION

In this study, we investigated ITH patterns in lung squamous cell carcinoma at levels of genomic, transcriptomic, and tumor-immune cell interactions and histological characteristics by multiregion profiling and single-cell data analysis to compare ITH at different levels. Despite the high somatic mutation burden, data from others and us showed that the spatial pattern of genetic ITH in lung squamous cell carcinoma is moderate, which might be due to spatial mixing of tumor subclones. In contrast, trancriptomic and immune ITH were higher, indicating non-genetic sources of variation (Sharma et al., 2018), and they impacted key cancer-related pathways, subtype characteristics, and proliferative potential. Even though the extent of overall ITH differed between genetic and non-genetic levels, regional differences were biologically consistent at the meta-heterogeneity level, as evident from copy-number-mediated expression changes, association between TERT expression and telomere length, and immune-mediated selection on neo-epitope burden. Notably, phenotypic heterogeneity in terms of histopathological characteristics was not effectively captured at any one level; rather, ITH at different levels synergistically impacted tumor histological features.

Multi-level ITH assessment found multiple lines of evidence for non-neutral tumor evolution. For instance, proliferation rate varied between subclones and also within the same subclone depending on the tumor microenvironment in different tumors. Intra-tumor immune heterogeneity patterns suggested that tumor-immune cell interactions impose negative selection by pruning the tumor cells carrying neo-epitopes that elicit a strong immune response. Immunological pruning of tumor cell populations by neo-epitope depletion likely imposes negative selection on genetic variations during tumor evolution (Jia et al., 2018; Zhang et al., 2018). Regional differences in subclonal growth characteristics due to non-genetic heterogeneity can influence the overall clonal makeup of a tumor with time. Overall, our study shows that despite having coherent patterns of ITH at different genetic and non-genetic levels, a multi-level assessment of heterogeneity is necessary to identify the determinants of phenotypic heterogeneity in clinically relevant characteristics and to appreciate the roles the microenvironment plays in influencing the mode of tumor evolution.

STAR★METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new reagents. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Subhajyoti De (sd948@cinj.rutgers.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human samples—We obtained surgically resected, fresh frozen tumor and matched normal tissue specimens from 9 de-identified, stage I-III lung squamous cell carcinoma patients under an IRB approved protocol (CINJ # 001709; approval date: 4/2017; PI: Subhajyoti De; Institution: Rutgers Cancer Institute of New Jersey). Details of the samples

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

are provided as in the table below. For each tumor 3-6 different regions that were geographically distant were identified, and biopsied to obtain tissue sections. When possible, H&E staining of tumors was done to identify tumor rich regions in order to take biopsies. In total, 42 biopsies (10 normal and 32 tumor) were processed in this study. Pathological tumor purity was 40%-80%. DNA and RNA were co-extracted from 29 biopsies (6 normal and 23 tumors) using TRizol method and were further processed for exome and RNA sequencing. For 8 other biopsies (2 normal and 6 tumors – from 2 patients) only DNA was isolated using QIAGEN DNeasy kit and exome sequencing was performed. For 8 other samples (1 normal and 3 tumor regions each from 2 patients) only DNA was isolated and targeted panel sequencing (Agilent SureSelect run on a HiSeq) was performed for 257 genes. Details of all samples are provided in the table below. Exome and RNA sequencing (rRNA depletion) were performed on Illumina HiSeq 2000 using 150bp paired-end protocol. Single cell RNaseq data for 5 non-small cell lung tumors were obtained from the study published by Lambrechts et al. (2018) (E-MTAB-6149 and E-MTAB-6653). Tumor and non-tumor cells from the core, middle, and edge regions from each tumor were profiled using 10X Genomics single cell RNaseq platform.

Summary of the human samples

Sample	Gender	Age	Smoker	Grade	Stage	Metastasis	Survival (days)	No. of tumor biopsies	Data	Tumor
P1	-	-	Non-smoker	-	-	-	NA	3	Exome + RNA	LUSC
P2	-	-	-	-	-	-	NA	3	Exome + RNA	LUSC
P3	Male	69	Ex-smoker	2	IIIB	-	NA	4	Exome + RNA	LUSC
P4	Male	56	Ex-smoker	2	IIIA	Yes	274	5	Exome + RNA	LUSC
P5	Female	69	Current Smoker	2	IIIA	-	129	4	Exome + RNA	LUSC
P6	Male	61	Ex-smoker	2	IIIA	-	161	4	Exome + RNA	LUSC
P7	Female	85	-	Grade2-moderately differentiated	T2a	No	NA	3	Exome	LUSC
P8	Male	61	-	Grade3-poorly differentiated	T2a	No	NA	3	Exome + targeted panel sequencing	LUSC
P9	Female	85	-	Moderately differentiated	T2a	No	NA	3	Targeted panel sequencing	LUSC
scP1	Female	70	Current Smoker	pT2bN0M0	IlA	-	-	3	scRNaseq	LUSC
scP2	Male	80	Ex-smoker	pT2bN0M0	lB	-	-	3	scRNaseq	LUSC
scP3	Male	68	Ex-smoker	pT4N2M0	lllB	-	-	3	scRNaseq	LUAD
scP4	Female	64	Ex-smoker	pT2aN1M0	llB	-	-	3	scRNaseq	LUAD

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Sample	Gender	Age	Smoker	Grade	Stage	Metastasis	Survival (days)	No. of tumor biopsies	Data	Tumor
scP5	Male	60	Ex-smoker	pT1cN0M0	IA3	-	-	3	scRNasea	Large cell

METHOD DETAILS

Genomic data analysis

FastQC (v0.11.7) was used for initial quality checks, and low-quality reads and PCR duplicates were removed. Next, we used BWA-mem (Li and Durbin, 2009) (v0.7.17-r1188) to map the reads onto human genome (GRCh38), and call variants using var-Scan2 (Koboldt et al., 2012) (mapping quality > 40, base quality > 20). The average sequencing depth ranged from ~50-191x for exome-seq. Variants present in dbSNP or those with strand bias were excluded, and only 'high confidence' somatic variants with tumor allele frequency > 5% at least in one tumor region and normal allele frequency < 1% were selected. A vast majority of the reported somatic variants had no read support in matched normal tissues. For each somatic variant deemed as high confidence variant in at least one tumor region, we queried the corresponding base-position in other tumor regions in that tumor specimen, and if the variant allele was supported by reads with mapping quality > 20 and base quality > 25 and variant allele frequency > 2%, it was included. All identified somatic mutations were annotated with SnpEff (Cingolani et al., 2012) (v4.3t). Missense, nonsense, frameshift, or splicing mutations in known COSMIC cancer genes with high-predicted impact were marked.

We used desonstructSigs (Rosenthal et al., 2016) to identify patterns of mutational signatures on somatic variants. Mutational signatures were inferred for ubiquitous and non-ubiquitous (shared + unique) variations separately, also all variations together (ubiquitous + shared + unique). Also, analysis was done both using all 30 signatures as well as selected signatures relevant to lung cancer biology (Alexandrov et al., 2016).

Copy number variation analysis was done using FACETS (Shen and Seshan, 2016) based on exome-sequencing data. Reads with mapping quality > 15 and base quality > 20 were considered for CNV analysis, and genomic regions with inferred duplication, deletion, or LOH were identified. Whole genome duplication and ploidy-level changes were inferred based on chromosome-wide, haplotype-aware copy number inferences.

