Comparative genomic profiling identifies targetable brain metastasis drivers in non-small cell lung cancer

Marcin Nicos
Medical University of Lublin
Luuk Harbers
Karolinska Institute https://orcid.org/0000-0003-3910-6497
Enrico Patrucco
University of Turin
Maximilian Kramer-Drauberg
University of Turin
Xiaolu Zhang
School of Basic Medicine, Shandong University https://orcid.org/0000-0002-9350-2880
Claudia Voena
University of Turin
Anna Kowalczyk
Medical University of Gdansk
Aleksandra Bozyk
Medical University of Lublin
Rafal Peksa
Medical University of Gdansk https://orcid.org/0000-0002-4904-7059
Bozena Jarosz
Medical University of Lublin
Justyna Szumilo
Medical University of Lublin
Michele Simonetti
Karolinska Institutet https://orcid.org/0000-0003-3322-1697
Monika Zuk
Medical University of Gdansk
Bartosz Wasag
Medical University of Gdansk https://orcid.org/0000-0002-3634-7562
Katarzyna Reszka
Medical University of Lublin
Renata Duchnowska

Medical University of Gdansk https://orcid.org/0000-0002-9272-3462
Janusz Milanowski
Medical University of Lublin
Roberto Chiarle
University of Turin
Magda Bienko
Karolinska Institutet https://orcid.org/0000-0002-6499-9082
Pawel Krawczyk
Medical University of Lublin
Jacek Jassem
Medical University of Gdańsk https://orcid.org/0000-0002-8875-6747
Chiara Ambrogio
https://orcid.org/0000-0003-4122-701X
Nicola Crosetto (**nicola.crosetto@scilifelab.se**)
Karolinska Institutet https://orcid.org/0000-0002-3019-6978

Article

Keywords:

Posted Date: February 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1320380/v1

Abstract

Brain metastases (BM) severely impact the prognosis and quality of life of patients with non-small cell lung cancer (NSCLC). To identify targetable drivers of NSCLC-BM, we profiled somatic copy number alterations (SCNAs) in 51 matched pairs of primary NSCLC and BM samples from 33 patients with lung adenocarcinoma (LUAD) and 18 patients with lung squamous cell carcinoma (LUSC). BM consistently had a higher burden of SCNAs compared to the matched primary tumors, and SCNAs were typically homogeneously distributed within BM, as revealed by multi-region SCNA profiling in 15 BM samples, suggesting that BM do not undergo extensive evolution once formed. By comparing focal SCNAs in matched NSCLC-BM pairs, we identified BM driving alterations affecting multiple cancer genes, including several targetable genes such as CDK12, DDR2, ERBB2, and NTRK1, which we validated in an independent cohort of 84 BM samples. We explored the metastatic potential of CDK12 and DDR2 in vitro and in vivo and found that overexpression of either gene alone in murine lung cancer cells causes the induction of key genes involved in epithelial-mesenchymal transition. Finally, we performed whole-exome sequencing of 40 NSCLC-BM pairs and identified BM driver mutations in multiple cancer genes, including several genes involved in epigenome editing and 3D genome organization, such as EP300, CTCF, and STAG2, which we validated by targeted sequencing of an independent cohort of 115 BM samples. Our study represents the most comprehensive genomic characterization of LUAD and LUSC BM available to date, uncovering potentially targetable NSCLC-BM drivers and inspiring the design of novel precision treatment strategies for NSCLC patients.

Main

Brain metastases (BM) are detected in 20–40% of non-small cell lung cancer (NSCLC) patients at the time of diagnosis and eventually 50% of the patients succumb to them[1]. Due to limited blood-brain barrier permeability, NSCLC-BM are refractory to most cytotoxic agents and immunotherapeutics[2,3]. Next-generation EGFR and ALK inhibitors as well as ROS1, RET, MET, and NTRK inhibitors that penetrate the blood-brain barrier have shown promising activity against BM in NSCLC patients whose primary tumors carry mutations in these genes[4,5]. However, it remains critical to identify specific drivers of NSCLC-BM that can provide novel and more effective therapeutic targets.

Relatively few studies have mapped the genomic landscape of BM in general and, specifically, of NSCLC-BM. In a large-scale pan-cancer study, 2,583 metastases from 20 different tumor types were profiled by whole-genome sequencing[6]. However, only one sample was a NSCLC-BM (see Supplementary Table 2 in ref. [6]). In another study, 73 pairs of lung cancer adenocarcinoma (LUAD) and matched BM were profiled by whole-exome sequencing (WES) and low-pass whole-genome sequencing (WGS) for calling somatic copy number alterations (SCNAs)[7]. This study led to the identification of *MYC*, *YAP1*, and *MMP13* amplifications and *CDKN2A/B* deletions as LUAD-BM drivers[7]. More recently, 12 pairs of primary NSCLC and matched BM were profiled by WES, leading to the identification of several BM-associated mutations

in known cancer genes, including *AHNAK2*, *ANKRD36C*, *BAGE2*, *KMT2C*, and *PDE4DIP*[8]. However, none of these studies identified NSCLC-BM drivers that can be targeted with available drugs. Therefore, further investigations aimed at identifying targetable NCSCL-BM drivers are needed.

To this end, here we performed a comparative genomic characterization of 51 pairs of primary NSCLC samples encompassing the two main histological subtypes (LUAD and LUSC) and matched BM. This led us to identify several plausible NSCLC-BM driving genomic events and potentially targetable genes, including *CDK12*, *DDR2*, *ERBB2*, and *NTRK1*, which we validated in an independent cohort of 84 (for SCNAs) and 115 (for mutations) BM samples. We investigated the metastatic potential of two putative BM driving events—amplification of *CDK12* and *DDR2*—by overexpressing these genes in mouse NSCLC cell lines and assessing the induction of epithelial-mesenchymal transition *in vitro* and the formation of BM *in vivo*. Our results provide the most comprehensive compendium of targetable NSCLC-BM genetic drivers to date, uncovering many NSCLC-BM driving genetic events and several potentially targetable genes that are frequently altered in BM.

Results

NSCLC-BM harbor more copy number alterations compared to matched primary tumors

To identify novel genetic drivers of NSCLC-BM, we retrospectively collected 51 pairs of formalin-fixed paraffin-embedded (FFPE) primary NSCLC and BM samples from 33 patients with LUAD and 18 patients with LUSC (discovery cohort) (**Fig. 1a, Supplementary Table 1 and Methods**). We first profiled SCNAs in these samples by performing low-pass WGS and determining DNA copy number levels using two different callers (QDNAseq[9] and CNVkit[10]) (**Supplementary Table 2 and Methods**). Since both callers yielded similar copy number profiles (mean Pearson's correlation coefficient, PCC: 0.92 at 50 kilobase, kb resolution), we decided to conduct all subsequent analyses using QDNAseq given that this caller is tailored for FFPE samples[9] (**Supplementary Fig. 1a, b**). Visual inspection of the copy number profiles revealed that BM had consistently more SCNAs than the corresponding primary tumors (**Fig. 1b**). Accordingly, the SCNA burden (*i.e.*, the fraction of the genome amplified or deleted) was significantly higher in BM, independently of the primary tumor histology (**Fig. 1c, d, Supplementary Fig. 1c-f, and Supplementary Table 3**). Most SCNAs were medium-sized (1–10 megabases, Mb), followed by focal (<1 Mb) and large (>10 Mb) alterations (**Fig. 1e**). BM had significantly more large alterations than primary tumors, whereas the latter harbored significantly more focal SCNAs (**Fig. 1e**). Amplifications were more frequent along the q-arm of chromosome (chr) 1, 3, 8, and 17 and along the p-arm of chr5, 7, and 20, whereas deletions were more frequent on chr3p, 4p, 5q, 8p, 9p, and 18q (**Fig. 1f-h**). Recurrent BM-associated amplifications on chr5p and 8q were also previously identified in a small cohort of 7 primary lung cancer and BM pairs profiled by WES[11].

To corroborate these results, we applied an orthogonal non-sequencing-based method (NanoString[12]) to quantify the copy number of 87 genes frequently amplified or deleted in human cancers, in 4 NSCLC-BM pairs and one additional BM sample in our discovery cohort, for which we had enough genomic DNA left (**Supplementary Table 4 and Methods**). We found a strong correlation (PCC: 0.62) between the NanoString counts and the expected *vs.* observed read number log2 ratio of the genomic bins encompassing the genes assayed by Nanostring (**Supplementary Fig. 1g**). Notably, the NanoString counts were significantly higher ($P < 2.210^{-16}$, Wilcoxon test, two-tailed) for amplified genes compared to genes with a neutral copy number. In turn, the latter had significantly higher ($P = 1.110^{-7}$, Wilcoxon test, two-tailed) counts compared to deleted genes (**Supplementary Fig. 1h**). Together, these results indicate that NSCLC-BM tend to harbor substantially more SCNAs compared to their matched primary tumors, unlike for example in colorectal cancer, where SCNAs are largely conserved between primary and metastatic lesions[13–15].

