Hypothesis

Data-driven Understanding On Cancer Lineage Plasticity: Examples and Chances

Longjin Zeng *

Posted Date: 30 December 2024

doi: 10.20944/preprints202412.2473.v1

Keywords: neuroendocrine; squamous; lung; machine learning; overview

Hypothesis

Data-Driven Understanding on Cancer Lineage Plasticity: Examples and Chances

Longjin Zeng

Department of Basic Medicine, Army Medical University, Chongqing 400038, P.R. China; zlj2036360@163.com

Abstract: Advancements in omics technologies have promoted the development of precision oncology. Lineage plasticity, a hallmark of cancer, incorporates molecular and histological aspects. Histological differentiation of adenocarcinoma, neuroendocrine, and squamous characteristics occurs in different anatomic locations. Lung cancer, which is highly heterogeneous, encompasses these differentiations, and therefore serves as a model for exploration. Data-driven understanding is critical in cancer differentiation research, with the two major differentiation pathways, squamous and neuroendocrine, supported by omics data. Here, genetic and non-genetic profiles are reviewed based on patient datasets, and shareable molecular features are described. This paper mainly discusses machine learning approaches to feature selection, where network modeling is effective for designing programmable differentiation. All methods are presented within the context of cancer lineage plasticity along with examples and hypotheses. It emphasizes that selected patient datasets combined with methods will ultimately lead to actionable cancer lineage. Chances for clinical translation are in the spotlight, including biomarkers, molecular subtypes, and targeted therapies.

Keywords: neuroendocrine; squamous; lung; machine learning; overview

Introduction & Background

Conventional cancer therapies target organs or tissues. However, the advent of precision oncology has revolutionized treatment strategies through molecular targeting.[1] In theory, cancer cells deceive normal cells through similar molecules, or manipulate normal cells via specific molecules, i.e., any gene can potentially drive cancer. [2] Trans-differentiation pathways, including neuroendocrine and squamous, are supported by omics data and extend beyond the cellular level. [3] The concept of lineage plasticity, which merges molecular and histologic aspects, can guide drug discovery. [4] Nonetheless, targeting oncogenic lineage-restricted transcription factors (TFs) indiscriminately may lead to severe toxicity and failure to achieve desired tissue specificity. [5]

The accumulation of big data in cancer research has enabled the direct identification of lineage-restricted molecules. Omics studies aim to identify molecular subtypes with biological significance, revealing, for instance, distinct clusters of squamous carcinomas across different tissues. The TCGA consortium's integration of 33 cancer types has facilitated cross-tissue analyses, exploring anatomical systems and original lineages such as the pan-squamous phenotype. [6] Previous studies have strongly demonstrated the molecular and cellular aspects of the pan-squamous phenotype. [7]

Unlike the well-documented pan-squamous phenotype, the neuroendocrine phenotype remains relatively rare. Mechanisms underlying neuroendocrine phenotype lack consensus evidence and are often described as systemic disorders. Current diagnosis of neuroendocrine neoplasms (NENs) relies heavily on neuroendocrine markers, predominantly hormones and bioactive peptides (e.g., *CHGA* and *SYP*). Due to the occurrence of neuroendocrine phenotype in diverse anatomical locations, there remains an unmet need for reliable biomarkers. Progress in this field hinges on developing biomarkers through integrative approaches combining knowledge and omics analyses. [8,9]

The patient-centric model presents tailored opportunities to address rare diseases, a particularly pressing challenge given the large population of China. Data from cell lines or animal models may

 doi:10.20944/preprints202412.2473.v1

not always translate directly to human contexts, highlighting the importance of focusing on patient-derived data and utilizing machine learning with illustrative examples to generate initial hypotheses. The structure of this paper is outlined in **Figure 1**. It is anticipated that data providing access to key features will aid in the development of targeted therapies, and this mini-review aims to discuss the methodologies and their application to relevant diseases.

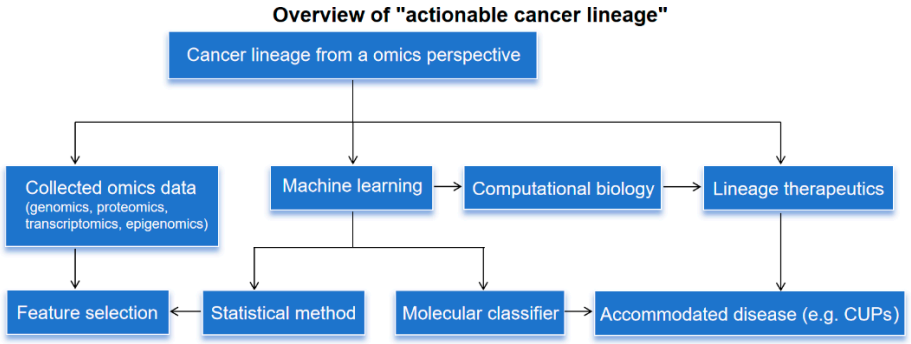

Figure 1. Overview of the "actionable cancer lineage," encompassing sections on omics insights, machine learning, and clinical applications. Cancer data can be used to identify lineage-related molecules as central hubs through feature selection, providing a theoretical foundation for therapeutic strategies and disease understanding. Abbreviation: CUP (cancer of unknown primary).

