Subtyping of Small Cell Lung Cancer using plasma cell-free nucleosomes

Gavriel Fialkoff[1,2,*], Nobuyoki Takahashi[3,4,*], Israa Sharkia[1,2], Jenia Gutin[1,2], Lorinc Pongor[3], Arun Rajan[5], Samantha Nichols[3], Linda Sciuto[3], Rasa Vilimas[3], Chante Graham[3], Parth Desai[3], Miriam Maoz[6], Yakir Rottenberg[6], Nevo Shoshan[6], Hovav Nechushtan[6], Tamar Peretz[6], Diana Roame[7], Paula Carter[7], Rajesh Kumar[3], Ayala Hubert[6], Jonathan E Cohen[6], Deep Upadhyay[3], Azzam Salah[6], Mark Temper[6], Albert Grinshpun[6], William Douglas Figg Sr[7], Aviad Zick[6], Ronen Sadeh[1,2], Nir Friedman[1,2,**], Anish Thomas[3,**]

[1] School of Computer Science and Engineering, Hebrew University of Jerusalem

[2] Institute for Medical Research Israel-Canada, The Hebrew University-Hadassah Medical School, Jerusalem, Israel

[3] Developmental Therapeutics Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD

[4] Department of Medical Oncology, National Cancer Center East Hospital, Chiba, Japan

[5] Thoracic and GI Malignancies Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD

[6] Sharett Institute of Oncology, Hadassah-Hebrew University Medical Center, Jerusalem, Israel

[7] Genitourinary Malignancies Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD

* These authors contributed equally

** Corresponding authors

Abstract

Small cell lung cancer (SCLC) is an aggressive malignancy with exceptionally poor prognosis and limited therapeutic advances in the past few decades. Although SCLCs are treated as a single disease entity in clinic, emerging data support subtypes of SCLC driven by expression of distinct transcription regulators, which engender unique therapeutic vulnerabilities. However, the translational potential of these observations is limited by access to tumor biopsies. Here, we apply chromatin immunoprecipitation of cell-free nucleosomes carrying active chromatin modifications followed by sequencing (cfChIP-seq) to 299 plasma samples from patients with advanced SCLC, non-SCLC cancers, and healthy adults. We find that cfChIP-seq reliably estimates SCLC circulating free DNA (cfDNA) tumor fraction, recovers the unique epigenetic states of SCLC tissue and cell of origin, and importantly tumor gene expression. Comparison of cfChIP-seq signals to matched tumor transcriptome shows genome-wide concordance particularly in SCLC lineage-defining transcription factors,

presenting a direct link between gene expression in the tumor and plasma cell-free nucleosomes. This work sets the stage to non-invasively profile SCLC transcriptomes using plasma cfDNA histone modifications.

Introduction

Small cell lung cancer (SCLC) is a neuroendocrine lung cancer that is highly aggressive and metastatic with dismal patient outcomes, accounting for approximately 15% of all lung cancers[1]. SCLC sheds the largest amount of cfDNA of all solid tumors. Prior studies have identified cfDNA mutations in more than 80% of SCLC patients[2–11], but recurrent targetable mutations in known oncogenes, such as those seen in the kinases that comprise targetable drivers in lung adenocarcinoma, are rare in SCLC[12]. Recurrent mutations also do not demonstrate consistent co-occurrence or mutual exclusivity, and thus do not define SCLC subtypes.

SCLCs exhibit high expression of neuronal and neuroendocrine transcription factors and *MYC* paralogs that drive a broad range of genes related to cell proliferation and growth signaling[12,13]. Importantly, SCLC subtypes driven by distinct transcription factors have unique therapeutic vulnerabilities[14–18]. However, identification of SCLC transcriptomic subtypes and their application in the context of subtype-specific therapies has proven challenging due to limited access to tumor specimens. The majority of SCLC patients do not undergo surgical resection as their disease is detected after it has spread beyond the primary site[19]. Moreover, patients with relapsed disease generally deteriorate quickly, and recurrence suspected on imaging is typically followed by immediate treatment without biopsies. Highlighting this challenge, SCLC is represented in none of the large sequencing initiatives like The Cancer Genome Atlas and Pan-cancer Analysis of Whole Genomes.

Identifying tumor-specific alterations in cell free DNA (cfDNA) presents a powerful opportunity to reduce cancer morbidity and mortality[20,21]. Most of the current clinical applications of cfDNA are centered around interrogating the mutational landscape, and as such are of limited utility in defining transcriptomic subtypes. We recently reported chromatin immunoprecipitation and sequencing of cell-free nucleosomes from human plasma (cfChIP-seq) to infer the transcriptional programs by genome-wide mapping of plasma cell free-nucleosomes carrying specific histone modifications[22]. Specifically, tri-methylation of histone 3 lysine 4 (H3K4me3) is a well characterized histone modification, marking transcription start

sites (TSS) of genes that are poised or actively transcribed, and predictive of gene expression[22–25].

The translational potential of the newly described SCLC transcriptional phenotypes and their associated vulnerabilities is limited by access to tumor biopsies. We hypothesized that plasma histone modifications may non-invasively track tumor gene expression programs, including lineage-defining transcription factors, opening opportunities for subtype-directed therapies for SCLC, currently treated as a single disease. To this end, we applied cfChIP-seq on 286 plasma samples from self-reported healthy subjects (n=33), patients with non-small cell lung cancer (NSCLC, n=14), colorectal cancer (CRC, n=38), and SCLC patients (n=82) who had 141 plasma samples collected at multiple timepoints during their treatment time-course. The plasma-based gene expression programs were benchmarked against the time-point matched tumor transcriptome in 53 patients with SCLC (**Fig. 1A; Supplementary Table 1**).

Results

SCLCs have distinct cfChIP-seq signals that track tumor burden and prognosis

Plasma samples were collected, processed, and H3K4me3 ChIP-seq performed directly on ~1ml of plasma (Methods, [22]) with a median of 3.1 million unique reads sequenced per sample (**Supplementary Table 2**). For every gene, the number of normalized reads mapping to its respective TSS regions was computed, resulting in gene counts resembling transcription counts in RNA-seq data. Comparing the gene counts in SCLC plasma samples to healthy reference samples (Methods), we found significantly elevated counts in hundreds to thousands of genes in many samples (**Fig. 1B, S1A**).

cfDNA of cancer patients consists of DNA fragments originating from tumor cells and DNA released by normal cells, commonly of the hematopoietic lineage. The tumor fraction can vary substantially, depending on a number of variables including the extent of disease involvement, tumor burden and activity[26]. To account for this variability, we estimated an ‘SCLC-score’ reflecting the proportion of tumor-derived cfDNA. The score is based on a linear regression that matches the observed profile (gene counts) in a sample to a weighted mixture of reference cfChIP-seq profiles. Reference consisted of samples from healthy individuals and a SCLC archetype based on the 10% of SCLC samples (n = 15) with the highest number of differential genes compared with healthy samples (**Fig. S1A**, Methods). The SCLC-score ranged from 0 (‘healthy like’) to 1 (‘SCLC like’), with a median of 0.32 and 0.05 in SCLC samples collected before and after treatment, respectively, suggesting a

decrease of SCLC tumor fraction by the therapeutic interventions. In contrast, plasma from healthy subjects and patients with non-SCLC cancers displayed absent or very low SCLC-scores (median of 0 in healthy and NSCLC, 0.1 in CRC; anova $p < 10^{-15}$. **Fig. 1C**). To assess our estimation in an unsupervised manner, we performed principal component analysis (PCA) on the gene counts of healthy controls and SCLC samples (Methods). Superimposing the estimated SCLC-score over the 2-dimensional PCA plot, we found that PCA1 - the axis contributing most to sample variability - correlated almost perfectly with estimated tumor load (r 0.99, $p < 10^{-15}$ **Fig. 1D, S1B-C**).

Importantly, cfChIP-seq SCLC-scores were significantly correlated with multiple other measures of tumor fraction including somatic copy number alteration-based estimates from ultra-low pass whole genome sequencing[27], circulating tumor cell (CTC) counts, total cfDNA concentrations (Methods. Pearson r 0.77, 0.43 and 0.62; $p < 3e^{-4}$, 0.01, and 0.03, respectively), computerized tomography scan-based volumetric tumor assessments, and standardized unidimensional tumor measurements[28] (Pearson r and p: 0.57 and 0.61; <0.008 and $<2e^{-4}$, respectively **Fig. 1E** and **S1D-E**). Furthermore, cfChIP-seq SCLC-scores tracked radiographic tumor burden through the treatment time course (**Fig. 1F**), and predicted treatment response and overall survival (**Fig. 1G** and **S1F-G**).

Taken together, these results demonstrate the potential utility of cfChIP-seq to non-invasively assess cfDNA tumor fraction in patients with advanced SCLC.