To infer telomere length in the tumor regions, we used TelSeq (Ding et al., 2014), which considers the reads with > = 7 TTAGGG/CCCTAA repeats to provide an estimate of telomere length. Telomere length was estimated for both tumor regions (TTL: Tumor Telomere Length) as well as normal (NTL: Normal Telomere Length). We reported telomere length for different tumor regions relative to their matched normal region from the same donor as $\log_2$ of fold change between tumor and normal [$\log_2$ (TTL/NTL)].

Targeted sequencing was done for two samples (P8 and P9) for 257 genes (ABL1, AKT1, AKT2, AKT3, ALK, APC, AR, ARAF, ARID1A, ARID2, ATM, ATR, ATRX, AURKA,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

AURKB, AXL, BAP1, BARD1, BCCIP, BCL2, BCL2L2, BCL6, BCOR, BCR, BLM, BRAF, BRCA1, BRCA2, BRIP1, BTK, CBFB, CCND1, CCND2, CCND3, CCNE1, CDC73, CDH1, CDK12, CDKN1B, CDKN2a, CDKN2B, CDKN2C, CHEK1, CHEK2, CHD1L, CHD4, CIC, CREBBP, CRKL, CRLF2, CSF1R, CTCF, CTNNA1, CTNNB1, DAXX, DDR2, DNMT3A, DNMT3B, DNMT1, DOT1L, EGFR, EMSY- c11orf30, EP300, EPHA3, EPHA5, EPHB1, ERBB2, ERBB3, ERBB4, ERG, ESR1, ETV1, ETV4, ETV5, ETV6, EWSR1, EZH2, AMER1, FAM46C, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCL, FBXW7, FGF10, FGF14, FGF19, FGF23, FGF3, FGF4, FGF6, FGFR1, FGFR2, FGFR3, FGFR4, FLT1, FLT3, FLT4, FOXL2, GATA1, GATA2, GATA3, GID4, GNA11, GNA12, GNAQ, GNAS, GPR124, GRIA1, GRIN2A, GRM1, GRM2, GRM3, GSK3B, HGF, HRAS, HSP90AA1, IDH1, IDH2, IGF1R, IKBKE, IKZF1, IL7R, INHBA, IRF4, IRS2, JAK1, JAK3, JAK3, JUN, KAT6A, KAT5, KDM5A, KDM5C, KDM6A, KDR, KEAP1, KIT, KLHL6, KRAS, LRP1B, LYN, MAP2K1, MAP2K2, MAP2K4, MAP3K1, MCL1, MDM2, MDM4, MED12, MEF2BNB-MEF2B, MEN1, MET, MITF, MLH1, KMT2A, KMT2D, MPL, MRE11A, MSH2, MSh6, MTAP, MTOR, MUTYH, MYC, MYCL1, MYCN, MYD88, NF1, NF2, NFE2L2, NFKBIA, NKX2-1, NOTCH1, NOTCH2, NPM1, NRAS, NTRK1, NTRK2, NTRK3, NUP93, PAK3, PALB2, PAX5, PBRM1, PDGFRA, PDGFRB, PDK1, PIK3CA, PIK3CG, PIK3R1, PIK3R2, PPM1D, PPP2R1A, PRDM1, PRKAR1A, PRKDC, PTCH1, PTEN, PAXIP1, RAD51C, PTPN11, RAF1, RARA, RB1, RET, RICTOR, RNF43, ROS1, RPTOR, RUNX1, SETD2, SETD3, SF3B1, SMAD2, SMAD4, SMARCA4, SMARCA5, SMARCB1, SMO, SOCS1, SOX10, SOX2, SOX9, SPEN, SPOP, SRC, STAG2, STAT3, STAT4, STK11, SUFU, TERT, TET2, TGFBR2, TMPRSS2, TNFAIP3, TNFRSF14, TOP1, TP53, TP53BP1, TRIM24, TRIM28, TRIM33, TSC1, TSC2, TSHR, VHL, WISP3, XPO1, YES1, ZNF217, ZNF703). Somatic mutations were called in different regions as described above. Heatmaps and dendrograms were made as described below.

Heatmaps, genetic dendrograms, and heterogeneity estimates—Heatmaps depicting regional abundance of somatic SNVs were drawn using variant allele frequency (vaf) data for somatic SNVs across all regions of a tumor. Pairwise distance between the regions d_{ij} in a patient was computed as distance in terms of difference in tumor-purity-adjusted variant allele frequency (vaf) for all detected SNVs, as $d_{ij} = S_{kçv}|S_{ik}\text{-}S_{jk}|/V$ where s_{ik} is the vaf of k^{th} variant in i-th region in a tumor. Then multiregional tumor trees (unrooted dendrograms) were drawn using this distance metric using neighbor joining. The dendrograms were bootstrapped using *boot.phylo* function in the *ape* R package. Distances between pairs of regions in the trees represent the extent similarity between different regions of the same tumor at the genomic SNV level. We computed $= S_i \neq {}_B\, l_i/l_i = {}_B$ where $l_i = {}_B$ is the branch length of the phylogram from the healthy tissue to its nearest node, providing an estimate of the ratio of intra-patient regional diversity to tumor-non-malignant tissue divergence.

For each tumor specimen, based on the catalog of somatic mutations and copy number data across multiple regions variant clusters were identified using PyClone (v0.13.1) (Roth et al., 2014). Default parameter setting was used to identify somatic mutation clusters. For downstream analyses we only analyzed mutation clusters of size ≥ 3 variants. Key

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

conclusions were unaffected by the choice of threshold. Intra-region diversity and richness at genetic levels were computed using Shannon's and Gini-Simpson's indices on the relative abundances of the mutation clusters in each tumor region.

Identification of subclones and evolutionary dynamics parameters utilizing Bayesian framework—We also used a Bayesian approach developed by Williams et al. (2018), as implemented in their SubClonalSelection.jl to draw clonal inferences from variant allele frequencies. Subclonal architecture of each sample was detected and the selective advantage (s) and time of appearance of each subclone (t) were measured simultaneously, as recommended by the authors. Input parameters for SubClonalSelection.jl were set as default and fitted with specific sample (read depth, minimum/Maximum range of VAF, minimum cellularity), specifically, mutation rate as the input parameter was the effective mutation rate per tumor doubling. All simulations were run for a 10^6 iterations.

Transcriptomic data analysis—After initial quality-checks of the raw RNA sequencing reads using FastQC (v0.11.7), and removal of any low quality reads, STAR aligner (v 2.6.0c) (Dobin et al., 2013) was used to map the remaining reads onto human genome (GRCh38). RSEM (v1.3.1) (Li and Dewey, 2011) was used for transcript quantification, and $\log_2$ (TPM +1) (Transcripts per million) values were reported for different tumor regions and also matched non-malignant regions. Principal Component Analysis (PCA) and hierarchical clustering was done on $\log_2$ (TPM+1) using FactoMineR package. ESTIMATE (Yoshihara et al., 2013) was used for predicting tumor purity, and the presence of stromal/immune cells in tumor tissues.