SCNAs are spatially homogeneous in NSCLC-BM

Primary tumors, including NSCLC, are typically characterized by high spatial heterogeneity of SCNAs and mutations, as demonstrated by multi-region sequencing performed in different cancer types[16–19]. In contrast, metastases seem to be less spatially heterogeneous than the corresponding primary tumors[20,21]. Low spatial heterogeneity of SCNAs or mutations might be explained by a scenario in which metastases form relatively late during tumor evolution, leaving a short time for genetic drift and spatial diversification to occur in metastases. To test how SCNAs are spatially distributed in NSCLC-BM, we profiled SCNAs in 2–3 relatively large (25–50 mm^2) regions in individual FFPE tissue sections from 15 NSCLC-BM samples in the discovery cohort, leveraging the CUTseq method that we previously developed for multi-region SCNA profiling[22] (**Supplementary Fig. 2, Supplementary Table 2 and Methods**). The SCNA profiles of all regions in the same tissue section clustered together with the corresponding SCNA profile obtained from genomic DNA (gDNA) extracted from an adjacent tissue section (**Fig. 2a**). Accordingly, the pairwise correlations between the SCNA profiles of all the regions in the same tissue section were typically very high (PCC > 0.90), except for 3 BM samples (MN41B, MN47B, and MN54B) in which some of the pairwise correlations dropped (**Fig. 2b**). Visual inspection of the SCNA profiles in these cases revealed that, indeed, some SCNA events were private to some regions and were not detected in the gDNA extracted from other regions in the same section or from an adjacent tissue section (**Fig. 2c**). These findings suggest that SCNAs might form relatively late during NSCLC evolution, in contrast with other tumor types, such as breast and prostate cancer, where SCNAs have been proposed to originate in punctuated bursts in the early stages of tumor evolution and tend to be conserved between primary tumors and metastases[23,24].

Identification of NSCLC-BM driving SCNAs

We then sought to identify SCNAs that could drive NSCLC-BM by amplifying or deleting known cancer genes. We first manually curated a list of cancer-associated genes comprising all 720 genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC)[25] and 58 additional genes listed in the Integrative Genomics Viewer (IGV)[26] database that had been previously associated with NSCLC (**Supplementary Table 5**). We herein refer to this gene list as 'COSMICplus'. Among these genes, the five most frequently amplified genes in the discovery cohort were *ARNT* (58% of all the samples), *MLLT11* (58%), *TRIO* (53%), and *THEM4* (52%), whereas *ARHGEF10* (31%), *CDKN2A/B* (20%), *DCC* (17%) and *PHLPP1* (15%) were the most frequently deleted genes (**Fig. 3a**). To identify plausible SCNA drivers in our cohort, we reasoned that cancer genes that are significantly more frequently amplified or deleted in BM than in the corresponding primary tumors could be considered as candidate NSCLC-BM drivers. To identify such genes, we applied Genomic Identification of Significant Targets in Cancer (GISTIC)[27] to identify genomic regions significantly focally amplified or deleted across the 51 primary tumor and BM samples in the discovery cohort (**Methods**). Intersection of the identified regions with the COSMICplus cancer gene list described above (**Supplementary Table 5**) revealed 45 and 12 genes that were significantly focally amplified and deleted, respectively, in BM but not in the corresponding primary tumors (**Fig. 3b and Supplementary Table 6**). The same genes were altered at low frequency (comparable to the alteration frequency in primary tumors in the discovery cohort) in 1,017 primary NSCLC tumors sequenced in The Cancer Genome Atlas (TCGA) (**Fig. 3c**), further indicating that their alteration is specific to BM. The three most significantly amplified genes in BM were *MLLT11* (60.1% of BM vs. 54.9% of NSCLC), *TP63* (47.1% vs. 25.5%), and *MYC* (49% vs. 27.5%), whereas the three most significantly deleted genes in BM were *CDNK2A/B* (23.5% vs. 11.8%), *DCC* (31.4% vs. 0%), and *TBX3* (13.8% vs. 0%). Importantly, *MYC* amplifications and *CDKN2A/B* deletions were previously identified as putative BM drivers by applying GISTIC to a cohort of 73 matched BM and primary LUAD samples[7], highlighting the validity of our approach. Notably, among genes significantly more frequently amplified in BM we found several genes that could be directly targeted with available drugs, including *CDK12* (amplified in 25.5% of BM *vs.* 5.9% of primary tumors), *DDR2* (43.1% *vs.* 17.6%), *ERBB2* (25.5% *vs.* 5.9%), and *NTRK1* (52.9% *vs.* 21.6%).

To corroborate these findings, we profiled SCNAs in an independent cohort of 84 BM samples from 23 LUSC and 61 LUAD patients (validation cohort), using the CUTseq method that we previously developed[22] (**Supplementary Table 2 and 7, and Methods**). In agreement with our previous findings in the discovery cohort, amplifications were also significantly more frequent than deletions also in the validation cohort, independently of the primary tumor histology (**Supplementary Fig. 3a, b**). Moreover, the percentage of samples carrying amplifications or deletions of COSMICplus genes was strongly correlated (PCC: 0.95) between the discovery and validation cohort (**Supplementary Fig. 3c**). The correlation was even higher (PCC: 0.98) for BM driving genes that were significantly more frequently amplified or deleted in BM compared to their matched primary tumors (**Supplementary Fig. 3d**). Altogether, these results indicate that SCNAs affecting specific cancer genes may drive the formation or progression of BM in NSCLC patients and that some of these genes might be targeted with available drugs.

CDK12 and *DDR2* overexpression triggers the epithelial-mesenchymal transition process

Next, we sought to investigate whether the cancer genes that we identified as significantly more frequently amplified in BM than in primary tumors can indeed act as metastasis drivers. To this end, we overexpressed two of the potentially targetable NSCLC-BM drivers that we identified—*CDK12* and *DDR2*—and assessed their metastatic potential *in vitro* and *in vivo* (**Methods**). When separately overexpressed in mouse LUAD cell lines, both genes potently instigated the induction of key mouse genes involved in the epithelial-mesenchymal transition (EMT) process[28], including *Lef1*, *Slug*, *Snail*, *Twist1*, *Zeb1*, and *Vimentin* (**Fig. 3d-i and Methods**). However, subcutaneous injection of the same cells in mice did not result in overt BM formation, and only in 2 out of 6 mice injected with *CDK12* overexpressing cells we detected a macroscopic submeningeal metastasis (**Supplementary Fig. 4a-c and Methods**). Altogether, these results suggest that amplification and, consequently, overexpression of these genes could prime NSCLC cells to metastasize by inducing an EMT, but these events alone may not be sufficient to drive full BM formation *in vivo*.

Identification of NSCLC-BM driving mutations

Lastly, we turned our attention to single-nucleotide variants (SNVs) and small insertions and deletions (indels), aiming at identifying mutations that might drive NSCLC-BM in addition to SCNAs. To this end, we performed whole-exome sequencing (WES) on 40 of the 51 NSCLC-BM sample pairs in the discovery cohort for which we had enough gDNA left (**Supplementary Table 2 and Methods**). Since germline gDNA had not been collected from these patients, we used the primary tumor samples as a reference for calling SNVs and indels in the corresponding BM samples (**Methods**). Using this approach, the mutations identified can be considered as BM-specific events. We found that the metastasis mutation burden was comparable between LUAD and LUSC samples (**Fig. 4a**). Late-onset BM (*i.e.*, metastases detected >12 months after diagnosis of the primary tumor) had a significantly lower mutation burden than synchronous (<2 months) or early-onset (2–12 months) BM (**Fig. 4b**), suggesting that early and late metastases might be driven by different mutational processes. Most of the mutations identified were missense substitutions (mainly C>T and C>A), and in a few cases (MN1B, MN14B, MN30B, MN40B) frameshift indels represented the dominant mutation type (**Fig. 4c and Supplementary Fig. 5a**). The most common mutational signatures[29] were signature 1, 3, 4, and 6 (**Fig. 4d, Supplementary Fig. 5b, and Methods**). As expected, signature 4, which is associated with tobacco exposure, was overrepresented in patients with smoking history (**Supplementary Fig. 5c**). We then assessed which genes, among those listed in the COSMICplus list (**Supplementary Table 5**), were commonly mutated in the 40 NSCLC-BM samples sequenced by WES. Among the most frequently mutated genes, we identified several genes directly or indirectly implicated in cell motility. These included *MUC16* (82% of the samples), which encodes a large cell surface protein and is best known as a biomarker (CA-125) for ovarian cancer[30],

MACF1 (60%), which encodes a large intracellular protein that bridges actin and microtubule filaments[31], and *FAT1* (57%), *FAT3* (55%) and *FAT4*, which encode three members of the cadherin family of transmembrane proteins previously implicated in the regulation of cell motility[32] (**Fig. 4d**). Other frequently mutated genes included several genes encoding DNA or histone-modifying enzymes, such as *KMT2A* (38%), *KMT2C* (52%), *KMT2D* (45%), *SETD2* (40%), and *SETD1B* (32%), which encode various histone methyltransferases[33], *EP300* (35%), which encodes a histone acetyltransferase playing a pivotal role in transcription regulation[34], and *KAT6B*, also encoding a histone acetyltransferase (**Fig. 4d**). Notably, *KMT2C* mutations were previously detected in 50% of BM samples in a Chinese cohort of 12 matched NSCLC and BM profiled by WES[8], in line with the frequency measured in our cohort. Another gene frequently mutated in BM reported in the same study was *PDE4DIP* (25% of BM samples), which encodes a phosphodiesterase 4D interacting protein and was mutated in 28% of BM in our cohort.