Review

Omics insights for neuroendocrine and squamous phenotypes

Currently, the characterization of the cancer genome, particularly with regard to coding gene drivers, is well-established. [10,11] Two key frameworks—evolutionary shaping and tissue specificity—have been highlighted in genomic studies. [5,10] This review focuses on shared characteristics between neuroendocrine and squamous phenotypes. First, somatic mutations and copy number variations were analyzed in parallel. Chromosome 3q amplification (such as *TP63* and *SOX2*) was a significant feature of the squamous lineage, and cell cycle dysregulation including *TP53* and *CDKN2A* mutations was also prevalent. [7] Compared to the squamous lineage, NENs are less well understood. [8,9] Recently, NENs have been divided into neuroendocrine tumors (NETs) and neuroendocrine carcinomas (NECs) based on the proliferative index, with point mutations showing tissue-specificity in NETs. [5] *MEN1* is the most significantly mutated gene in NETs, while *TP53* and *RB1* are predominant drivers in NECs. [12–15] These results indicate that genetic events can drive phenotypic changes and influence tumor evolution. [10] Meanwhile, the experimental literature supported these mentioned driver mutations, which will not be expanded in detail here. Beyond that, germline mutations are noteworthy in NETs. And rare variants are expected to be captured in large datasets such as the AACR-GENIE, MSK series, Chinese-OrigiMed, and UK 100,000 genomes projects. Analyzing genomic data retrospectively in combination with NCG annotation can fully unlock the potential of gene therapeutics. [11]

The transcriptome and proteome are commonly applied for quantitative studies from a non-genetic perspective. To reduce off-target effects, widely expressed targets in tumors are preferred in proteome studies, focusing on essential genes rather than lineage specificity. [16–18] Cross-ethnic studies help obtain conservative results, with substantial data accumulated on lung adenocarcinoma. [19] While proteome studies offer the advantage of targetability, detecting low-abundance proteins (e.g., cell surface proteins) remains challenging. These proteins serve as markers for cellular differentiation and lineage. In contrast, the transcriptome does reveal convergent pathways. Neuroendocrine lineage based solely on transcriptome was not favored due to genetic discrepancies (e.g., *RB1* mutations). However, recent data collection has allowed the molecular depiction of the pan-neuroendocrine landscape. [20,21] Meanwhile, convergent biologies between neuroendocrine

 doi:10.20944/preprints202412.2473.v1

and squamous phenotypes are also recognized by epigenetic inheritance. [22,23] Overall, non-genetic components reveal more shareable characteristics compared to genomics.

Despite the existence of convergent pathways against neuroendocrine and squamous phenotypes, there is room for improvement—one direction in neuroendocrine studies, such as small-cell cytology. An interesting finding is that small-cell carcinoma of the bladder is more similar to bladder carcinoma than to small-cell lung cancer, challenging the neuroendocrine phenotype concept. [24] Beyond neuroendocrine markers, the original lineage is the driver of neuroendocrine phenotype. Opinions on transcriptome suggest that neuroendocrine cells originate from the central and peripheral nervous systems, with shared transcription circuits (e.g., *ASCL1* and *NEUROD1* crosstalk) supporting this view. [8,25] In this regard, high-throughput screenings assist in the search for molecules, driven by data.

A bold hypothesis is that neuroendocrine and squamous molecules may regulate each other, possibly due to their mutually exclusive expression profiles. For example, smokers often exhibit neuroendocrine or squamous features in lung cancer. Recent studies also suggest that in advanced metastatic cancers of the bladder and colon, the proportion of neuroendocrine-like subtypes may be increased. [26,27] Another important aspect is that algorithms analyzing the data suggest that localized lineage molecules could represent the broader molecular profile. Given these complexities, discussing neuroendocrine and squamous lineages in isolation, based solely on transcriptomic data, may not be advisable.