We measured the extent of perturbation in cellular pathways using a metric called tITH calculated using network-based Jensen-Shannon Divergence (Park et al., 2016). nJSD was applied as a distance measure between two network states. For each pathway, tITH was defined based on two distance values, distance from tumor to non-malignant tissue (*NT*) and distance from tumor to maximally ambiguous network (*TA*) into a single entropy-based metric of pathway-level dysregulation as follows:

$$\mathrm{tITH} = \mathrm{NT} \,/\, \mathrm{NT} + \mathrm{TA}$$

This metric was calculated for relevant pathways in each tumor region across all patient samples. Same score was used in assessing multi-drug resistance pathway activity.

We computed epithelial versus mesenchymal (EM) characteristics scores based on a published approach (Ramaker et al., 2017) using expression signature of 76 genes relevant for epithelial and mesenchymal characteristics with minor modification. Of the gene signatures, we could estimate gene expression [$\log_2$(TPM+1)] for 49 epithelial related genes (e-genes) and 11 mesenchymal related genes (m-genes) in our samples. For each gene, we estimate its mean (μ) and standard deviation (σ) in the 6 non-malignant tissue samples, and then in *i*-th tumor region, we computed the Z-score (Z_i) corresponding to its expression, $e_i$as $Z_i = (e_i - \mu)/\sigma$. Then, for each tumor region, EM score was calculated as:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

EM score = av. Z-score of m-genes − av. z-score of e-genes

EM score was compared between regions from the same tumor. Proliferation (PI) and apoptotic (AI) indices were calculated using a similar approach using 124 proliferation-associated genes (Wilkerson et al., 2012)and 6 apoptosis-related genes. For each gene, Z-scores corresponding to its expression in tumor regions were calculated based on the mean and standard deviation of its expression in the non-malignant samples, and then proliferation index (PI) and apoptosis index (AI) were defined as mean z-scores of all proliferation and apoptosis genes, respectively. Similar strategy was followed for calculating scores for hypoxia signature (Buffa et al., 2010) and Pemetrexed resistance signature.

mRNA expression based subgroups were also inferred for different regions of the same tumor across all samples. Predictor centroid data for different subgroups (primitive, classical, secretory and basal) was downloaded from Wilkerson et al., (2012). For the same set of genes (as centroids), Spearman correlation coefficient between expression data of samples and centroids was calculated. Subgroup showing the highest correlation coefficient (with pval < 0.01) was assigned to each sample.

Multiregional tumor trees at transcriptomic levels were constructed for every patient using RNA expression data for all genes across different regions, using an approach similar to that used at genomic level. Manhattan distance was computed between all regions of a patient sample using $\log_2$ (TPM+1) and then unrooted dendrograms were drawn using this distance metric.

Immune signature analysis—Immune cell infiltrations were inferred from molecular signatures of immune cell types. ESTIMATE (v1.0.13) (Yoshihara et al., 2013) was used for predicting the level of immune infiltration in tumor tissues. CIBERSORT (Gentles et al., 2015) was used to estimate abundance of different immune cell populations from expression data. Standard LM22 signature gene file, and 1000 permutations were used to calculate deconvolution p values.

Class I and class II HLA types were predicted from both DNA and RNA data using HLAminer (Warren et al., 2012). Only most likely HLA class I alleles (Confidence (−10 * log10(Eval)) > = 30 and Score > = 500) were considered. HLA identified at both DNA and RNA levels were selected for further analysis.

MuPeXI (Bjerregaard et al., 2017) was used for predicting potential neo-epitopes (neo-epitopes) using exome and RNA sequencing data. It uses somatic mutation calls (SNVs and InDels), a list of HLA types, and a gene expression profile to predict tumor specific peptides. For any neoantigen to be classified as potential neo-epitope, it should be expressed, and have a high affinity to its respective expressed HLA alleles. All predicted neo-epitopes were classified into two categories based on their binding affinity to predicted HLA types; weak binders and strong binders. Strong binders were defined as: expression > 1, mutRank < 0.5, Priority score > 0 and difference between normal and mutRank >0.5 and weak binders as: mutRank <2. mutRANK is % Rank of prediction score for mutant peptides as defined by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

netMHCpan 4.0 (Jurtz et al., 2017), which is used at backhand in MuPeXI. Based on this criteria weak binders and strong binders were predicted for all tumor regions across all samples. Further these neo-epitopes were classified into three categories based on their regional abundances; ubiquitous, shared and unique. This classification was done for both weak and strong binders separately.

TCR repertoire predictions were done on both DNA and RNA data using MiXCR (Bolotin et al., 2015) and VDJtools (Shugay et al., 2015). Paired-end RNA and exome fastq files for each sample were provided for MiXCR analysis. The exome data was analyzed using the default MiXCR pipeline parameters, and the RNA-seq files were analyzed using the pipeline and parameters recommended for RNA-seq data in the MiXCR documentation. For each sample, the clonotypes with the highest count (supporting reads) were selected from the exome data (> 3 supporting reads) and the RNA data (> 24 supporting reads). The count, fraction, and sequence of these clonotypes were then compared to other clonotypes from the same tumor to look for intratumor heterogeneity. In addition, the MiXCR output files were then converted to vdjtools format and analyzed using vdjtools to get basic statistics, spectratypes, segment usage graphs, and diversity statistics for each sample.

Anti-PD1 favor score was calculated for each tumor region using gene expression data [$\log_2$ (TPM+1)] for 28 genes that comprise of anti-PD1 favor signature (Gibney et al., 2016). For each gene, Z-scores corresponding to its expression in tumor regions were calculated based on the mean and standard deviation of its expression in the non-malignant samples, and then anti-PD1 favor score was defined as mean z-scores of the genes involved.

Similar to genomic and transcriptomic data, multiregional tumor trees were also made for every patient to infer immunogenic heterogeneity (iITH). In this case trees were made using two different datasets; i) immune cell proportions from CIBERSORT (Gentles et al., 2015) and ii) expression of neo-epitopes predicted from MuPeXI (Bjerregaard et al., 2017). Expression of predicted neo-epitopes was considered zero in a region if the neo-epitope is not detected in that region. Manhattan distance was computed between all regions of a patient sample using both the datasets independently and separate unrooted dendrograms were drawn using respective distance metrics.

Single Cell RNA sequencing data analysis—We obtained single cell RNaseq data for 5 non-small cell lung tumors from the dataset published by Lambrechts et al., (2018) (E-MTAB-6149 and E-MTAB-6653). For each tumor core, middle, and edge regions were profiled on 10X platform, and tumor, immune, and stromal cells were annotated. For individual tumor cells, we used the log2CPM values to calculated proliferation, apoptosis, hypoxia and EMT scores as described above, after taking Alveolar Type 2 (AT2) cells from the same patients as control.

We then used single cell gene expression data to infer copy number and clonal clusters using HoneyBadger (Fan et al., 2018). In brief, the inference is drawn by considering that copy number gain or loss would result in systematic increase or decrease in expression of genes that are co-localized in the genome, and cells that are within the same subclonal clusters would show more similarities than those that are distant in the clonal phylogeny. Under

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

default setting HoneyBadger first explicitly infers per-cell copy number gain or loss from expression-based raw copy number log2 ratios, before making clonal architecture inference, which is unsuitable for large datasets. So we used raw copy number log2 ratios inferred from gene expression data by HoneyBadger directly to make clonal architecture inference, bypassing per-cell explicit copy number calling. Single cell profiling allows, in principle, complete phylogenetic reconstruction, but in practice, sparse single cell RNaseq data makes deep branching inferences unreliable. So, we compared top 4 major subclonal clusters down the phylogenetic hierarchy for systematic differences in proliferation and other cancer-related pathways.