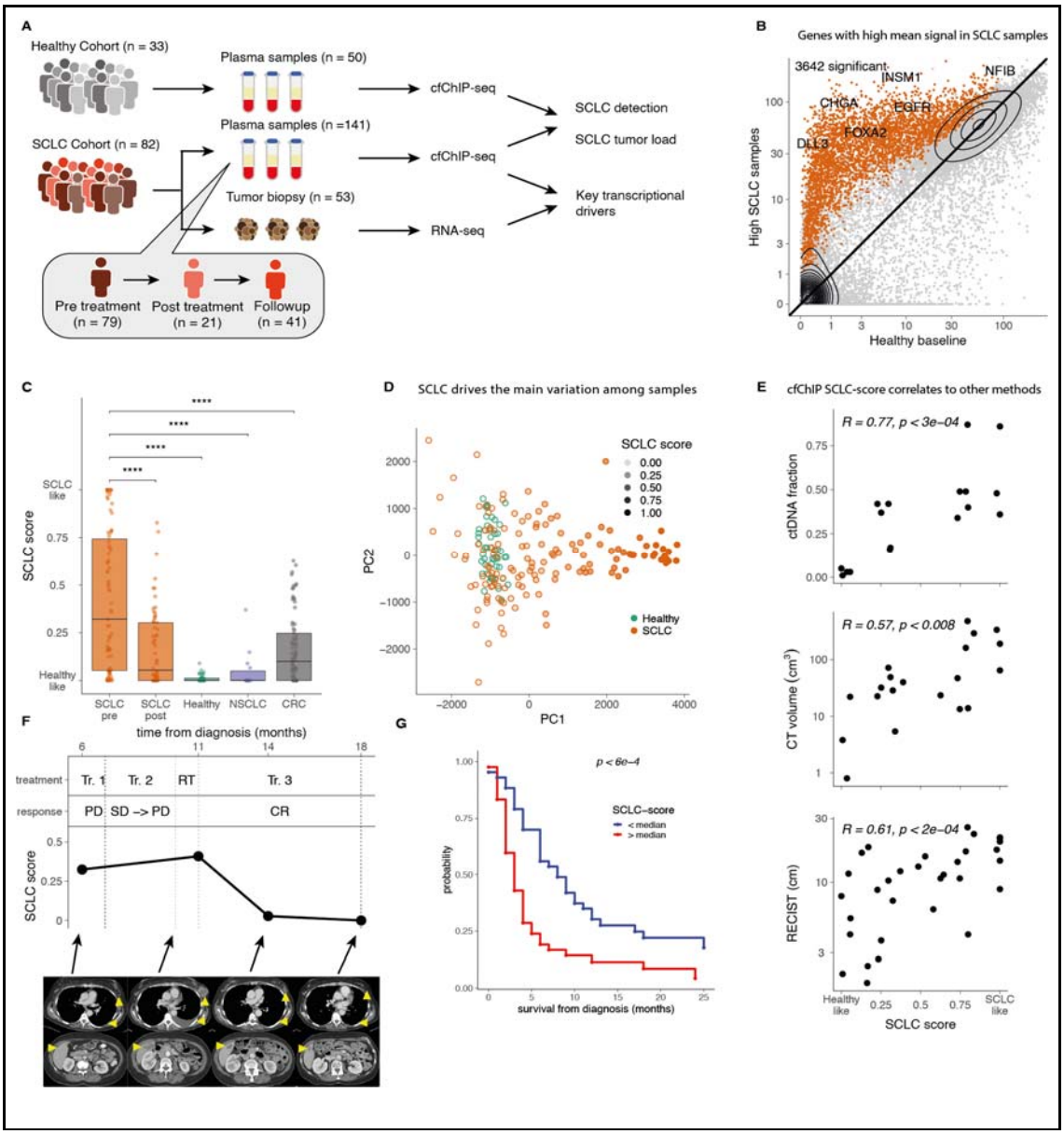
A
Healthy Cohort (n = 33)
Plasma samples (n = 50)
cfChIP-seq
SCLC detection
SCLC tumor load
Plasma samples (n =141)
SCLC Cohort (n = 82)
cfChIP-seq
Key transcriptional drivers
Tumor biopsy (n = 53)
RNA-seq
Pre treatment (n = 79)
Post treatment (n = 21)
Followup (n = 41)
B
Genes with high mean signal in SCLC samples
3642 significant
INSM1
NFIB
CHGA
EGFR
FOXA2
DLL3
High SCLC samples
Healthy baseline
C
SCLC score
SCLC like
Healthy like
SCLC pre
SCLC post
Healthy
NSCLC
CRC
D
SCLC drives the main variation among samples
SCLC score
PC2
PC1
Healthy
SCLC
E
cfChIP SCLC-score correlates to other methods
R = 0.77, p < 3e−04
ctDNA fraction
R = 0.57, p < 0.008
CT volume (cm³)
R = 0.61, p < 2e−04
RECIST (cm)
SCLC score
F
time from diagnosis (months)
treatment
Tr. 1
Tr. 2
RT
Tr. 3
response
PD
SD –> PD
CR
SCLC score
G
p < 6e−4
SCLC−score
< median
> median
probability
survival from diagnosis (months)

Fig. 1: SCLC plasma samples exhibit distinct cfChIP-seq signals that correlate with tumor burden

A. Study outline. Plasma samples were collected from a cohort of healthy individuals and patients with SCLC at various time points during treatment. cfChIP-seq was performed on plasma samples and tumor biopsy samples - many of them time-matched with plasma draw - profiled using RNA-seq. Plasma cfChIP-seq-based gene expression profiles were benchmarked against tumor RNA-seq, and key SCLC transcription regulators examined.
B. Genes with significantly high coverage in selected SCLC (n=15) samples (Methods) compared to healthy baseline. For each gene we compare mean normalized coverage in the SCLC samples (y-axis) to a healthy cohort reference (x-axis). 3642 genes with significantly higher coverage ($q<0.001$) in at least 3 SCLC samples are marked (**Supplementary Table 3)**.
C. Median SCLC-score in different cohorts as calculated by linear regression (Methods). Each dot is a sample, and boxplots summarizes distribution of SCLC-scores in each group. Pre/post refers to the time point along treatment. ****: $P < 0.0001$.
D. Principal components 1 and 2 of SCLC (orange) and healthy (green) samples. Transparency indicates SCLC-score as in C. Principal components analysis was done using all genes (~40,000).
E. Correlation of SCLC-score with other plasma and imaging-based measures of tumor burden. Abbreviations: RECIST: Response Evaluation Criteria in Solid Tumors.
F. Changes of SCLC-score and radiological tumor burden in a patient with SCLC over the treatment time-course (CL0191). Abbreviations: Tr. 1 - durvalumab & olaparib; Tr. 2 - Topotecan & M6620; RT - radiation; Tr. 3 - investigational therapy; CR - complete response; SD - stable disease; PD - progressive disease.
G. Kaplan Meier curve of overall survival in patients with SCLC-score above (red; median survival: 3 month) or below (blue; median survival: 8 month) median. *p* calculated by log-rank test.

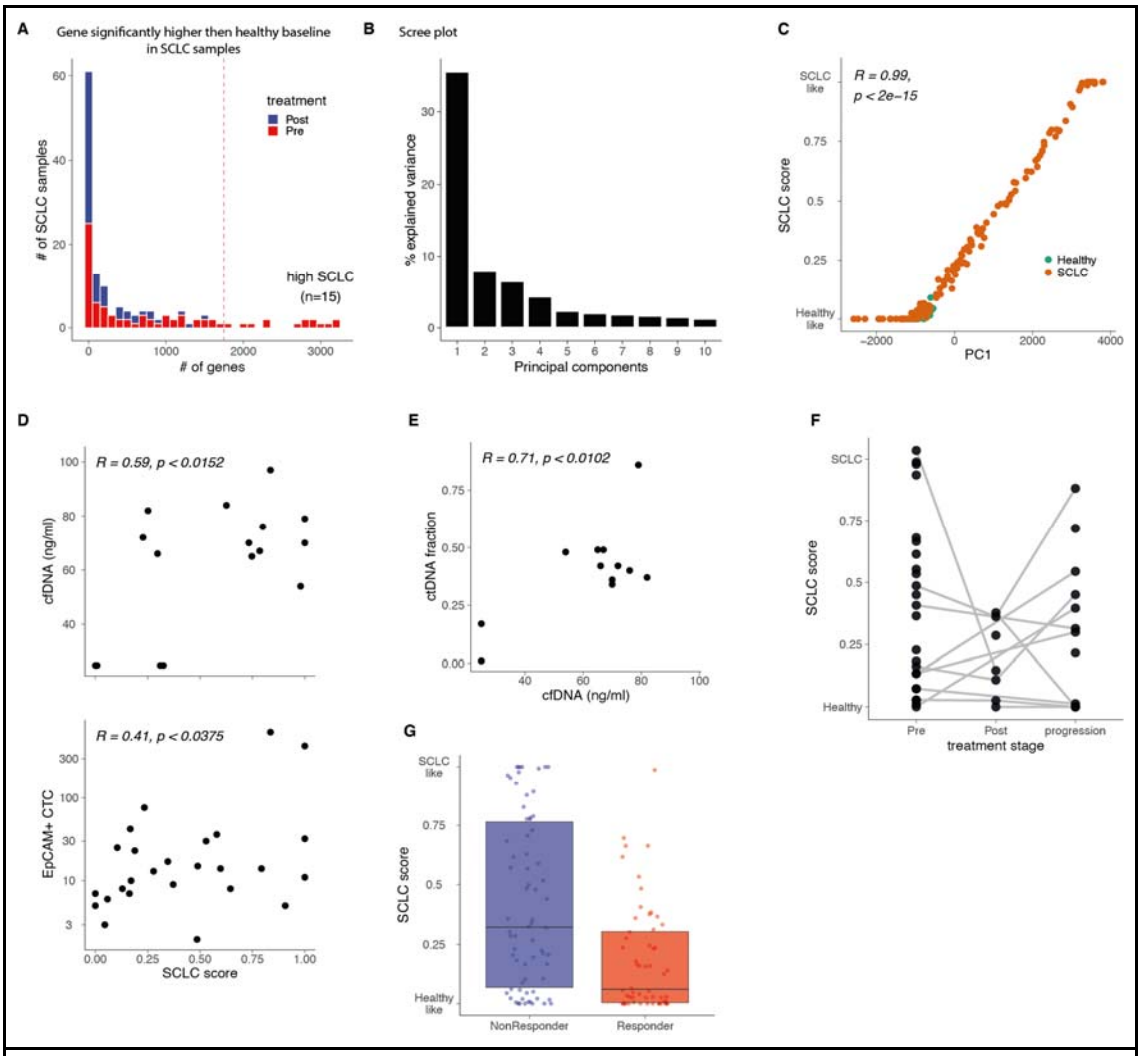

Fig. S1: SCLC-score correlates with tumor load and predicts response to treatment

