Whole-exome sequencing identifies somatic mutations associated with lung cancer metastasis to the brain

Zhenghao Liu[1#], Meiguang Zheng[1#], Bingxi Lei[1#], Zhiwei Zhou[1], Yutao Huang[1], Wenpeng Li[1], Qinbiao Chen[1], Pengcheng Li[2], Yuefei Deng[1]

[1]Department of Neurosurgery, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China; [2]Department of Thoracic Oncology, Cancer Center of Wuhan Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
Contributions: (I) Conception and design: P Li, Y Deng; (II) Administrative support: P Li, Y Deng; (III) Provision of study materials or patients: B Lei, Z Zhou, Y Huang; (IV) Collection and assembly of data: W Li, Q Chen; (V) Data analysis and interpretation: Z Liu, M Zheng; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.
[#]These authors contributed equally to this work.
Correspondence to: Prof. Yuefei Deng. Department of Neurosurgery, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, 107 Yanjiang West Road, Guangzhou 510120, China. Email: dengyuefei@sysu.edu.cn; Prof. Pengcheng Li. Department of Thoracic Oncology, Wuhan Union hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China. Email: lipc1981@163.com.

Background: Lung cancer is the most aggressive cancer, resulting in one-quarter of all cancer-related deaths, and its metastatic spread accounts for >70% of these deaths, especially metastasis to the brain. Metastasis-associated mutations are important biomarkers for metastasis prediction and outcome improvement.
Methods: In this study, we applied whole-exome sequencing (WES) to identify potential metastasis-related mutations in 12 paired lung cancer and brain metastasis samples.
Results: We identified 1,702 single nucleotide variants (SNVs) and 6,131 mutation events among 1,220 genes. Furthermore, we identified several lung cancer metastases associated genes (*KMT2C*, *AHNAK2*). A mean of 3.1 driver gene mutation events per tumor with the dN/dS (non-synonymous substitution rate/ synonymous substitution rate) of 2.13 indicating a significant enrichment for cancer driver gene mutations. Mutation spectrum analysis found lung-brain metastasis samples have a more similar Ti/Tv (transition/ transversion) profile with brain cancer in which C to T transitions are more frequent while lung cancer has more C to A transversion. We also found the most important tumor onset and metastasis pathways, such as chronic myeloid leukemia, ErbB signaling pathway, and glioma pathway. Finally, we identified a significant survival associated mutation gene *ERF* in both The Cancer Genome Atlas (TCGA) (P=0.01) and our dataset (P=0.012).
Conclusions: In summary, we conducted a pairwise lung-brain metastasis based exome-wide sequencing and identified some novel metastasis-related mutations which provided potential biomarkers for prognosis and targeted therapeutics.

Keywords: Whole-exome sequencing (WES); lung cancer (LC); somatic mutation; brain metastasis

Page 2 of 12

Liu et al. WES identifies lung cancer metastasis to the brain

Introduction

Lung cancer (LC) is the leading cause of cancer death in both men and women, accounting for one-quarter of all cancer deaths (1). The 5-year survival rate has failed to improve significantly over the last 30 years and remains at a mere 19%, due to recurrence and metastasis. Metastasis accounts for about 90% of cancer-related deaths and is the inevitable outcome of most human tumors. The most common site of lung cancer metastasis is the brain, and about 50% of all lung cancers develop into brain metastasis (BM) during the disease process (2,3). It has been reported that the rate of brain metastasis from lung cancer has recently increased, placing a great burden on public health services.

The 'seed-and-soil' hypothesis, which is the most widely accepted hypothesis for the formation of metastasis, denotes that the growth of metastatic cancer cells depends on the intrinsic abilities of the cancer cells themselves ('seeds') and the target organ microenvironment ('soil') (4). The cancer cell population has multiple genetically heterogeneous subpopulations (5). Metastasis is a Darwinian natural selection process in which 'seeds', with distinct metastatic traits that enable them to obtain metastatic advantage, are selected from a genetically- and epigenetically-heterogeneous tumor cell subpopulation (6,7). The advantageous lung 'seeds' proliferate in brain 'soil' that provides a congenial ground and form metastatic brain tumors, whose genetic landscape is reshaped (8).