Cancer Lineage Plasticity Guided by Machine Learning

Semi-supervised and unsupervised clustering methods

Most statistical methods assume linear fitting, and comparative studies screen to identify conservative features. [28,29] In this field, semi-supervised approaches have been extensively utilized in prognostic markers. The data format can be continuous or binary variables, and the patient's prognosis including follow-up time with survival events is also required. Generally used analyses include Cox regression and the shrinkage method, the latter of which minimize unimportant variables from high-dimensional data. Interestingly, lineage-restricted molecules are often accompanied by prognostic significance (e.g., *NKX2-1*). Apart from this, correlation analyses compare known lineage-restricted molecules with unknown ones, common in feature selection. [5]

Compared with semi-supervised methods, unsupervised clustering is less affected by feature selection. [30] Unsupervised clustering aims to uncover biological signals through dimensionality reduction, without requiring prior knowledge. Importantly, clustering is a population-based design that divides whole data into several groups, which can be complemented with studies of global correlation. How to combine omics and machine learning methods has been described in detail elsewhere, and will not be discussed here. Multi-omics may retain strong signals better than individual omics, as opposed to individual omics, which is suitable for external validation. Given the higher weight of underlying biological pathways, transcriptome data was often preferred in multi-omics analysis due to its high variability. [31] However, using TCGA consortium data alone underestimates protein levels, addressed by the CPTAC consortium. [17]

Subtype classifiers: training and external validation

Subclass classifiers are a method for solving classification issues that usually require given sample labels. The principle of "subtype classifiers" is training molecules concerning sample labels, and then using them for external validation. This methodology reduces the number of variables, and the assignment probability of individual samples is determined by a small number of feature comparisons. [32] Importantly, differences in technology must be considered (e.g., RNA sequencing vs. Microarrays). In cancer, molecular classifiers are mainly tissue-based. [33] Moreover, a new study suggests that original lineage is also practical by trained classifiers. [34] In classification tasks, DNA methylation arrays are more sensitive than RNA sequencing but struggle to detect detailed cluster information. [35] The available evidence suggests that both DNA methylation and transcriptome are feasible for classification, with transcriptome performing well in clustering scenarios.

 doi:10.20944/preprints202412.2473.v1

Computational biology: focus on quantitative & qualitative

Compared to the other two approaches, computational biology can use high-throughput data combined with cellular experimental data. Key challenges in cancer include that lineage-restricted molecules might be undruggable and play essential roles in normal tissues. [36] Thus, constructing regulatory networks or targeting operative molecules is warranted. Network modeling to design gene circuits is gaining attention, with cell line data supporting signal transduction studies. Previous studies proved that programmable gene editing led to direct differentiation. [37] In lung cancer, genetic circuits can be engineered to block triple differentiation (adenocarcinoma-neuroendocrine-squamous). Candidates perhaps are the KRAB zinc finger protein family, which may persistently inhibit the differentiation program due to the fact that this family often exhibits transcriptional repression. [38] Furthermore, there may be some families or molecules that regulate differentiation in future studies.

Genetic circuits are designed as pre-simulated molecular networks that are later validated by specific molecules. The most classical approach is to combine network simulations and high-throughput screening to identify factors of direct differentiation. In large-scale level studies, it is preferable to study the entire gene family rather than a single member, emphasizing family characteristics. If this factor is untargetable, then the alternative strategy suggests to change to its collaborators. This review outlines the blocking of multiple differentiation through transcriptional repression and the establishment of regulatory networks for each lineage. As with all, computational biology is at the root of genetic circuits.

Computational biology embraced network modeling, which required a combination of quantitative and qualitative approaches, usually in the context of cell lines. Quantitative research found global transcription rates can be inferred using the network modeling method named GENIE3. [39] It can rapidly calculate linkages between genes, and tools like it include ARACNe, CellNet and Mogrify. Compared to quantitative approaches, Boolean modeling is a classic tool in qualitative studies. [40,41] Boolean models require existing phenotypes, such as taking *ΔNp63* (Truncated isoform of *TP63*) as an input and squamous markers as an output. The relationship between genes is characterized as binary variables (activation or inhibition) based on experimental data, with middle regulations being molecules of interest. These can be kinases and chromatin regulatory factors, and there are already clinical trials underway. [3,4,16]

Clinical Applications and Future Directions

Lung cancer - a molecular subtype model

Lineage-restricted TFs were widely recognized diagnostic markers used in immunohistochemistry to distinguish histology. In contrast, molecular subtypes are emerging as an important focus. While molecular subtypes can be validated using clinical samples, translating these findings into routine clinical practice remains challenging. Each subtype may have a biological basis, but conserved signals across subtypes hold greater value. [2,15] The ideal subtype derived from a single tissue type should be mappable to anatomical systems or original lineages, meaning it can be applied in both horizontal and longitudinal comparative studies. For instance, overexpressed molecules associated with the three histological types of lung cancer can be replicated in lung adenocarcinoma (~10% Jaccard index, using GSE94601 as reference data; **Supplementary Table S1**). Similar findings were observed in Lund's advanced bladder cohort. [42] These results suggest that convergent pathways can be identified across different organs, potentially serving as candidate metagenes (**Supplementary Table S2**). It could be argued that the molecular drivers behind subtypes are more important than the subtypes themselves.