A. Distribution of the number of genes with significantly higher coverage in SCLC samples compared to healthy baseline (Methods). Dashed line represents the 90th percentile.
B. Scree plot of PCA (fig. 1D).
C. Correlation of PC1 and SCLC-score (presented in fig 1C-D). High correlation indicates that the main variation in the data is driven by the SCLC contribution to the cfDNA.
D. Correlation of SCLC-score and other plasma-based methods for tumor load estimations (cfDNA concentration, circulating tumor cells).
E. Correlation of cfDNA concentration and ctDNA fraction.
F. Dynamics of SCLC-score during treatments of topotecan and berzosertib (an ATR inhibitor) (ClinicalTrial.gov identifier NCT02487095). Gray lines indicate multiple timepoints of the same individual.
G. Median SCLC-score in patients that did or did not respond to various investigational treatments: combination of topotecan and berzosertib (NCT02487095, NCT03896503); olaparib and durvalumab (NCT02484404); nanoparticle camptothecin (CRLX101) and olaparib (NCT 02769962); M7824 (PD-1 inhibitor and TGF-B trapping) and topotecan or temozolomide (NCT03554473). Each dot is a sample, and boxplots summarize distribution of values in each group. Abbreviations: ATR: ataxia telangiectasia and Rad3-related.

cfChIP-seq recovers SCLC tissue and cellular origins

We next explored whether plasma cfChIP-seq signals might profile the epigenetic state of tissue and cell of origin of SCLC tumors. We found that many of the more than 3500 genes with significantly elevated coverage in SCLC samples (**Fig. 1B**, methods) were elevated specifically in the SCLC samples compared with healthy controls or other cancers (**Fig. 2A-B**). SCLC-signature genes were highly enriched for genes expressed in SCLC cell lines[29] and pulmonary neuroendocrine cells[30] (gene overlap of 277/465 and 53/92; $p < 4.710^{-97}$ and $< 3.510^{-15}$ respectively[31,32]). Specifically, SCLC samples displayed high counts of canonical genes such as *DLL3, INSM1, CHGA* and *CRMP1* - significantly higher than in healthy samples (**Fig. 2C-D)**.

In order to characterize the tissue-specific contributions to cfDNA, we defined a set of tissue-specific genomic loci from tissue ChIP-seq reference data[33,34](Methods). Using this approach, we verified that signals in healthy samples are mainly derived from neutrophils, megakaryocytes, and monocytes [22,35]. In contrast, in SCLC samples we identified signals from lung, brain, and B-cells, suggesting contributions from cells of pulmonary, neuronal, and lymphocytes lineages (**Fig. 2E and S2A**). As expected, ChIP-seq signatures for lung, brain, and B-cell were positively correlated with the SCLC-score (**Fig. S2B**).

To further test whether cfChIP-seq can provide clues to the cell-of-origin, we used pulmonary cell-type-specific signatures derived from lung scRNA-seq[36]. Across the ~50 cell-types examined, spanning lung epithelial, endothelial, stromal, and immune cells, we found remarkable enrichment of the neuroendocrine cell type in SCLC samples compared to healthy samples (**Fig. 2F**). This result was especially striking since neuroendocrine cells constitute only 0.13% of a healthy lung tissue[36]. Neuroendocrine cell signature was significantly elevated in SCLC samples compared with ChIP-seq of normal lung (**Fig. 2G**). Signals of other lung cells including ciliated cells and alveolar epithelial type 1 cells were also higher in SCLC samples compared to healthy donors (**Fig. 2F, S2C and S2D**), hinting at the possibility of SCLCs arising from non-neuroendocrine cells-of-origin as previously described[37] or injury to the specific cells.

Together, these findings demonstrate that cfChIP-seq recovers the unique epigenetic states of tissue and cell of origin of SCLC tumors.

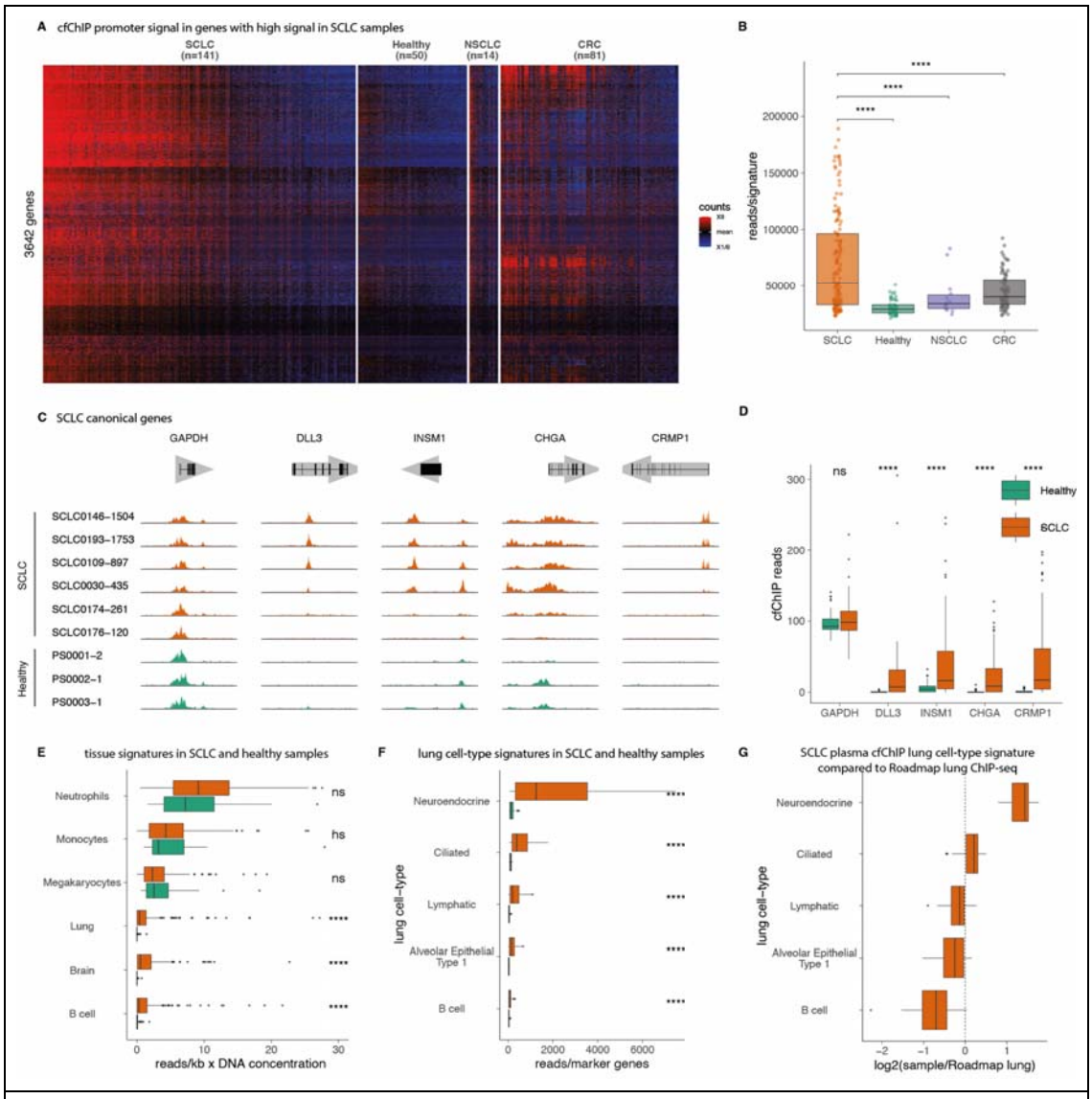

Fig. 2: cfChIP-seq recovers SCLC tissue and cellular origins

A. Heatmap showing patterns of the relative cfChIP-seq coverage of SCLC signature genes. The normalized coverage on the gene promoter was log-transformed (log2(1+coverage)) and adjusted to zero mean for each gene across the samples.
B. Median and distribution of SCLC and healthy sample coverage on genes shown in A. ****: P < 0.0001
C. Genome browser view of cfChIP-seq signal in canonical SCLC genes (*DLL3*, *INSM1, CHGA, SYP*) and *GAPDH* as control. Orange and green tracks represent SCLC and healthy samples respectively.
D. Median and distribution of the cumulative cfChIP-seq coverage over the SCLC-signature genes.
E. Cell and tissue of origin signatures in healthy and SCLC samples. x-axis values indicate absolute contribution of signature (normalized reads/kb corrected by estimated cfDNA concentration; methods) Neutrophils, monocytes and megakaryocytes are observed in both groups, while lung, brain and B-cells are observed only in SCLC samples.
F. Single cell RNA-seq-derived lung cell-type signatures in healthy and SCLC samples. Values (x-axis) indicate the sum of normalized reads in the marker genes of every tissue. An increased signal is observed in many of these cell-types in SCLC samples only.
G. Ratio for signal in cell-types shown in F of high-score SCLC samples (n=31) compared to Roadmap lung ChIP-seq. A significant elevation is observed in neuroendocrine and ciliated cells.