To identify high-confidence NSCLC-BM driving mutations, we then applied cancer-specific high-throughput annotation of somatic mutations (CHASM)[35] to the list of identified mutations (**Methods**). In 26 out of 40 (65%) NSCLC-BM samples sequenced, we identified driving mutations in at least one gene (**Fig 4e**). Among 50 high-confidence NSCLC-BM drivers identified by CHASM, there were several genes involved in chromatin editing/remodeling and three-dimensional (3D) genome organization. These included *EP300* (driving mutations detected in 7 out of 40 (17.5%) BM samples), *PBMR1* (17.5%), which encodes a component of the SWI/SNF chromatin remodeling complex[36], *KDM6A* (15%), *CREBBP1* (10%), which encodes a master transcription coactivator frequently mutated in leukemias,[37] *KMT2A* (5%), *CTCF* (5%), which encodes a transcription factor crucial for structuring chromatin loops along the genome[38], and *STAG2* (5%), which encodes a subunit of the cohesin complex that is also crucial for shaping chromatin loops and is mutated in many cancer types[39] (**Fig 4e**). Accordingly, gene ontology analysis revealed significant enrichment of terms related to transcription and chromatin remodeling, but also of terms associated with cytoskeleton and cell motility (**Supplementary Fig. 5d**). Notably, 4 out of 40 (10%) BM samples contained driving mutations in *CDK12*, further implicating this potentially targetable gene in the pathogenesis of NSCLC-BM.

Lastly, to corroborate these findings, we performed targeted sequencing of 115 additional BM samples (validation cohort) using the Glasgow Cancer Core Panel covering 174 genes frequently mutated in solid tumors (**Supplementary Table 7 and 8, and Methods**). Among frequently mutated genes in the validation cohort, we again found multiple genes encoding chromatin-modifying enzymes, including *KMT2A* (mutated in 30% of the samples), *EP300* (24%), *SETD2* (17%), *CREBBP1* (19%), *STAG1* (11%) and *STAG2* (13%) (**Supplementary Fig. 6a**). Importantly, we observed a strong correlation (PCC: 0.82) between the frequency of mutation in genes covered by the panel in the discovery and validation cohort (**Supplementary Fig. 6b**). The correlation was even stronger (PCC: 0.89) for the subset of high-confidence

NSCLC-BM driving genes identified by CHASM, which were also covered by the gene panel (**Supplementary Fig. 6c**). Altogether, these results implicate mutations in multiple genes involved in chromatin editing and genome architecture in the pathogenesis of BM in a subset of NSCLC patients. Future studies are therefore needed to dissect the functional consequences of NSCLC-BM driving mutations and test the efficacy of chromatin-modifying drugs in these patients.

Discussion

Although BM remain a major cause of morbidity and mortality in patients with NSCLC, the genomic landscape of NSCLC-BM has not been thoroughly characterized. A previous study conducted on 73 pairs of LUAD and BM samples identified amplifications of *MYC*, *YAP1*, and *MMP13*, and deletions of *CDKN2A* as putative LUAD-BM drivers[7]. However, that study did not include LUSC-BM and did not identify targetable BM drivers. In contrast, here we identified several plausible NSCLC-BM driving genomic events affecting potentially targetable genes, such as *CDK12*, *DDR2*, *ERBB2*, and *NTRK1*. The products of these genes can be targeted, respectively, with dinaciclib; dasatinib; lapatinib, afatinib, dacomitinib, neratinib, or pyrotinib; and entrectinib, laroterctinib, or repotrectinib. Hence, clinical trials could be readily designed to assess the efficacy of these agents in NSCLC patients with BM harboring *CDK12*, *DDR2*, *ERBB2*, or NTRK1 driving amplifications or mutations. Of note, DDR2, ERBB2 and NTRK1 inhibitors have proven effective against primary NSCLC tumors harbouring *DDR2* or *ERBB2* mutations or *NTRK1* rearrangements[40]. However, their activity against NSCLC-BM with amplifications of these genes has not been assessed.

For two NSCLC-BM drivers that are targetable by existing drugs (*CDK12* and *DDR2*), we sought to assess their metastatic potential *in vitro* and *in vivo*. Overexpression of these genes in murine LUAD cell lines—to mimic the most likely functional consequence of the amplification of these genes—led to potent induction of key genes involved in EMT but did not result in significant brain metastasis formation when these cells were injected in mice. This could be explained by the fact that metastatization is a multifactorial process, therefore, even though alterations in single genes can prime EMT, on their own they might not be capable of inducing the formation of metastases.

In addition to identifying NSCLC-BM drivers that can be targeted by existing drugs, our study also revealed a remarkable enrichment of SCNAs and mutations in genes encoding for histone methyltransferases (*KMT2A*, *KMT2C*, *KMT2D*, *SETD2*, *SETD1B*), histone acetyltransferases (*EP300*, *KAT6B*), chromatin remodeling factors (*PBMR1*), and 3D genome shapers (*CTCF*, *STAG1/2*) in BM compared to primary tumors. Alterations in these genes may contribute to activating pro-metastatic gene expression programs in NSCLC cells by rewiring 3D genome domains, such as chromatin loops[41] and topologically associating domains (TADs)[42], or by rewiring enhancer-promoter interactions and transcriptional complexes. Alternatively, mutations in those genes might affect the global arrangement of

chromatin in the nucleus, altering its mechanical properties such as stiffness, which in turn might prime cells to metastasize. Future studies are needed to chart the structural and functional 3D genome landscape in NSCLC-BM cells, aiming at uncovering potential vulnerabilities and identifying novel therapeutic targets.

Our study also contributes to shedding light on the evolutionary history of BM in NSCLC patients. By comparing SCNA profiles in matched NSCLC and BM samples, we found that the latter harbor a significantly higher burden of SCNAs, with amplifications dominating the genomic landscape of NSCLC-BM. Furthermore, by profiling SCNAs across multiple tissue regions in individual BM lesions, we showed that the SCNA profiles measured at different locations within the same metastasis are largely similar. This observation is consistent with the previous finding that spatially and temporally separated BM are genetically homogenous[43]. Together, our observations are compatible with a tumor evolution model in which SCNAs form in a single burst in one or few tumor cells within the primary tumor mass, which then metastasize to the brain where they expand without further SCNA diversification, explaining the absence of SCNA spatial heterogeneity within BM. However, the latter could also be attributed to a rapid outgrowth of a single metastatic clone, leaving little time for new SCNAs to emerge in the population.

In our study, all samples were obtained from Caucasian individuals from a homogenous ethnic group (European Poles). Thus, larger multi-centric studies including more diverse populations are needed to uncover the full spectrum of NSCLC-BM genomic alterations and potential therapeutic targets.

In conclusion, our study provides the most comprehensive compendium of presumptive NSCLC-BM genetic drivers to date, highlighting the importance of comparative genomic profiling of matched primary and metastatic tumors to identify metastasis drivers. The list of NSCLC-BM drivers identified in this study represent a powerful resource for designing future studies exploring the molecular pathogenesis and targetability of NSCLC-BM.

Methods

EXPERIMENTAL METHODS

Samples

The collection of all tumor samples described in this study was approved by the Ethics Committee of the Medical University of Lublin, Poland under ethical permit no. KE-0254/235/2016.

Discovery cohort. We retrieved archival formalin-fixed paraffin-embedded (FFPE) tissue samples from 51 pairs of NSCLC and matched brain metastases (BM). The samples included surgical and diagnostic (biopsy) specimens that had been previously collected at the Medical University of Lublin, Poland and at the Medical University of Gdansk, Poland. At the time of biopsy or surgical resection of the primary tumor, all patients were chemo-, radio-, immune- and molecularly targeted therapy naïve. BM material was obtained from neurosurgery. We defined BM in the discovery cohort as synchronous if they were detected between 0 and 2 months, early if they were detected between 3 and 12 months, and late if they were detected after 12 months since the primary diagnosis of lung cancer.

