medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Transcriptomic Characterization Reveals Blood-based Molecular Signatures of NSCLC Patients in Response to Anti-PD-1 Therapy Combined with Chemotherapy

Xi Tom Zhang[1,2*#] Rui Aaron Chen[1*], Wenqing Sherry Li[1], Runpeng Harris Han[1], Guodong Gordon Su1[1], Wuhao Huang[3], Yuru Freya Liu[1], Yu Yvonne Cai[1], Yuyan Xiong[4], Shengguang Wang[3#]

1 AImmue Lab, School of Medicine, Northwest University, Xi'an, Shaanxi 710069, China

2 Data Science Program, Northwestern University, Evanston, IL 60208, US.

3 Department of Lung Cancer, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin Lung Cancer Center, Tianjin 300060, China.

4 Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, School of Medicine, Northwest University, Xi'an, Shaanxi 710069, China

* These authors contributed equally

Corresponding authors

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Background: For patients with non-small cell lung cancer (NSCLC), the PD-1/PD-L1 blockade treatment were incorporated into first-line treatment commonly. Despite the improved survival observed in PD-1 blockade treatment, a large proportion of patients do not respond while others actually progress during treatment.

Method: Transcriptomic profiling was performed on whole blood samples from 30 patients received anti-PD-1 (Tislelizumab) plus chemotherapy. Expression levels of differentially expressed genes (DEGs) identified from two comparisons (post-and pre-treatment, responders and non-responders) were validated by real-time quantitative PCR, analyzed within tissue database and meta-analysis database, then followed by enrichment analysis in high-level representations and *in silico* leukocyte deconvolution.

Results: A panel of blood-based gene signatures (FDR $p<0.05$, fold change<-2 or >2) were identified (DEG n=155 and 112 in two comparisons) and validated that not only differentially expressed between post- and pre- treatment or responders and non-responders but also in tissue samples between normal and tumor. Genes DLG5 and H3C10 were found negatively associated with overall survival ($p<0.05$). Enrichment of immunological and metabolism pathways and gene sets indicating activated circulating leukocytes were observed.

Conclusion: The molecular and cellular signatures characterized in this study may provide potential blood-based predictors of the response to PD-1 blockade treatment in NSCLC patients.

Key words: non-small cell lung cancer; anti-PD-1 therapy; responder; molecular signatures; biomarker

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Given the significant benefit checkpoint inhibitors (ICIs) immunotherapy has achieved for various types of cancer in the past decade, it is emerging as a "common denominator" and even in the neoadjuvant (presurgical) setting for cancer therapy [1]. In many countries, ICIs targeting the PD-1/PD-L1 (programmed death-1/programmed death ligand-1) pathway have been gradually approved as monotherapy or in combinations with chemotherapy [2, 3]. Non-small cell lung cancer (NSCLC) as the most common subtype of lung cancer presents a significant health concern with its diagnosis rate and burden highest for male, ever-smoker and the elderly population [4, 5]. Advanced NSCLC was one of the first pioneers becoming a common therapeutic focus of PD-1-targeted immunotherapy as a first-line therapy [6, 7]. It is proved that anti-PD-1 agents with or without chemotherapy could restore anti-tumor activities of multiple immune cell subsets, together leading to an increased overall survival compared with sole conventional chemotherapy [8, 9].

To date, several anti-PD-1 or anti-PD-L1 antibodies have been approved by regulatory authorities covering almost all the common types of cancer indications as monotherapy or in combination with other drugs [10]. Despite the impressive success of PD-1 blockade therapy in cancer, it seems the clinical benefit of this treatment is always and only limited to a small subset of patients [11]. Taking one of the approved anti-PD-1 antibodies Pembrolizumab as an example, the objective response rate for the unselected population with NSCLC was only 19% and the median overall survival was 12 months [12]. All approved PD-1/PD-L1 blockade displayed similar efficacy while combinations of anti-PD-1 therapy with chemotherapy have showed more encouraging results in the up-front treatment of NSCLC [13]. NSCLC as well as other lung cancers are characterized by high degree molecular heterogeneity associated to high level of

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

genetic and phenotypic variations, co-evolving with the complex nature of immune microenvironment [14-16]. The outcome of anti-PD-1 treatment for NSCLC patents is associated with many factors, such as PD-1/PD-L1 expression level, tumor mutation burden, mRNA expression of certain biomarker genes, neo-antigens, and the composition of tumor-infiltrating immune cells [17]. Many clinical studies as well as review articles suggested PD-L1 expression testing should become (if not yet) routine for patients with newly diagnosed NSCLC, while other biomarkers should be evaluated in the future [18-21].

The low response rate to PD-1/PD-L1 blockade treatment suggests better understanding the molecular signatures responding to the therapy is important in the search for potentially reliable biomarkers. Lately, a number of new biomarkers predictive of tumor response to ICIs were proposed, including DNA variant of NK cells [22], CD28 levels in tumor-infiltrating CD8-T cells [23], metabolite from on-treatment serum [24], local or peripheral immune cell clusters [25, 26], or combination of T cell properties and energy metabolism markers in blood [27]. In the present study, comprehensive computational analyses were performed to investigate the molecular signatures and pathway representations in peripheral blood as potential predictors of the response and outcome in NSCLC patients who received PD-1 inhibitor plus chemotherapy. As several innovative *in silico* deconvolution methods were developed for inference of tumor-infiltrating immune cells in tissue samples [28, 29], we adapted from these new tools and updated a different immune profiling approach from our previous study [30] targeting blood samples. Here, we enumerated a set of key immune cell repertoires in NSCLC patients and responders as we hypothesized that the circuiting immune cell subsets in peripheral blood may ultimately allow discrimination between patients that derive benefit from PD-1 blockade.

Result

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Study design and overall analytical workflow

In order to identify the blood-based molecular signature associated with PD-1 blockade treatments in cancer, whole blood samples were collected from 30 NSCLC patients who underwent anti-PD-1 therapy plus chemotherapy. The majority of these patients were male (86.7%), over 60-year-old (66.7%), ever-smokers (80%), and classified as lung squamous cell carcinoma (SqCC) (76.7%) (**Table 1**). Paired blood samples were collected from 12 patients to construct subgroups for post- and pre-treatment comparisons. A little less than 50% of the patients showed pathologic complete response and were considered as responders in the following analysis.

Blood-based DEGs changed upon combined treatment abnormally expressed in lung cancer tissue

Standard bioinformatics analysis was applied to process RNA-seq reads from 5 pairs of blood samples (pre v post) to identify the genes that differentially expressed in response to anti-PD-1 treatment. Differential expression analysis revealed that a total of 155 protein-coding genes out of almost 30,000 genes (total n =29,195, FDR p <0.05 and fold change >2 or <0.5) were differentially expressed between pre- and post-treated blood samples (**Figure 2A**). Utilizing the overall RNA-seq profiles, principal components analysis (PCA) constructed complete separation of pre and post sample groups as showed by biplots (**Supplementary Figure S1**). Hierarchical clustering using the 155 DEGs organized the pre-treatment samples into a distinct group from the post-treatment samples (**Figure 2B** and **Supplementary Table S1a**). The mRNA expression difference of 10 top-ranked genes were examined in an independent patient group (n=7 v 7) by qRT-PCR and found all of them were differentially expressed in same patterns (**Figure 2C**). Out of 10, 8 were confirmed significantly (all p<0.05) changed in pairwise comparison.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Next, we sought to investigate the expression profiles of these top-ranked genes in cancer tissues. Firstly, we searched human tissue database GENT2 that generated from NCBI GEO databases that contains microarray data from >499 lung cancer samples and >1117 normal lung samples. Interestingly, the transcriptional changes of 8 genes (out of top 10 DEGs) from pre to post were consistently seen from tumor to normal lung tissue ($p<0.01$) (**Figure 2D, Supplementary Figure S2** and **Supplementary Table S2a**). Then we extended our search to a larger lung cancer database that was built from 56 datasets (over 6700 patients) and allows meta-analysis of the gene expression profiles as well as the association between expression and overall survival. Combining statistical power from multiple datasets, a strong agreement to the data above was observed that a consistent tumor-to-normal change was significantly detected in 6 out of the 8 genes identified above (all $p<0.05$) (**Figure 3, Supplementary Figure S3** and **Supplementary Table S2a)**.

DEGs down-regulated in responders were over-expressed in lung cancer and associated with poor overall survival

We hypothesized that the molecular signatures correlated with complete responses may indicate the prognosis of NSCLC patients. RNA-seq data from responders (res) and non-responders (non-res) was collected and submitted for PCA analysis. Several combinations of key principal components (PCs), such as PC1 + PC3, successfully separated res and non-res samples (**Figure 4A,** yellow highlighted). A total of 112 DEGs as identified by comparing res and non-res subjects and utilized to construct hierarchical clustered heatmap where res and non-res subjects were grouped distinctly (**Figure 4B** & **4C** and **Supplementary Table S1b**). Our qRT-PCR results confirmed in an independent cohort that the expression differences of 7 top-ranked DEGs

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(out of 10) were consistently seen between res and non-res subjects (n=7 v 7) (**Supplementary Table S1b**).

Again, we investigate the tissue expression profiles of the top 10 DEGs in GENT2 and LCE databases. Seven genes were found differently expressed between normal lung and lung cancer tissue according to GENT2 database (**Figure 4E** and **Supplementary Table S2b**) and 4 of them showed consistent differences in meta-analysis from LCE databases (**Figure 5A** and **Supplementary Table S2b**). For instance, genes DLG5 and H3C10 were down-regulated in responder groups as detected by RNAseq and qRT-PCR (**Figure 4B & 4D)**, and down-regulated in normal lung tissue as compared with tumor tissue (**Figure 4E & 5A)**. Notably, unlike the fact that survival meta-analyses in LCE database provided no significant association found for DEGs from pre v post comparison, our search for top DEGs from res v non-res comparison revealed two cases of negative expression-survival association in lung cancer patients ($p<0.01$ for DLG5 and $p=0.04$ for H3C10) (**Figure 5B** and **5C**).

Immunological and metabolism pathways and leukocyte abundances are key signatures in responders after treatment

Pathway and gene ontology (GO) overrepresentation analysis provided an answer about which regulatory mechanisms are statistically significantly associated with the combinational treatment or the treatment responsiveness. KOBAS-i is a newly released tool utilizing machine learning algorithms and integrating functional class scoring approaches into a single ensemble score, to optimize the prioritization of biologically relevant pathways from expression data. The results indicated the co-modulated DEGs from post v pre comparison were significantly enriched in mechanism clusters that commonly connected to each other (C1-C3, C5& C6) (**Figure 6A**), covering all top-ranked pathways (**Figure 6B**). Differentially, DEGs from res v non-res

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

comparison were enriched in separated mechanism (**Figure 6C**). Interestingly, the top 10 overrepresented mechanisms (ranked by corrected p-value) in two comparisons shared pathways such as "Immune System" and "Metabolism of protein" and gene ontologies such as "Protein binding" (**Figure 6D**). The difference was the significant mechanisms in comparison of post- v pre focused immunological and metabolism pathways while the significant mechanisms in comparison of res v non-res are more diverse.

Interestingly, GSEA analyses focused on immunologic signature gene sets (collection C7, in total 5279 gene sets) identified 52 down-regulated and 14 up-regulated gene sets from comparison of post- and pre-treatment (**Figure 7A,** pink). Zero down-regulated and 116 up-regulated immune gene sets were recognized using our DEG data from res v non-res comparison (**Figure 7A,** blue). Overlapped up-regulated gene sets (n=9) of two comparisons indicated a broad spectrum of the major leukocyte subsets were activated in responding patients after treatment, including monocyte, Dendritic cells (DCs), T helper 2 (TH2) cells and B cells (**Figure 7A,** purple). Moreover, the shared gene sets contained gene set GSE1889 whereas genes are up-regulated after stimulated by TNF as comparing control cells when versus T-reg cells. This finding suggested, in addition to the activated immune cell subsets, inactivation of T-reg cells was also triggered in a successful treatment (**Figure 7A,** purple).

Finally, an *in silico* leukocyte deconvolution approach (AImmune) for inference of key circulating immune cells components was modified from our previous study [30] with inclusion of new gene markers adapted from single cell RNAseq data obtained from peripheral blood mononuclear cells (PBMCs) of NSCLC patients and healthy donors. To assess deconvolution performance, the relative expressions and pairwise correlations of these new signature genes across 7 key circulating immune cells were illustrated by heatmap and similarity matrix,

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

respectively (**Figure 7B** and **Supplementary Figure S4**). Immune cell scores were generated by AImmune approach utilizing the RNAseq profiles in res and non-res subjects and no significant difference was observed (**Figure 7B**). We also calculated and compared the scores for a simulated cohort (n=20 each group) based on the individual profiles (**Supplementary Figure S5**). The results indicated monocytes and both CD4+ and CD8+ T lymphocytes showed significant increase in res subjects as compared to non-res subjects.