Modified subtypes in lung adenocarcinoma and lung squamous carcinoma can accomplish the above promises. This hypothesis is that lung adenocarcinoma is a subset of lung cancer, and lung squamous carcinoma is the pan-squamous miniature.

Assuming that genes with a high mRNA-protein correlation were taken as an assessment criterion for being activated, RNA processing and extracellular matrix pathways differed between

 doi:10.20944/preprints202412.2473.v1

subtypes in lung squamous carcinoma. [43] This observation aligns with findings in pan-squamous Chinese proteomic analysis. [44] Given the large differences in numbers between subtypes (LSQ1: 1090 vs LSQ2: 299, based on mRNA-protein correlation; $|\Delta_r| > 0.3$), this may be caused by post-transcriptional regulation. Regarding post-transcriptional regulation, the coefficient of variation of five molecules associated with RNA processing, RBM10, SF1, CPSF6, SLTM and DDX5, were lower than those of KRT5. It is suggested that they may be less prone to off-target effects. [16] If this assumption is plausible, then extracellular matrix pathways should play an activated role in the cancer epithelium. Further, the understanding of epithelial mechanotransduction can be aided by high-dimensional analysis at spatial resolution. [45]

Cell-of-origin pairing for cancers of unknown origin and rare diseases

This section named "Cell-of-origin pairing for cancers of unknown origin and rare diseases" has some conceptual overlap with the "subtype classifiers" described above, as both involve scenario simulations of classification issues. Clinical translation of cell-of-origin pairing is appropriate in cancers of unknown origin and rare diseases. Consider platforms for sequencing, DNA methylation and transcriptome analyses performed well for categorization in both primary and unknown origin cancers. [28,46] A lack of approved therapies is a common problem in this field, including cancers of unknown origin and rare diseases. For this, tissue-agnostic therapies, such as genomics-guided and lineage-based therapies, are suggested in emerging viewpoints. [1] Lineage-based therapies allow for simultaneous consideration of diagnosis and treatment, offering broad applications. For some diseases, ethnicity differences are determinative; for instance, esophageal squamous carcinoma is highly prevalent in China, while esophageal adenocarcinoma is more prevalent in Caucasians. [9] This is why exploring the differences between adenocarcinoma and squamous carcinoma makes sense.

Lineage-based concepts also guide drug discovery in rare diseases. One hopeful chance is that olfactory neuroblastoma mimics small-cell lung cancer. Clustering analysis also appreciates the rationality of this measure because of convergent signaling. For rare diseases, the greatest problem is the lack of appropriate control samples, and outlier analysis may bring in targets although the efficacy needs to be improved. Theoretically, population-based N-of-1 trials can yield fast-track molecular insights. To support these, the Treehouse group proposes outlier analysis in pediatric cancer based on N-of-1 design. Firstly, the background data should be established, and individual patients were then compared against it to identify dysregulated molecules. [47] As an example, if a new patient has squamous cytology and has completed mass spectrometry, the aberrant targets can be recognized using the Chinese pan-squamous proteome. [44]

Targeted therapy design

Lineage-based targeted therapies have currently reached the preclinical stage, with early interception and refractory metastasis being prioritized. [3,4] Carcinoma in situ was a pre-invasive state that could further progress to invasive cancer, and early intervention may prevent cancer progression. Pre-invasive genetic drivers are few, but once formed, such conditions are often "irreversible." The key challenge is that directly targeting TFs may lead to de novo trans-differentiation, such as *NKX2-1* in lung adenocarcinoma and *ΔNp63* in lung squamous carcinoma. For trans-differentiation studies mainly including Pre- and post-transformation paired and direct collection of post-transformation samples, the former is temporarily limited by the lack of sufficient sample sizes. Refractory tumors collected directly, such as lung adenosquamous cell carcinoma, require pathological evidence, which has been well-studied by Ji's laboratory. [48] Early intervention is particularly emphasized, and lineage-restricted molecules serve as diagnostic markers. Furthermore, the therapeutic window needs comprehensive assessment, considering factors like the interaction of environment and genes. [49] Because disease regression took a long time, the results may not be sufficient for replication. Importantly,

squamous differentiation should intervention could be inferred from GSE108082 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108082) and remained a lineage feature of squamous cell carcinoma. [50] It is desirable to obtain molecules that arrest the

 doi:10.20944/preprints202412.2473.v1

squamous differentiation and influence disease regression. In addition, identifying highly selective targets by gene circuits, such as Boolean modeling, is suggested. The choice of targeting design is crucial, and in addition, it should prioritize highly heterogeneous or transcription-driven cancers, such as some pediatric cancers.