Validation cohort. We retrieved archival FFPE tissue sections from 115 NSCLC-BM surgically resected at the same institutions where the discovery cohort samples were collected. Since only fine-needle aspiration biopsies were performed on the primary tumors, we could not obtain enough gDNA to genomically profile both primary tumors and matched BM.

gDNA extraction and sonication

We cut two consecutive sections (4 mm and 8 mm thick, respectively) from each FFPE tissue block in the discovery and validation cohort and used the 4 mm thick section for hematoxylin-eosin (H&E) staining and the 8 mm thick section for genomic DNA (gDNA) extraction. To extract gDNA we used the QIAamp DNA FFPE Tissue Kit (Qiagen, cat. no. 56404) following the manufacturer's protocol. We assessed the quality and quantity of the gDNA samples using a NanoDrop 2000 (Thermo Fisher Scientific, cat. no. ND-2000) and a Qubit 3.0 fluorimeter (Thermo Fisher Scientific, cat. no. Q33216) and retained only samples that had a range of light absorption (A_{260}/A_{280}) comprised between 1.8 and 2.0. We sheared 6–10 ng of each purified gDNA in 100 mL of Nuclease-Free Water (Thermo Fisher Scientific, cat. no. 4387936) using a Bioraptor Plus (Diagenode, cat. no. B01020001) with cycling conditions optimized to achieve a mean target size of 150–200 base-pairs (bp) (30 sec ON, 90 sec OFF, 40 cycles). We evaluated the distribution of the sheared gDNA on a Bioanalyzer 2100 (Agilent Technologies, cat. no. G2943CA) using a High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626). We re-measured the quantity of the sheared gDNA on a Qubit fluorometer and stored the samples at –20 °C until we prepared sequencing libraries. The amount of gDNA that we could extract varied quite substantially from sample to sample (range: 2.3–55 ng; mean ± s.d.: 21 ± 12.5,) and was substantially higher in the case of metastatic samples.

SCNA profiling in matched NSCLC primary and BM samples

To profile SCNAs in our discovery cohort, we prepared individual sequencing libraries from each of the 51 pairs of NSCLC primary and BM samples in the cohort using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, cat. no. E7645) and corresponding NEBNext Multiplex Oligos for Illumina

(New England Biolabs, cat. no. E7335) following the manufacturer's instructions. We assessed the size distribution, the quality, and the quantity of each library on a Bioanalyzer 2100 (Agilent Technologies, cat. no. G2943CA) using a High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626). We sequenced all libraries shallowly on a NextSeq 500 system (Illumina) using a NextSeq 500/550 High Output v2 kit 75 cycles (Illumina, cat. no. FC-404-2005) aiming at generating 30 million reads per sample.

SCNA validation on nanoString

To validate SCNAs identified in the discovery cohort, we applied the nCounter v2 Cancer CN Assay (nanoString, cat. no. CNV-CAN2-12) to four NSCLC-BM pairs and one additional BM sample for which we had enough (150–200 ng) gDNA. We re-quantified the gDNA concentration of each sample on a Qubit fluorimeter (Thermo Fisher Scientific) and then handled it to the Karolinska Institutet Gene facility (Stockholm, Sweden), which performed probe hybridization, automatic purification, and immobilization on an nCounter Prep Station (nanoString) following the manufacturer's instruction. Data collection was performed on an nCounter Digital Analyzer (nanoString) at the same facility. We analyzed the resulting data using the freely available nCounter Analysis Software (nanoString) (see **Supplementary Table 4**).

Multi-region SCNA profiling

To assess how the SCNAs identified in the discovery cohort were spatially distributed in the BM samples, we applied the CUTseq method, which we previously described[22], to gDNA extracted from 2–4 macroscopic regions per section in 15 BM FFPE tissue sections that were sufficiently large (>1 cm^2) to capture DNA from multiple regions (see **Supplementary Fig. 2**). To extract gDNA from each region, we applied the PinPoint Solution (Zymo research, cat. no. D3001) onto each region and let it solidify for 30 min at room temperature. We then scraped the solidified solution, encapsulating the underlying tissue, with an insulin needle and transferred the scraped tissue into a 1.5 mL tube. We lysed the tissue using a buffer containing 10 mM Tris HCl/100 mM NaCl/50 mM EDTA/1% SDS/19 mg/mL Proteinase K (New England Biolabs, cat. no. P8107S) pH 7.5 and incubated the solution overnight at 55 °C on a thermomixer, shaking at 800 rpm. The following day, we purified each sample using a standard phenol-chloroform extraction protocol and quantified the gDNA samples using a Qubit 3.0 fluorimeter (Thermo Fisher Scientific, cat. no. Q33216). We applied CUTseq[22] to prepare multiplexed sequencing libraries by pooling the gDNAs extracted from up to 71 different regions together. We assessed the size distribution, quality, and quantity of the libraries on a Bioanalyzer 2100 (Agilent Technologies, cat. no. G2943CA) using a High Sensitivity DNA kit (Agilent Technologies, cat. no. 5067–4626) and sequenced them on a NextSeq 500 system (Illumina) using a NextSeq 500/550 High Output v2 kit (75 cycles) (Illumina, cat. no. FC-404-2005).

Validation of BM-driving SCNAs

To validate BM-driving SCNAs identified in the discovery cohort, we applied CUTseq[22] to 84 BM samples in the validation cohort, for which we had enough gDNA (see **Supplementary Table 7**). We performed CUTseq in 96-well plates using a low-volume contactless liquid-dispensing device (I.DOT One, Dispendix GmbH, Germany) to dispense the CUTseq digestion and ligation mix. We then pooled the samples (24 samples per pool) in 1.5 mL DNA Lo-Bind tubes (Eppendorf, cat. no. 0030108051) before proceeding to *in vitro* transcription and preparing a multiplexed sequencing library from each pool using the CUTseq protocol[22]. We assessed the quality of each library using a Bioanalyzer 2100 (Agilent Technologies, cat. no. G2943CA) and a High Sensitivity DNA kit (Agilent Technologies, cat. no. 5067–4626). We sequenced all libraries on a NextSeq 500 system (Illumina) using a NextSeq 500/550 High Output v2 kit (75 cycles) (Illumina, cat. no. FC-404-2005).

Whole exome sequencing

To identify BM-driving mutations, we performed whole-exome sequencing (WES) in 40 NSCLC primary and BM sample pairs in the discovery cohort, for which we had enough gDNA. We first prepared a sequencing library from each sample using the SureSelect XT HS Kit (Agilent Technologies, cat. no. G9704A) following the manufacturer's instructions. We assessed the size distribution, quality, and quantity of each library on a Bioanalyzer 2100 (Agilent Technologies) using a High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626). To reach the recommended input for exome capture (500–1,500 ng), we pooled up to 8 libraries together and concentrated the pools using the Savant SpeedVac DNA 130 Integrated Vacuum Concentrator System (Thermo Fisher Scientific, cat. no. DNA130-230) using the standard heating mode, until all the solutions were entirely evaporated. We resuspended each pool in 12 mL of Nuclease-Free Water (Thermo Fisher Scientific, cat. no. 4387936) and performed exome capture using the SureSelect XT HS Target Enrichment kit and SureSelect Human All Exon v6 baits (Agilent Technologies, cat no. G9704K) following the manufacturer's protocol. We again assessed the size distribution, quality, and quantity of all captured libraries on a Bioanalyzer 2100 and Qubit and sequenced all libraries on a NovaSeq 6000 system (Illumina) at the National Genomics Infrastructure (NGI Stockholm, Sweden) using a 2150 bp flowcell S4 (Illumina, cat. no. 20012866).

Validation of BM-driving mutations

To validate BM-driving mutations identified in the discovery cohort, we performed targeted gene sequencing of all the 115 BM samples in the validation cohort (see **Supplementary Table 8**). We first prepared a DNA sequencing library from each sample using the SureSelect XT HS Kit (Agilent Technologies, cat. no. G9704A), following the manufacturer's instructions. We assessed the size distribution, quality, and quantity of each library on a Bioanalyzer 2100 (Agilent Technologies) using a

High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626). To reach the recommended gDNA input for targeted gene capture (500–1,500 ng), we pooled up to 8 libraries together and concentrated each pool using the Savant SpeedVac DNA 130 Integrated Vacuum Concentrator System (Thermo Fisher Scientific, cat. no. DNA130-230) on standard heating mode, until all solutions were entirely evaporated. We resuspended each pool in 12 µL of Nuclease-Free Water (Thermo Fisher Scientific, cat. no. 4387936) and captured 174 cancer-associated genes using the SureSelect CD Glasgow Cancer Core Panel (Agilent Technologies, cat no. 5191-6736) following the manufacturer's instructions. We again assessed the size distribution, quality, and quantity of all captured libraries using Bioanalyzer 2100 and Qubit and then sequenced all libraries on a NovaSeq 6000 platform (Illumina) at the National Genomics Infrastructure (NGI Stockholm, Sweden) using a 2150 bp flowcell S4 (Illumina, cat. no. 20012866).