The extent of immune infiltration was calculated based on the number of different types of immune cells in each region (core, middle, edge) as annotated by Lambrechts et al. (2018). Nested anova was used to test the statistical) significance of the difference in growth rates and other tumor associated features in both clusters and regions.

Histopathological screening of H&E stained slides—High-resolution digital images of Hematoxylin and Eosin (H&E) stained sections were evaluated by a board certified pathologist blinded to genomic data. Each image was scored for percent tumor nuclei and for the presence of lymphocytes on a scale of 1 (minimal) to 3 (robust) over the entire section.

MetaITH pipeline—metaITH - a computational pipeline (https://github.com/sjdlabgroup/metaITH) provides an interface to perform meta-analysis of intra-tumor heterogeneity patterns across different levels, and for estimating transcriptomic signatures of disease-relevant biological processes. Heterogeneity-related utilities include heatmaps, dendrograms, and measures of intra-tumor divergence and diversity at different levels. Signature-related utilities include geneset signatures for proliferation, apoptosis, hypoxia, multi-drug resistance, anti-PD1 favor, but those could be also used to estimate other user-defined signatures as well.

QUANTIFICATION AND STATISTICAL ANALYSIS

All graphs and statistical analyses, exome sequencing, RNaseq, scRNaseq data analysis were performed using R version 3.4.0. Statistical significance was assessed using paired Wilcoxon rank sum test and nested ANOVA test as applicable.

DATA AND CODE AVAILABILITY

The Sequence Read Archive (SRA) accession number for the raw data from the exome sequencing and RNaseq datasets reported in this paper is PRJNA574648. metaITH - a computational framework generated in this study has been made available at: https://github.com/sjdlabgroup/metaITH.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ACKNOWLEDGMENTS

The authors acknowledge financial support from R01GM129066, P30CA072720, and Robert Wood Johnson Foundation (S.D.). The authors thank the other members of the laboratories of S.D. and the Center for Systems and Computational Biology at Rutgers Cancer Institute, especially Dr. Saurabh V. Laddha, for helpful discussions.

REFERENCES

Aghili L, Foo J, DeGregori J, and De S (2014). Patterns of somatically acquired amplifications and deletions in apparently normal tissues of ovarian cancer patients. Cell Rep. 7, 1310–1319. [PubMed: 24794429]

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature 500, 415–421. [PubMed: 23945592]

Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science 354, 618–622. [PubMed: 27811275]

Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, Ji HP, and Maley CC (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat. Med 22, 105–113. [PubMed: 26618723]

Ardeshir-Larijani F, Bhateja P, Lipka MB, Sharma N, Fu P, and Dowlati A (2018). KMT2D Mutation Is Associated With Poor Prognosis in Non-Small-Cell Lung Cancer. Clin. Lung Cancer 19, e489–e501. [PubMed: 29627316]

Bjerregaard A-M, Nielsen M, Hadrup SR, Szallasi Z, and Eklund AC (2017). MuPeXI: prediction of neo-epitopes from tumor sequencing data. Cancer Immunol. Immunother 66, 1123–1130. [PubMed: 28429069]

Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, and Chudakov DM (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods 12, 380–381. [PubMed: 25924071]

Brock A, Chang H, and Huang S (2009). Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours. Nat. Rev. Genet 10, 336–342. [PubMed: 19337290]

Buffa FM, Harris AL, West CM, and Miller CJ (2010). Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. Br. J. Cancer 102, 428–435. [PubMed: 20087356]

Cingolani P, Platts A, Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, and Ruden DM (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92. [PubMed: 22728672]

Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, Sher X, Liu XQ, Lu H, Nebozhyn M, et al. (2018). Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science 362, eaar3593. [PubMed: 30309915]

Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, et al. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat. Biotechnol 29, 1120–1127. [PubMed: 22081019]

De S (2011). Somatic mosaicism in healthy human tissues. Trends Genet. 27, 217–223. [PubMed: 21496937]

De S, and Ganesan S (2017). Looking beyond drivers and passengers in cancer genome sequencing data. Ann. Oncol 28, 938–945. [PubMed: 27998972]

Ding Z, Mangino M, Aviv A, Spector T, and Durbin R; UK10K Consortium (2014). Estimating telomere length from whole genome sequence data. Nucleic Acids Res. 42, e75. [PubMed: 24609383]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. [PubMed: 23104886]

Fan J, Lee H-O, Lee S, Ryu D-E, Lee S, Xue C, Kim SJ, Kim K, Barkas N, Park PJ, et al. (2018). Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 28, 1217–1227. [PubMed: 29898899]

Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, Nair VS, Xu Y, Khuong A, Hoang CD, et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat. Med 21, 938–945. [PubMed: 26193342]

Gibney GT, Weiner LM, and Atkins MB (2016). Predictive biomarkers for checkpoint inhibitor-based immunotherapy. Lancet Oncol. 17, e542–e551. [PubMed: 27924752]

Hou J, Lambers M, den Hamer B, den Bakker MA, Hoogsteden HC, Grosveld F, Hegmans J, Aerts J, and Philipsen S (2012). Expression profiling-based subtyping identifies novel non-small cell lung cancer subgroups and implicates putative resistance to pemetrexed therapy. J. Thorac. Oncol 7, 105–114. [PubMed: 22134068]

Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, Shafi S, Johnson DH, Mitter R, Rosenthal R, et al.; TRACERx Consortium (2017). Tracking the Evolution of Non-Small-Cell Lung Cancer. N. Engl. J. Med 376, 2109–2121. [PubMed: 28445112]

Jia Q, Wu W, Wang Y, Alexander PB, Sun C, Gong Z, Cheng J-N, Sun H, Guan Y, Xia X, et al. (2018). Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. Nat. Commun 9, 5361. [PubMed: 30560866]

Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, and Nielsen M (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J. Immunol 199, 3360–3368. [PubMed: 28978689]

Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, and Ellisen LW (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat. Commun 9, 3588. [PubMed: 30181541]

Kim W, Ludlow AT, Min J, Robin JD, Stadler G, Mender I, Lai T-P, Zhang N, Wright WE, and Shay JW (2016). Regulation of the Human Telomerase Gene TERT by Telomere Position Effect-Over Long Distances (TPE-OLD): Implications for Aging and Cancer. PLoS Biol. 14, e2000016. [PubMed: 27977688]

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, and Wilson RK (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576. [PubMed: 22300766]

Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, Bassez A, Decaluwé H, Pircher A, Van den Eynde K, et al. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med 24, 1277–1289. [PubMed: 29988129]

Li B, and Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. [PubMed: 21816040]

Li H, and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. [PubMed: 19451168]

Li S, Garrett-Bakelman FE, Chung SS, Sanders MA, Hricik T, Rapaport F, Patel J, Dillon R, Vijay P, Brown AL, et al. (2016). Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nat. Med 22, 792–799. [PubMed: 27322744]

Macheret M, and Halazonetis TD (2015). DNA replication stress as a hallmark of cancer. Annu. Rev. Pathol 10, 425–448. [PubMed: 25621662]

Marusyk A, Almendro V, and Polyak K (2012). Intra-tumour heterogeneity: a looking glass for cancer? Nat. Rev. Cancer 12, 323–334. [PubMed: 22513401]