Lung cancer is highly heterogeneous, consisting mainly of adenocarcinoma, squamous and neuroendocrine lineage, and small cell lung cancer being the most aggressive and representative type of neuroendocrine carcinoma. Lineage-based targeted therapies are particularly indicated for highly heterogeneous carcinoma in the precancerous stage. Precancerous patients benefit from early detection and intervention, which can boost survival or lower morbidity. In this regard, three therapeutic options have been proposed: 1) prevention of high-risk precancerous populations by block differentiation, available for instance in the KRAB zinc finger family; 2) direct targeting of TFs in the pre-invasive or refractory state by library screening, such as chemical probes and loss or gain of function; 3) programmable circuit design for highly selective targeting from downstream, upstream or co-factors. This is a universal design of lineage-restricted TFs, straightforward targeting. Exceptions include nuclear signaling receptors like *AR, ER, MYC*, etc. [37,51,52] For these, three possible options for targeting were summarised in **Figure 2**.

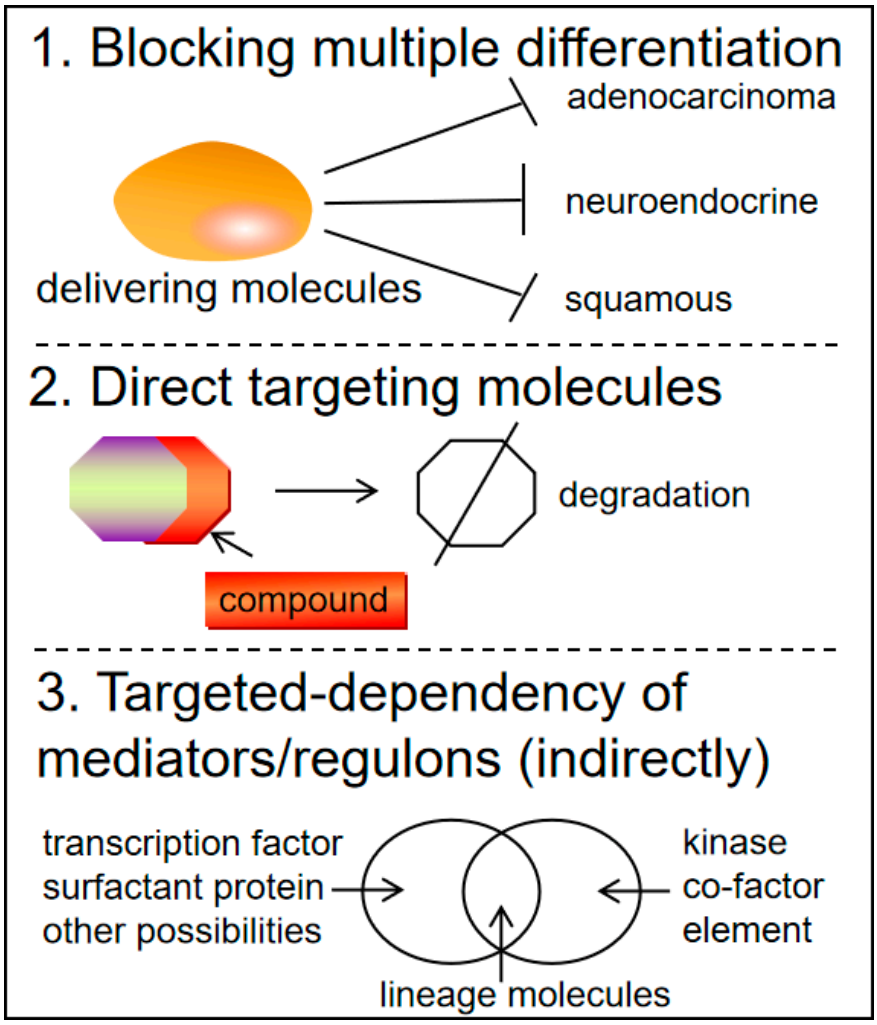

Figure 2. Exploration of lineage-based therapeutics targeting both multiple and single differentiation pathways. One approach is based on the hypothesis that histologic phenotypes serve as a background. The other focuses on directly targeting compounds within monodifferentiation pathways or on mediators that regulate downstream molecules. For downstream markers, the keratin family may be involved. The key drivers with the greatest potential are transcription factors and cell surface proteins.

 doi:10.20944/preprints202412.2473.v1

In all, the above analyses require identifying molecules with prognostic significance, incorporating convergent subtypes, etc., all of which require machine learning as a foundation. Data-driven initial understanding of molecular properties, such as whether they are associated with the cancer lineage, and from genes to diseases is the call of precision medicine.