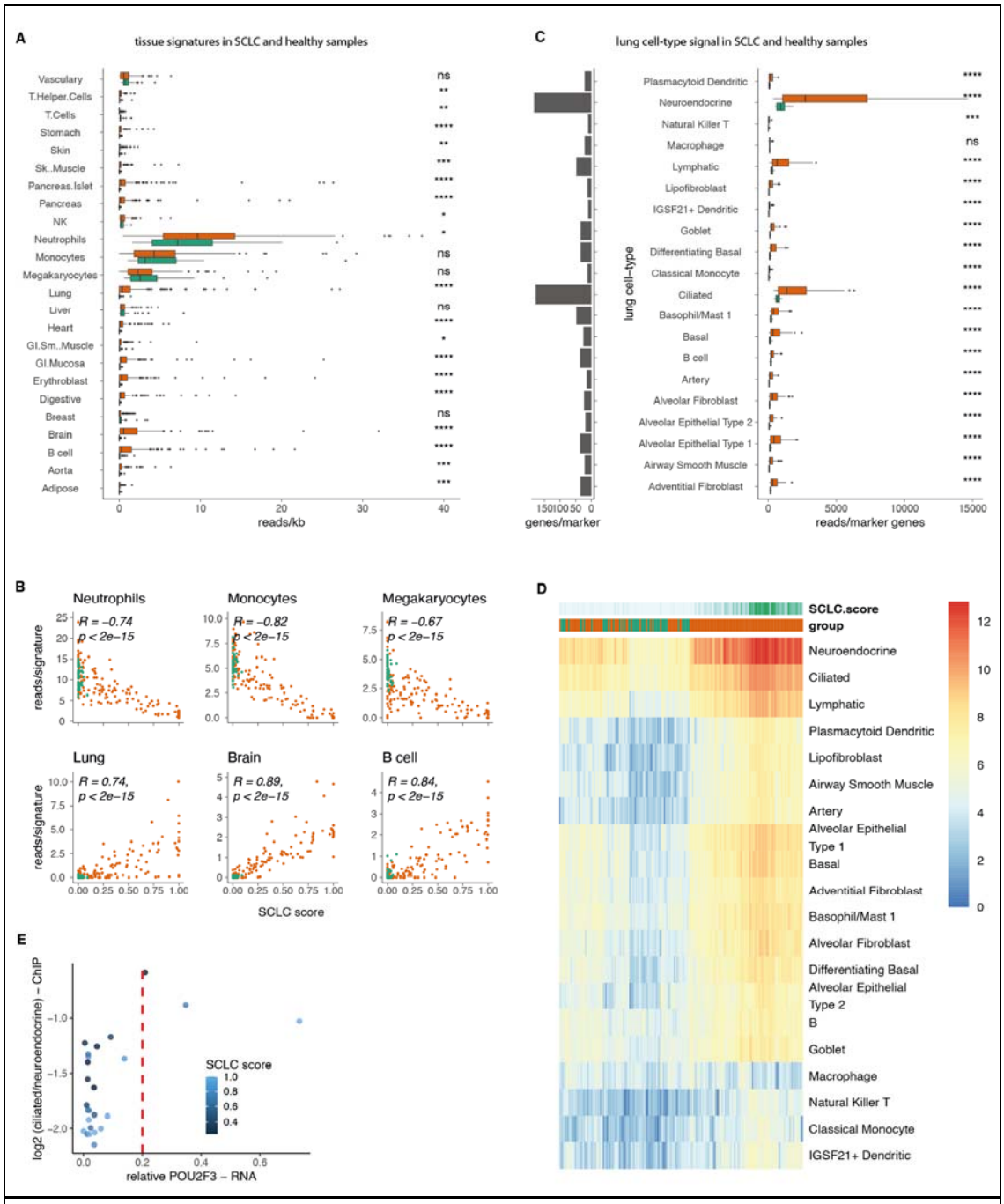

Fig. S2: SCLC tissue and cell of origin

A. Distribution of signal for cell-type signatures (same as fig. 2E).
B. Correlation of cell-type signature (shown in fig 2E) and SCLC-score. Cell-types observed only in SCLC samples are positively correlated with SCLC-score, while cell-types observed in healthy samples are negatively correlated to SCLC-score.
C. Left: number of marker genes used for every cell-type (Methods; **Supplementary Table 4**). Right: Distribution of signal for lung cell-type marker genes (same as fig. 2F). **** and ***: P < 0.0001 and < 0.001, respectively
D. Heatmap showing patterns of signal in lung cell-type markers in the SCLC and healthy samples. Color represents log2(1+cumulative coverage in promoters of marker genes).

Plasma cfChIP-seq informs tumor gene expression

Recent studies reveal that SCLC tumors are transcriptionally heterogeneous[18,38,39]. We sought to understand whether cfChIP-seq reflects gene expression patterns of SCLC tumors.

As an example, among the high SCLC-score samples, one sample (SCLC030.435) in particular exhibited markedly higher signal at several genes (e.g., *POU2F3* and *BCL2*) compared to the other high SCLC-score samples **(Fig. 3B)**. Examining time-point matched tumor RNA-seq of these samples revealed differential expression of many of the same genes, indicating that differential cfChIP-seq signals indeed reflect tumor gene expression (**Fig. 3B and S3A**).

To examine the relationship between circulating chromatin state and tumor gene expression, we computed for each gene the correlation between cfChIP-seq counts and RNA-seq TMM-normalized FPKM values in samples with available paired plasma and tumor (n=53). Excluding genes with low dynamic range in cfChIP-seq or RNA-seq (Methods), a significant positive correlation was found in more than 25% of the genes (623 genes with $q < 0.05$; Pearson $0.33 < r < 0.93$). cfDNA tumor fraction may confound this correlation, with high tumor fraction plasma samples expected to have a better correlation with tumor gene expression. Indeed, when comparing only the high SCLC-score samples with matched tumor (n=22), a similar number of significant genes was observed (527 genes with $q < 0.05$, 416 of them overlap with the significant genes in the full dataset) at higher correlation (Pearson $0.55 < r < 0.98$. **Fig. 3C and S3B**) despite the reduced statistical power. In particular, a high positive correlation was observed between cfChIP-seq and RNA-seq read counts for several important SCLC oncogenes such as *BCL2, NFIB, SOX2* and *FOXA2* (**Fig. 3D**).

To better understand the concordance between tumor RNA-seq and cfChIP-seq, we examined reasons why some genes, and not others show a positive correlation. A basic difference between chromatin and transcription is that, at the single cell level, an active chromatin mark is essentially binary (promoter marked or not marked), while RNA levels have a large dynamic range. Indeed, genes with cfChIP-seq high dynamic range tend to have higher correlation (**Fig. S3C**). Another possible source of divergence between the cfChIP-seq and tumor RNA-seq is the interference of the non-tumorous tissues contributing to each measurement (**Fig 3A** and **Fig S3D-F**, **Supplementary Table 5**).

Overall, these findings demonstrate that cfDNA chromatin state assessed by cfChIP-seq informs the tumor gene expression programs, especially in plasma samples with high

concentration of tumor, and reveal important variables that affect the concordance between cfDNA chromatin state and tumor gene expression.

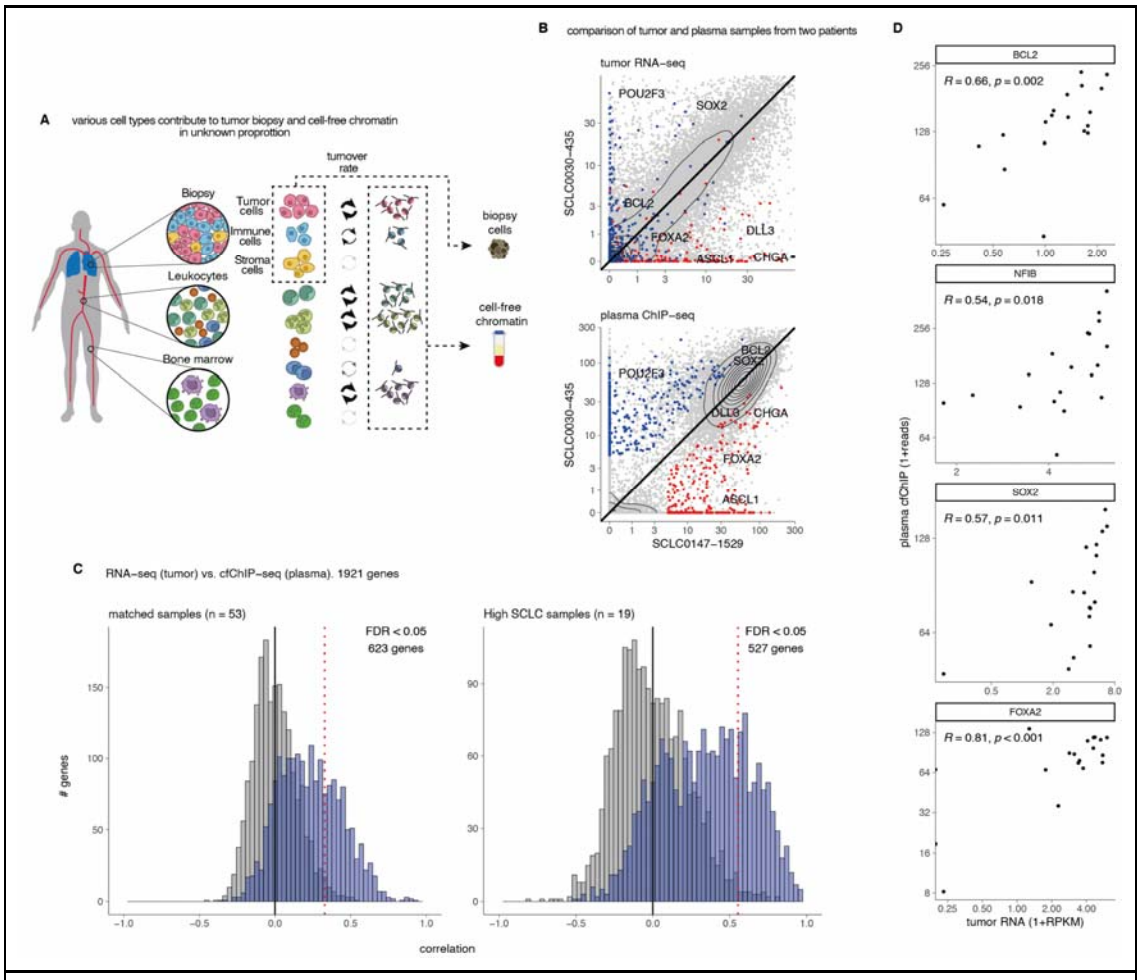

Fig. 3: Plasma cfChIP-seq informs tumor gene expression

A. Concept figure illustrating the various sources and proportions of cells sampled in plasma cfChIP-seq and tumor RNA-seq. The data obtained from these two assays is derived from a different unknown mixture of multiple types of cells.
B. Comparison of gene expression in two high SCLC-score samples. Top: TMM-normalized FPKM of tumor RNA-seq. Bottom: normalized gene counts of cfChIP-seq. Blue and red points indicate genes that had high cfChIP-seq gene counts in one sample compared to the other and low gene counts in healthy reference cfChIP.
C. Gene level analysis of the correlation between tumor gene expression and plasma cfChIP-seq coverage across individuals with matched tumor and plasma samples (left: all SCLC samples. right: high SCLC-score samples). For each gene we computed the Pearson correlation of its tumor expression and the normalized cfChIP-seq coverage across the samples. Shown is a histogram of the correlations on genes with high dynamic ranges (Methods). In gray is the histogram of a random permutation of the relation between tumor expression and plasma cfChIP.
D. Examples of correlation for several known SCLC oncogenes.