Mazor T, Pankov A, Johnson BE, Hong C, Hamilton EG, Bell RJA, Smirnov IV, Reis GF, Phillips JJ, Barnes MJ, et al. (2015). DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. Cancer Cell 28, 307–317. [PubMed: 26373278]

McGranahan N, and Swanton C (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell 168, 613–628. [PubMed: 28187284]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Park Y, Lim S, Nam J-W, and Kim S (2016). Measuring intratumor heterogeneity by network entropy using RNA-seq data. Sci. Rep 6, 37767. [PubMed: 27883053]

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, et al. (2014). Singlecell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401. [PubMed: 24925914]

Ramaker RC, Lasseigne BN, Hardigan AA, Palacio L, Gunther DS, Myers RM, and Cooper SJ (2017). RNA sequencing-based cell proliferation analysis across 19 cancers identifies a subset of proliferation-informative cancers with a common survival signature. Oncotarget 8, 38668–38681. [PubMed: 28454104]

Rosenthal R, McGranahan N, Herrero J, Taylor BS, and Swanton C (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 17, 31. [PubMed: 26899170]

Rosenthal R, Cadieux EL, Salgado R, Bakir MA, Moore DA, Hiley CT, Lund T, Tanić M, Reading JL, Joshi K, et al.; TRACERx consortium (2019). Neoantigen-directed immune escape in lung cancer evolution. Nature 567, 479–485. [PubMed: 30894752]

Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, and Shah SP (2014). PyClone: statistical inference of clonal population structure in cancer. Nat. Methods 11, 396–398. [PubMed: 24633410]

Sharma A, Jiang C, and De S (2018). Dissecting the sources of gene expression variation in a pan-cancer analysis identifies novel regulatory mutations. Nucleic Acids Res. 46, 4370–4381. [PubMed: 29672706]

Shen R, and Seshan VE (2016). FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 44, e131. [PubMed: 27270079]

Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, Pogorelyy MV, Nazarov VI, Zvyagin IV, Kirgizova VI, et al. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. PLoS Comput. Biol 11, e1004503. [PubMed: 26606115]

Suda K, Kim J, Murakami I, Rozeboom L, Shimoji M, Shimizu S, Rivard CJ, Mitsudomi T, Tan A-C, and Hirsch FR (2018). Innate Genetic Evolution of Lung Cancers and Spatial Heterogeneity: Analysis of Treatment-Naïve Lesions. J. Thorac. Oncol 13, 1496–1507. [PubMed: 29933065]

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al.; Cancer Genome Atlas Research Network (2018). The Immune Landscape of Cancer. Immunity 48, 812–830.e14. [PubMed: 29628290]

Tomasetti C, Vogelstein B, and Parmigiani G (2013). Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc. Natl. Acad. Sci. USA 110, 1999–2004. [PubMed: 23345422]

Wangari-Talbot J, and Hopper-Borge E (2013). Drug Resistance Mechanisms in Non-Small Cell Lung Carcinoma. J. Cancer Res. Updates 2, 265–282.

Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, and Holt RA (2012). Derivation of HLA types from shotgun sequence datasets. Genome Med. 4, 95. [PubMed: 23228053]

Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, Muldrew K, Miller CR, Randell SH, Socinski MA, et al. (2010). Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin. Cancer Res 16, 4864–4875. [PubMed: 20643781]

Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, Miller CR, Socinski MA, Parsons AM, Thorne LB, et al. (2012). Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS One 7, e36530. [PubMed: 22590557]

Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, and Graham TA (2018). Quantification of subclonal selection in cancer from bulk sequencing data. Nat. Genet 50, 895–903. [PubMed: 29808029]

Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun 4, 2612. [PubMed: 24113773]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Zhang AW, McPherson A, Milne K, Kroeger DR, Hamilton PT, Miranda A, Funnell T, Little N, de Souza CPE, Laan S, et al. (2018). Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. Cell 173, 1755–1769.e22. [PubMed: 29754820]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- Lung squamous cell carcinoma has a moderate level of intra-tumor genetic heterogeneity
- Transcriptomic heterogeneity impacts cancer pathways, driving phenotypic heterogeneity
- Neo-epitope burden negatively correlates with immune infiltration
- Non-genetic heterogeneity influences tumor evolutionary dynamics

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

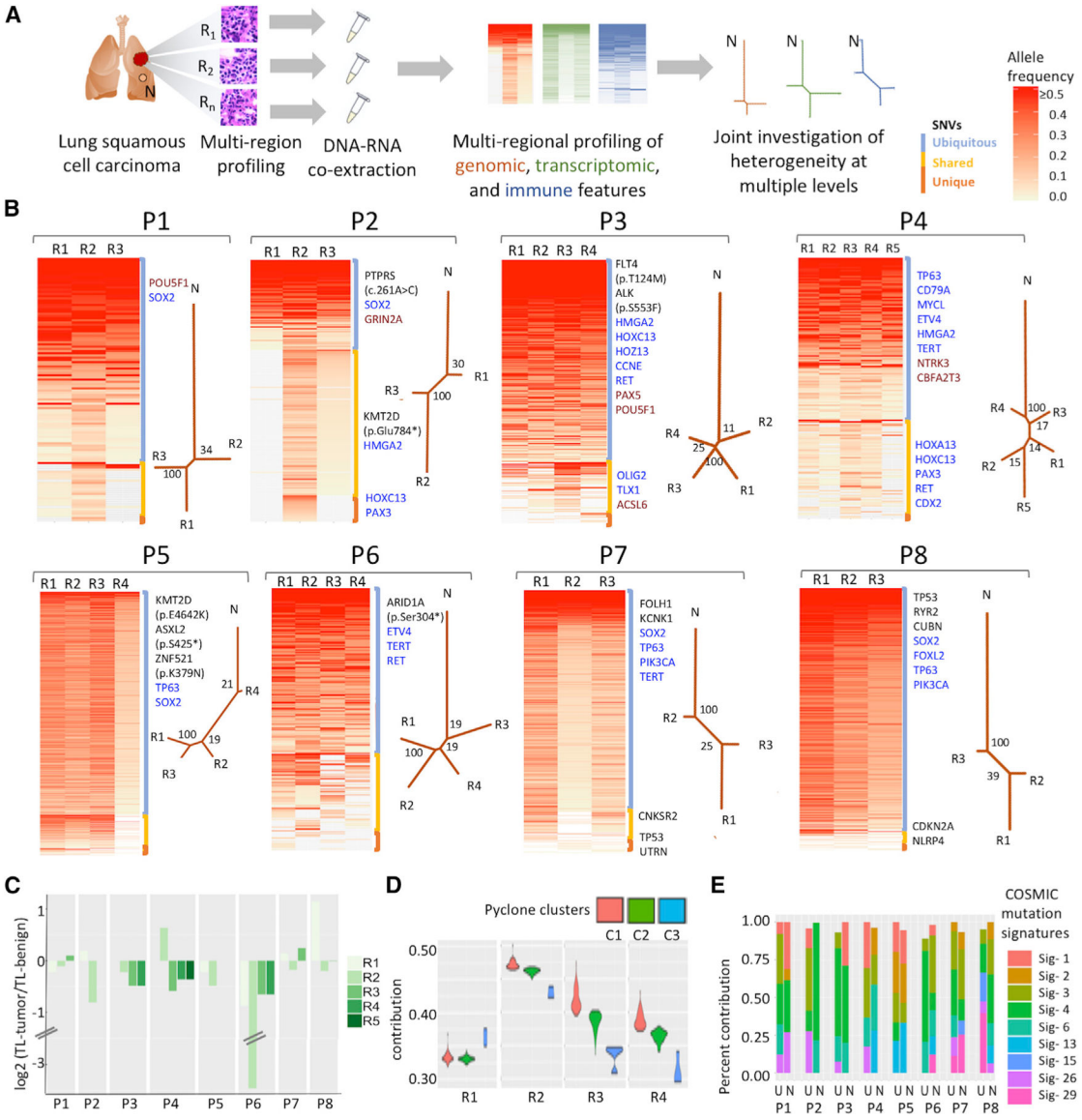


Figure 1. Assessment of Genetic Heterogeneity in Lung Squamous Cell Cancer

(A) Schematic representation of the study design showing multi-dimensional analysis of intra-tumor heterogeneity based on multi-region profiling of tumor specimens.