Discussion

Given the majority of tumor patients fail respond to anti-PD-1/PD-L1 therapy, a lot of effort made by clinical trials led to the identification of predictive biomarkers to select treatment responders and the discovery of underlying mechanisms during treatment for improving current treatment modalities [31, 32]. Although PD-1/L1 immunohistochemical expression has been widely evaluated as a predictive biomarker, it is not implemented as a routine diagnostic test for anti–PD-1/PD-L1 therapy worldwide, including China [33]. Part of the reason is lack of universal standard of antibody for detecting PD-L1 expression and no mention it is still controversial whether PD-1/L1 expression is an inadequate determinant of treatment benefit with anti-PD-1/L1 therapies.

The overall hypothesis of the present study is the blood-based molecular signatures that differently expressed in blood samples from responders versus non-responders may differ between normal and cancer in tissue as tumor-driving genes, and thus provide additive prognostic value in patient selection. Our integrated analysis workflow (**Figure 1**) also included innovative approaches as another layer of blood-based strategy to evaluate the leukocyte abundance as non-invasive biomarkers reflecting the host immunity, which ultimately will improve the clinical efficacy of PD-1 blockade therapy. By profiling the patients and responders

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

during the treatment, our data provided new evidence for targetable pathways, driving factors and molecular and cellular determinants of response that is correlated with the beneficial response and clinical efficacy.

At the beginning of this study, investigation of the top-ranked 10 DEGs found 6 genes (up: HBG1, HBG2, GATA2, IFI27; down: LHX4 and ANKRD22) that were not only differentially expressed in blood samples between post and pre subjects but also abnormally expressed in lung cancer tissue (**Figures 2C, 2D** and **Figure 3**). These findings were significant as they were, on one hand, validated by multiple approaches (RNA-seq, qRT-PCR and GENT2 microarray database); on the other hand, they were also confirmed by multiple datasets and diverse cohorts (LCE meta-analysis database). This gene signature group contains a modulator of innate immune responses (IFN-I signal-induced gene IFI27) [34, 35], a hematopoietic transcription regulator (GATA2) [36], a metabolism mediator (ANKRD22) [37, 38], and a transcription factor involved in differentiation control (LHX4) [39], which is consistent with the following findings that immunological and metabolism mechanisms are enriched in post-treatment and responder patients (**Figure 6 A & 6B**). Moreover, several of them were proven to regulate the oncogenesis of lung cancer or respiratory diseases. For instance, ANKRD22 as we found down-regulated in NSCLC patients after therapy, was reported to promote the progression of NSCLC by enhancing cell proliferation [40]. In the present study we showed HBG1/2, as coding genes for hemoglobin subunit gamma, were up-regulated after treatment and also enriched in normal tissue comparing to tumor tissue. This finding is supported by study that stated HBG1/2 expression is upregulated in normal lung tissue by smoking and is downregulated again when tumors form in smokers [41]. More importantly, a recent report also determined HBG1 was consistently upregulated in PBMCs of patients with stage III/IV NSCLC from the 2nd anti-PD-1 antibody treatment through

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

to the 5th treatment [42]. This observation seems dependent on stage and histology as both protein and mRNA levels of HBG1/2 were found down-regulated in stage I NSCLC adenocarcinomas [43]. Together, these findings suggest that the gene signature we identified and investigated in the present study could potentially act as a novel biomarker for effective patient selection.

Next round of comparison (res v non-res) in RNA-seq profiles identified DEGs with diverse function and roles. Three of the top 10 DEGs are histone family genes (H3C10/HIST1H3H, H2BC17/HIST1H2BO, H4C8/HIST1H4H) (**Figure 4B & Supplementary Table S1b**). A couple of histone genes have been reported to be hypermethylated in lung cancer and associated with oncogenesis in a wide range of cancers, suggesting histone gene could act as a potential universal-cancer-only marker [44]. H3C10 was previously reported up-regulated in breast cancer and its high expression was associated with poor overall survival [45]. H3C10 is rarely studied and its clinical relevance in lung cancer remains unknown. Here we searched two large tissue databases covering >50 studies and it turned out that genes H3C10 and DLG5 are both significantly decreased in normal tissue as compared to tumor tissue (**Figures 4D, 4E & 5A)**. Our survival meta-analysis also showed both genes are significantly associated with poor prognosis of lung cancer patients (**Figures 5B-C**). Notably, the result of DLG5 is not quite in agreement with an earlier study that claimed DLG5 is down-regulated in SqCC tissue and indicated a favorable prognosis [46]. Nevertheless, the observation that DLG5 is down-regulated in normal tissue is agreement with our original finding that DLG5 was down-regulated in responders as compared with non-responders (**Figures 4B**). Given the consistent results across multiple platforms and comparisons, we concluded that blood-based gene signatures may have potential predictive value in NSCLC.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In the last part of this study, we sought to understand the transcriptomic profiles from a different perspective that involves high-level representations, such as pathways, gene ontologies and gene sets. KOBAS-i as utilized in our analysis is an intelligent prioritization tool that covering multiple pathway databases (KEGG Pathway, Reactome and PANTHER) and GO databases [47]. It turned out the most top-ranked pathways retrieved from our DEG lists were mostly assigned to pathways collected in Reactome (**Figure 6B & 6D upper**). Reactome is known to be composed of less broad but more detailed terms as compared to KEGG [48]. In agreement with that, our over-representation analysis on post v pre comparison yielded a complex cluster network showing the significant enriched terms were connected and overlaid (**Figure 6A**). This result demonstrated the pathways or GO associated with immune system (C1 & C5) and metabolism (C2 & C3) were significantly enriched in post-treatment subjects. In contrast, analysis on res v non-res comparison displayed a lot more metabolism regulatory terms were enriched (such as "O2/CO2 exchange"), suggesting metabolic regulations or metabolic products may play key roles in driving the outcome of anti-PD-1 blockade treatment (**Figure 6C**). Our findings provided additional insights to the growing body of evidence documenting that combination of metabolic and immune biomarkers in lung cancer is quite valuable for responder prediction [27, 49]. These signaling cascades identified here are also of interest from a therapeutic perspective, combining PD-1 checkpoint inhibitors with a broad range of clinically active partners is a highlighted strategy with potential importance [50, 51].

One limitation of the pathway analysis was only minimum directional effect (absolute values for z-score<2) was observed in the individual pathway, which is probably due to the complex interactions between immune components and regulators within each pathway. To overcome this issue, quantitative analysis was performed to evaluate the directional impact on immunological

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

gene sets and circulating leukocytes. In consistent with the pathway analyses above, the gene set enrichment analysis also highlighted a couple of metabolism-related gene sets (GSE9006) uniquely in responders (**Figure 7A**, blue), in addition to gene sets representing immune cell subset (GSE22886, GSE3982 and GSE18893) (**Figure 7A**, purple). It is not surprising a broad-spectrum of leukocyte subsets (Monocytes, DCs, Th2 cells and B cells) were up-regulated or down-regulated (Treg cells) in the responder patients, which was also validated by our immune cell profiling analysis (**Figure 7C**). This finding is strongly supported by several recent studies and not limited to lung cancer [52-55]. Several key lymphocyte subsets (higher levels of CD4- and CD8-T cells; lower levels of NK cells) measured in blood samples at baseline were correlated with overall survival as well as clinical response in NSCLC patients received anti-PD-1 antibody treatment [53]. Similarly, the cellular signatures such as M1 macrophage and peripheral T cell showed better predictive performance than PD-L1 immunohistochemistry, tumor mutation burden, or tumor-infiltrating lymphocytes in NSCLC [54].

It is important to point out the *in silico* leukocyte deconvolution analysis (AImmune) presented here only strengthened the conclusions as drawn by other studies but also proposed a novel leukocyte enumeration approach that designed for circulating blood samples, instead of tissue samples as reported by other studies. We have previously used this approach in analyzing PBMC samples from patients with lung infection [30] and demonstrated the abundance of certain leukocyte subsets was lung disease severity (unpublished data). Similar immune profiling technologies have been employed to enumerate tumor-infiltrating immune cells (e.g. TIMER [56] and CIBERSORT [57]). These methods were recently upgraded taking advantage of the fast-growing single cell RNAseq data and well-developed modern machine learning algorithms (e.g. TIMER 2.0 [58]; CIBERSORTx[29]). Here for the first time, we reported an implementation of

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

modified AImmune approach. With the existing gene markers tested and validated in our previous study [30], new gene markers adapted from PBMC-based scRNAseq data of NSCLC patients and healthy donor are now included in AImmune (**Figure 7B** & **Supplementary Figure S5**). We believe our results suggest alternative technology option in terms of evaluating leukocytes as blood-based biomarkers for immunotherapy.

In summary, the integrated analyses employed in the current study identified blood-based molecular and cellular signatures that are associated with responsiveness of ICI plus chemotherapy as well as the overall survival. Furthermore, the present study demonstrated that immunological and metabolic pathways may prove a useful panel of targetable biomarkers to improve patient selection and prognosis prediction. Despites its insights, the retrospective and preliminary nature of this small cohort study should be highlighted. Given the high molecular heterogeneity among different NSCLC subtypes, the number of patients sequenced in this study might be too small to provide strong statistical power or drawn informative conclusions. It is encouraged that our current findings and results should be further confirmed by a larger cohort with long-term follow-up and time series analysis.

Method

Study cohort and sample collection

This study was approved by the local Ethics Committee and the Institutional Review Board of Northwest University (approval number: 200402001) and all patients provided written informed consent. This retrospective study collected 42 whole blood samples from 30 patients visited Tianjin Cancer Hospital from 2020 to 2021 (**Figure 1**). All patients were diagnosed with non-small cell lung cancer (**Table 1**) and under treatment of anti–PD-1 monoclonal antibody Tislelizumab [59, 60], in combination with standard chemotherapy according to the histology.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Overall study design and data analytical flow

Transcriptomic profiling (RNA-seq) was performed on mRNA samples in two comparison groups: 1) post-treatment (post) versus (v) pre-treatment (pre) and 2) responders versus non-responders (**Figure 1**). The data analytical flow started by multiple sequential analysis at gene level, such as RNA-seq processing and real-time quantitative PCR (qRT-PCR) validation, which were followed by investigation of differential expression genes (DEGs) in tissue database (Gene Expression database of Normal and Tumor tissues 2, GENT2) [61] and meta-analysis of differential gene expression and expression-survival association in lung cancer database (Lung Cancer Explorer, LCE) [41]. The DEG data produced from the gene-level analyses were then submitted for further computational analyses in higher-level representations, such as canonical pathways and functional gene ontologies (KEGG Orthology Based Annotation System-intelligent version, KOBAS-i) [47], immune-related gene sets (Gene Set Enrichment Analysis, GSEA) [62, 63] and immune cell subsets (AImmune).

AImmune analysis and computational analysis

A novel *in silico* leukocyte deconvolution method, named AImmune, is a computational approach developed by integrating our established immune cell profiling [30] with published single cell RNAseq data obtained from NSCLC PBMCs (1071 qualified cells from one patient, GSE127471) and healthy PBMCs (8369 and 7687 cells from two donors, 10X Genomics). Briefly, with the additional marker genes included, more than 30 candidate marker genes for each cell subsets in leukocytes (CD4-T cells, CD8-T cells, B cells, Monocytes, DCs, NK cells and NKT cells) were selected based on their expression patterns across immune cell subsets [57] (**Figure 7B**). The pairwise similarity statistic of all cell subsets (**Figure S4**) was computed between all pairs of the candidate marker genes within the normalized RNA-seq profiles (FPKM)

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

from whole blood samples. Using the criteria (average Pearson correlation factor >0.60, p<0.01), 10-20 selected marker genes were identified as our final marker genes. The raw cell abundance score was calculated as the sum of the simple averages of the marker genes' log2 expression, which allows comparison of cell composition across subject groups. Monte Carlo simulation, as previously used in proteomics analysis and produce scores for peptide [64], was employed here to estimate the relation between the raw cell abundance score and generated a new dataset (n=20 in each group) with the same distributions and correlations as the raw scores. This approach also tested a novel deconvolution model (unpublished) built by DNN (deep neural networks) algorithms and then trained by pseudo-bulk samples obtained by randomly subsampling of published scRNA-seq data [65]. Machine learning-based feature extraction (marker gene selection) was integrated for model optimization. Most of the computational analysis procedure was coded by common Python packages, Monte Carlo simulation was performed by R package faux; scRNA-seq data was processed by R package scanpy; machine learning model was developed and tested with Python library Tensorflow. All computational analysis were performed and visualized using R version 3.6.1 or Python version 3.7.9.

RNA isolation, transcriptomic profiling, and qRT-PCR

Total RNA was isolated using Trizol (Invitrogen, Waltham, MA, USA) and the purity and concentration were verified using a NanoDrop ND-1000 instrument (ThermoFisher Scientific, Waltham, MA, USA). The integrity of the RNA was assessed by a 2100 Bioanalyzer gel image analysis system (Agilent, Santa Clara, CA, USA). Only RNA samples with an integrity number of 7 or higher were processed for RNA-seq. At least 100 ng mRNA per sample was submitted for library preparation.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Qualified RNA samples were then enriched and synthesized into two strand cDNA for library preparation. RNA-seq libraries were constructed using the TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA, USA). The libraries from qualified RNA samples were sequenced in the 150 nt paired-end mode on an Illumina HiSeq 6000 platform at Novogene Bioinformatic Technology (Tianjin, China). After quality filtering (FastQC, quality value >5), over 30 billion clean reads were obtained in each library and then used for down-stream analysis.