A large number of studies have attempted to weigh BM relevant factors in lung cancer, revealing factors that include young age (<60 years) (9,10), non-squamous cell carcinoma (10,11), and the presence of clinically bulky mediastinal lymph nodes (≥2 cm) (9) are associated with a high BM rate. However, other studies have reported conflicting results (12,13). Many candidate metastasis genes have also been found to be involved in metastasis through changes in gene expression levels (14). The expression levels of E-cadherin, N- cadherin, KIFC1, and FALZ may be used to identify patients at high risk of lung-brain metastasis (15,16). However, the molecular basis of metastatic gene expression remains largely unknown, and the genetic profiles of brain metastases from LC might give us a close insight into tumor initiation, dissemination, and local progression (14). It is universally acknowledged that the development of lung cancer determined by gene mutations and these oncogenes play as critical drivers in tumor metastasis. The identification of driver oncogenes and the development of targeted therapies have improved lung cancer patients' outcomes. WES provide a convenient approach for understanding tumor mutation, mutational burden and further guide therapeutic decision making. To reveal the molecular mechanisms and the genetic alterations involved in metastasis of lung tumors to the brain, we carried out whole-exome sequencing (WES) of the primary tumors and their corresponding BMs from 12 patients with metastatic non-small-cell lung carcinoma. Our study might be instrumental for the identification of new genetic targets which may provide new therapeutic strategies for the design of drug intervention to improve the severity of the disease.

We present the following article in accordance with the STREGA reporting checklist (available at http://dx.doi.org/10.21037/atm-21-1555).

Methods

Patients and specimens

The pairwise lung-brain tumor samples and adjacent histologically normal tissue samples from 12 patients were collected at the Sun Yat-sen Memorial Hospital from 2010 to 2015. All procedures performed in this study involving human participants were in accordance with the Declaration of Helsinki (as revised in 2013). The Ethics Committee at Sun Yat-sen Memorial Hospital approved the utilization of samples. All informed consents were taken from 12 LC patients. A total of 12 paired samples were subjected to hematoxylin and eosin (HE) staining, then the normal cells were isolated from the tumor cells through histopathological examination. The normal cells were regarded as matched normal controls. The total amount of DNA extracted from the archived formalin-fixed paraffin-embedded (FFPE) samples of tumor tissue was up to standard and qualified.

Next-generation sequencing, variant calling, and annotation

The genomic DNAs were exacted and sonicated to an average size of 200 bp. The targeted DNA fragments were captured and pulldown and exon-wide libraries were created using the Roche SeqCap EZ Exome V3 (Roche Diagnostics, Basel, Switzerland) and TruePrep DNA Library Prep Kit V2 for Illumina (#TD501, Vazyme, Nanjing, China), and paired-end sequence data were generated using Illumina HiSeq machines (Illumina, Inc., San Diego, CA, USA).

The sequence data were aligned to the human reference genome (NCBI build 37) using BWA (Burrows-Wheeler Aligner) (17) and sorted and removed polymerase chain reaction (PCR) duplication using GATK 4.1.2.0 (https://gatk.broadinstitute.org/) (18). Somatic mutation calling was performed using Mutect1, Mutect2 (18), and VarDict (19). Somatic mutations existing in at least 2 of the results of the 3 software were selected as high confident mutations and to be involved in the further bioinformatics and bio-statistical analysis. Copy number variants (CNVs) from WES data were detected by CNVKIT (Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing) (20). The dN/dS (non-synonymous substitution rate/synonymous substitution rate) ratios of each domain were calculated via The DiversiTools according to Xia's description (21). The GISTIC2.0 (Genomic Identification of Significant Targets in Cancer) algorithm was used to identify regions of the genome that are significantly amplified or deleted across a set of samples (22). Somatic variants were annotated by Ensembl Variant Effect Predictor (23). Transition (Ti) and transversion (Tv) ratios were applied to measure the selection in cancer genomes and to show mutation characteristics between different cancer types.

Spectrum, signatures of somatic mutations, and subclonal architecture inference

Non-negative matrix factorization and model-selection were applied to delineate mutational processes underlying genome-wide single nucleotide variants (SNVs) and to identify the major mutational signatures (24). In the cohort, 1 sample was considered as a strong association with 1 mutational signature if the proportion of the contribution >20% with MutationalPatterns (version 1.10) (25) and deconstructSigs (26). The sub-clonal architecture of tumors was inferred by sciClone (27) and clonevol (28).

Survival analysis, enrichment analysis

Pathway analysis was based on the Database for Annotation, Visualization and Integrated Discovery (DAVID) bioinformatics (29), and WebGestalt (30) with significant Benjamini adjusted P value (P<0.05). The co-mutation profile was prepared with R package ComplexHeatmap (31). The Cancer Genome Atlas (TCGA) mutation and survival data were downloaded from the Genomic Data Commons (GDC) database (https://portal.gdc.cancer.gov/exploration).