In vitro and *in vivo* assessment of metastatic potential

We purchased from Addgene lentiviral expression constructs for the human *CDK12* (pHAGE-GFP-CDK12, cat. no. 116723) and *DDR2* (pHAGE-PURO-DDR2, cat. no. 116729) genes as well as for *GFP* to serve as negative control (pHAGE-GFP, cat. no. 106281). We produced lentiviral particles in HEK293T cells (ATCC, cat. no. CRL-3216). We then used the lentiviruses to infect TECLA-1 or ChA9.6 mouse LUAD cancer cell lines, which we previously established[44,45]. We grew both cell lines in DMEM medium (Gibco, cat. no. 10566016) supplemented with 10% Fetal Bovine Serum (Gibco) and 2% Penicillin, 5 mg/mL Streptomycin (Gibco) and selected successfully transduced cells either based on GFP expression (control cells) or by adding 1 mg/mL Puromycin (Invitrogen) to the growth medium.

To assess the induction of EMT genes, we isolated total RNA from TECLA-1 or ChA9.6 cells overexpressing the human *CDK12* or *DDR2* gene using the TRIzol reagent (Thermo Fisher Scientific, cat. no. 15596018), and then reverse transcribed it with random primers using the RevertAid RT Kit (Thermo Fisher Scientific, cat. no. 00940535) following the manufacturer's instructions. We performed real-time PCR on the obtained cDNA with the PowerTrack SYBR Green Master Mix (Applied Biosystems, cat. no. 00864923) and PCR primers targeting the mouse genes listed in **Supplementary Table 10**, using a 7500 Fast Real-Time PCR System (Applied Biosystems).

To assess the metastatic potential of *CDK12* or *DDR2* overexpression *in vivo*, we injected TECLA-1 or ChA9.6 cells overexpressing *CDK12*, *DDR2* or green fluorescent protein (GFP, negative control) intravenously in isogenic Balb/C mice (8 mice per group, 210^5 cells injected per mouse). We monitored the injected mice for 15 days and then euthanized them at the appearance of signs of breathing difficulties. One animal in the GFP group died before the euthanasia and therefore we excluded it from the study. After euthanasia, we collected various organs (Brain, lungs, kidney, spleen, liver) from each mouse,

fixed them in formalin, and embedded them in paraffin following standard procedures. We cut each brain block at different levels into multiple 3 mm-thick sections, which we stained with hematoxylin-eosin following a standard protocol and examined under an optical microscope at 4 magnification. In two mice injected with TECLA-1 cells overexpressing *CDK12* cells, we detected a submeningeal lesion that stained positive for cytokeratin 14 following immunohistochemistry.

COMPUTATIONAL METHODS

Sequencing data processing

We demultiplexed all raw sequence reads to FASTQ files using the BaseSpace Sequence Hub cloud service of Illumina. For CUTseq data, we further demultiplexed the reads to sample specific FASTQ files using a custom Python script available at https://github.com/ljwharbers/metastatic_lungcancer. Next, we aligned all reads to the GRCh37/hg19 reference genome using *BWA* (v0.7.17-r1188)[46] and then sorted and indexed them using *SAMtools* (v1.10)[47]. For CUTseq libraries, which included unique molecular identifiers (UMIs), we deduplicated the reads using *UMI-tools* (v1.1.1)[48]. For libraries prepared for targeted gene sequencing, we deduplicated the reads using the *Agilent Genomics NextGen Toolkit* (AGeNT) (v2.0.5). For all other libraries, we deduplicated the reads using the *MarkDuplicates* tool in the *Genome Analysis ToolKit* (GATK, v4.1.4.1)[49]. Finally, for libraries used for single-nucleotide variant (SNV) calling, we recalibrated the base scores using the *BQSRPipelineSpark* command in GATK.

Copy number calling

To determine DNA copy number levels, we used the R package *QDNAseq*[9], which is optimized for FFPE samples, and *CNVkit*[10]. Unless otherwise specified, we binned the genome in 50 kb windows. We plotted genome-wide copy number profiles by aggregating together the profiles of individual chromosomes using custom scripts in R available at https://github.com/ljwharbers/metastatic_lungcancer. We called amplifications (AMP) and deletions (DEL) using a threshold of the log2 ratio equal to 0.32 and −0.42, respectively. To determine the fraction of the genome with AMP or DEL, we calculated the percentage of 50 kb genomic windows that were called either as amplified or deleted using custom scripts in R available at https://github.com/ljwharbers/metastatic_lungcancer. To determine alterations in cancer-related genes, we overlapped amplified and deleted genomic regions with the COSMICplus gene list available in **Supplementary Table 5**. To determine significant focal amplification and deletion events, we used *GISTIC2*[27] with default settings.

SNVs and indels calling

We analyzed WES and targeted sequencing data using *GATK* (v4.1.4.1) following the Broad Institute's best practices. In the case of WES data, we used *GATK Mutect2* with primary tumor samples as a reference to call BM-specific SNVs and indels. We then removed false positive calls due to sequence artifacts and contamination with the following sequence of GATK commands: *LearnReadOrientationModel*; *GetPileupSummaries*; *CalculateContamination*; and *FilterMutectCalls*. Lastly, we annotated the filtered calls using the *GATK* command *Funcotator*. To determine potential driver mutations, we used CHASMplus with the following two NSCLC annotators and two brain tumor annotators, namely: *chasmplus_GBM*, *chasmplus_LGG*, *chasmplus_LUAD* and *chasmplus_LUSC*.

Code availability

All custom code used to process and analyze the sequencing data is available at the following GitHub link: https://github.com/ljwharbers/metastatic_lungcancer.

Data availability

Because of privacy regulations, the raw sequencing data generated in this study cannot be made publicly available. A detailed summary of sequencing statistics for each dataset is available in **Supplementary Table 2**.

Declarations

Acknowledgements

The two first authors reserve the right to cite the manuscript as either Nicoś et al or Harbers et al to reflect their equal contribution. We thank Britta Bouwman (Bienko-Crosetto lab) for critically reading the manuscript, Adrian Perdyan (visiting students at the Department of Pathology, Medical University of Gdansk) for helping with clinical data curation, and Gabriel Felicio dos Santos and Ricardo Daniel Estrada Moreno (visiting students in the Bienko-Crosetto lab) for helping with genomic DNA extraction from the validation cohort. We acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. This work was supported by a Mobilność Plus scholarship from the Polish Ministry of Science and High Education (1622/MOB/V/2017/0), a START scholarship from the Foundation for Polish Science (FNP), and a grant from the Polish National Science Center (UMO- 2016/23/D/NZ2/02890) to M.N.; by a grant from the H2020-MSCA-ITN-2018 Marie Sklodowska-Curie Actions Innovative Training Networks ('aDDRess', grant no. 812829) to N.C. funding L.H.; by funding from the National Cancer Institute (grant no. 1R01CA222598) to R.C.; by funding from the Science for Life Laboratory supporting

the purchase of the NextSeq 500 used in this study to M.B.; by funding from the Medical University of Gdańsk (statutory grant ST-23, 02-0023/07) to J.J.; by funding from the Giovanni Armenise–Harvard Foundation, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. 101001288) and AIRC (IG 2021 - ID. 25737) to C.A.; and by funding from the Swedish Research Council (grant no. 521-2014-2866), the Swedish Cancer Research Foundation (grant no. CAN 2015/585), the Ragnar Söderberg Foundation, the Swedish Foundation for Strategic Research (grant no. BD15-0095), and the Strategic Research Programme in Cancer (StratCan) at Karolinska Institutet to N.C.

Author contributions

Conceptualization: R.D., J.M., P.K., J.J., M.N., N.C. *Clinical samples*: A.K., A.B. *Pathological annotation*: R.P., B.J., J.S. *Methodology*: M.N. *Investigation*: M.N., X.Z., M.S. *Data curation*: M.N., L.H. *Formal analysis*: L.H. *In vitro and in vivo models*: E.P., M.K-D., C.V., R.C., C.A. *Validation*: M.N., M.Z., B.W. *Funding acquisition*: N.C., M.N., R.C., M.B., J.J., C.A. *Project administration*: M.N., N.C. *Software*: L.H. *Supervision*: N.C., C.A. *Visualization*: L.H., N.C. *Writing*: N.C. and M.N. with contributions from all Authors.