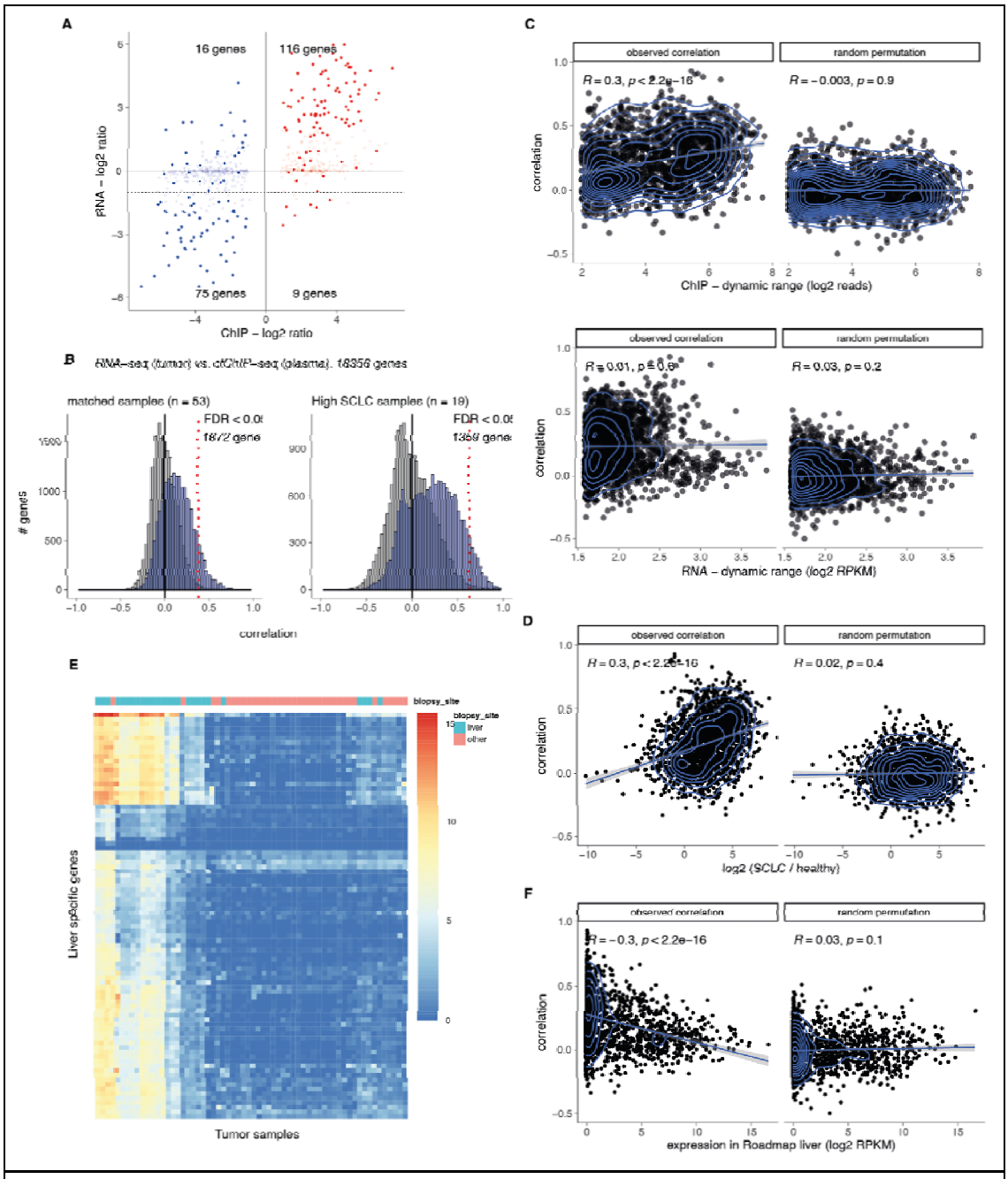

Fig S3: cfChIP-seq RNA correlation confounders

A. Comparison of signal ratio of the two individuals shown in Fig. 3A in tumor RNA and plasma cfChIP. Blue and Red points correspond to genes with higher cfChIP-seq counts in every sample. Lighter points represent genes with low RNA counts (combined TMM-FPKM< 3).
B. Gene level analysis of the correlation in tumor expression level and plasma cfChIP-seq levels (same as in Fig. 3B for all Refseq genes).
C. Comparison of the effect of cfChIP-seq and RNA dynamic range (x-axis) on cfChIP-RNA correlation (y-axis) shown in Fig. 3B. Left panel for the observed correlation and right panel for a random permutation (grey histogram of Fig. 3B). Results indicate that genes with low ChIP dynamic range tend to have lower correlation.
D. Same as C for SCLC/healthy ratio (x-axis). The ratio was computed using the mean gene counts of high SCLC samples and healthy samples.
E. log2(TMM-FPKM) of tumor samples in liver specific genes. High expression is observed in samples where biopsy was obtained from the liver.
F. Same as C for gene expression in the liver (x-axis).

Plasma cfChIP-seq predicts tumor gene expression of SCLC lineage-defining transcription factors

Given the concordance between tumor gene expression and plasma chromatin state, we sought to examine the possibility of subtyping SCLC tumors directly from plasma using cfChIP-seq.

Transcriptional signatures of SCLC heterogeneity converge on two major cell states, namely neuroendocrine (NE) and non-neuroendocrine (non-NE)[40,41]. Using a previously validated signature[40], we estimated neuroendocrine gene expression score in plasma and tumor, and found a significant positive correlation between scores derived from cfChIP-seq and tumor RNA-seq (**Fig. 4A**; Pearson r 0.76 and $p < 0.001$).

The SCLC neuroendocrine cell states are further characterized by expression of key lineage-defining transcription factors, *ASCL1* and *NEUROD1* defining NE cell states and *POU2F3* defining non-NE cell states. A fourth subgroup has been characterized by expression of YAP1[42,43] or low expression of all three transcription factors accompanied by an inflamed gene signature[16]. Most tumors in our cohort had high expression of NE-lineage defining genes ASCL1 and NEUROD1, with co-expression of both genes in some cases (**Fig. 4B**). Expression of POU2F3, YAP1, and a newly described subtype ATOH1[14,44] was seen in ~7% of tumors and were largely mutually exclusive.

We sought to examine whether the expression of SCLC lineage-defining transcription factors in the tumor is reflected in the cfDNA. We found significantly elevated counts of *ASCL1*, *NEUROD1*, and *POU2F3* in SCLC samples in contrast to barely detectable levels in healthy plasma. In contrast, *YAP1* counts were similar among healthy and SCLC plasma samples (**Fig. 3C** and **S4A**), likely due to *YAP1* activity in normal tissue contributing to cfDNA, as suggested by H3K4me3 of *YAP1* promoter regions in normal tissue[33].

To evaluate whether these patterns are indicative of tumor gene expression in individual patients, we assessed the concordance between high SCLC-score plasma cfChIP-seq counts and gene expression levels derived from time-point matched tumor transcriptome (n=18). A strong correlation between cfChIP-seq and tumor RNA-seq was observed in three of the transcription factors (*ASCL1*, *NEUROD1* and *POU2F3*) (Pearson r 0.93, 0.85, and 0.97; $p < 1*10^{-5}$,$<7*10^{-5}$, and $<1*10^{-5}$ respectively), which were also absent from healthy control cfDNA. A similarly high positive correlation was observed between plasma cfChIP-seq and tumor RNA-seq of *ATOH1* (atonal homolog 1) (Pearson r 0.89; $p < 1*10^{-5}$; **Fig. 4D**).

Notably, the TSS of *POU2F3* and *ATOH1* is marked by H3K4me3 in many of the SCLC samples, however, only in a small subset of them do the signals span beyond the TSS region to a wide region of approximately 10KB. We find these additional regions especially informative as they correlate best with gene expression in the tumor (**Fig. S4B-C**).

Revisiting the cell-type source of the cfChIP-seq samples, we noticed that the three samples with high relative *POU2F3* expression in the respective tumors, had high ciliated to neuroendocrine cell ratio (**Fig. S4D**). This is in accordance with their classification as non-NE tumors and *POU2F3* is normally selectively expressed in tuft cells, a rare chemosensory cell type in the pulmonary epithelium[45,46].

These findings suggest that cfChIP-seq can fill an unmet need of molecularly classifying SCLC into transcriptomic subsets of potential therapeutic relevance in a limited-invasive manner directly from plasma.