In the validation study to *ERF*, we downloaded the expression and survival data of *ERF* mutation and expression from TCGA project, and Cox-regression was conducted to binary gene expression data dichotomized by median expression level for TCGA dataset.

Statistical analysis

Cox regression was applied for survival analysis between mutation and overall survival (OS) time and the Kaplan-Meier plot were used to show the difference between survival time among different groups. Since our research involved a small sample size of participants, all the clinically-related statistical analysis was considered to be significant when P<0.05 without multiple correction testing. Data is shown as the mean ± standard deviation (SD), and a two-tailed P value <0.05 was considered statistically significant.

Results

Landscape of somatic mutations in primary LC and BM

We collected and quantified DNA from the original 12 non-small cell lung cancer (NSCLC) patient FFPE samples and the matched BM samples. The average coverage depth for the tumor cells and normal cells were 194× and 120×, respectively. Detailed clinicopathological information is summarized in Table S1. We identified 1,702 SNVs and 6,131 mutation events in 1,220 genes from 12 paired LC and BM, including the most frequent LC driver gene mutations such as *TP53*, *EGFR*, *BRCA1*, *BRCA2*, and *BRAF*. We identified a mean of 3.1 driver gene mutation events per tumor with the dN/dS of 2.13 which is slightly higher than non-metastatic LC samples in TCGA, indicating a significant enrichment for the cancer driver gene mutations. We did not find any difference in the dN/dS ratio between primary tumor (dN/dS =2.20), BM tumor (dN/dS =2.06), and shared mutations between LC and BM (dN/dS =2.25).

We identified several LC metastases-associated genes (*KMT2C*, *AHNAK2*, *PDE4DIP*, *ANKRD36C*, and *BAGE2*), and the mutations of these genes showed distribution diversity among the LC and BM samples (*Figure 1A,B*). Mutations of *KMT2C* were found in 25% of LC samples; however, the mutation frequency in BM was up to 50%, indicating the positive selection of *KMT2C* mutations during metastasis. According to TCGA dataset *AHNAK2* have significant enrichment in LC, with a mutation ratio of 18.8% in LC while 9.98% in pan-cancer (P=7.2×10^{-9}, Chi-

Page 4 of 12

Liu et al. WES identifies lung cancer metastasis to the brain
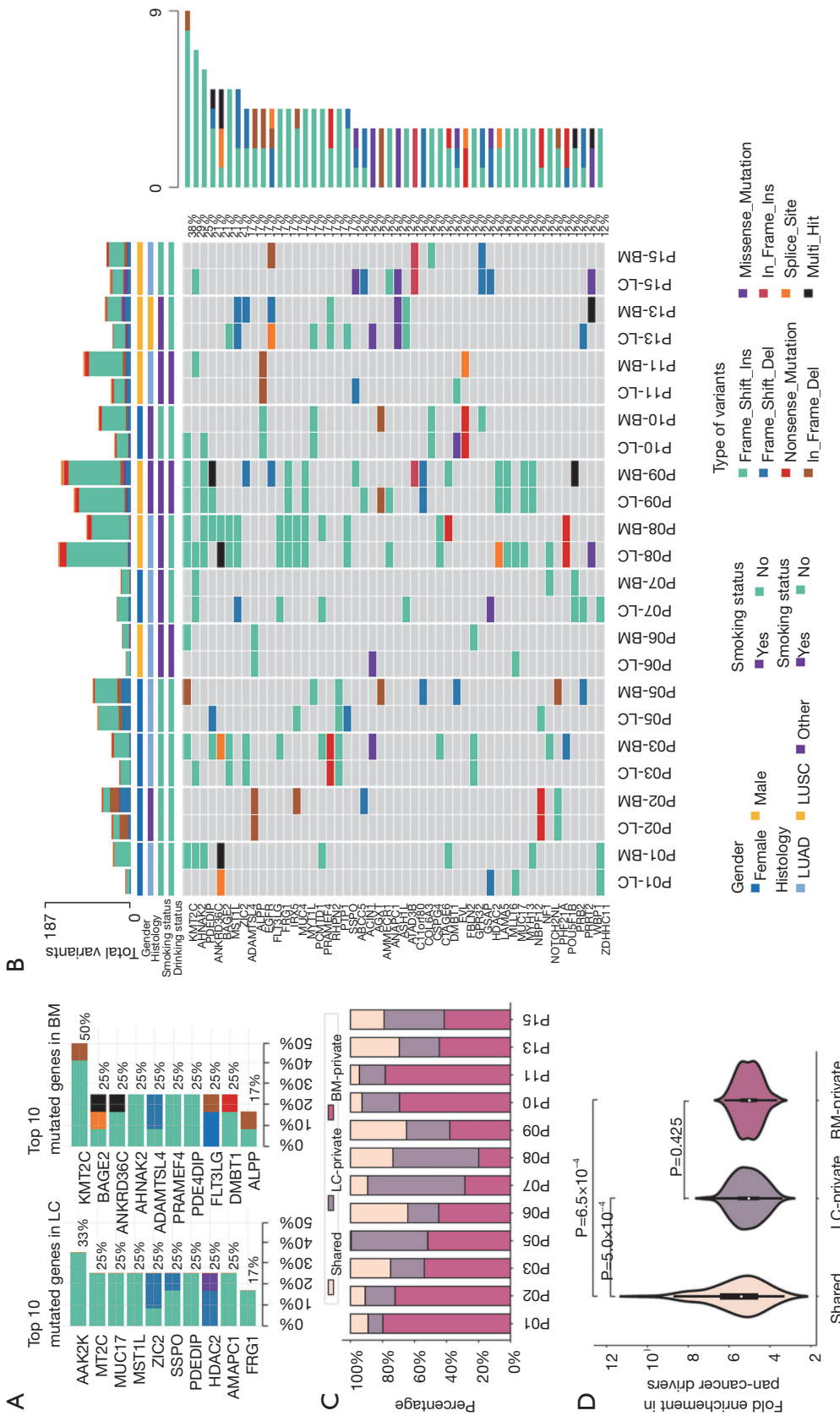
Figure 1 The mutational landscape and patterns of genetic divergence in paired primary lung cancer and brain metastases. (A) Top mutated genes in LC and brain cancer samples; (B) tumor-specific mutations and lung-brain shared mutations; (C) drive mutation genes in tumor specific mutations and shared mutations; (D) oncoprint of lung and brain mutations in all the samples. LC, lung cancer.