Competing interests

The authors declare no competing interests.

References

1. An, N. *et al.* Risk factors for brain metastases in patients with non-small-cell lung cancer. *Cancer Med.* **7**, 6357–6364 (2018).

2. Chamberlain, M. C., Baik, C. S., Gadi, V. K., Bhatia, S. & Chow, L. Q. M. Systemic therapy of brain metastases: non-small cell lung cancer, breast cancer, and melanoma. *Neuro-Oncol.* **19**, i1–i24 (2017).

3. Kim, M. *et al.* Barriers to Effective Drug Treatment for Brain Metastases: A Multifactorial Problem in the Delivery of Precision Medicine. *Pharm. Res.* **35**, 177 (2018).

4. Nishino, M., Soejima, K. & Mitsudomi, T. Brain metastases in oncogene-driven non-small cell lung cancer. *Transl. Lung Cancer Res.* **8**, S298–S307 (2019).

5. Tan, A. C., Itchins, M. & Khasraw, M. Brain Metastases in Lung Cancers with Emerging Targetable Fusion Drivers. *Int. J. Mol. Sci.* **21**, E1416 (2020).

6. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).

7. Shih, D. J. H. *et al.* Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma. *Nat. Genet.* **52**, 371–377 (2020).

8. Liu, Z. *et al.* Whole-exome sequencing identifies somatic mutations associated with lung cancer metastasis to the brain. *Ann. Transl. Med.* **9**, 694 (2021).

9. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.* **24**, 2022–2032 (2014).

10. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).

11. Tomasini, P. *et al.* Comparative genomic analysis of primary tumors and paired brain metastases in lung cancer patients by whole exome sequencing: a pilot study. *Oncotarget* **11**, 4648–4654 (2020).

12. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.* **26**, 317–325 (2008).

13. Brannon, A. R. *et al.* Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol.* **15**, 454 (2014).

14. Lee, S. Y. *et al.* Comparative genomic analysis of primary and synchronous metastatic colorectal cancers. *PloS One* **9**, e90459 (2014).

15. Vignot, S. *et al.* Comparative analysis of primary tumour and matched metastases in colorectal cancer patients: evaluation of concordance between genomic and transcriptional profiles. *Eur. J. Cancer Oxf. Engl. 1990* **51**, 791–799 (2015).

16. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

17. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256–259 (2014).

18. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).

19. Yan, T. *et al.* Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma. *Nat. Commun.* **10**, 1670 (2019).

20. Wang, D. *et al.* Multiregion Sequencing Reveals the Genetic Heterogeneity and Evolutionary History of Osteosarcoma and Matched Pulmonary Metastases. *Cancer Res.* **79**, 7–20 (2019).

21. Wei, Q. *et al.* Multiregion whole-exome sequencing of matched primary and metastatic tumors revealed genomic heterogeneity and suggested polyclonal seeding in colorectal cancer metastasis. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **28**, 2135–2141 (2017).

22. Zhang, X. *et al.* CUTseq is a versatile method for preparing multiplexed DNA sequencing libraries from low-input samples. *Nat. Commun.* **10**, 4732 (2019).

23. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).

24. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).

25. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

26. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

27. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

28. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **15**, 178–196 (2014).

29. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

30. Felder, M. *et al.* MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. *Mol. Cancer* **13**, 129 (2014).

31. Cusseddu, R., Robert, A. & Côté, J.-F. Strength Through Unity: The Power of the Mega-Scaffold MACF1. *Front. Cell Dev. Biol.* **9**, 641727 (2021).

32. Zhang, X. *et al.* History and progression of Fat cadherins in health and disease. *OncoTargets Ther.* **9**, 7337–7343 (2016).

33. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13**, 343–357 (2012).

34. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**, 953–959 (1996).

35. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–6667 (2009).

36. Centore, R. C., Sandoval, G. J., Soares, L. M. M., Kadoch, C. & Chan, H. M. Mammalian SWI/SNF Chromatin Remodeling Complexes: Emerging Mechanisms and Therapeutic Strategies. *Trends Genet. TIG* **36**, 936–950 (2020).

37. Mullighan, C. G. *et al.* CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature* **471**, 235–239 (2011).

38. Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).

39. Waldman, T. Emerging themes in cohesin cancer biology. *Nat. Rev. Cancer* **20**, 504–515 (2020).

40. Guo, Y. *et al.* Recent Progress in Rare Oncogenic Drivers and Targeted Therapy For Non-Small Cell Lung Cancer. *OncoTargets Ther.* **12**, 10343–10360 (2019).

41. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

42. Szabo, Q. *et al.* TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Sci. Adv.* **4**, eaar8082 (2018).

43. Brastianos, P. K. *et al.* Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov.* **5**, 1164–1177 (2015).

44. Blasco, R. B., Patrucco, E., Mota, I., Tai, W.-T. & Chiarle, R. Comment on 'ALK is a therapeutic target for lethal sepsis'. *Sci. Transl. Med.* **10**, eaar4321 (2018).

45. Ambrogio, C. *et al.* Modeling lung cancer evolution and preclinical response by orthotopic mouse allografts. *Cancer Res.* **74**, 5978–5988 (2014).

46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

48. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

49. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

50. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat. Oxf. Engl.* **5**, 557–572 (2004).

51. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

Figures

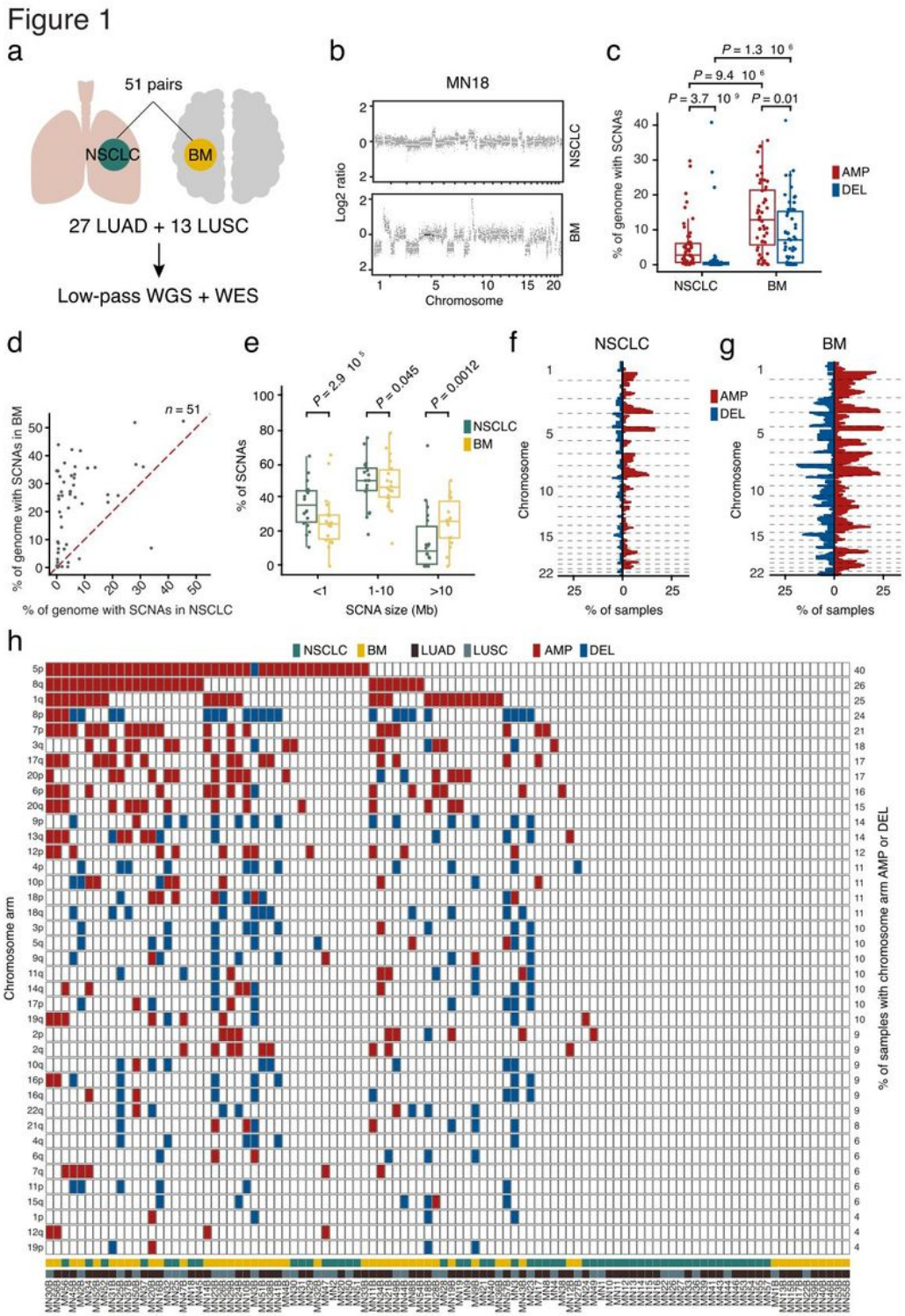

Figure 1

NSCLC-BM harbor significantly more SCNAs compared to primary tumor samples. (**a**) Schematic representation of the discovery cohort and of the genomic assays performed on it. WGS, whole-genome sequencing. WES, whole-exome sequencing. (**b**) Example of genome wide SCNA profiles (50 kilobase, kb resolution) for one NSCLC-BM pair from one patient (MN18) in the discovery cohort. Each grey dot represents a 50 kb genomic bin. Black dots mark copy number levels determined by circular binary segmentation[50]. (**c**) Percentage of the genome amplified (AMP) or deleted (DEL) in each primary tumor and BM sample in the discovery cohort. Each dot represents one sample. *P*, Wilcoxon test, two-tailed. (**d**) Correlation of the percentage of the genome either amplified or deleted between the 51 (*n*) pairs of primary NSCLC and BM samples in the discovery cohort. Each dot represents one sample pair. The dashed red line represents the bisector of the angle between the x- and y-axis. (**e**) Distributions of the length of genomic segments amplified or deleted in each primary tumor and BM sample in the discovery cohort, separately for focal (<1 megabase, Mb), medium (1–10 Mb) and large (>10 Mb) SCNAs. *P*, Wilcoxon test, two-tailed. (**f**) Percentage of primary tumor samples (*n*) in the discovery cohort having consecutive 50 kb genomic bins (vertical axis) amplified (AMP) or deleted (DEL). (**g**) Same as in (f) but for the corresponding BM samples. (**h**) Percentage of samples in the discovery cohort harboring at least one amplification (AMP) or deletion (DEL) along the chromosome arm indicated on the left. Sample IDs are the same as in **Supplementary Table 1**. LUAD, lung adenocarcinoma. LUSC, lung squamous cell carcinoma. In all boxplots in (c) and (e), each box ranges from the 25th to the 75th percentile, the horizontal line marks the median value, and the whiskers span from the minimum to the maximum value.

Figure 2

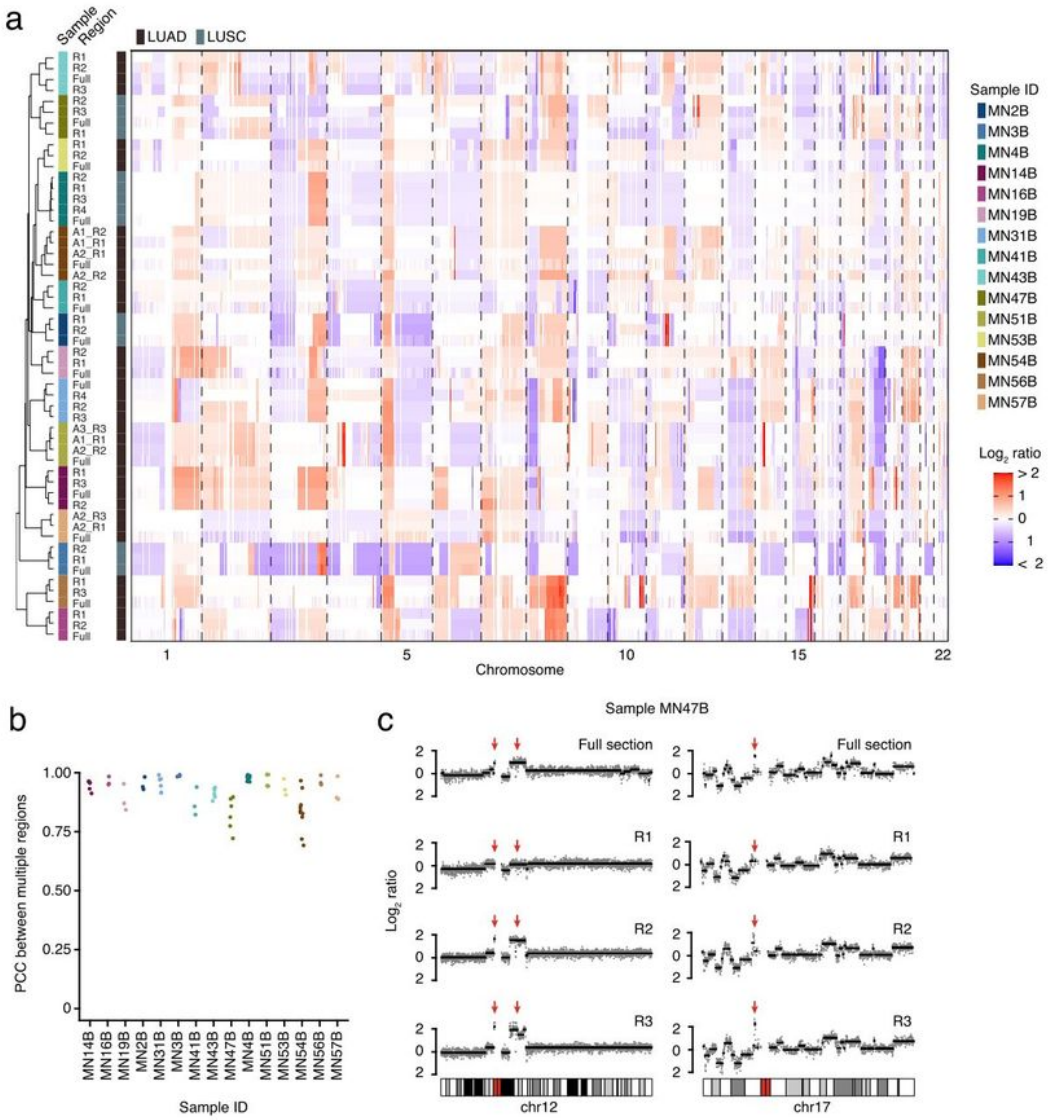

Figure 2

SCNAs are spatially homogeneous across NSCLC-BM. (**a**) Clustered heatmap of the observed vs. expected read count log2 ratio in consecutive 50 kb genomic bins for multiple tissue regions in 15 BM samples from the discovery cohort (see **Supplementary Fig. 2** for a map of the regions profiled). Individual regions profiled in each sample are labeled as '_R'. 'A1' and 'A2' refer to spatially distinct metastatic lesions. Sample IDs are the same as in **Supplementary Table 1**. (**b**) Pearson's correlation

coefficient (PCC) of the genome-wide copy number levels between all possible pairs of tissue regions profiled in each BM sample shown in (a). (**c**) Example of copy number profiles along chromosome (chr) 12 and 17, in three tissue regions (R) profiled by CUTseq[22] inside a single FFPE tissue section from BM sample MN47B (see **Supplementary Table 1**). The profiles on the top row correspond to a full FFPE tissue section adjacent to the one in which gDNA was extracted from the regions displayed. Overall, the profiles are largely conserved between regions. The red arrows indicate clear differences between regions.