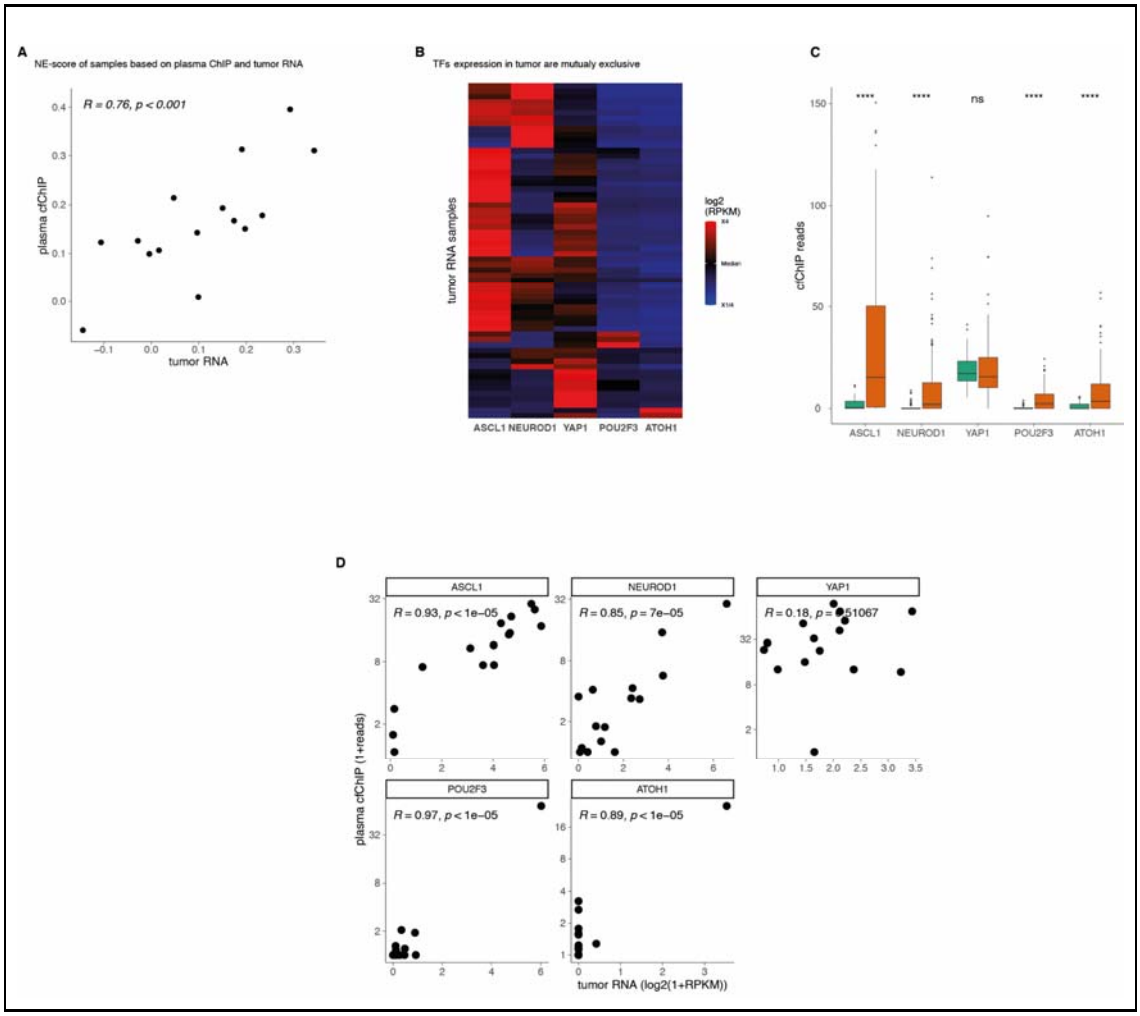

Fig. 4: cfChIP-seq displays differential expression of SCLC transcription drivers and can be used for tumor classification

A. Correlation of NE-score computed based on plasma cfChIP-seq and tumor RNA-seq.
B. Heatmap showing relative expression levels of 4 canonical SCLC transcription drivers. Values are presented in log2(1+TMM normalized-FPKM) and adjusted to zero mean for each sample across the 5 genes.
C. Median and distribution of SCLC and healthy cfChIP-seq plasma sample coverage on the gene shown in B with the addition of *ATOH1*. ***** P < 0.0001
D. Correlation of plasma cfChIP-seq coverage and tumor RNA-seq on the genes shown in C.

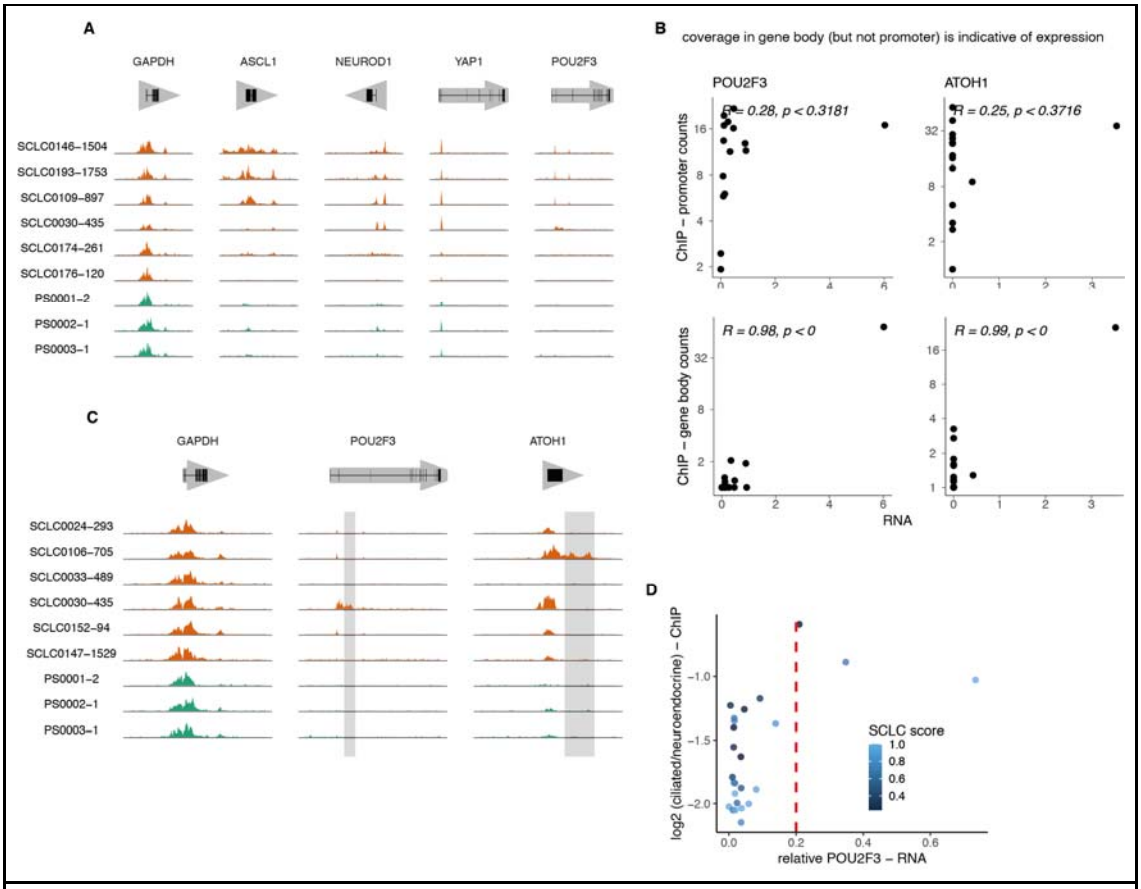

Fig. S4: chromatin specific TF signatures

A. Genome browser view of cfChIP-seq signal in SCLC key transcriptional regulators.
B. Correlation of plasma cfChIP-seq coverage in promoter and gene body and tumor RNA-seq of the genes *POU2F3* and *ATOH1*. Noticeably, in these genes, the correlation to expression is higher in the gene body.
C. Genome browser view of cfChIP-seq signal in panel D. Orange and green tracks represent SCLC and healthy samples respectively. While a signal in the gene promoter is observed in many samples, only in specific samples there is a significant coverage also in the gene body.
D. Relation of the ratio of signal in markers of ciliated and neuroendocrine plasma ChIP samples (y-axis) and relative RNA expression of *POU2F3* in the tumor biopsy (x-axis). Relative RNA *POU2F3* was calculated by log2(*POU2F3*)/sum(log2(ASCL1,NEUROD1, YAP1, POU2F3)) where all RNA values are TMM-FPKM normalized.

Discussion

Cancer master transcription factors control global gene expression programs[14,47–49], and represent attractive targets to classify and treat cancer given their essential role in driving distinctive cell identities[50,51] and the tendency for cancer cells to become highly dependent on their sustained and high-level expression[52]. However, transcription factor profiling in clinical tumor samples is challenging, particularly for cancers with limited access to tumor specimens. Recent studies have revealed subtypes of SCLC, the most lethal type of lung cancer, defined by expression of lineage-defining transcription factors and their unique therapeutic vulnerabilities[14–18]. Yet, further testing and broad clinical application of these findings is limited by the availability of tumor tissue, explained by its widely metastatic presentation and aggressive clinical course which precludes surgical resection at diagnosis and tumor biopsies at relapse. As such, all patients with SCLC receive the same treatments. Here, we apply plasma cfChIP-seq which reports the promoter state of cell-free chromatin in circulation[21] to SCLC samples. We find that cfChIP-seq recovers the unique epigenetic states of SCLC tissue and cell of origin, and importantly tumor gene expression, particularly SCLC lineage-defining transcription factors, providing a systematic view of tumor state, opening the possibility of molecularly classifying SCLC directly from 1ml of plasma.

SCLC has a distinct cell-free chromatin signature, which can be detected in plasma of patients using cfChIP-seq and used to differentiate SCLC from other cancers and healthy controls. The SCLC signature can be detected using this approach even when it has low representation in the plasma (e.g., after therapy), and is highly correlated with serologic and radiological estimations of tumor burden and prognosis. In matched plasma and tumor biopsy samples, we show for the first time, the concordance of gene expression inferred from plasma cell-free chromatin and tumor transcriptome at the level of the individual patient. Importantly, cfChIP-seq profiles identify activity of key SCLC transcriptional drivers, including *ASCL1* and *NEUROD1* that drive NE phenotypes and *POU2F3* that drives non-NE phenotype[16,38,39,41]. These results raise the possibility of molecular profile-based treatments for patients with SCLC, which might be more effective than the current one-size fits all approach. Firmly establishing the clinical performance of cfChIP-seq will require a subsequent study involving more patients and prospective independent validation.