square test). However, the mutation frequency of *AHNAK2* in our dataset is as high as 26.9% which is 1.43 times that of the LC population in TCGA dataset (P=0.02, Chi-square test). We also observed all *EGFR* mutations were shared between LC and BM, suggesting that *EGFR* mutations are drivers and likely to be an early event before BM.

We found more somatic mutations in BM lesions (median 71, range 23–180) than in LC lesions (median 48.5, range 13–187), while the difference was not statistically significant (P=0.069, Student's *t*-test) (Figure S1A,B). High correlation between TMB (tumor mutation burden) of LC and TMB of BM were confirmed by Pearson coefficient 0.65 (P=0.02) (Figure S1C,D), indicating the TMB of BM can be estimated by that of primary LC when brain tissue is not available, so as to screen patients who will most likely benefit from programmed death-ligand 1 (PD-L1) immunotherapy. A sum of 18.2% (0.5–35.9%) of all mutations were shared between LC and BM, clearly suggesting a common ancestral truncal clone with 30.0% (9.3–60.8%) LC-specific and 51.8% (19.9–79.7%) BM-specific, respectively (*Figure 1C*). Although metastases had more private mutations than the primary tumor, they were not enriched for the pan-cancer driver genes (32) (*Figure 1D*). This suggested that few additional private genomic driver genes were required for metastasis when the primary cancer was already advanced.

In order to provide more landscape for the mutations identified in our study, we conducted a pathway analysis to the most frequently mutated genes (mutation frequency >5%) (Table S2). We found the most important tumor pathways were chronic myeloid leukemia (P=0.002), ErbB signaling pathway (P=0.0014), and glioma pathway (P=0.05). Keyword enrichment indicated important metabolic abnormalities for the lung-metastasis cancers including EGF-like domain and tyrosine-specific phosphatase (Table S3).

Copy number variations (CNVs)

To further explore BM-related molecular events, genomic CNVs were analyzed: 8q21.2, 6p22.1, 12p13.33, and 5q35.3 were the most common chromosome deleted regions in both LC and BM, and 8q24.13 were the most common regions with gain copy numbers in both LC and BM (Figure S2A). Loss of 6p22.1, which harbors *HLA-A*, *HLA-G*, and *HLA-H*, was most frequent in both the LC and BM of participant 9 and in BM of participant 11 (Figure S2B). Interestingly, these samples also had high

TMB (Figure S1C). This may be due to the loss of *HLA* function associated with higher overall mutation burden and a larger fraction of HLA-binding neoantigens (33). The recurrent deletion of *HLA* was detected as the early events, indicating the important role of the immune system in LCBM, and that these patients may benefit from immunotherapy.