Precision medicine vision

As interest and efforts in precision oncology escalate, recognizing the importance of biomarkers and their use in developing targeted therapies in clinical research is indispensable. Significant methodological advancements in genomics-guided clinical trial designs, such as basket and umbrella trials within the master protocol framework, have been made. However, umbrella trials like NCT02154490 and NCT03292250 have shown unsatisfactory results. [53] In contrast to genomics-guided clinical trial design, lineage-based concepts hold promise for enhancing outcomes for shared targets, with *DLL3* as a successful example. Data-driven analysis not only enhances drug discovery for targeted therapies but also holds equal importance for immunotherapy and traditional treatment modalities. Lung cancer, being the most prevalent and fatal malignancy, already possesses a considerable amount of data that could be leveraged for clinical translation. [54]

To enhance the operational effectiveness of cancer lineage plasticity, patient-centric datasets, and illustrative examples have been curated. The utilization of freely available public resources, where academic advancements surpass financial incentives, remains crucial. Transcriptome has become routine in disease studies, with TFs likely playing a key role in driving cellular fate. In this context, the development of dedicated TFs platform (e.g., TFome™) may prove beneficial. Certain TFs exhibit cancer-specific expression patterns, which can be analyzed through regulated networks. [55] Up to now, clinical translation has successfully incorporated RNA-related products (e.g., Oncotype DX® and CancerTYPE ID®) primarily for prognostic stratification and origin classification. Looking ahead, liquid biopsies show promise in determining tumor origins compared to traditional tissue specimens. Additionally, integrating TF data with DNA methylation profiles represents a promising frontier. High-resolution data continues to enrich our understanding of life sciences, yet it's important to acknowledge limitations such as oversimplified assumptions, binary classifications, and the exclusive focus on TFs and transcriptomics.

Conclusions

Population-based data analysis enables the identification of cancer lineage factors, emphasizing the importance of sharing over differentiation. In genetics, the notable feature of the pan-squamous phenotype is chromosome 3q amplification. While mutations *MEN1* and *TP53- RB1* are predominantly enriched in neuroendocrine tumors and neuroendocrine carcinomas, respectively. The non-genetic part of the collection mainly consists of the transcriptome, proteome, and epigenome all associated with cancer lineage. For precision therapy, one possible approach is to design gene circuits to reduce tissue toxicity and induce direct differentiation. Furthermore, original lineage therapy should focus on early interception and provide insights into pre-invasive, refractory, rare, and unknown primary cancers. By the way, data collection involves assumptions beneath it; for example, follow-up data in cancer are often predicated on the use of radiochemotherapies, and aiming to develop novel therapies demands prospective validation.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author contributions: Longjin Zeng wrote, conceptualized, and interpreted the manuscript.

Acknowledgments: Only articles with usable data or illustrations are cited; I sincerely apologize for not being able to discuss all authors and their respective studies. In addition, thanks to neuronal perspectives from Julien Sage, and Vân Anh Huynh-Thu, co-author of gene regulatory network tools. I am also deeply grateful for the support of familiar colleagues, and the inspiration drawn from Lund University. Most importantly, thanks to the data contributors.

 doi:10.20944/preprints202412.2473.v1

Conflict of interest statement: No conflict of interest was declared.

Data availability statement: Data sharing is not applicable as no new data was generated.