This study also provides a broadly applicable framework for benchmarking features of cfDNA against the tumor molecular profile. While plasma cfChIP-seq shows good agreement with tumor RNA-seq, the correspondence is imperfect due to multiple factors. First, there are inherent differences between chromatin state, which is to a large extent on or off, and gene expression, which has a large dynamic range[24,57,58]. In addition, several factors confound our estimations of both the plasma and tissue compartments. Plasma cell-free chromatin reflects contributions from multiple sources, which also include the tumors. Tumor RNA-seq contains contributions from multiple sources in addition to the tumor including tissue-infiltrating immune cells, stromal cells, endothelial cells, and more. Thus, to understand the plasma-tumor correspondence, we need to account for the different cell-types contributing to each compartment. By explicitly accounting for differences in SCLC-score we could extract tumor-specific and tumor-extrinsic features even when the tumor contribution is low. Ideally, we would deconvolve the signal from these different sources and extract from plasma samples information about the state of multiple cell types (e.g., tumors, immune cells, stroma cells). Establishing reliable deconvolution requires better references of the ChIP-seq profiles of tumor cells and biopsies, together with baseline estimates of their fractions in each sample[35].

Overall, our study opens a unique opportunity to bridge the gap between molecular studies and patient treatment in SCLC, an exceptionally lethal malignancy, which to date is treated as a homogenous disease with identical treatments for all patients. Moreover, this work suggests the applicability of cfChIP-seq to a wider context to profile and subtype tumors, in a way that can be transformative for patient care.

Materials and Methods

Patients

We undertook an observational study using plasma collected from patients with small cell cancer who received care at the National Cancer Institute (NCI). Patients were enrolled in therapeutic clinical trials (ClinicalTrials.gov #NCT02484404; NCI protocol #15-C-0145; ClinicalTrials.gov #NCT02487095; NCI protocol #15-C-0150; ClinicalTrials.gov #NCT02769962; NCI protocol #16-C-0107; ClinicalTrials.gov # NCT03554473; NCI protocol #18-C-0110; and ClinicalTrials.gov # NCT03896503; NCI protocol #20-C-0009). We also collected samples from small cell cancer patients who were enrolled in the NCI thoracic malignancies natural history protocol (ClinicalTrials.gov # NCT02146170; NCI protocol #14-C-0105). If tumor samples were available, we also sequenced their RNA at matched or different time points of when the plasma was collected. Human subjects committee at NCI approved the studies; all patients provided written informed consent for plasma, tumor, and matched normal sample sequencing.

Tumor RNA sequencing

Formalin-Fixed, Paraffin-Embedded (FFPE) tumor tissue samples or frozen tumor samples in selected samples were prepared for RNA-seq. RNA enrichment was performed using TruSeq RNA Exome Library Prep according to manufacturer's instructions (Illumina, San Diego). Paired-end sequencing (2 x 75 bp) was performed on an Illumina NextSeq500 instrument. RNA was extracted from FFPE tumor cores using RNeasy FFPE kits according to the manufacturer's protocol (QIAGEN, Germantown, MD). RNA-seq libraries were generated using TruSeq RNA Access Library Prep Kits (TruSeq RNA Exome kits; Illumina) and sequenced on NextSeq500 sequencers using 75bp paired-end sequencing method (Illumina, San Diego, CA). For transcriptomic analyses, raw RNA-Seq count data were normalized for inter-gene/sample comparison using TMM-FPKM, followed by log2(x+1) transformation, as implemented in the edgeR R/Bioconductor package[61].

Tumor volumetric estimation

cfDNA Tfx

The DSP Circulating DNA kit from Qiagen was utilized to extract cell-free DNA from aliquots of plasma which were eluted into 40-80 uL of re-suspension buffer using the Qiagen

Circulating DNA kit on the QIAsymphony liquid handling system. Library preparation utilized the Kapa Hyper Prep kit with custom adapters (IDT and Broad Institute). Samples were sequenced to meet a goal of 0.1x mean coverage utilizing the Laboratory Picard bioinformatics pipeline, and Illumina instruments were used for all of cfDNA sequencing with 150 bp and paired-end sequencing. Library construction was performed as previously described[62]. Kapa HyperPrep reagents in 96-reaction kit format were used for end repair/A-tailing, adapter ligation, and library enrichment polymerase chain reaction (PCR). After library construction, hybridization and capture were performed using the relevant components of Illumina's Nextera Exome Kit and following the manufacturer's suggested protocol. Cluster amplification of DNA libraries was performed according to the manufacturer's protocol (Illumina) using exclusion amplification chemistry and flowcells. Flowcells were sequenced utilizing Sequencing-by-Synthesis chemistry. Each pool of whole genome libraries was sequenced on paired 76 cycle runs with two 8 cycle index reads across the number of lanes needed to meet coverage for all libraries in the pool. Somatic copy number calls were identified using CNVkit[63] (version 0.9.9) with default parameters. Tumor purity and ploidy were estimated by sclust[63,64] and sequenza[65]. cfDNA Tfx was estimated based on the somatic copy number alteration profiles using *ichorCNA*, a previously validated analytical approach[27,65].

CTCs

CTCs were detected from 10 mL of peripheral blood drawn into EDTA tubes. Epithelial cell adhesion molecule (EpCAM)-positive CTCs were isolated using magnetic pre-enrichment and quantified using multiparameter flow cytometry. CTCs were identified as viable, nucleated, EpCAM+ cells that did not express the common leukocyte antigen CD45, as described previously[17].

Radiological volumetric segmentation

We performed volumetric segmentation, a three-dimensional assessment of computed tomography as previously described [66]. that may more accurately predict clinical outcomes than conventional evaluation by Response Evaluation Criteria in Solid Tumors (RECIST). Briefly, experienced radiologists reviewed the computed tomography sequences to determine the best ones to use for segmentation using the lesion management application

within PACS (Vue PACS v 12.0, Carestream Health, Rochester, NY). We also assessed sizes of lesions by RECIST v.1.1.

Plasma cfChIP-seq

Sample collection

Plasma for cfDNA from EDTA. Sample should be processed within 2 hours. Centrifuge at 4C at 1500xg for 10 minutes. Transfer plasma equally into Eppendorf tubes being careful not to disturb the leukocyte layer. Centrifuge a second time at 4C at 10,000xg for 10 minutes. Without disturbing pellet, transfer plasma into standard cryovials. Barcode as plasma and enter sample note: ctDNA. Store at -80C

Bead preparation

50µg of antibody was conjugated to 5mg of epoxy M270 Dynabeads (Invitrogen) according to manufacturer instructions. The antibody-beads complexes were kept at 4ºC in PBS, 0.02% azide solution.

Immunoprecipitation, NGS library preparation, and sequencing

Immunoprecipitation, library preparation and sequencing were performed by Senseera LTD. as previously reported[22], with certain modifications that increase capture and signal to background ratio. Briefly, ChIP antibodies were covalently immobilized to paramagnetic beads and incubated with plasma. Barcoded sequencing adaptors were ligated to chromatin fragments and DNA was isolated and next-generation sequenced.

Sequencing Analysis

Reads were aligned to the human genome (hg19) using bowtie2 (2.3.4.3) with 'no-mixed' and 'no-discordant' flags. We discarded fragments reads with low alignment scores (-q 2) and duplicate fragments. See **Supplementary Table 1** for read number, alignment statistics, and numbers of unique fragments for each sample.

Preprocessing of sequencing data was performed as previously described. Briefly, the human genome was segmented into windows representing TSS, flanking to TSS, and background (rest of the windows). The fragments covering each of these regions were quantified and used for further analysis. Non-specific fragments were estimated per sample and extracted resulting in the specific signal in every window. Counts were normalized and scaled to 1 million reads in healthy reference accounting for sequencing depth differences. Detailed information regarding these steps can be found at supplementary note in [22]

Statistical analysis

SCLC score

To estimate SCLC-score reflecting tumor-related fraction in samples, we performed a leave-one-out non-negative least square using the 'nnls' R package (1.4).

Denoting by G the number of gene promoter, and by S the number of samples in the reference cohort, a matrix X_{GXS+1} is composed containing the counts per promoters in the S healthy samples, with the addition of a 'SCLC prototype' composed on the mean promoter counts of the 10% SCLC samples with most genes significantly higher than healthy (n=15). For every vector Y_{1XG} of counts per promoter in the SCLC samples, we estimate the non-negative coefficient □ by computing $argmax_{\beta}\ ||X\beta - Y||^2$ subject to □ ≥ 0. A similar process is computed for healthy samples with the exception that for the i_{th}sample we use the matrix X_{-i} where the column containing that sample is eliminated. The estimated □ is normalized to 1 by computing □ $= \frac{\beta}{\sum_i \quad \beta_i}$. The tumor-related fraction is defined as the value of □$_{SCLC}$ - the fraction assigned to the SCLC prototype. The healthy fraction of the i_{th} sample is defined as $\sum_{i \neq SCLC} \quad \beta_i$ - the sum of fractions assigned to all other healthy components.

An alternative approach was tested using a reference atlas of roadmap tissue data. In this approach the healthy fraction is set to be the sum of fractions assigned to common tissues observed typically in healthy samples (neutrophils, monocytes, megakaryocytes) and the tumor-related fraction is defined as 1 - the tumor related fraction. The two approaches resulted in very similar estimations for the vast majority of samples (not shown).

We compute the principle component analysis (PCA) of all gene counts in all SCLC and healthy samples as implemented in the R 'prcomp' function. Scree plot was generated using the 'fviz_eig' function of the R 'factoextra' package (1.0.7).

Tissue signatures

Genomic regions selected for tissue specific markers are as described in [22]. cfDNA of SCLC patients consists of tumor related cfDNA above the hematopoietic derived cfDNA observed in healthy individuals. Estimation of the absolute contribution of tissues to the cfDNA pool was calculated by multiplying the normalized signal per signature and the estimated number of reads in the sequencing library size [22] which approximates the cfDNA concentration.