The differences include gains of chromosomes 7q35 and loss of 7q22.1 and 7q36.3, which were more frequent in metastasis samples, and gains of chromosomes 11q13.2 and losses of 7q11.23 and 2q13, which were less frequent. Most recurrent CNV regions were shared in LC and BM samples, indicating that CNVs are early molecular events in tumorigenesis and metastasis.
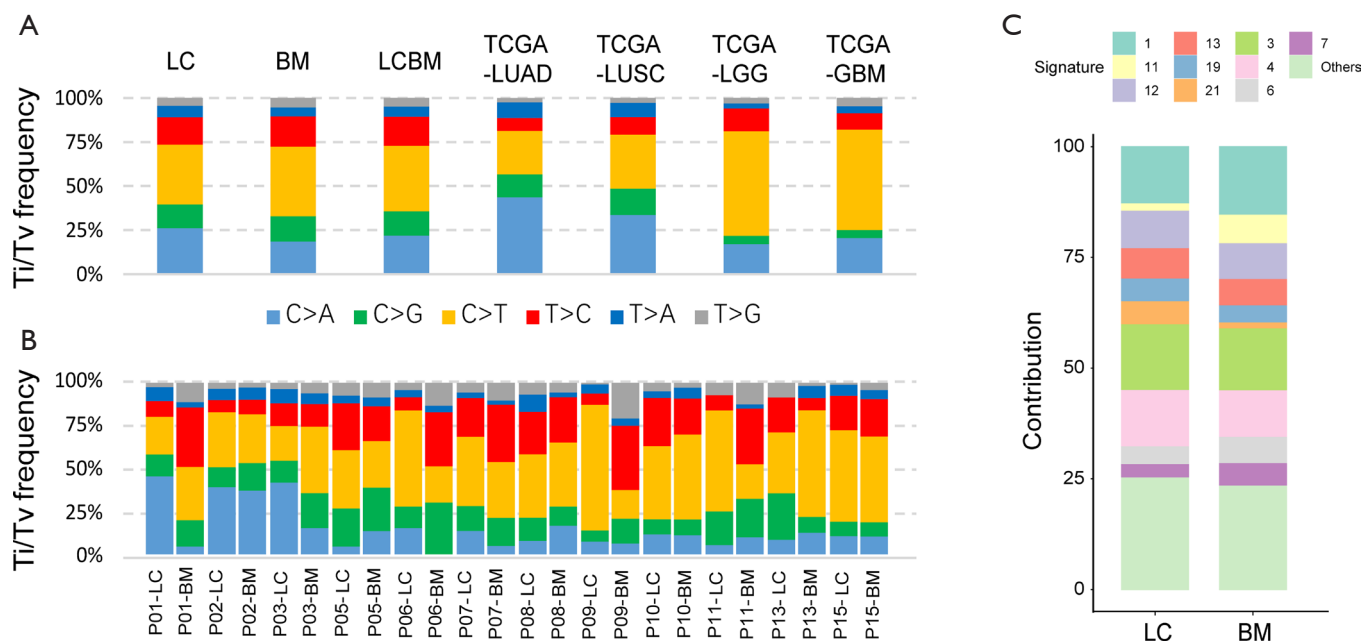
Spectrum and signatures

To determine the relationships between the mutational spectra and tumor organ sites, we analyzed the spectra of LC and BM from our study and primary lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), low grade glioma (LGG), and glioblastoma multiforme (GBM) from TCGA dataset. The C to T substitution was the most common transition in our LC and BM samples, which was much closer to primary brain cancer (LGG and GBM) but significantly different from primary lung cancer (LUAD and LUSC) which has higher C to A transversion (*Figure 2A*). These evidences were consistent with our hypothesis that the mutations identified in our study have a higher probability of being associated with BM. The mutational spectrum of LC and BM samples from the same individuals are more similar to each other than those from different patients, implying that different mutational processes were involved during the development of metastasis between the different patients (*Figure 2B*).

We further analyzed mutational signatures in BM and LC and signatures 1, 3, and 4, which have been linked to aging, BRCA1/2 mutations, and smoking, respectively, were identified as dominant in either BM or LC samples (34) (*Figure 2C*). There was no significant difference between the signature levels of LC and BM tissues (Wilcoxon rank sum test, P>0.05), indicating the change of mutational signatures happened before metastasis and may not lead to their great difference between the two groups.