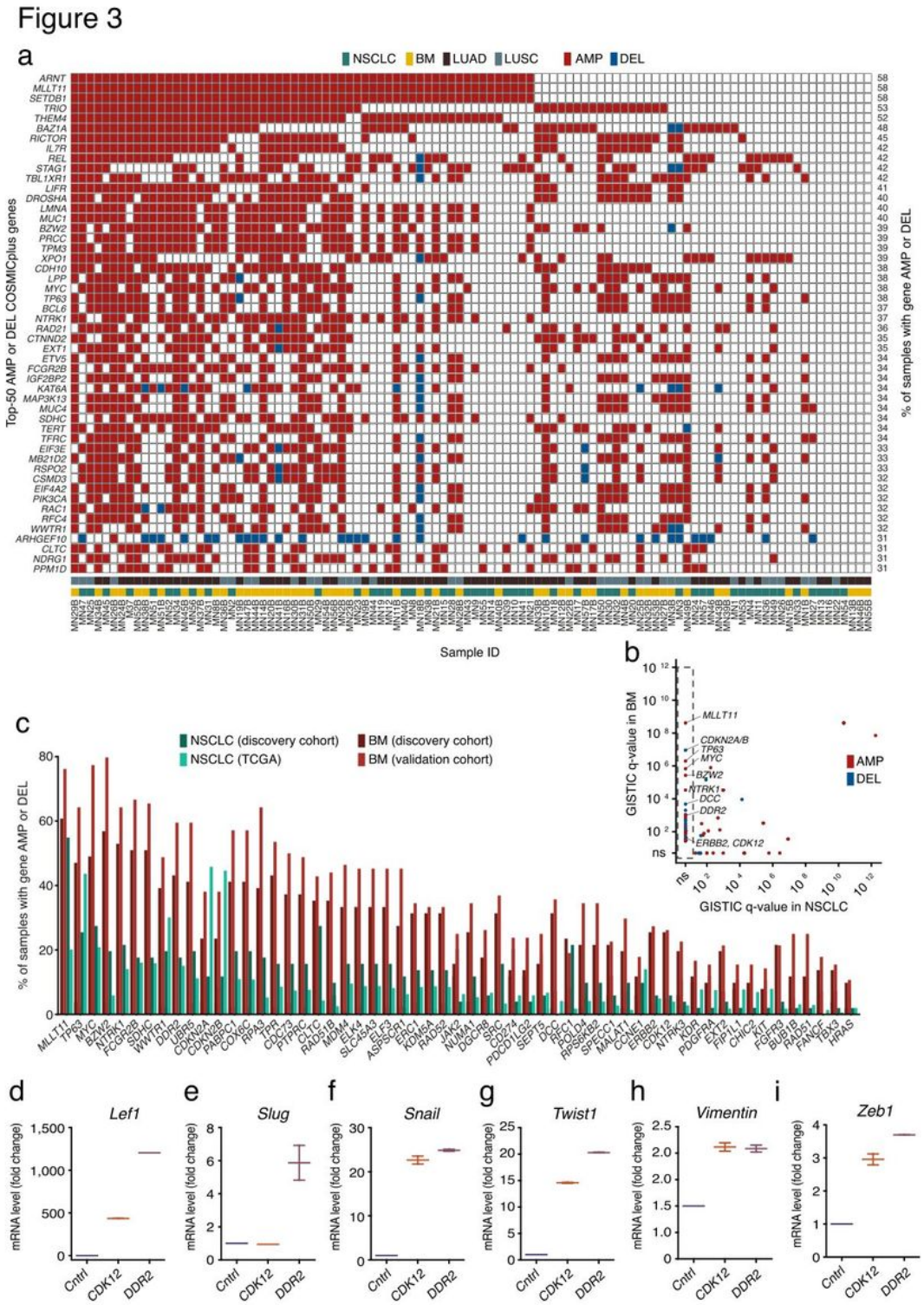

Figure 3

Comparison of SCNA profiles between matched primary tumor and BM samples allows identifying NSCLC-BM driving genes. (**a**) Copy number status of the 50 most frequently amplified (AMP) or deleted (DEL) genes in the COSMICplus list (see **Supplementary Table 3**) across all the samples included in the discovery cohort. Sample IDs are the same as in **Supplementary Table 1**. LUAD, lung adenocarcinoma. LUSC, lung squamous cell carcinoma. (**b**) Q-value assigned by GISTIC[51] to each gene in the COSMICplus

list (see **Supplementary Table 3**) in each pair of primary tumor and BM samples in the discovery cohort. The genes inside the dashed rectangle on the left have a non-significant (ns) q-value in NSCLC and a significant q-value in the BM samples, indicating that they are significantly more amplified (AMP) or deleted (DEL) in BM than in the primary tumor. Only COSMICplus genes that have a significant q-value in at least one sample type are shown. The three most frequently amplified and deleted genes in BM as well as actionable genes significantly amplified in BM are indicated. (**c**) Percentage of samples in which the indicated genes are amplified or deleted, in the discovery cohort, in the validation cohort (84 BM samples) profiled by CUTseq[22] (see **Methods**) and in TCGA. (**d-i**) Real-time PCR quantification of the relative expression of key genes (plot titles) implicated in the epithelial-mesenchymal transition (EMT) process in TECLA-1 mouse LUAD cells transduced with *GFP* (Cntr) or human *CDK12* or *DDR2* genes. The y-axis indicates the fold change compared to a reference gene (*GAPDH*) ENRICO. Horizontal colored bars represent the mean of three technical replicates. Error bars indicate ± standard deviation.

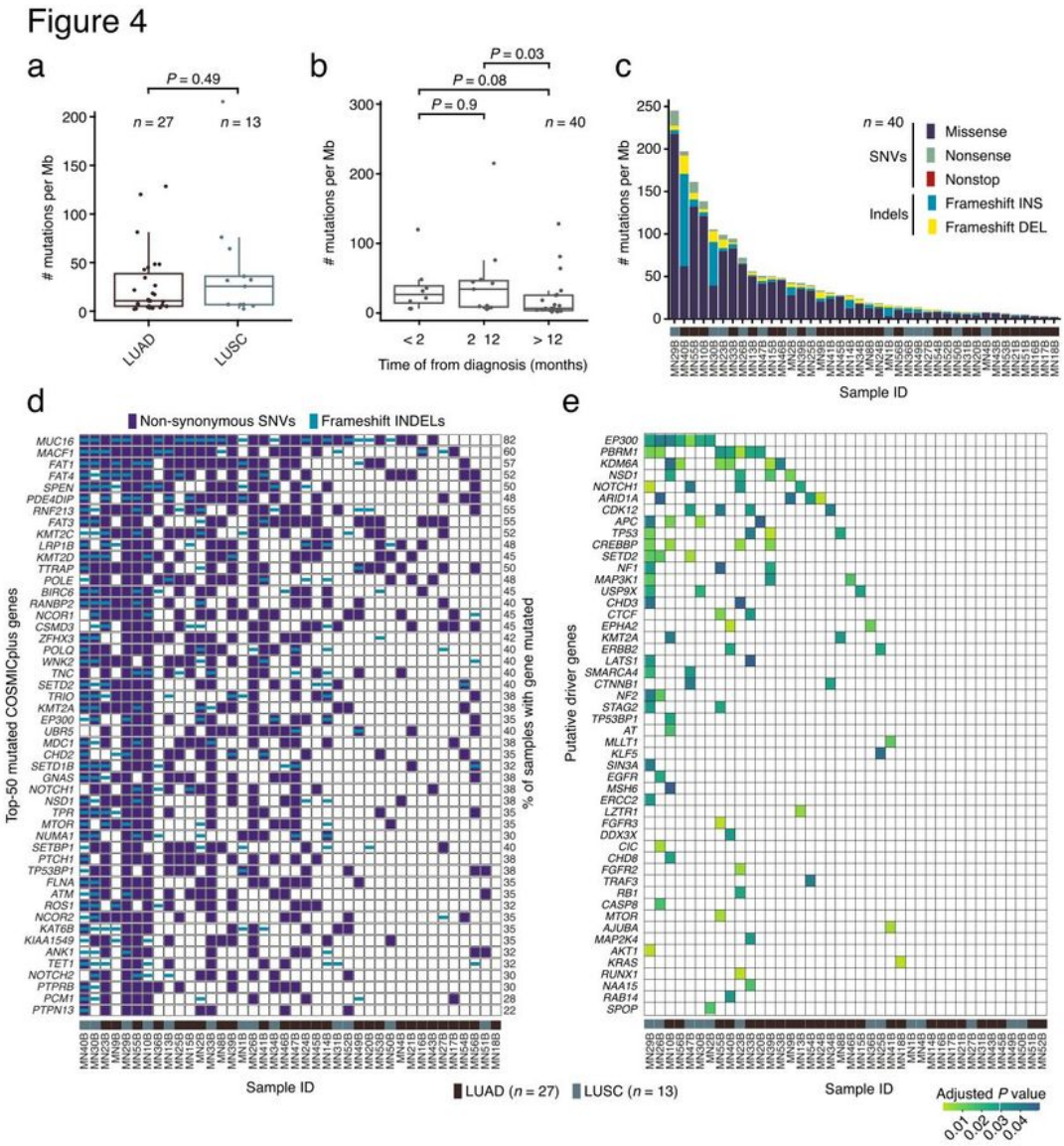

Figure 4

WES of matched primary tumor and BM samples allows identifying NSCLC-BM driving mutations. (**a**) Tumor mutation burden in 40 BM samples in the discovery cohort profiled by whole-exome sequencing (WES). LUAD, lung adenocarcinoma. LUSC, lung squamous cellular carcinoma. *P*, Wilcoxon test, two-tailed. (**b**) Same as in (a) but stratifying patients based on the timing of BM diagnosis. (**c**) Number of different mutation types in each of the 40 BM samples profiled by WES. SNVs, single-nucleotide variants.

Indels, small insertions and deletions. (**d**) Top-50 COSMICplus (see **Supplementary Table 5**) mutated genes in the 40 BM samples in the discovery cohort profiled by WES. (**e**) COSMICplus genes annotated as high-confidence BM drivers based on (CHASM)[35] . The sample IDs in (c) and (d) are the same as in **Supplementary Table 1**.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- SupplementaryTable1.xlsx
- SupplementaryTable10.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable5.xlsx
- SupplementaryTable6.xlsx
- SupplementaryTable7.xlsx
- SupplementaryTable8.xlsx
- SupplementaryTable9.xlsx