PCR primers were designed top-ranked genes (10 for each comparison) as obtained by differential gene analysis (**Supplemental Table S3**). Real-time quantitative PCR was performed in real-time PCR systems (Bio-Rad, Hercules, CA, USA). The 2–ΔΔCT method for relative gene expression analysis was performed to calculate the mRNA levels of individual gene. Two or three replicates were measured for each sample.

Bioinformatics and statistics analyses

All raw RNA-seq reads were filtered by R package trim_fastq to remove adapters, rRNA and low-quality reads. The QC criteria included: removing bases below Phred quality 20, containing over two "N", or shorter than 75. The output reads were then indexed by aligner STAR and mapped to reference genome by BAM. Normalized read counts were generated and compared between groups to generate DEGs using R package DESeq2. Another R package countToFPKM was employed to produce FPKM for AImmune analysis. Genes that displayed at least two-fold difference in gene expression between comparison groups (false discovery rate [FDR]<0.05) were considered significant DEGs and carried forward in the analysis. Differentially expressed RNAs were illustrated as a volcano plot. Hierarchical clustering was performed to show the gene expression patterns and similarities among samples. PCA was performed to cluster subjects based on the differentially expressed transcripts. GSEA analysis was carried out by searching the

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

established MSigDB gene-set collections (C7). Differences of mRNA levels and cell composition scores were evaluated using independent *t*-tests or paired *t*-tests if pairwise samples were given. A p value of <0.05 was considered statistically significant. Bioinformatics and statistics analyses were performed and visualized using R version 3.6.1 or Python version 3.7.9.

Funding

This work was supported by National Natural Science Foundation of China (82071863).

Author contributions

Conceptualization, Xi Zhang; Data curation, Xi Zhang, Rui Chen, Wenqing Li, Guodong Su, Wuhao Huang and Shengguang Wang; Formal analysis, Xi Zhang, Rui Chen, Wenqing Li and Runpeng Han; Funding acquisition, Xi Zhang; Investigation, Xi Zhang, Rui Chen, Wenqing Li and Shengguang Wang; Methodology, Xi Zhang, Runpeng Han, Yuyan Xiong and Shengguang Wang; Project administration, Xi Zhang, Wenqing Li and Shengguang Wang; Resources, Xi Zhang, Rui Chen, Runpeng Han, Wuhao Huang and Shengguang Wang; Software, Rui Chen; Supervision, Xi Zhang, Yuyan Xiong and Shengguang Wang; Validation, Xi Zhang; Visualization, Xi Zhang; Writing – original draft, Xi Zhang; Writing – review & editing, Xi Zhang, Yuru Liu, Yu Cai, Yuyan Xiong and Shengguang Wang.

Acknowledgments

We thank all patients who donated blood.

Conflicts of Interest Statement

The authors have declared that no conflict of interest exists.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Reference

1. Topalian, S.L., J.M. Taube, and D.M. Pardoll, *Neoadjuvant checkpoint blockade for cancer immunotherapy.* Science, 2020. 367(6477).
2. Dong, Y., et al., *Recent Advances and Future Prospects in Immune Checkpoint (ICI)-Based Combination Therapy for Advanced HCC.* Cancers (Basel), 2021. 13(8).
3. Hargadon, K.M., C.E. Johnson, and C.J. Williams, *Immune checkpoint blockade therapy for cancer: An overview of FDA-approved immune checkpoint inhibitors.* Int Immunopharmacol, 2018. 62: p. 29-39.
4. Barta, J.A., C.A. Powell, and J.P. Wisnivesky, *Global Epidemiology of Lung Cancer.* Ann Glob Health, 2019. 85(1).
5. Rodak, O., et al., *Current Landscape of Non-Small Cell Lung Cancer: Epidemiology, Histological Classification, Targeted Therapies, and Immunotherapy.* Cancers (Basel), 2021. 13(18).
6. De Giglio, A., et al., *The Landscape of Immunotherapy in Advanced NSCLC: Driving Beyond PD-1/PD-L1 Inhibitors (CTLA-4, LAG3, IDO, OX40, TIGIT, Vaccines).* Curr Oncol Rep, 2021. 23(11): p. 126.
7. Sgambato, A., et al., *Anti PD-1 and PDL-1 Immunotherapy in the Treatment of Advanced Non-Small Cell Lung Cancer (NSCLC): A Review on Toxicity Profile and its Management.* Curr Drug Saf, 2016. 11(1): p. 62-8.
8. Rivera Vargas, T. and L. Apetoh, *Can Immunogenic Chemotherapies Relieve Cancer Cell Resistance to Immune Checkpoint Inhibitors?* Front Immunol, 2019. 10: p. 1181.
9. Murciano-Goroff, Y.R., A.B. Warner, and J.D. Wolchok, *The future of cancer immunotherapy: microenvironment-targeting combinations.* Cell Res, 2020. 30(6): p. 507-519.
10. Twomey, J.D. and B. Zhang, *Cancer Immunotherapy Update: FDA-Approved Checkpoint Inhibitors and Companion Diagnostics.* AAPS J, 2021. 23(2): p. 39.
11. Havel, J.J., D. Chowell, and T.A. Chan, *The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy.* Nat Rev Cancer, 2019. 19(3): p. 133-150.
12. Peters, S., K.M. Kerr, and R. Stahel, *PD-1 blockade in advanced NSCLC: A focus on pembrolizumab.* Cancer Treat Rev, 2018. 62: p. 39-49.
13. Mathew, M., et al., *Combining chemotherapy with PD-1 blockade in NSCLC.* Pharmacol Ther, 2018. 186: p. 130-137.
14. Bassanelli, M., et al., *Heterogeneity of PD-L1 Expression and Relationship with Biology of NSCLC.* Anticancer Res, 2018. 38(7): p. 3789-3796.
15. Nicos, M., et al., *The Role of Intratumor Heterogeneity in the Response of Metastatic Non-Small Cell Lung Cancer to Immune Checkpoint Inhibitors.* Front Oncol, 2020. 10: p. 569202.
16. Wang, D.C., et al., *Lung Cancer Heterogeneity and New Strategies for Drug Therapy.* Annu Rev Pharmacol Toxicol, 2018. 58: p. 531-546.
17. Santini, F.C. and M.D. Hellmann, *PD-1/PD-L1 Axis in Lung Cancer.* Cancer J, 2018. 24(1): p. 15-19.
18. Teixido, C., et al., *PD-L1 expression testing in non-small cell lung cancer.* Ther Adv Med Oncol, 2018. 10: p. 1758835918763493.
19. Yang, Q., et al., *Multimodal detection of PD-L1: reasonable biomarkers for immune checkpoint inhibitor.* Am J Cancer Res, 2018. 8(9): p. 1689-1696.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

20. Grossman, J.E., et al., *Is PD-L1 a consistent biomarker for anti-PD-1 therapy? The model of balstilimab in a virally-driven tumor.* Oncogene, 2021. 40(8): p. 1393-1395.
21. Vanderwalde, A., et al., *Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11,348 patients.* Cancer Med, 2018. 7(3): p. 746-756.
22. Trefny, M.P., et al., *A Variant of a Killer Cell Immunoglobulin-like Receptor Is Associated with Resistance to PD-1 Blockade in Lung Cancer.* Clin Cancer Res, 2019. 25(10): p. 3026-3034.
23. Kim, K.H., et al., *PD-1 blockade-unresponsive human tumor-infiltrating CD8(+) T cells are marked by loss of CD28 expression and rescued by IL-15.* Cell Mol Immunol, 2021. 18(2): p. 385-397.
24. Nie, X., et al., *Serum Metabolite Biomarkers Predictive of Response to PD-1 Blockade Therapy in Non-Small Cell Lung Cancer.* Front Mol Biosci, 2021. 8: p. 678753.
25. Araujo, B.d.L.V., et al., *Immune Cell Profiling of Peripheral Blood as Signature for Response During Checkpoint Inhibition Across Cancer Types.* Front Oncol, 2021. 11: p. 558248.
26. Frigola, J., et al., *Molecular profiling of long-term responders to immune checkpoint inhibitors in advanced non-small cell lung cancer.* Mol Oncol, 2021. 15(4): p. 887-900.
27. Hatae, R., et al., *Combination of host immune metabolic biomarkers for the PD-1 blockade cancer immunotherapy.* JCI Insight, 2020. 5(2).
28. Li, B., et al., *Computational Deconvolution of Tumor-Infiltrating Immune Components with Bulk Tumor Gene Expression Data.* Methods Mol Biol, 2020. 2120: p. 249-262.
29. Newman, A.M., et al., *Determining cell type abundance and expression from bulk tissues with digital cytometry.* Nat Biotechnol, 2019. 37(7): p. 773-782.
30. Zhang, X., et al., *Cystic Fibrosis Plasma Blunts the Immune Response to Bacterial Infection.* Am J Respir Cell Mol Biol, 2019. 61(3): p. 301-311.
31. Tunger, A., et al., *The Evolving Landscape of Biomarkers for Anti-PD-1 or Anti-PD-L1 Therapy.* J Clin Med, 2019. 8(10).
32. Niu, M., et al., *Predictive biomarkers of anti-PD-1/PD-L1 therapy in NSCLC.* Exp Hematol Oncol, 2021. 10(1): p. 18.
33. Ma, Y., et al., *Tumor Mutational Burden and PD-L1 Expression in Non-Small-Cell Lung Cancer (NSCLC) in Southwestern China.* Onco Targets Ther, 2020. 13: p. 5191-5198.
34. Tang, B.M., et al., *A novel immune biomarker IFI27 discriminates between influenza and bacteria in patients with suspected respiratory infection.* Eur Respir J, 2017. 49(6).
35. Gao, J., et al., *IFI27 may predict and evaluate the severity of respiratory syncytial virus infection in preterm infants.* Hereditas, 2021. 158(1): p. 3.
36. Castano, J., et al., *GATA2 Promotes Hematopoietic Development and Represses Cardiac Differentiation of Human Mesoderm.* Stem Cell Reports, 2019. 13(3): p. 515-529.
37. Pan, T., et al., *ANKRD22, a novel tumor microenvironment-induced mitochondrial protein promotes metabolic reprogramming of colorectal cancer cells.* Theranostics, 2020. 10(2): p. 516-536.
38. Qiu, Y., et al., *ANKRD22 is involved in the progression of prostate cancer.* Oncol Lett, 2019. 18(4): p. 4106-4113.
39. Dong, X., et al., *LIM-Homeodomain Transcription Factor LHX4 Is Required for the Differentiation of Retinal Rod Bipolar Cells and OFF-Cone Bipolar Subtypes.* Cell Rep, 2020. 32(11): p. 108144.
40. Yin, J., et al., *ANKRD22 promotes progression of non-small cell lung cancer through transcriptional up-regulation of E2F1.* Sci Rep, 2017. 7(1): p. 4430.
41. Cai, L., et al., *LCE: an open web portal to explore gene expression and clinical associations in lung cancer.* Oncogene, 2019. 38(14): p. 2551-2564.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