Clonal evolution during the development of LCBM

Phylogenetic trees give clear overviews of the order

Page 6 of 12

Liu et al. WES identifies lung cancer metastasis to the brain

Figure 2 Spectrum and signatures. (A) Mutation spectrum of lung tumor specific mutations, brain tumor specific mutations, shared mutations, and TCGA lung and brain mutations; (B) mutation spectrum of all the samples in our cohort; (C) mutation signatures of mutations for lung tumor samples and brain tumor samples. TCGA, The Cancer Genome Atlas.

of mutation events, allowing the track of emergence and movement of clones from LC to BM (*Figure 3*). Phylogenetic trees of the 9 participants showed that mutations on the trunk were probably earlier genetic alteration events, followed by mutations on the branch which occurred later during tumorigenesis and BM development. Clonal evolution analyses revealed that LC and BM tumors had the same evolutionary process in 3 participants (P02, P07, and P10), LC tumors harbored a cluster of LC-private clones in another 3 participants (P05 and P08), and BM tumors harbored clones that were nonexistent in matched LC tumors in 3 other participants (P06, P13, and P15), indicating the mutations on BM-private clones may contribute to metastatic progression.

OS by genotype

In order to identify independent prediction factors for outcomes, we conducted survival analysis on several potential factors. Participants with aberrations of 18 genes in LC (Figure S3) and 15 genes in BM (Figure S4) had significantly worse OS than those without these aberrations (P<0.05). Of these genes, we identified a significant survival-associated mutation gene, *ERF*, which was confirmed

by both TCGA (P=0.01) and our dataset (P=0.012) (*Figure 4A,B,C*). Additionally, in order to show the prognostic roles of *ERF*, we also found that high expression of *ERF* genes in TCGA was a significant risk factor for the OS time [hazard ratio (HR) =1.46, P<1.2×10^{-22}, *Figure 4D*]. Taken together, our findings revealed an important role for ERF in prognostic prediction of LC.

Multivariate analysis demonstrated gender (P=2.02× 10^{-119}), smoking status (P=1.21×10^{-269}), metastatic tumor size (P=0), and the ratio of shared mutations in lung and brain cancers (P=0.019) were significantly associated with OS time, while no significant associations were found in drinking status (P=0.996), the number of metastatic tumors (P=0.746), mutation numbers of primary tumor (P=0.840), or metastatic tumor (P=0.248) (*Figure 4D*).

Discussion

The metastatic cascade involves multiple steps, including invasion, entry into the circulation from the primary tumor, systemic dissemination, arrest and extravasation in secondary organs, settlement into latency, reactivation, outgrowth, and potential seeding of tertiary metastasis (35). Genetic and epigenetic changes accumulating among
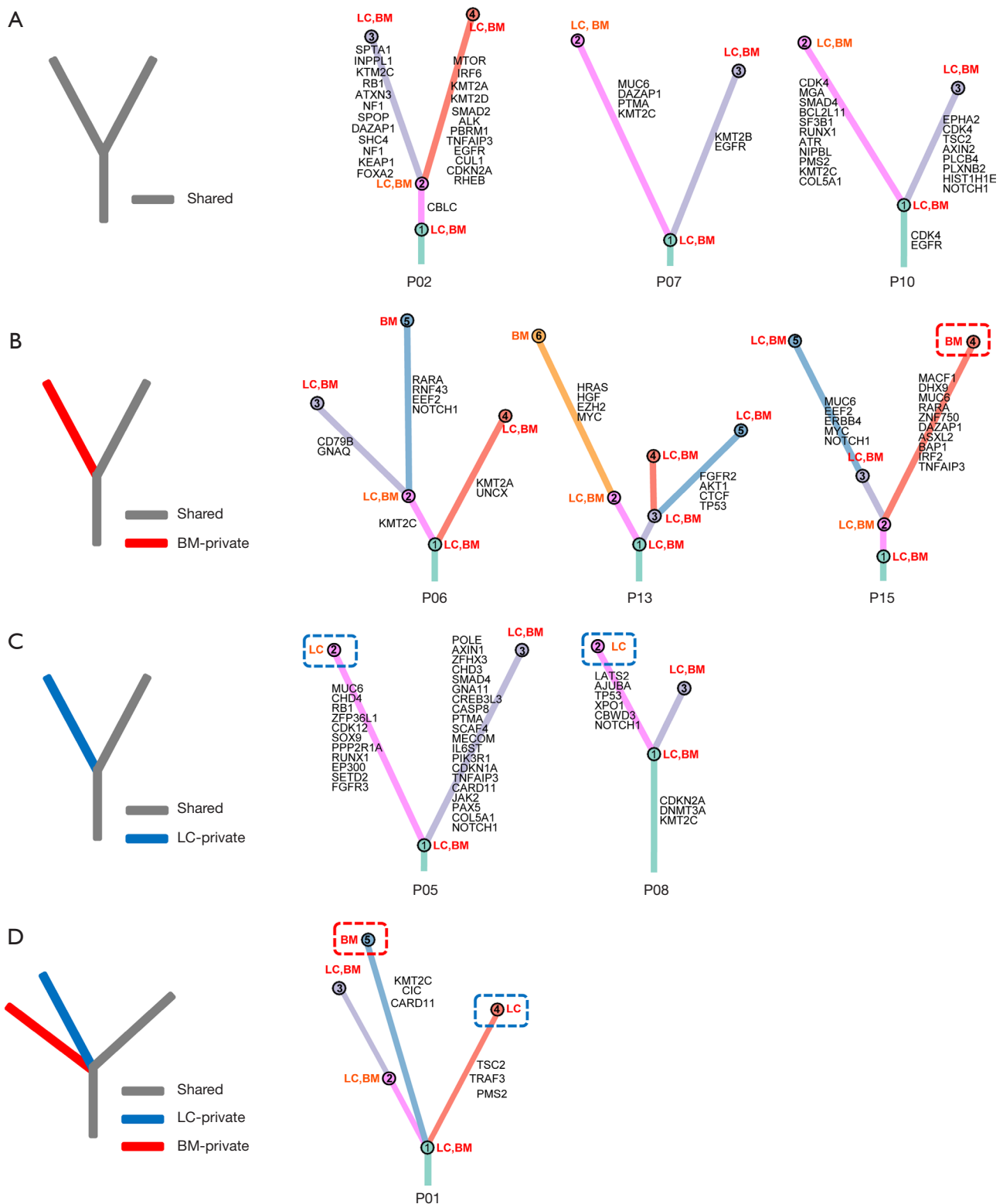
Figure 3 Phylogenetic trees indicating 4 types of evolution. (A) LC tumors and BM tumors had the same evolutionary process; (B) BM tumors harbor a cluster of BM-private clones; (C) LC tumors harbor a cluster of LC-private clones; (D) LC and BM tumors harbor a cluster of LC-private clones and a cluster of BM-private clones respectively. LC, lung cancer; BM, brain metastasis.