References

1. Subbiah V, Gouda MA, Ryll B, Burris HA, 3rd, et al. 2024. The evolving landscape of tissue-agnostic therapies in precision oncology. CA: a cancer journal for clinicians.
2. de Magalhães JP. 2022. Every gene can (and possibly will) be associated with cancer. Trends in genetics : TIG 38: 216-7.
3. Davies A, Zoubeidi A, Beltran H, Selth LA. 2023. The Transcriptional and Epigenetic Landscape of Cancer Cell Lineage Plasticity. Cancer Discov 13: 1771-88.
4. Fujii M, Sekine S, Sato T. 2024. Decoding the basis of histological variation in human cancer. Nature reviews Cancer 24: 141-58.
5. Haigis KM, Cichowski K, Elledge SJ. 2019. Tissue-specificity in cancer: The rule, not the exception. Science (New York, NY) 363: 1150-1.
6. Hoadley KA, Yau C, Hinoue T, Wolf DM, et al. 2018. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 173: 291-304.e6.
7. Guan Y, Wang G, Fails D, Nagarajan P, et al. 2020. Unraveling cancer lineage drivers in squamous cell carcinomas. Pharmacology & therapeutics 206: 107448.
8. Rindi G, Inzani F. 2020. Neuroendocrine neoplasm update: toward universal nomenclature. Endocrine-related cancer 27: R211-r8.
9. Xue J, Lyu Q. 2024. Challenges and opportunities in rare cancer research in China. Science China Life sciences 67: 274-85.
10. Martincorena I, Raine KM, Gerstung M, Dawson KJ, et al. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 17A1: 1029-41.e21.
11. Dressler L, Bortolomeazzi M, Keddar MR, Misetic H, et al. 2022. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. Genome biology 23: 35.
12. Cao Y, Zhou W, Li L, Wang J, et al. 2018. Pan-cancer analysis of somatic mutations across 21 neuroendocrine tumor types. Cell research 28: 601-4.
13. Wu H, Yu Z, Liu Y, Guo L, et al. 2022. Genomic characterization reveals distinct mutation landscapes and therapeutic implications in neuroendocrine carcinomas of the gastrointestinal tract. Cancer communications (London, England) 42: 1367-86.
14. van Riet J, van de Werken HJG, Cuppen E, Eskens F, et al. 2021. The genomic landscape of 85 advanced neuroendocrine neoplasms reveals subtype-heterogeneity and potential therapeutic targets. Nature communications 12: 4612.
15. Yachida S, Totoki Y, Noë M, Nakatani Y, et al. 2022. Comprehensive Genomic Profiling of Neuroendocrine Carcinomas of the Gastrointestinal System. Cancer discovery 12: 692-711.
16. Chang L, Ruiz P, Ito T, Sellers WR. 2021. Targeting pan-essential genes in cancer: Challenges and opportunities. Cancer Cell 39: 466-79.
17. Savage SR, Yi X, Lei JT, Wen B, et al. 2024. Pan-cancer proteogenomics expands the landscape of therapeutic targets. Cell 187: 4389-407.e15.
18. Zhou Y, Lih TM, Pan J, Höti N, et al. 2020. Proteomic signatures of 16 major types of human cancer reveal universal and cancer-type-specific proteins for the identification of potential therapeutic targets. Journal of hematology & oncology 13: 170.

 doi:10.20944/preprints202412.2473.v1

19. Vavilis T, Petre ML, Vatsellas G, Ainatzoglou A, et al. 2024. Lung Cancer Proteogenomics: Shaping the Future of Clinical Investigation. Cancers 16.
20. Chen F, Zhang Y, Gibbons DL, Deneen B, et al. 2018. Pan-Cancer Molecular Classes Transcending Tumor Lineage Across 32 Cancer Types, Multiple Data Platforms, and over 10,000 Cases. Clin Cancer Res 24: 2182-93.
21. Wang Z, Liu C, Zheng S, Yao Y, et al. 2024. Molecular subtypes of neuroendocrine carcinomas: A cross-tissue classification framework based on five transcriptional regulators. Cancer Cell 42: 1106-25.e8.
22. Cejas P, Xie Y, Font-Tello A, Lim K, et al. 2021. Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer. Nature communications 12: 5775.
23. Terekhanova NV, Karpova A, Liang WW, Strzalkowski A, et al. 2023. Epigenetic regulation during cancer transitions across 11 tumour types. Nature 623: 432-41.
24. Chang MT, Penson A, Desai NB, Socci ND, et al. 2018. Small-Cell Carcinomas of the Bladder and Lung Are Characterized by a Convergent but Distinct Pathogenesis. Clin Cancer Res 24: 1965-73.
25. Tsunemoto R, Lee S, Szűcs A, Chubukov P, et al. 2018. Diverse reprogramming codes for neuronal identity. Nature 557: 375-80.
26. Loriot Y, Kamal M, Syx L, Nicolle R, et al. 2024. The genomic and transcriptomic landscape of metastastic urothelial cancer. Nature communications 15: 8603.
27. Moorman AR, Benitez EK, Cambuli F, Jiang Q, et al. 2024. Progressive plasticity during colorectal cancer metastasis. Nature.
28. Heinze G, Wallisch C, Dunkler D. 2018. Variable selection - A review and recommendations for the practicing statistician. Biometrical journal Biometrische Zeitschrift 60: 431-49.
29. Zhang JZ, Wang C. 2023. A comparative study of clustering methods on gene expression data for lung cancer prognosis. BMC research notes 16: 319.
30. Chalise P, Kwon D, Fridley BL, Mo Q. 2023. Statistical Methods for Integrative Clustering of Multi-omics Data. Methods in molecular biology (Clifton, NJ) 2629: 73-93.
31. Mamatjan Y, Agnihotri S, Goldenberg A, Tonge P, et al. 2017. Molecular Signatures for Tumor Classification: An Analysis of The Cancer Genome Atlas Data. The Journal of molecular diagnostics : JMD 19: 881-91.
32. Wu Y, Huang HC, Qin LX. 2021. Making External Validation Valid for Molecular Classifier Development. JCO precision oncology 5.
33. Xu Q, Chen J, Ni S, Tan C, et al. 2016. Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 29: 546-56.
34. Rydzewski NR, Shi Y, Li C, Chrostek MR, et al. 2024. A platform-independent AI tumor lineage and site (ATLAS) classifier. Communications biology 7: 314.
35. Xia D, Leon AJ, Cabanero M, Pugh TJ, et al. 2020. Minimalist approaches to cancer tissue-of-origin classification by DNA methylation. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 33: 1874-88.
36. Cruz FD, Matushansky I. 2012. Solid tumor differentiation therapy - is it possible? Oncotarget 3: 559-67.
37. Prasad K, Cross RS, Jenkins MR. 2023. Synthetic biology, genetic circuits and machine learning: a new age of cancer therapy. Molecular oncology 17: 946-9.
38. Cappelluti MA, Mollica Poeta V, Valsoni S, Quarato P, et al. 2024. Durable and efficient gene silencing in vivo by hit-and-run epigenome editing. Nature 627: 416-23.