(B)Heatmap shows regional variation in allele frequency of somatic single nucleotide variants (SNVs) for 8 patient samples. Somatic variants below detection threshold (<2% allele frequency) are marked in gray. Variations are categorized into three categories: ubiquitous, present in all regions (blue); shared, present in multiple regions but not all (yellow); and unique, present in single region (orange). Dendrograms represent genetic similarity (represented by branch length) between different regions within a tumor for each patient sample based on variant allele frequency of somatic SNVs. Numbers on the nodes represents bootstrap values.

(C) Regional differences in inferred telomere length in tumor regions, relative to matched normal tissue.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(D) Relative abundance of somatic mutation clusters identified in different tumor regions for P6.

(E) COSMIC mutational signatures inferred from somatic base changes in tumors corresponding to the mutations that are ubiquitous (U) and non-ubiquitous (NU; i.e., shared and unique). Signature 1, mutational process initiated by spontaneous deamination of 5-methyl cytosine; signature 2, mutational process due to APOBEC activity; signature 3, mutational process due to homologous recombination defect; signature 4, mutational process due to smoking; signature 6, mutational process due to defective DNA mismatch repair; signature 13, mutational process due to APOBEC activity; signature 15, mutational process due to defective DNA mismatch repair; signature 26, mutational process due to defective DNA mismatch repair; and signature 29, mutational process due to tobacco chewing habit.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

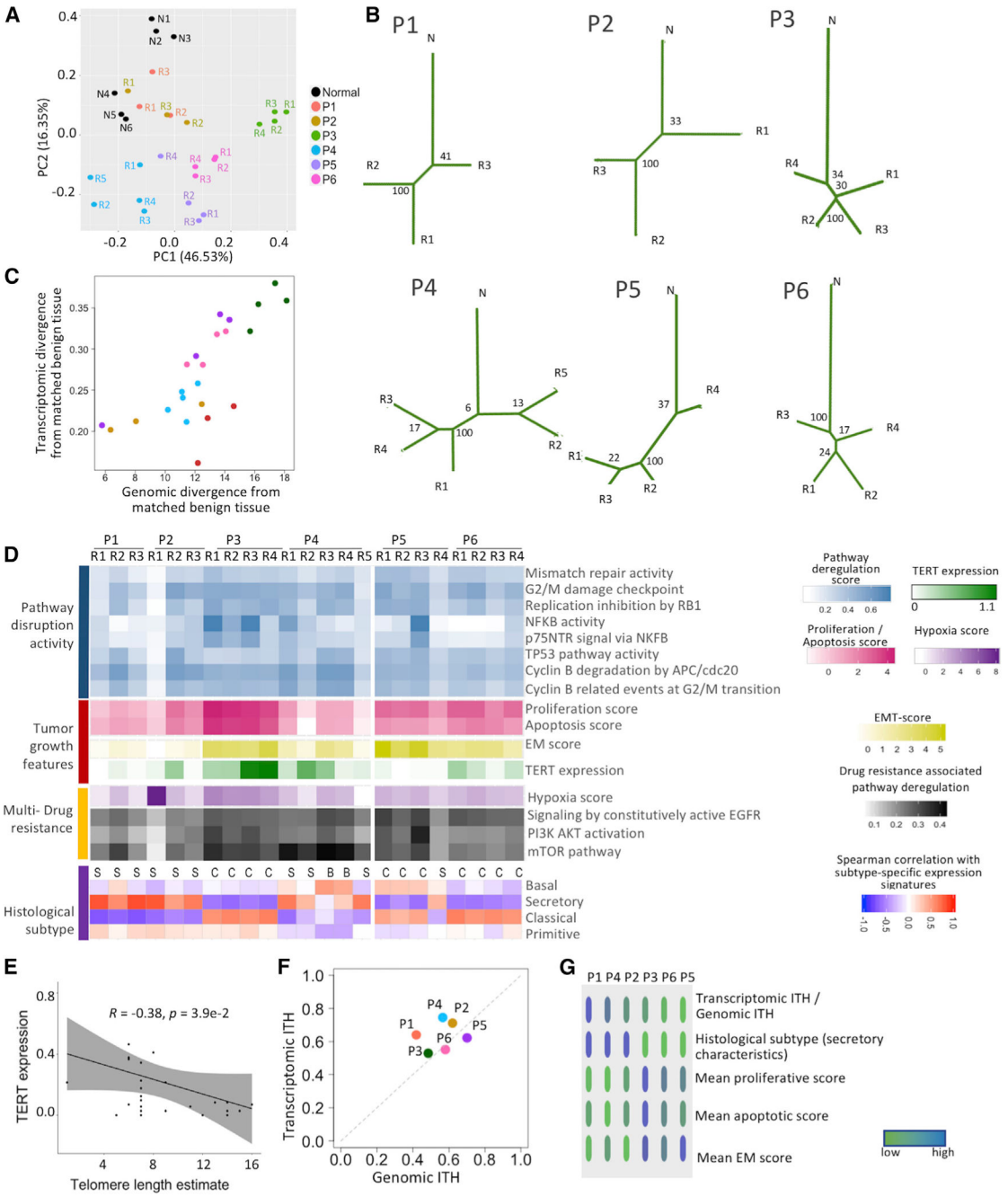


Figure 2. Assessment of Intra-tumor Transcriptomic Heterogeneity in Lung Squamous Cell Cancer

(A) Principal-component analysis (PCA) plot showing the extent of transcriptomic variation within and across tumor and non-malignant tissue regions from the patients. Non-malignant tissues from different patients are shown in black, whereas tumor tissues from different patients are shown in other colors. The proportion of variation explained by the first and second principal components are 46.53% and 16.35%, respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(B) Dendrograms represent similarity (represented by branch length) between different regions for each patient sample based on gene expression profiles for all genes. Numbers on the nodes represent bootstrap values.
(C) Scatterplot comparing the extent of transcriptomic and genomic divergence for different tumor regions from their respective matched non-malignant tissues. Color codes are same as in (A).
(D) Heatmaps showing the extent of pathway disruption activity measured using nJSD, proliferative score, apoptosis score, epithelial-mesenchymal score, TERT expression, hypoxia score, multi-drug-resistant pathway disruption activity, and histological subtype score based on expression of published biomarker genes for different tumor regions.
(E) Scatterplot showing inverse association between TERT expression and telomere length estimates. Spearman correlation coefficient and p value are shown at the top.
(F) Scatterplot comparing the extent of transcriptomic and genomic intra-tumor heterogeneity for the tumor samples. The color code is same as in (A).
(G) Oncoprint plot showing histological subtype (secretory characteristics), proliferative score, apoptotic score, and epithelial versus mesenchymal characteristics score for the tumor samples ranked according to their ratio of transcriptomic ITH over genetic ITH.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