Page 8 of 12

Liu et al. WES identifies lung cancer metastasis to the brain
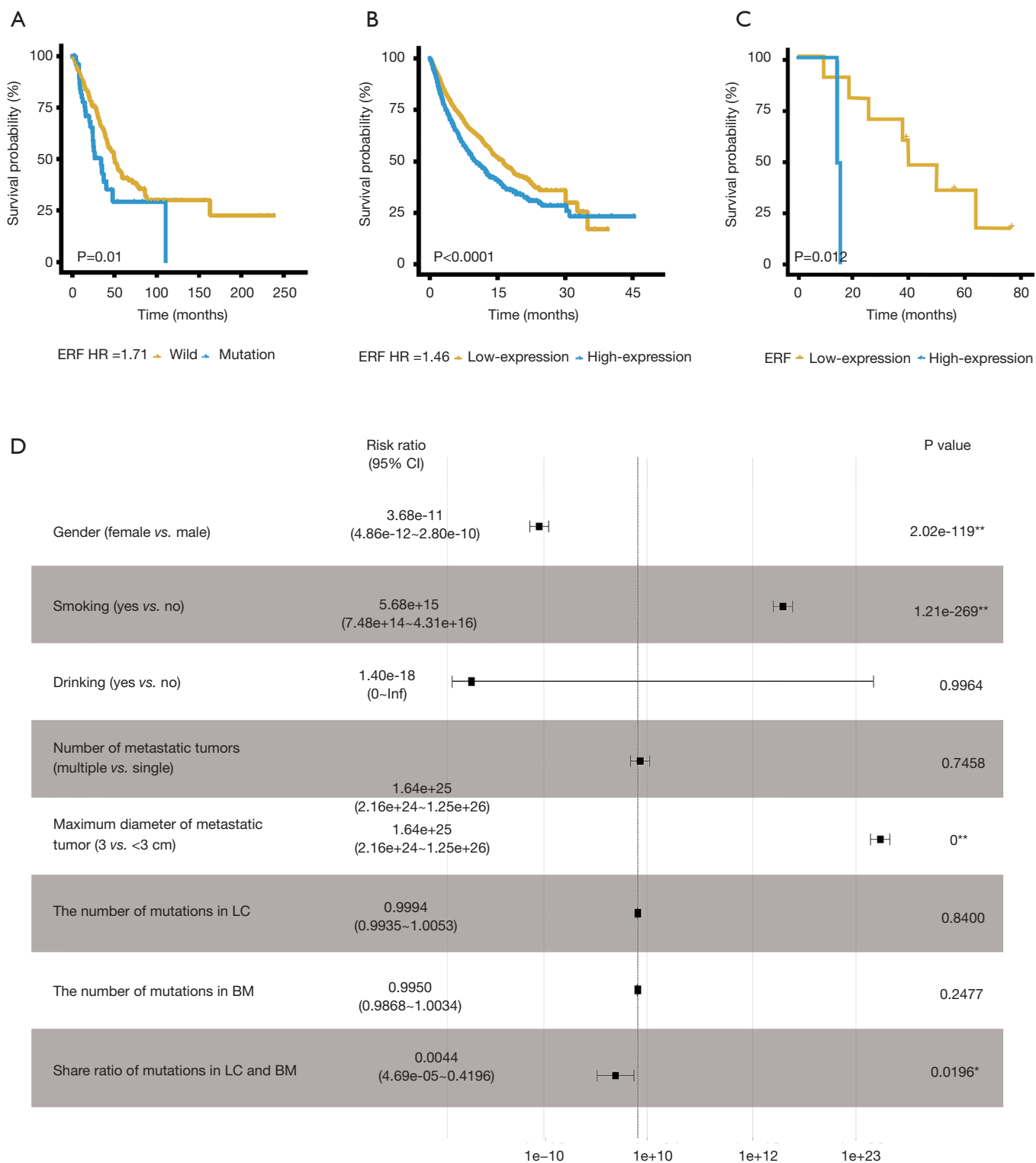
Figure 4 Overall survival analysis. (A) Survival analysis of several potential factors; (B) survival analysis between ERF mutation and OS times in TCGA; (C) survival analysis between ERF expression and OS times in TCGA; (D) survival analysis between ERF mutation and OS times in our cohort. *, P<0.05; **, P<0.01. OS, overall survival; TCGA, The Cancer Genome Atlas.