42. Wu, J., et al., *Gene expression profiles of peripheral blood mononuclear cells from patients with advanced non-small cell lung cancer treated with anti-PD-1 monoclonal antibodies.* Journal of Clinical Oncology, 2019. 37(15_suppl): p. e14107-e14107.
43. Fahrmann, J.F., et al., *Proteomic profiling of lung adenocarcinoma indicates heightened DNA repair, antioxidant mechanisms and identifies LASP1 as a potential negative predictor of survival.* Clin Proteomics, 2016. 13: p. 31.
44. Dong, S., et al., *Histone-Related Genes Are Hypermethylated in Lung Cancer and Hypermethylated HIST1H4F Could Serve as a Pan-Cancer Biomarker.* Cancer Res, 2019. 79(24): p. 6101-6112.
45. Xie, W., et al., *Expression and potential prognostic value of histone family gene signature in breast cancer.* Exp Ther Med, 2019. 18(6): p. 4893-4903.
46. Che, J., et al., *Decreased expression of Dlg5 is associated with a poor prognosis and epithelial-mesenchymal transition in squamous cell lung cancer.* J Thorac Dis, 2021. 13(5): p. 3115-3125.
47. Bu, D., et al., *KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis.* Nucleic Acids Res, 2021. 49(W1): p. W317-W325.
48. Jassal, B., et al., *The reactome pathway knowledgebase.* Nucleic Acids Res, 2020. 48(D1): p. D498-D503.
49. Zhang, D., et al., *Profiling changes in metabolism and the immune microenvironment in lung tumorigenesis.* Ann Transl Med, 2019. 7(Suppl 3): p. S90.
50. Kudo, M., *Scientific Rationale for Combination Immunotherapy of Hepatocellular Carcinoma with Anti-PD-1/PD-L1 and Anti-CTLA-4 Antibodies.* Liver Cancer, 2019. 8(6): p. 413-426.
51. Schmidt, E.V., *Developing combination strategies using PD-1 checkpoint inhibitors to treat cancer.* Semin Immunopathol, 2019. 41(1): p. 21-30.
52. Simon, S., et al., *PD-1 and TIGIT coexpression identifies a circulating CD8 T cell subset predictive of response to anti-PD-1 therapy.* J Immunother Cancer, 2020. 8(2).
53. Ottonello, S., et al., *Association Between Response to Nivolumab Treatment and Peripheral Blood Lymphocyte Subsets in Patients With Non-small Cell Lung Cancer.* Front Immunol, 2020. 11: p. 125.
54. Hwang, S., et al., *Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer.* Sci Rep, 2020. 10(1): p. 643.
55. Gide, T.N., et al., *Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy.* Cancer Cell, 2019. 35(2): p. 238-255 e6.
56. Li, T., et al., *TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells.* Cancer Res, 2017. 77(21): p. e108-e110.
57. Newman, A.M., et al., *Robust enumeration of cell subsets from tissue expression profiles.* Nat Methods, 2015. 12(5): p. 453-7.
58. Li, T., et al., *TIMER2.0 for analysis of tumor-infiltrating immune cells.* Nucleic Acids Res, 2020. 48(W1): p. W509-W514.
59. Liu, S.Y. and Y.L. Wu, *Tislelizumab: an investigational anti-PD-1 antibody for the treatment of advanced non-small cell lung cancer (NSCLC).* Expert Opin Investig Drugs, 2020. 29(12): p. 1355-1364.
60. Desai, J., et al., *Phase IA/IB study of single-agent tislelizumab, an investigational anti-PD-1 antibody, in solid tumors.* J Immunother Cancer, 2020. 8(1).
61. Park, S.J., et al., *GENT2: an updated gene expression database for normal and tumor tissues.* BMC Med Genomics, 2019. 12(Suppl 5): p. 101.
62. Subramanian, A., et al., *GSEA-P: a desktop application for Gene Set Enrichment Analysis.* Bioinformatics, 2007. 23(23): p. 3251-3.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

63. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

64. Xu, H. and M.A. Freitas, *Monte carlo simulation-based algorithms for analysis of shotgun proteomic data.* J Proteome Res, 2008. 7(7): p. 2605-15.

65. Menden, K., et al., *Deep learning-based cell composition analysis from tissue expression profiles.* Sci Adv, 2020. 6(30): p. eaba2619.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1. General characteristics of the study subjects (N=30)

Characteristics	Count (percentage)
Gender	
Male	26 (86.7%)
Female	4 (13.3%)
Age	
≥60	20 (66.7%)
<60	10 (33.3%)
Smoking history	
Yes	24 (80%)
No	6 (20%)
Histology	
Squamous (SqCC)	23 (76.7%)
Adenocarcinoma (ADC)	6 (20%)
Large cell carcinoma (LCA)	1 (3.3%)
Stage	
IA	8 (26.7%)
IB	4 (13.3%)
IIA	1 (3.3%)
IIB	3 (10%)
IIIA	6 (20%)
IIIB	5 (16.7%)
IVA	1 (3.3%)
Unable to stage	2 (6.7%)
Pre- and post-treatment samples	
Available	12 (40%)
Not available	18 (60%)
Responder status	
Responder (pCR)	14 (46.6%)
Non-responder (non-pCR)	16 (53.3%)

pCR, pathologic complete response

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

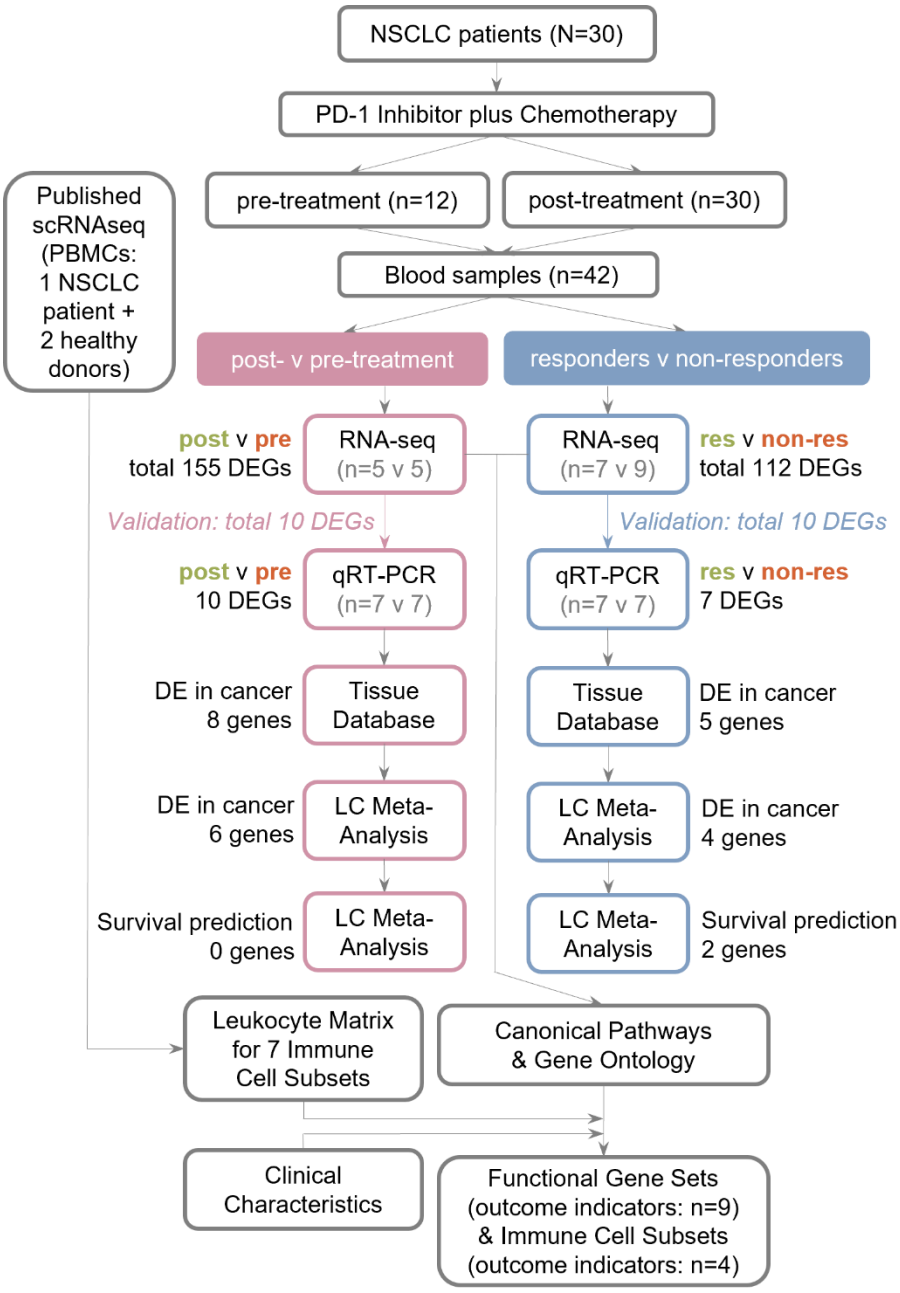


Figure 1. Workflow diagram of the study. Data collection and analytical workflow depicting the overall study design, profiling and validation scheme and computational analyses in multiple representation levels. NSCLC, non-small cell lung cancer; PBMC, peripheral blood mononuclear

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

cell; v, versus; scRNA-seq, single cell RNA-seq; res, responders; non-res, non-responders; DEGs or DE, differently expressed (genes); LC, lung cancer.

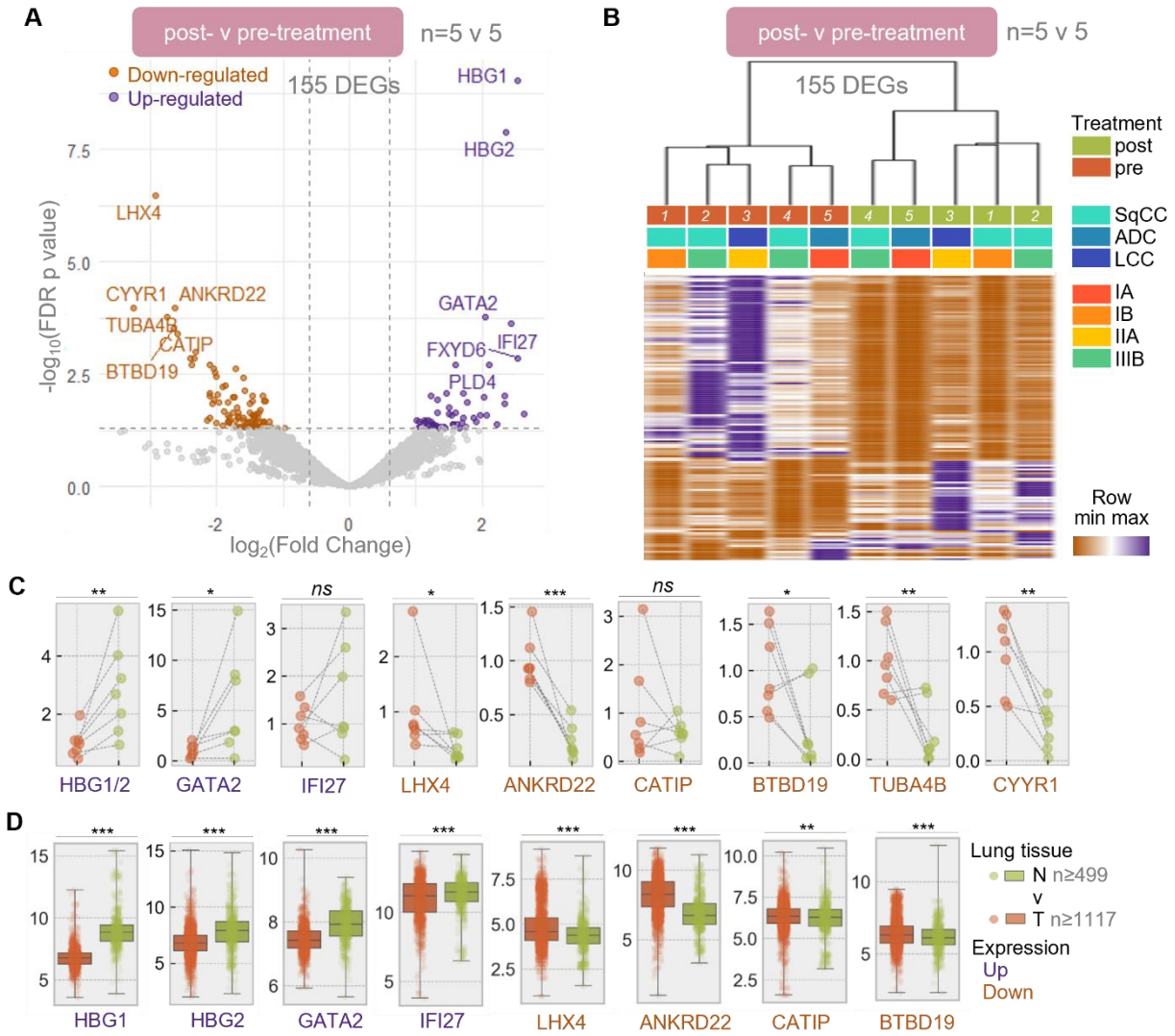


Figure 2. Blood-based DEGs responding to treatment expressed abnormally in lung cancer tissues. A) Volcano plots of total differentially expressed gens identified in pairwise comparisons of 5 patients and colored by regulation directions (up and down). B) The hierarchical clustering of the study subjects annotated by patient numbers and the heatmap representing the relative expression level of each DEG over two subject groups. Pre- and post-

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

treatment status, histology and stage information were color-coded and labelled. C) Relative mRNA levels of the top 10 genes (ranked by FDR p value) as measured in an independent patients set (pair n=7) for pairwise comparison. D) Relative mRNA levels of 8 genes as reported in normal lung and lung cancer tissues (data from GENT2 database). Genes HBG1 and HBG2 were combined because they were detected by same primer set. Dot plots and box plots, if shown, represent the individual values and the average values of the mRNA levels. N, normal lung tissue; T, lung cancer tissue; v, versus; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