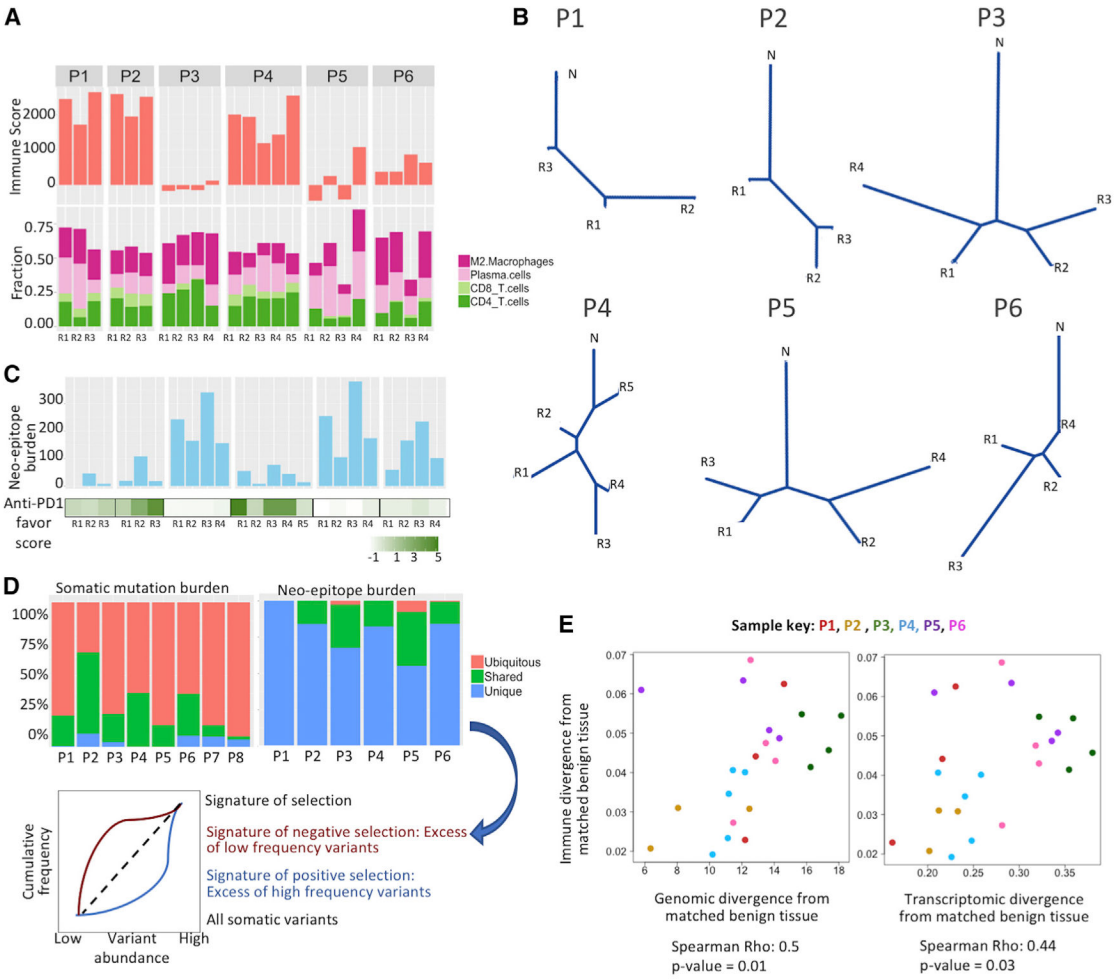


Figure 3. Assessment of Intra-tumor Immune Heterogeneity in Lung Squamous Cell Cancer

(A) Estimated immune score (top panel), relative proportion of different immune cell types including M2 macrophages, plasma cells, CD4, and CD8 T cells (bottom panel).

(B) Dendrograms represent similarity (represented by branch length) between different regions for each patient sample based on estimated immune cell fractions in different regions.

(C) Neo-epitope (total) burden, as predicted using mutation and expression data, in different regions of all patient samples (top panel), and anti-PD1 favor score, a measure of responsiveness against anti-PD1 therapy, in different regions of all patient samples (bottom panel).

(D) Somatic mutations in the tumor samples and those that are designated as neo-epitopes are grouped as ubiquitous, shared, and unique depending on their regional presentation. An excess of unique epitopes relative to the patterns observed for all somatic variants is indicative of negative selection on the neo-epitopes.

(E) Scatterplot comparing the extent of immune and genetic divergence (left panel) and the extent of immune and transcriptomic divergence (right panel) for different tumor regions from their respective matched non-malignant tissues. Color codes are same as Figure 2A.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

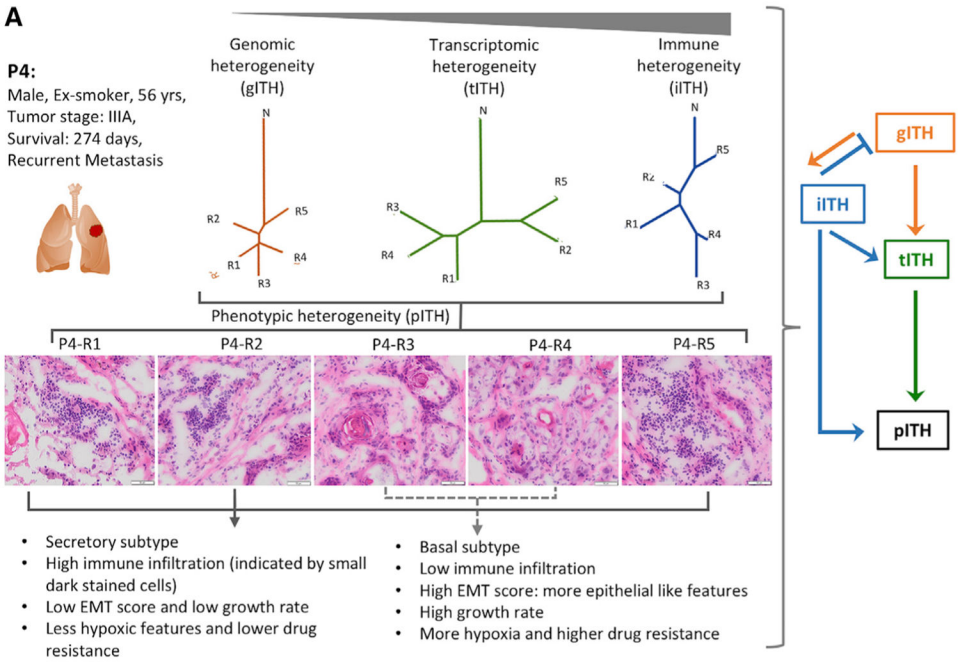


Figure 4. Impact of Multi-level Intra-tumor Heterogeneity

(A) As an example, multi-level intra-tumor heterogeneity in patient P4, a 56-year-old male exsmoker patient with stage IIIA tumor, is presented. Regional variations in histological characteristics and immune cell infiltration correlate with predicted subtype characteristics and immune scores. Furthermore, the tumor shows regional variations in proliferation and apoptosis scores, indicating coherence in multi-level intra-tumor heterogeneity. H&E stained slides for different regions of P4 are shown with scale bars of 50 μm at bottom right corners.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