 doi:10.20944/preprints202412.2473.v1

39. Zhang J, Han X, Ma L, Xu S, et al. 2023. Deciphering a global source of non-genetic heterogeneity in cancer cells. Nucleic Acids Res 51: 9019-38.
40. Kim N, Hwang CY, Kim T, Kim H, et al. 2023. A Cell-Fate Reprogramming Strategy Reverses Epithelial-to-Mesenchymal Transition of Lung Cancer Cells While Avoiding Hybrid States. Cancer research 83: 956-70.
41. Montagud A, Béal J, Tobalina L, Traynard P, et al. 2022. Patient-specific Boolean models of signalling networks guide personalised treatments. eLife 11.
42. Sjödahl G, Eriksson P, Liedberg F, Höglund M. 2017. Molecular classification of urothelial carcinoma: global mRNA classification versus tumour-cell phenotype classification. The Journal of pathology 242: 113-25.
43. Arad G, Geiger T. 2023. Functional Impact of Protein-RNA Variation in Clinical Cancer Analyses. Molecular & cellular proteomics : MCP 22: 100587.
44. Song Q, Yang Y, Jiang D, Qin Z, et al. 2022. Proteomic analysis reveals key differences between squamous cell carcinomas and adenocarcinomas across multiple tissues. Nature communications 13: 4167.
45. Li J, Ma J, Zhang Q, Gong H, et al. 2022. Spatially resolved proteomic map shows that extracellular matrix regulates epidermal growth. Nature communications 13: 4012.
46. Möhrmann L, Werner M, Oleś M, Mock A, et al. 2022. Comprehensive genomic and epigenomic analysis in cancer of unknown primary guides molecularly-informed therapies despite heterogeneity. Nature communications 13: 4485.
47. Vivian J, Eizenga JM, Beale HC, Vaske OM, et al. 2020. Bayesian Framework for Detecting Gene Expression Outliers in Individual Samples. JCO clinical cancer informatics 4: 160-70.
48. Tang S, Xue Y, Qin Z, Fang Z, et al. 2023. Counteracting lineage-specific transcription factor network finely tunes lung adeno-to-squamous transdifferentiation through remodeling tumor immune microenvironment. National science review 10: nwad028.
49. Stanton SE, Castle PE, Finn OJ, Sei S, et al. 2024. Advances and challenges in cancer immunoprevention and immune interception. Journal for immunotherapy of cancer 12.
50. Teixeira VH, Pipinikas CP, Pennycuick A, Lee-Six H, et al. 2019. Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. Nature medicine 25: 517-25.
51. Chen A, Koehler AN. 2020. Transcription Factor Inhibition: Lessons Learned and Emerging Targets. Trends in molecular medicine 26: 508-18.
52. Shin HY. 2018. Targeting Super-Enhancers for Disease Treatment and Diagnosis. Molecules and cells 41: 506-14.
53. Hayes DN, Oluoha O, Schwartz DL. 2024. For Squamous Cancers, the Streetlamps Shine on Occasional Keys, Most Baskets Are Empty, and the Umbrellas Cannot Keep Us Dry: A Call for New Models in Precision Oncology. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 42: 487-90.
54. Cai L, Xiao G, Gerber D, J DM, et al. 2022. Lung Cancer Computational Biology and Resources. Cold Spring Harbor perspectives in medicine 12.
55. Huang X, Song C, Zhang G, Li Y, et al. 2024. scGRN: a comprehensive single-cell gene regulatory network platform of human and mouse. Nucleic Acids Res 52: D293-d303.