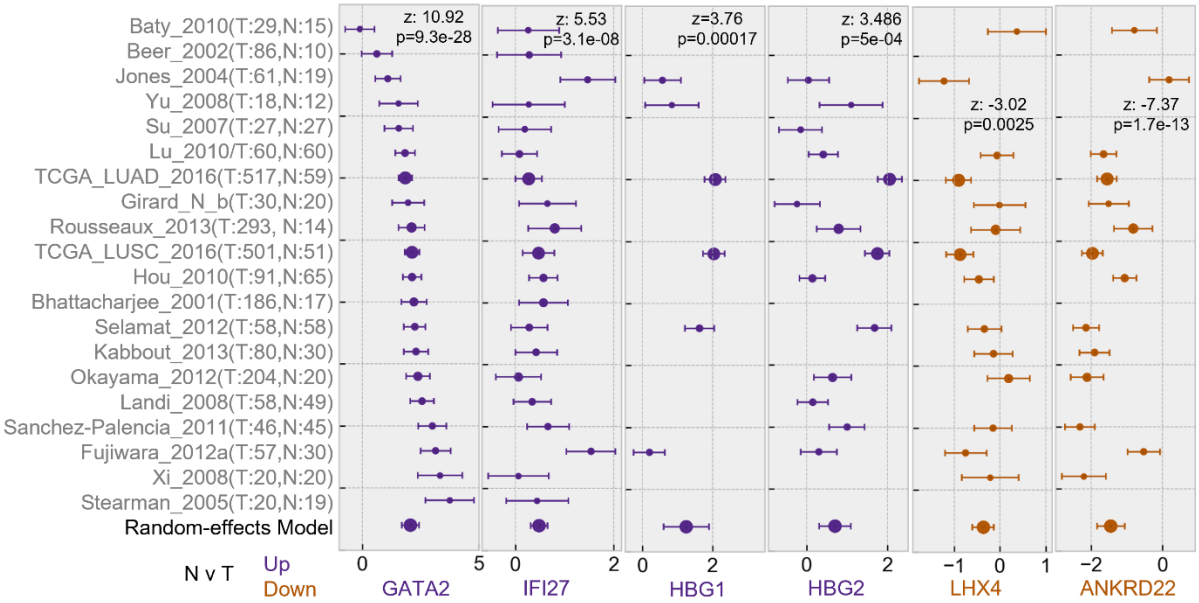


Figure 3. Meta-analyses confirmed the DEGs were differentially expressed between normal and cancer tissues. Forest plots show the standardized mean of gene expression difference between normal and tumor tissue as estimated from multiple studies (data from LCE database). The leftmost column shows the included studies by the first author's name and publication year and followed by the cohort size. The circles lined up in each column represent the effect estimates from individual studies and the very bottom circles show the pooled result for each gene as annotated. The size of each circle indicates the cohort size of individual study. The horizontal lines through the boxes illustrate the length of the 95% confidence interval in both positive and negative sides. Random-effects model was utilized to evaluate the overall effect as described by z-score and p value. N, normal lung tissue; T, lung cancer tissue; v, versus; scRNA-seq, single cell RNA-seq.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

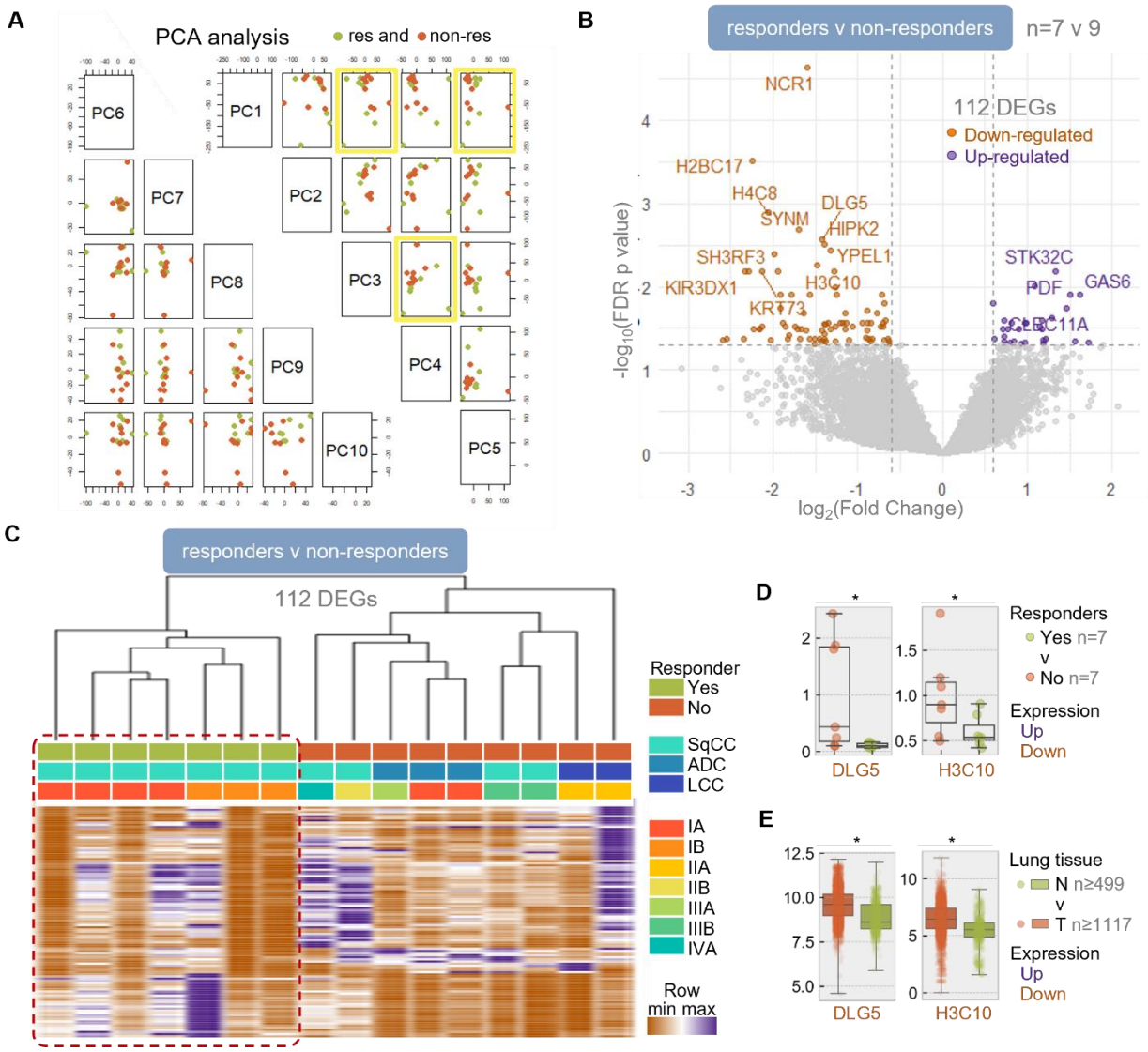


Figure 4. Blood-based DEGs downregulated in responders were enriched in lung cancer tissues. A) PCA biplots illustrating two groups of samples color-coded (res v non-res). Yellow boxes indicate combination of principal components that can distinguish two groups. B) Volcano plots of total differentially expressed gens identified in comparison of 7 res and 9 non-res patients and colored by regulation directions (up and down). C) The hierarchical clustering of the study subjects annotated by patient numbers and the heatmap representing the relative expression

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

level of each DEG over two subject groups. Responder status, histology and stage information were color-coded and labelled. D) Relative mRNA levels of genes DLG5 and H3C10 as measured in an independent patients set (7 res v 9 non-res). E) Relative mRNA levels of genes DLG5 and H3C10 as reported in normal lung and lung cancer tissues (data from GENT2 database). Dot plots and box plots, if shown, represent the individual values and the average values of the mRNA levels. res, responders; non-res, non-responders; N, normal lung tissue; T, lung cancer tissue; v, versus; *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

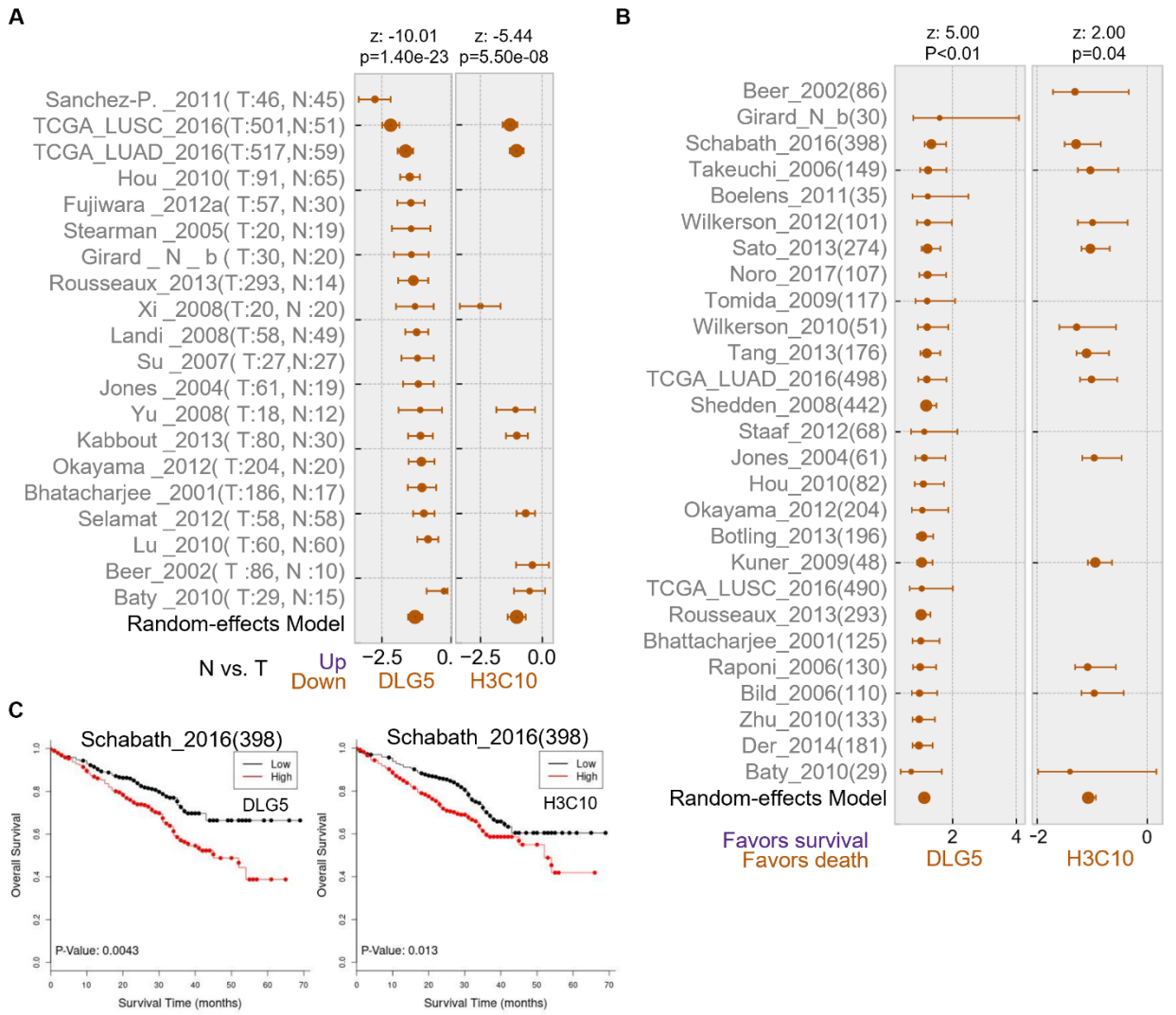


Figure 5. Meta-analyses suggested the expression profile and prognostic value of the DEGs identified from the responders. A-B) Forest plots show the standardized mean of gene expression difference between normal and tumor tissue (A) or the summary hazard ratio (B) as estimated from multiple studies (data from LCE database). The leftmost column shows the included studies by the first author's name and publication year and followed by the cohort size. The circles lined up in each column represent the effect estimates from individual studies and the very bottom circles show the pooled result for each gene as annotated. The size of each circle

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

indicates the cohort size (A) or weight (B) of individual study. The horizontal lines through the boxes illustrate the length of the 95% confidence interval in both positive and negative sides. Random-effects model was utilized to evaluate the overall effect as described by z-score and p value. C) Survival analysis of Schabath_2016 using groups defined by gene expressions of DLG5 (left) and H3C10 (right) with cutoff at median. res, responders; non-res, non-responders; v, versus.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 6. Immune-related and metabolic pathways were the most over-represented regulatory mechanisms

A & C) cirFunMap visualization of the over-representation analysis result in circular network view showing the enriched clusters from comparisons of pre v post (A) and res v non-res (C) (KOBAS-i). The top-ranked enrichment clusters (correlation > 0.35, top n = 7) were color-coded while node with bigger size stands for smaller p value. The remining lower ranked enrichment clusters were labelled in gray. B & D) Bubble plots of the top 10 pathways (upper) and the top 10 gene oncology (lower) enriched in comparisons of pre v post (B) and res v non-res (D).