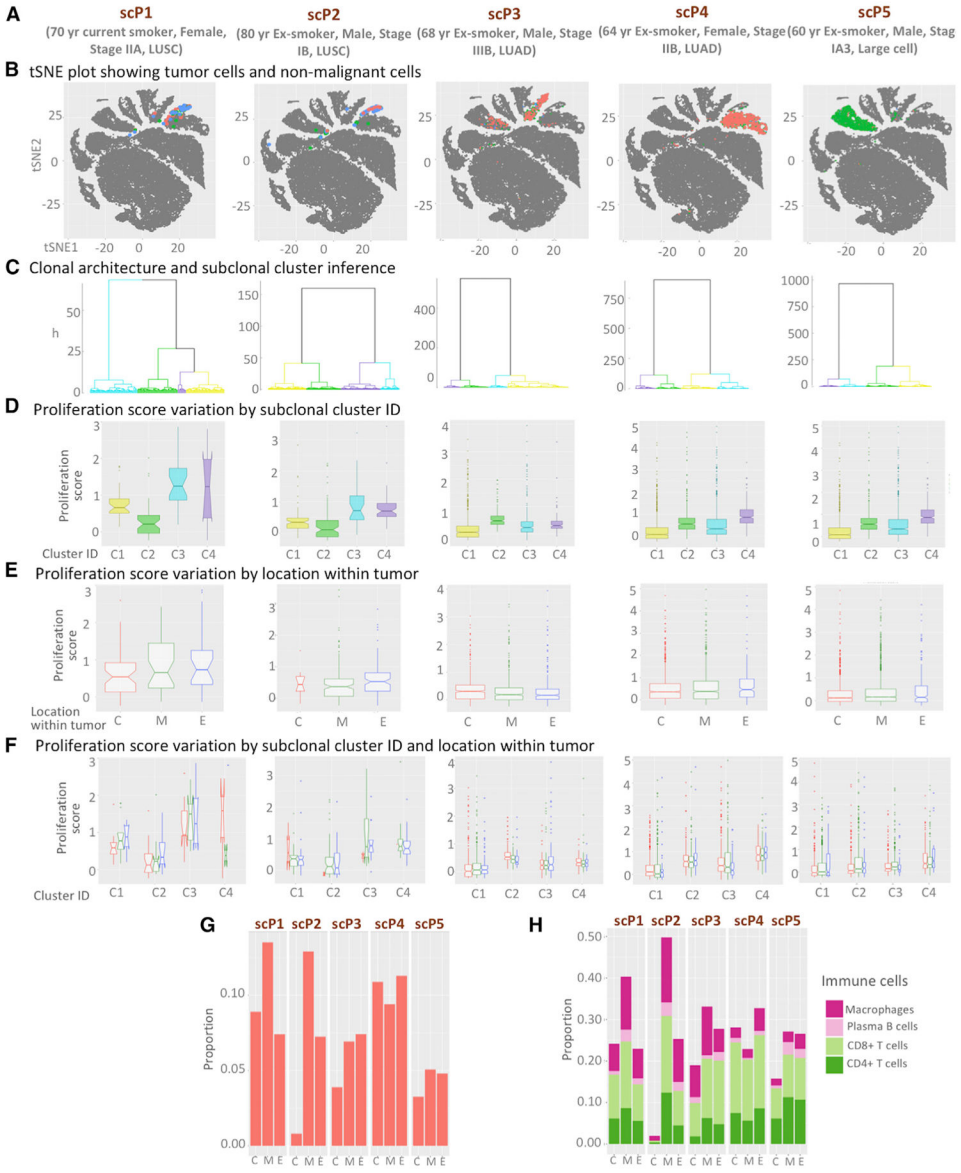


Figure 5. Intra-tumor Heterogeneity at Single-Cell Resolution

(A) Clinical information of 5 patients for which multi-region single-cell sequencing data was performed by Lambrechts et al. (2018).

(B) tSNE plots show the transcriptomic heterogeneity of the tumor cell populations and also non-malignant cell populations for reference. Tumor cells from different regions, namely, core, middle, and edge, of each tumor are colored with red, green, and blue, respectively, whereas all non-malignant cells are shown in gray.

(C) Tumor clonal architecture inferred from single cell RNA-seq data-guided copy number variation calls. For each tumor, 4 major subclonal clusters are numbered C1, C2, C3, and C4, marked with different colors.

(D) Boxplots showing distribution of proliferation scores of all tumor cells grouped by their subclonal cluster membership.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(E) Boxplots showing distribution of proliferation scores of all tumor cells grouped by different geographical regions—core (C), middle (M), and edge (E).
(F) Boxplots showing distribution of proliferation scores of all tumor cells grouped by different geographical regions and cluster identifier. Color codes are consistent between (B), (E), and (F) and also between (C) and (D).
(G) Proportion of immune cells out of total cells isolated for each patient.
(H) Proportion of different tumor relevant immune cell populations out of total immune cells for each tumor.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Human tumor and normal samples	Biospecimen Repository and Histopathology Service, Rutgers-CINJ and LCBRN - University of Virginia	https://www.cinj.org/research/biospecimen-repository-and-histopathology-service http://lungbio.sites.virginia.edu
Chemicals and tools		
TRIzol reagent	Life Technologies	Cat# 15596026
DNeasy Blood & Tissue Kits	QIAGEN	Cat# 69504
Brain Punch Tissue Set	Leica Biosystems	Cat# 39443001
Deposited data		
Exome sequencing files	This paper	PRJNA574648
RNaseq files	This paper	PRJNA574648
single cell RNaseq files	Lambrechts et al., 2018	E-MTAB-6149 and E-MTAB-6653
Software & Algorithms		
BWA v0.7.17-r1188	Li and Durbin, 2009	PMID: 19451168
VarScan2	Koboldt et al., 2012	PMID: 22300766
SnpEff v4.3t	Cingolani et al., 2012	PMID: 22728672
deconstructSig	Rosenthal et al., 2016	PMID: 26899170
FACETS	Shen and Seshan, 2016	PMID: 27270079
TelSeq	Ding et al., 2014	PMID: 24609383
PyClone V0.13.1	Roth et al., 2014	PMID: 24633410
STAR Aligner v 2.6.0c	Dobin et al., 2013	PMID: 23104886
RSEM V1.3.1	Li and Dewey, 2011	PMID: 21816040
ESTIMATE v1.0.13	Yoshihara et al., 2013	PMID: 24113773
nJSD	Park et al., 2016	PMID: 27883053
CIBERSORT	Gentles et al., 2015	PMID: 26193342
HLAminer	Warren et al., 2012	PMID: 23228053
MuPeXI	Bjerregaard et al., 2017	PMID: 28429069
MiXCR	Bolotin et al., 2015	PMID: 25924071
VDJtools	Shugay et al., 2015	PMID: 26606115
HoneyBADGER	Fan et al., 2018	PMID: 29898899