Lung single cell signatures

For the purpose of defining pulmonary cell-type specific signatures, we made use of published pulmonary scRNA-seq data [36]. In this publication, the authors defined cluster-specific marker gene sets, i.e. genes that are enriched for a specific cluster. To increase the specificity of cell-type signature in the cfDNA context, we include only genes that meet the following criteria:

a. The percent of cells outside the cluster that express the genes is less than 0.1

b. The average log fold-change expression in the cluster compared to other clusters is more then 2

c. The adjusted p-value of the gene is less than 0.1

d. The gene promoter read counts in cfDNA of healthy samples is less than 3

The last criteria is aimed to remove from the signature genes that are not lung-specific rather appear in circulation as part of normal cell turnover. A summary of the genes included in every lung cell-type signature can be found in **Supplementary Table 4**.

ChIP-RNA correlation

For examination of the tumor RNA-seq of tumor and plasma ChIP-seq relationship, we computed the Pearson correlation of all genes across the samples with matching tumor and plasma samples. Correlation was computed between 1+TSS counts of the ChIP-seq and log2(1+FPKM) and compared to the correlation achieved from a random permutation of the ChIP-seq samples. The correlation was computed for all samples with matching tumor-plasma (n=54) and separately for samples with SCLC-score > 0.7 (n=19) and matching tumor.

Dynamic range of ChIP-seq gene *g* was calculated by the 95th percentile - 5th percentile of the log2(1+normalized reads) across all plasma samples. High dynamic range genes were set to be genes with a dynamic range > 2.

The dynamic range of the RNA-seq of gene *g* was calculated similarly for log2(1+FPKM). High dynamic range genes were defined as genes with a 3 fold-change between high and low samples.

Acknowledgments

We would like to thank Alon Chappleboim and members of the Friedman lab for comments and suggestions. We would kindly like to thank Ahmad Shafiei and Mohammad Bagheri from Department of Radiology and Imaging Sciences, Center for Cancer Research, National Cancer Institute, Bethesda, USA for the volumetric measurements.

This work was supported in part by the European Research Council (ERC Adg #101019560 “cfChIP” awarded to N.F.).

Supplementary information

Supplementary Table 1 - SCLC Individuals and samples clinical information

Supplementary Table 2 - Sequencing statistics for samples sequenced in this study

Supplementary Table 3 - SCLC differential genes

Supplementary Table 4 - Lung cell-type marker genes used in figure 2F-G and S2C

Supplementary Table 5 - Liver specific genes used in figure S3E

References

1. Rudin, C. M., Brambilla, E., Faivre-Finn, C. & Sage, J. Small-cell lung cancer. *Nat Rev Dis Primers* **7**, 3 (2021).
2. Almodovar, K. *et al.* Longitudinal Cell-Free DNA Analysis in Patients with Small Cell Lung Cancer Reveals Dynamic Insights into Treatment Efficacy and Disease Relapse. *J. Thorac. Oncol.* **13**, 112–123 (2018).
3. Nong, J. *et al.* Circulating tumor DNA analysis depicts subclonal architecture and genomic evolution of small cell lung cancer. *Nat. Commun.* **9**, 3114 (2018).
4. Mohan, S. *et al.* Profiling of Circulating Free DNA Using Targeted and Genome-wide Sequencing in Patients with SCLC. *J. Thorac. Oncol.* **15**, 216–230 (2020).
5. Mondelo-Macía, P. *et al.* Current Status and Future Perspectives of Liquid Biopsy in Small Cell Lung Cancer. *Biomedicines* **9**, (2021).
6. Fernandez-Cuesta, L. *et al.* Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. *EBioMedicine* **10**, 117–123 (2016).
7. Devarakonda, S. *et al.* Circulating Tumor DNA Profiling in Small-Cell Lung Cancer Identifies Potentially Targetable Alterations. *Clin. Cancer Res.* **25**, 6119–6126 (2019).
8. Board, R. E. *et al.* Isolation and Extraction of Circulating Tumor DNA from Patients with Small Cell Lung Cancer. *Annals of the New York Academy of Sciences* vol. 1137 98–107 (2008).
9. Du, M. *et al.* Genomic alterations of plasma cell-free DNAs in small cell lung cancer and their clinical relevance. *Lung Cancer* **120**, 113–121 (2018).
10. Iams, W. T. *et al.* Longitudinal monitoring of circulating tumor DNA and peripheral T cell repertoire in patients with small cell lung cancer. *Journal of Clinical Oncology* vol. 35 8571–8571 (2017).
11. Hou, J.-M. *et al.* Clinical significance and molecular characteristics of circulating tumor cells and circulating tumor microemboli in patients with small-cell lung cancer. *J. Clin. Oncol.* **30**, 525–532 (2012).

12. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47–53 (2015).
13. Kim, D.-W., Kim, K.-C., Kim, K.-B., Dunn, C. T. & Park, K.-S. Transcriptional deregulation underlying the pathogenesis of small cell lung cancer. *Transl Lung Cancer Res* **7**, 4–20 (2018).
14. Rudin, C. M. *et al.* Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews Cancer* vol. 19 289–297 (2019).
15. Roper, N. *et al.* Notch signaling and efficacy of PD-1/PD-L1 blockade in relapsed small cell lung cancer. *Nature Communications* vol. 12 (2021).
16. Gay, C. M. *et al.* Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. *Cancer Cell* vol. 39 346–360.e7 (2021).
17. Thomas, A. *et al.* Therapeutic targeting of ATR yields durable regressions in small cell lung cancers with high replication stress. *Cancer Cell* **39**, 566–579.e7 (2021).
18. Lissa, D. *et al.* Heterogeneity of neuroendocrine transcriptional states in metastatic small cell lung cancers and patient-derived models. *Nat. Commun.* **13**, 2023 (2022).
19. Thomas, A., Pattanayak, P., Szabo, E. & Pinsky, P. Characteristics and Outcomes of Small Cell Lung Cancer Detected by CT Screening. *Chest* **154**, 1284–1290 (2018).
20. Corcoran, R. B. Circulating Tumor DNA: Clinical Monitoring and Early Detection. *Annual Review of Cancer Biology* vol. 3 187–201 (2019).
21. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences* vol. 112 E5503–E5512 (2015).
22. Sadeh, R. *et al.* ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nat. Biotechnol.* **39**, 586–598 (2021).
23. Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2926–2931 (2010).

24. Weiner, A. *et al.* High-resolution chromatin dynamics during a yeast stress response. *Mol. Cell* **58**, 371–386 (2015).
25. Liu, C. L. *et al.* Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol.* **3**, e328 (2005).
26. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
27. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
28. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
29. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
30. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
31. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* vol. 14 (2013).
32. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–7 (2016).
33. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
34. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1897 (2016).
35. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).
36. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
37. Ferone, G., Lee, M. C., Sage, J. & Berns, A. Cells of origin of lung cancers: lessons from mouse studies. *Genes Dev.* **34**, 1017–1032 (2020).

38. Stewart, C. A. *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat Cancer* **1**, 423–436 (2020).
39. Chan, J. M. *et al.* Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. *Cancer Cell* **39**, 1479–1496.e18 (2021).
40. Zhang, W. *et al.* Small cell lung cancer tumors and preclinical models display heterogeneity of neuroendocrine phenotypes. *Transl Lung Cancer Res* **7**, 32–49 (2018).
41. Rudin, C. M. *et al.* Author Correction: Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nat. Rev. Cancer* **19**, 415 (2019).
42. McColl, K. *et al.* Reciprocal expression of INSM1 and YAP1 defines subgroups in small cell lung cancer. *Oncotarget* vol. 8 73745–73756 (2017).
43. Qu, S. *et al.* Molecular Subtypes of Primary SCLC Tumors and Their Associations With Neuroendocrine and Therapeutic Markers. *Journal of Thoracic Oncology* vol. 17 141–153 (2022).
44. Simpson, K. L. *et al.* A biobank of small cell lung cancer CDX models elucidates inter- and intratumoral phenotypic heterogeneity. *Nature Cancer* vol. 1 437–451 (2020).
45. Yamashita, J., Ohmoto, M., Yamaguchi, T., Matsumoto, I. & Hirota, J. Skn-1a/Pou2f3 functions as a master regulator to generate Trpm5-expressing chemosensory cells in mice. *PLoS One* **12**, e0189340 (2017).
46. Huang, Y.-H. *et al.* POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. *Genes Dev.* **32**, 915–928 (2018).
47. Sanda, T. *et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* **22**, 209–221 (2012).
48. Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
49. Durbin, A. D. *et al.* Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry. *Nat. Genet.* **50**, 1240–1246 (2018).

50. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

51. D'Alessio, A. C. *et al.* A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. *Stem Cell Reports* **5**, 763–775 (2015).

52. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* **168**, 629–643 (2017).

53. Thomas, A. *et al.* Therapeutic targeting of ATR yields durable regressions in small cell lung cancers with high replication stress. *Cancer Cell* **39**, 566–579.e7 (2021).

54. Ulz, P. *et al.* Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat. Commun.* **10**, 4666 (2019).

55. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).

56. Esfahani, M. S. *et al.* Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* **40**, 585–597 (2022).

57. Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences* vol. 107 2926–2931 (2010).

58. Tippmann, S. C. *et al.* Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol. Syst. Biol.* **8**, 593 (2012).

59. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).

60. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).

61. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

62. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready

human exome targeted capture libraries. *Genome Biology* vol. 12 R1 (2011).

63. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
64. Cun, Y., Yang, T.-P., Achter, V., Lang, U. & Peifer, M. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nature Protocols* vol. 13 1488–1501 (2018).
65. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
66. Fenerty, K. E. *et al.* Predicting clinical outcomes in chordoma patients receiving immunotherapy: a comparison between volumetric segmentation and RECIST. *BMC Cancer* **16**, 672 (2016).