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Bubble with bigger size stands for smaller p value. Each term was colored following same code used to annotate the node clusters in A or B, respectively. Z-scores of individual terms were provided as calculated following Ingenuity pathway analysis manual. res, responders; non-res, non-responders; v, versus.

medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

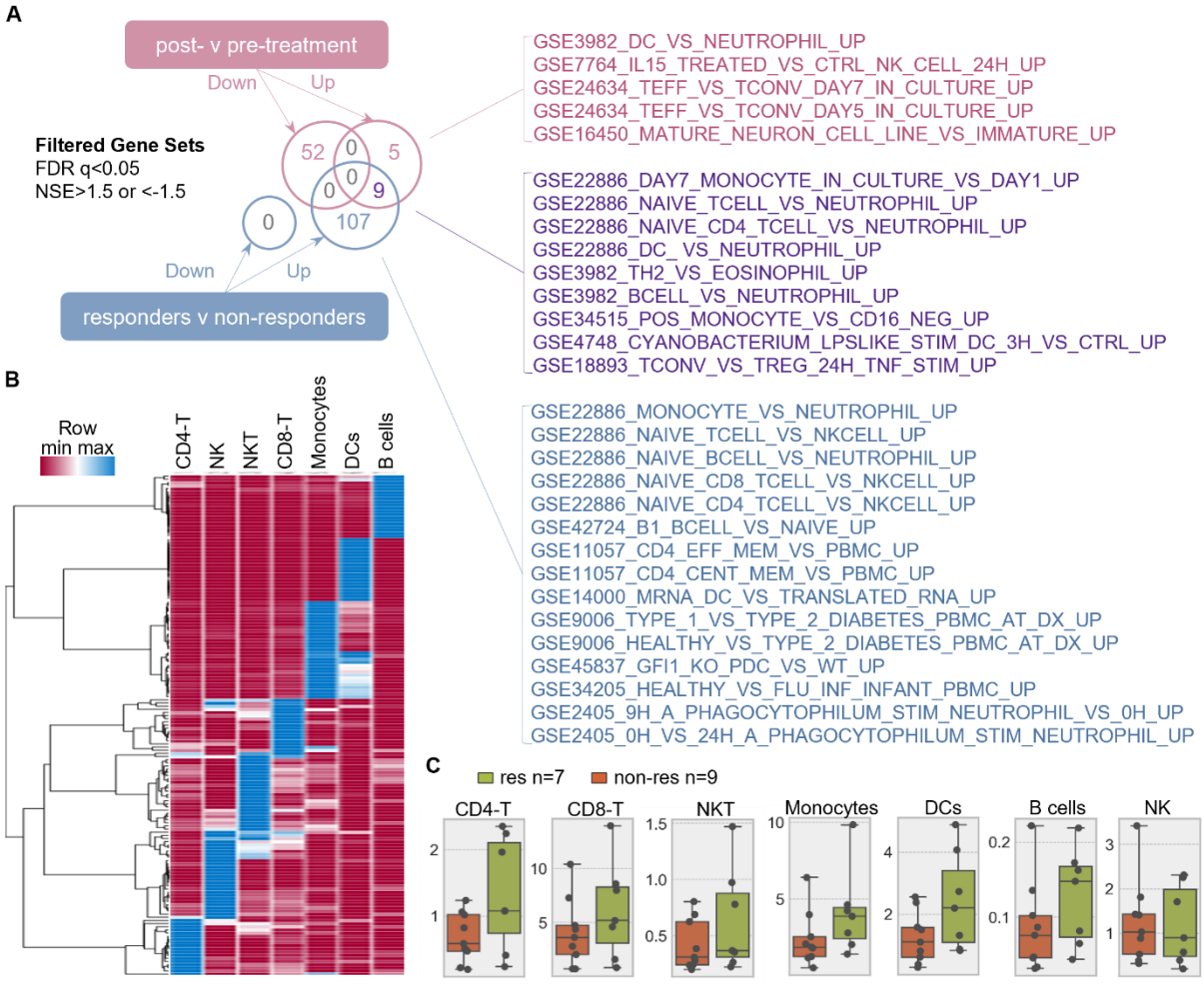
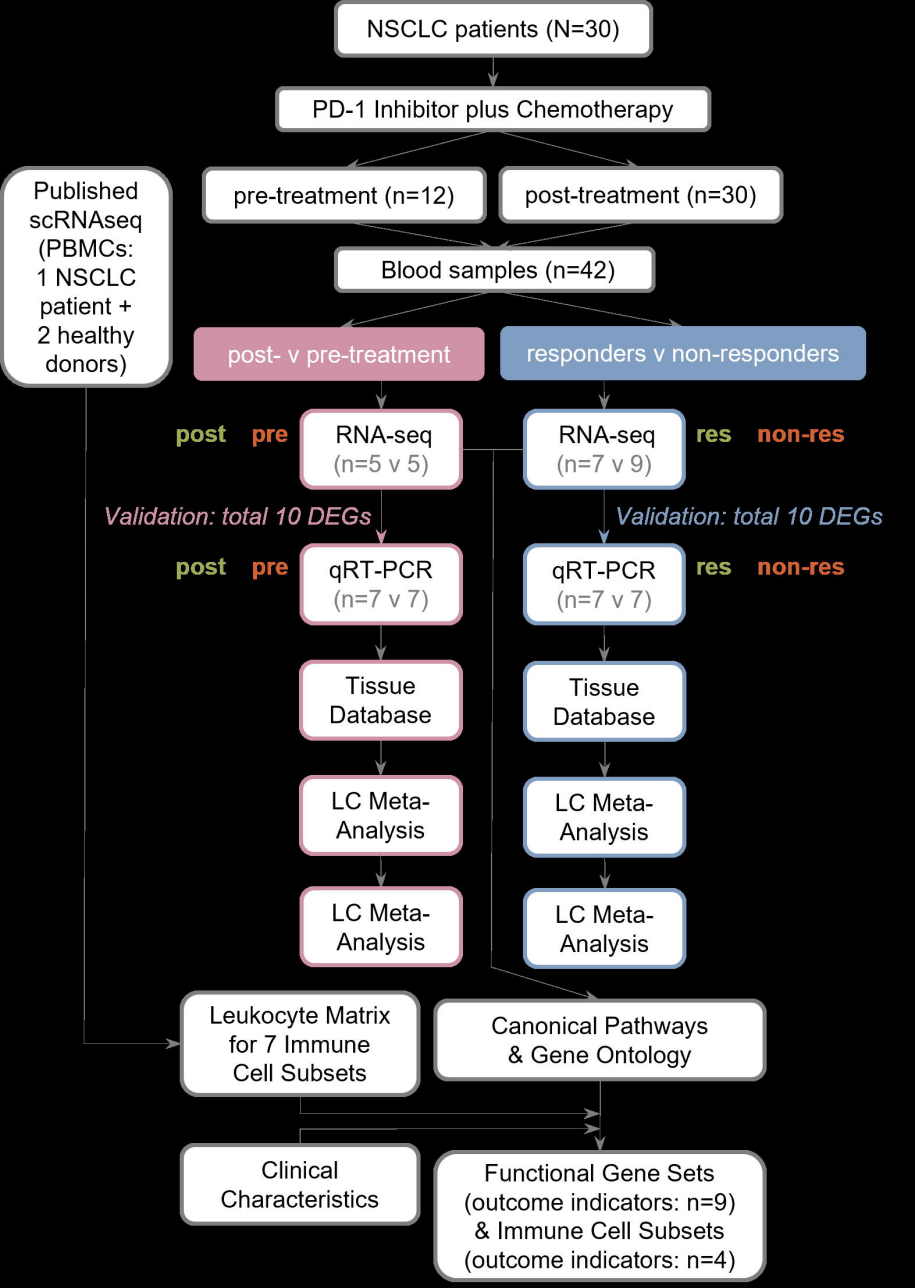


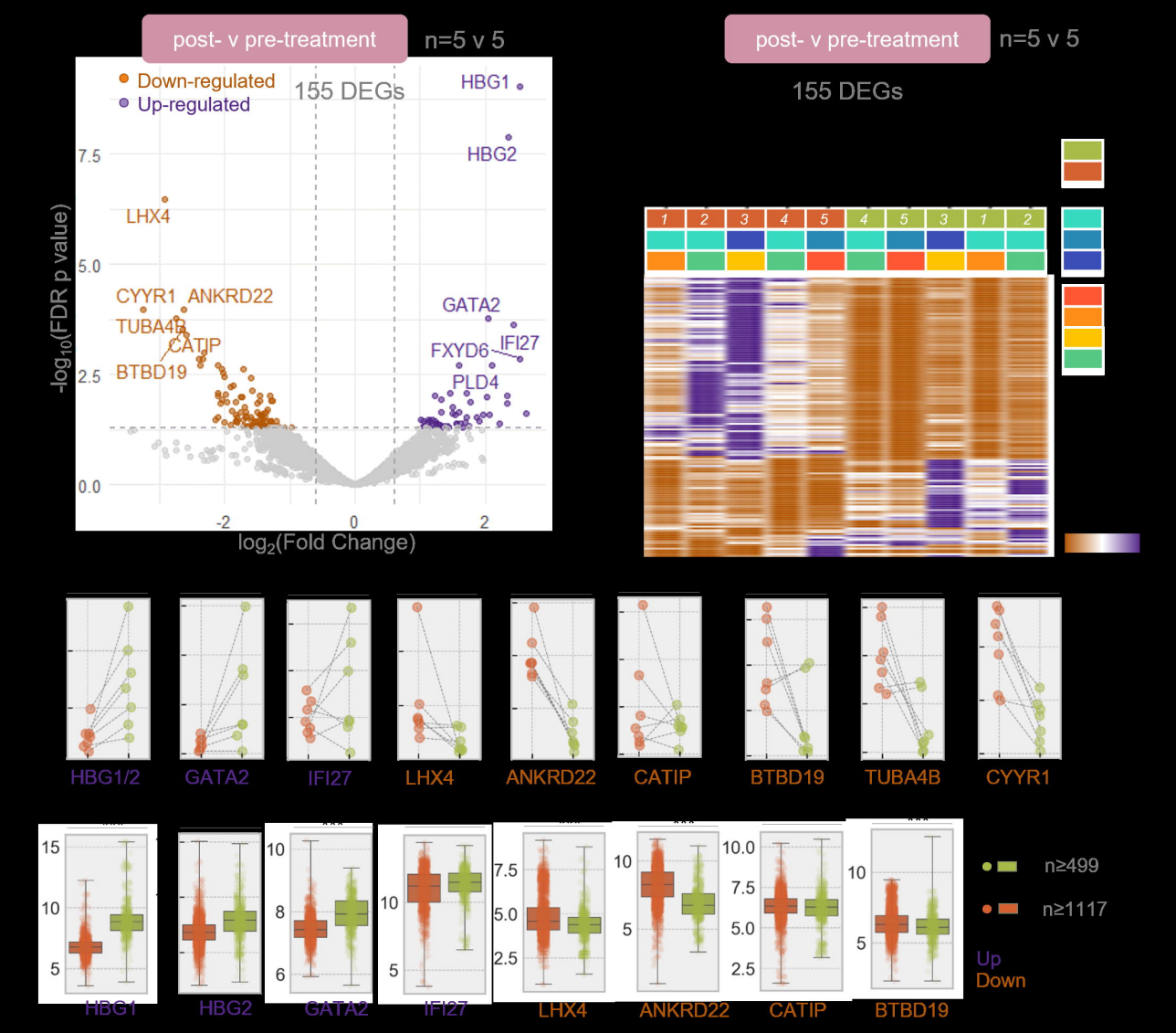
Figure 7. Integrated GSEA and immune profiling analysis revealed a broad-spectrum immune activation in responders received treatment

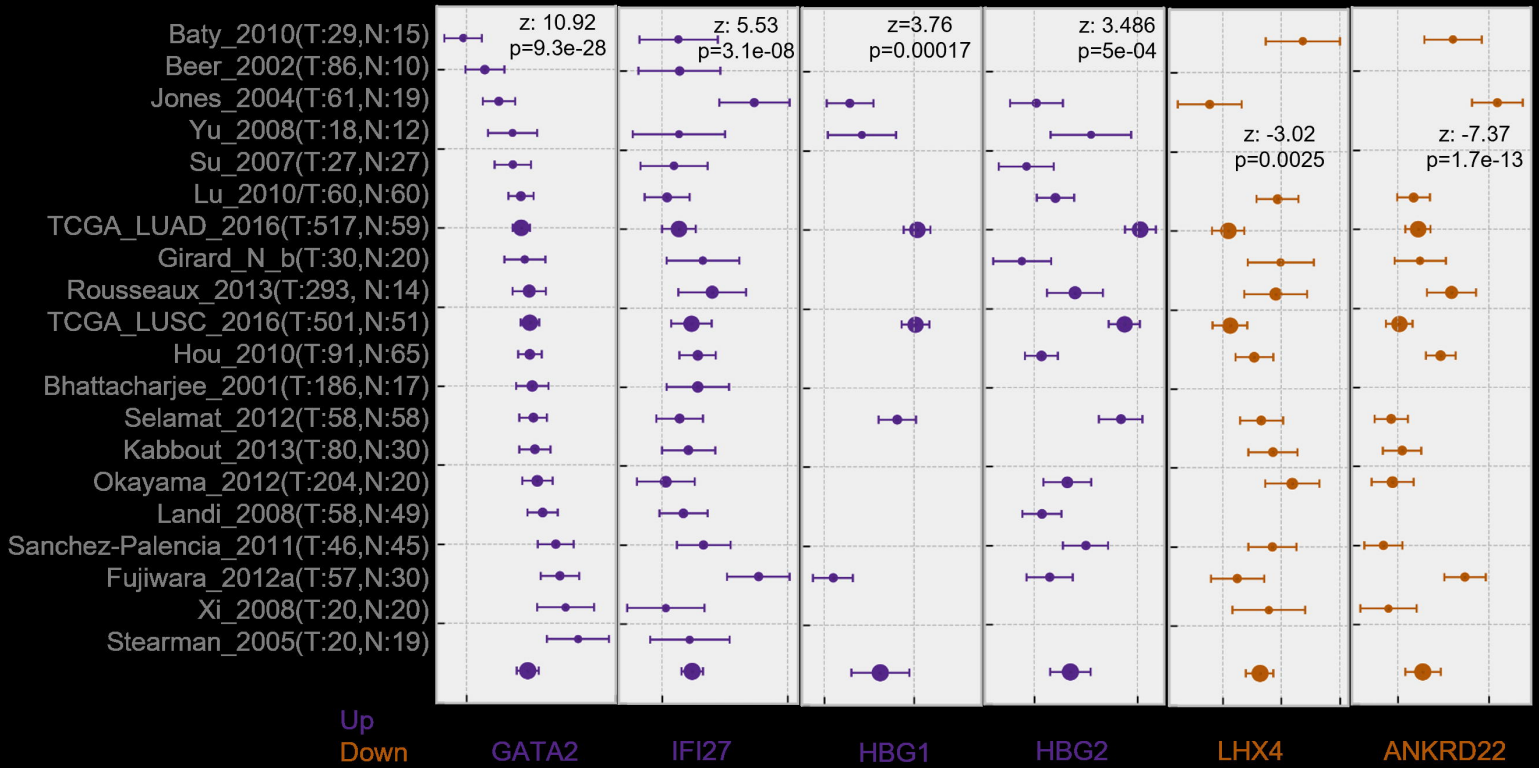
A) Integrated gene set enrichment analysis (GSEA) yielded quantitative measures of the overrepresentation of immunologic signature gene sets as regulated in either post-treatment subjects or the responder subjects. The overlapped or unique gene sets identified upregulated in each comparison are listed and annotated. All gene sets are presented in same direction (up-regulation) for convenience of interpretation. B) The hierarchical clustering of relative expression in new marker genes included in AImmune approach across 7 key immune cell

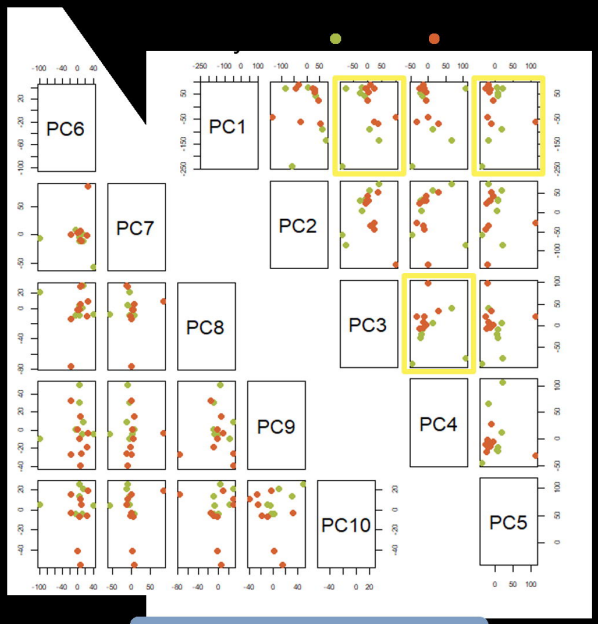
medRxiv preprint doi: https://doi.org/10.1101/2021.12.07.21267340; this version posted December 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

subsets. These expression profiles were adapted from scRNA-seq of NSCLC and healthy PBMCs. C) Immune cell scores of 7 cell subsets produced by updated AImmune approach, utilizing the transcriptomic profiles from res and non-res subjects (n=7 v 9). Dot plots and box plots show the individual values and the average value of the scores. res, responders; non-res, non-responders; v, versus; scRNA-seq, single cell RNA-seq.


NSCLC patients (N=30)
PD-1 Inhibitor plus Chemotherapy
pre-treatment (n=12)
post-treatment (n=30)
Blood samples (n=42)
Published scRNAseq (PBMCs: 1 NSCLC patient + 2 healthy donors)
post- v pre-treatment
responders v non-responders
post
pre
RNA-seq (n=5 v 5)
RNA-seq (n=7 v 9)
res
non-res
Validation: total 10 DEGs
Validation: total 10 DEGs
post
pre
qRT-PCR (n=7 v 7)
qRT-PCR (n=7 v 7)
res
non-res
Tissue Database
Tissue Database
LC Meta-Analysis
LC Meta-Analysis
LC Meta-Analysis
LC Meta-Analysis
Leukocyte Matrix for 7 Immune Cell Subsets
Canonical Pathways & Gene Ontology
Clinical Characteristics
Functional Gene Sets (outcome indicators: n=9) & Immune Cell Subsets (outcome indicators: n=4)


post- v pre-treatment
n=5 v 5
Down-regulated
Up-regulated
155 DEGs
HBG1
HBG2
LHX4
CYYR1
ANKRD22
TUBA4B
CATIP
BTBD19
GATA2
FXYD6
IFI27
PLD4
-log10(FDR p value)
log2(Fold Change)
post- v pre-treatment
n=5 v 5
155 DEGs
HBG1/2
GATA2
IFI27
LHX4
ANKRD22
CATIP
BTBD19
TUBA4B
CYYR1
HBG1
HBG2
GATA2
IFI27
LHX4
ANKRD22
CATIP
BTBD19
n≥499
n≥1117
Up
Down


Baty_2010(T:29,N:15)
Beer_2002(T:86,N:10)
Jones_2004(T:61,N:19)
Yu_2008(T:18,N:12)
Su_2007(T:27,N:27)
Lu_2010/T:60,N:60)
TCGA_LUAD_2016(T:517,N:59)
Girard_N_b(T:30,N:20)
Rousseaux_2013(T:293, N:14)
TCGA_LUSC_2016(T:501,N:51)
Hou_2010(T:91,N:65)
Bhattacharjee_2001(T:186,N:17)
Selamat_2012(T:58,N:58)
Kabbout_2013(T:80,N:30)
Okayama_2012(T:204,N:20)
Landi_2008(T:58,N:49)
Sanchez-Palencia_2011(T:46,N:45)
Fujiwara_2012a(T:57,N:30)
Xi_2008(T:20,N:20)
Stearman_2005(T:20,N:19)
z: 10.92
p=9.3e-28
z: 5.53
p=3.1e-08
z=3.76
p=0.00017
z: 3.486
p=5e-04
z: -3.02
p=0.0025
z: -7.37
p=1.7e-13
Up
Down
GATA2
IFI27
HBG1
HBG2
LHX4
ANKRD22

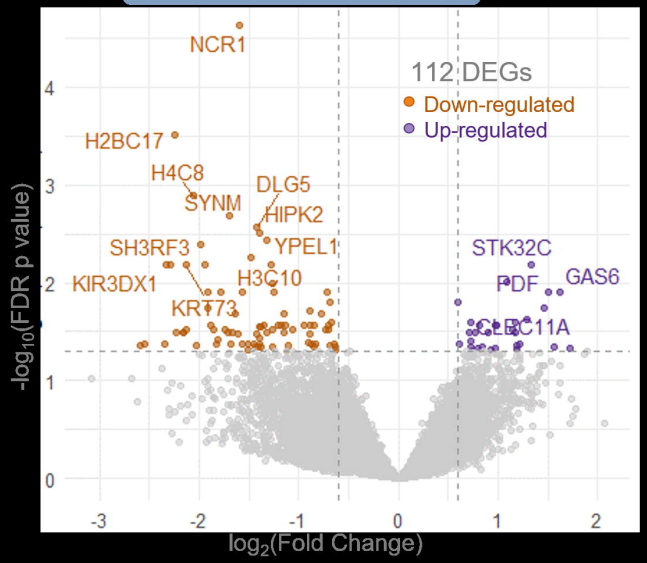


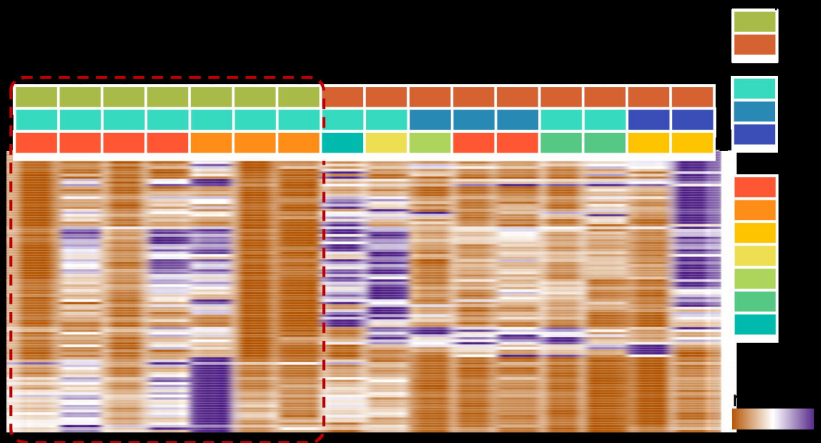
responders v non-responders

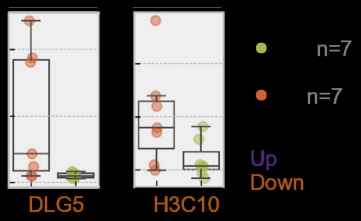
112 DEGs

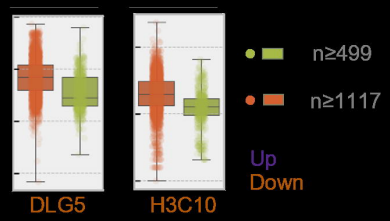
responders v non-responders

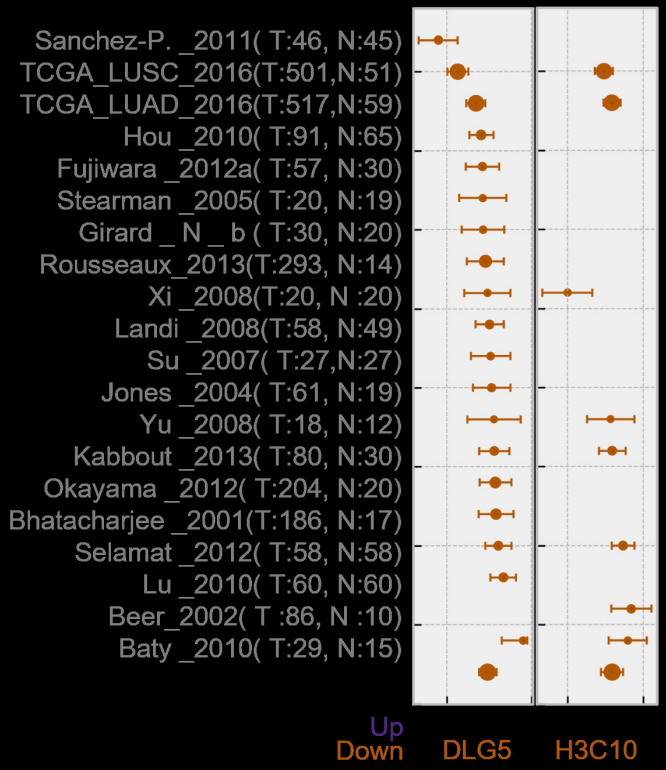
n=7 v 9

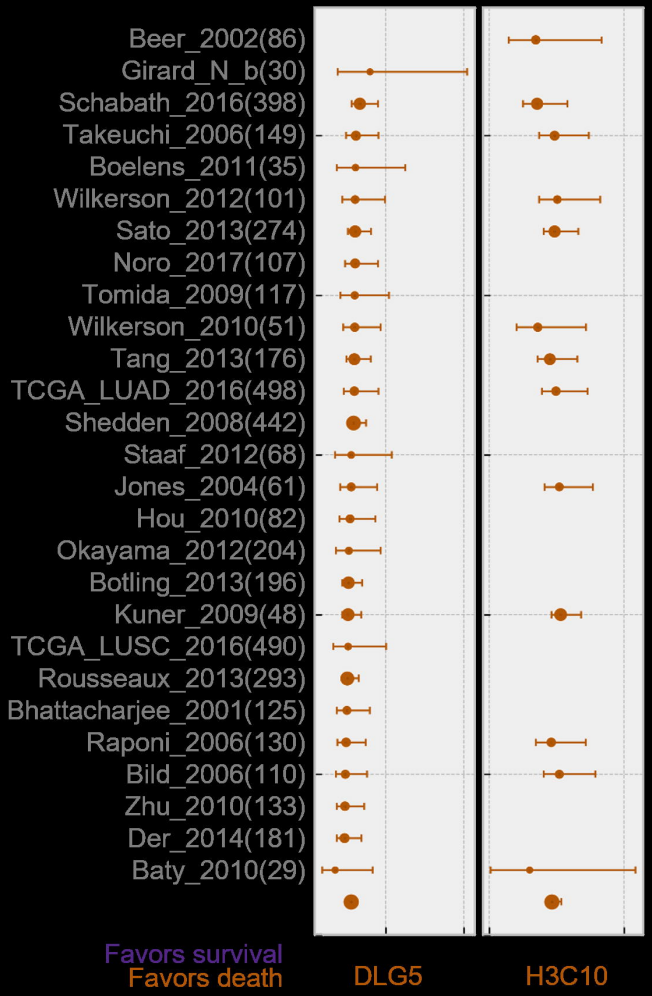


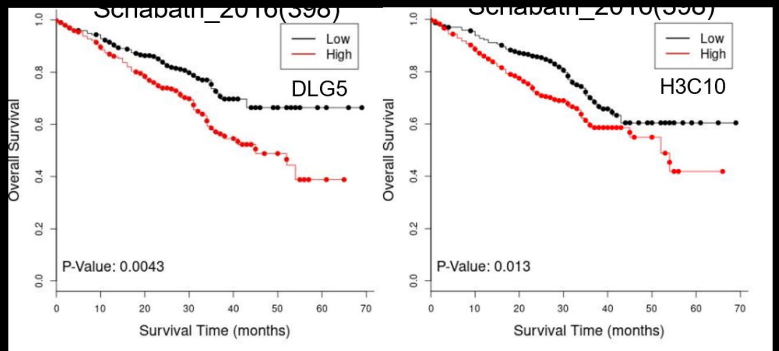


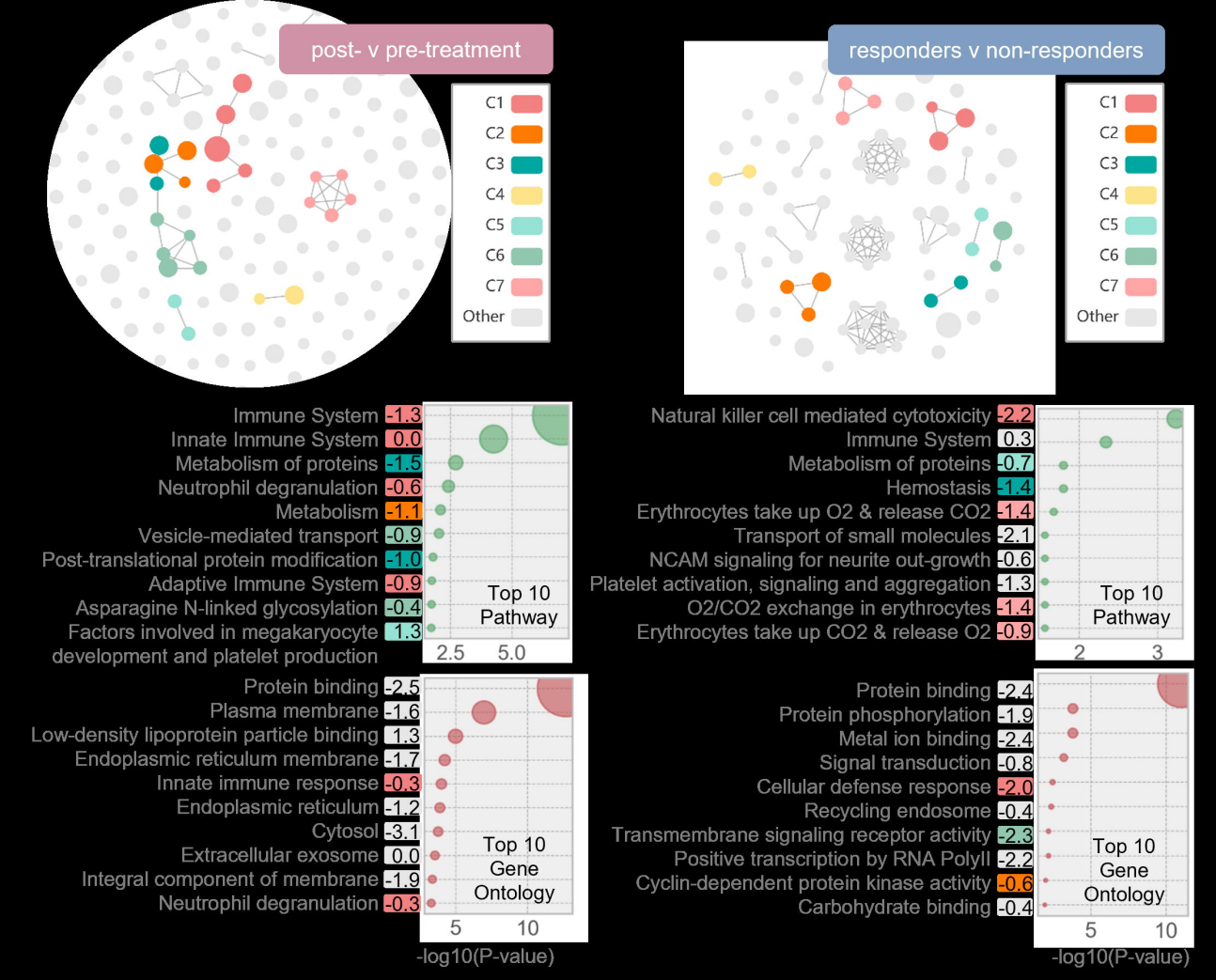


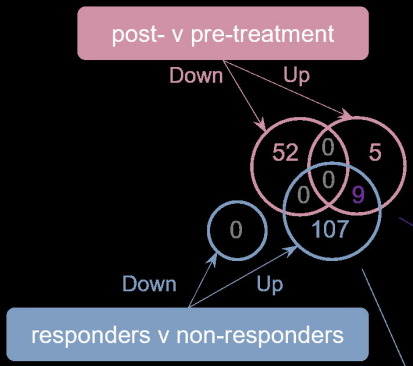



Sanchez-P. _2011(T:46, N:45)
TCGA_LUSC_2016(T:501,N:51)
TCGA_LUAD_2016(T:517,N:59)
Hou _2010(T:91, N:65)
Fujiwara _2012a(T:57, N:30)
Stearman _2005(T:20, N:19)
Girard _ N _ b (T:30, N:20)
Rousseaux_2013(T:293, N:14)
Xi _2008(T:20, N :20)
Landi _2008(T:58, N:49)
Su _2007(T:27,N:27)
Jones _2004(T:61, N:19)
Yu _2008(T:18, N:12)
Kabbout _2013(T:80, N:30)
Okayama _2012(T:204, N:20)
Bhatacharjee _2001(T:186, N:17)
Selamat _2012(T:58, N:58)
Lu _2010(T:60, N:60)
Beer_2002(T :86, N :10)
Baty _2010(T:29, N:15)
Up
Down
DLG5
H3C10

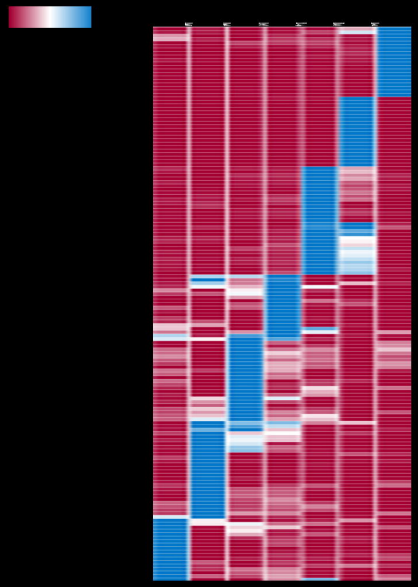
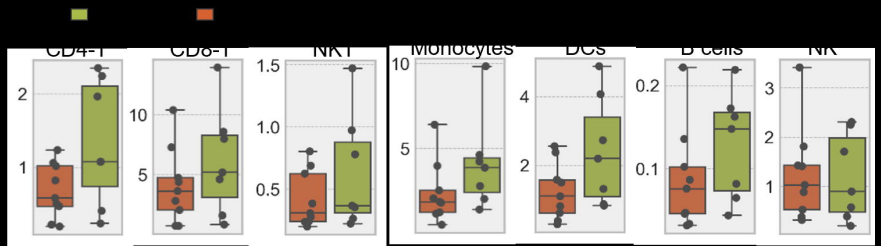

Beer_2002(86)
Girard_N_b(30)
Schabath_2016(398)
Takeuchi_2006(149)
Boelens_2011(35)
Wilkerson_2012(101)
Sato_2013(274)
Noro_2017(107)
Tomida_2009(117)
Wilkerson_2010(51)
Tang_2013(176)
TCGA_LUAD_2016(498)
Shedden_2008(442)
Staaf_2012(68)
Jones_2004(61)
Hou_2010(82)
Okayama_2012(204)
Botling_2013(196)
Kuner_2009(48)
TCGA_LUSC_2016(490)
Rousseaux_2013(293)
Bhattacharjee_2001(125)
Raponi_2006(130)
Bild_2006(110)
Zhu_2010(133)
Der_2014(181)
Baty_2010(29)
Favors survival
Favors death
DLG5
H3C10


Low
High
DLG5
Overall Survival
P-Value: 0.0043
Survival Time (months)
Low
High
H3C10
Overall Survival
P-Value: 0.013
Survival Time (months)


post- v pre-treatment
responders v non-responders
C1
C2
C3
C4
C5
C6
C7
Other
Immune System -1.3
Innate Immune System 0.0
Metabolism of proteins -1.5
Neutrophil degranulation -0.6
Metabolism -1.1
Vesicle-mediated transport -0.9
Post-translational protein modification -1.0
Adaptive Immune System -0.9
Asparagine N-linked glycosylation -0.4
Factors involved in megakaryocyte development and platelet production 1.3
Top 10 Pathway
2.5
5.0
Natural killer cell mediated cytotoxicity -2.2
Immune System 0.3
Metabolism of proteins -0.7
Hemostasis -1.4
Erythrocytes take up O2 & release CO2 -1.4
Transport of small molecules -2.1
NCAM signaling for neurite out-growth -0.6
Platelet activation, signaling and aggregation -1.3
O2/CO2 exchange in erythrocytes -1.4
Erythrocytes take up CO2 & release O2 -0.9
Top 10 Pathway
2
3
Protein binding -2.5
Plasma membrane -1.6
Low-density lipoprotein particle binding 1.3
Endoplasmic reticulum membrane -1.7
Innate immune response -0.3
Endoplasmic reticulum -1.2
Cytosol -3.1
Extracellular exosome 0.0
Integral component of membrane -1.9
Neutrophil degranulation -0.3
Top 10 Gene Ontology
5
10
-log10(P-value)
Protein binding -2.4
Protein phosphorylation -1.9
Metal ion binding -2.4
Signal transduction -0.8
Cellular defense response -2.0
Recycling endosome -0.4
Transmembrane signaling receptor activity -2.3
Positive transcription by RNA PolyII -2.2
Cyclin-dependent protein kinase activity -0.6
Carbohydrate binding -0.4
Top 10 Gene Ontology
5
10
-log10(P-value)



GSE3982_DC_VS_NEUTROPHIL_UP
GSE7764_IL15_TREATED_VS_CTRL_NK_CELL_24H_UP
GSE24634_TEFF_VS_TCONV_DAY7_IN_CULTURE_UP
GSE24634_TEFF_VS_TCONV_DAY5_IN_CULTURE_UP
GSE16450_MATURE_NEURON_CELL_LINE_VS_IMMATURE_UP

GSE22886_DAY7_MONOCYTE_IN_CULTURE_VS_DAY1_UP
GSE22886_NAIVE_TCELL_VS_NEUTROPHIL_UP
GSE22886_NAIVE_CD4_TCELL_VS_NEUTROPHIL_UP
GSE22886_DC_VS_NEUTROPHIL_UP
GSE3982_TH2_VS_EOSINOPHIL_UP
GSE3982_BCELL_VS_NEUTROPHIL_UP
GSE34515_POS_MONOCYTE_VS_CD16_NEG_UP
GSE4748_CYANOBACTERIUM_LPSLIKE_STIM_DC_3H_VS_CTRL_UP
GSE18893_TCONV_VS_TREG_24H_TNF_STIM_UP

GSE22886_MONOCYTE_VS_NEUTROPHIL_UP
GSE22886_NAIVE_TCELL_VS_NKCELL_UP
GSE22886_NAIVE_BCELL_VS_NEUTROPHIL_UP
GSE22886_NAIVE_CD8_TCELL_VS_NKCELL_UP
GSE22886_NAIVE_CD4_TCELL_VS_NKCELL_UP
GSE42724_B1_BCELL_VS_NAIVE_UP
GSE11057_CD4_EFF_MEM_VS_PBMC_UP
GSE11057_CD4_CENT_MEM_VS_PBMC_UP
GSE14000_MRNA_DC_VS_TRANSLATED_RNA_UP
GSE9006_TYPE_1_VS_TYPE_2_DIABETES_PBMC_AT_DX_UP
GSE9006_HEALTHY_VS_TYPE_2_DIABETES_PBMC_AT_DX_UP
GSE45837_GFI1_KO_PDC_VS_WT_UP
GSE34205_HEALTHY_VS_FLU_INF_INFANT_PBMC_UP
GSE2405_9H_A_PHAGOCYTOPHILUM_STIM_NEUTROPHIL_VS_0H_UP
GSE2405_0H_VS_24H_A_PHAGOCYTOPHILUM_STIM_NEUTROPHIL_UP