primary tumor cells and metastases may contribute to these multiple steps of the metastatic cascade (36,37). Whole-exome sequencing is a convenient to acquire comprehensive characterization of tumor mutations, from which neoantigens, tumor mutational burden (TMB), and clonality can be obtained. Even though there are previous study characterized the genomic difference between primary tumors and paired brain metastases in lung cancer patients by whole-exome sequencing, this study has larger sample size in comparison to previous study; survival analysis between mutation and overall survival (OS) time was conducted (38).

Spinal metastasis, hepatic metastasis, bone metastasis, pleural dissemination is the important modes of metastasis and a poor prognostic factor of lung cancer. Tang using SCIMET tool to value timing of lung cancer dissemination and found the metastatic seeding happened approximately 2.74 years before clinical detection and 87.5% non-lymph node metastases were mainly seeded by the primary tumors (39). Hence, it is necessary to collect a well-defined cohort of matched primary tumors and BM and perform comparative deep sequencing analyses to acquire some biomarkers of metastasis. However, most patients with brain metastasis of lung cancer (LCBM), which were in the late stage at diagnosis, are typically treated with palliative approaches such as chemotherapy, targeted therapy, and whole-brain radiotherapy instead of neurosurgical resection. Thus, researchers rarely have the opportunity to investigate matched primary-metastatic tumors in the mutation status of analyzed genes between tumor sites on a large-scale (40).

In this study, we collected 12 paired LC and BM samples and identified some genes associated with LCBM. The gene *KMT2C* (lysine-specific methyltransferase 2C, also known as MLL3), which belongs to the mixed-lineage leukemia (MLL) family of histone methyltransferases, was the most commonly mutated gene among our samples. Recent studies have revealed frequent mutations of *KMT2C* in several epithelial and myeloid cells, and it has been identified as a putative tumor suppressor (41,42). It has been proposed that in metastatic spread, circulating tumor cell (CTC) populations in the blood of carcinoma patients contain cells with the clonal capacity to initiate metastatic growth in distant organs, thus having similarity with the hematologic tumor. The involvement of *KMT2C* was originally identified as oncogenic fusions in leukemia (43), and the most enriched pathway in our study was chronic myeloid leukemia. Aljohani *et al.* reported that the *Keap1-Nrf2*-ARE pathway

was mutated in NSCLC patients that had metastasized to the brain and in CTC according to WGS (44). Recent reports have demonstrated that metastatic brain tumors of NSCLC and LUAC patients carried higher EGFR mutations rate than those without metastases (45,46). Our results further emphasized the association between the BM of LC and leukemia. The gene AHNAK nucleoprotein 2 (*AHNAK2*) is a prognostic marker and an oncogenic protein for clear cell renal cell carcinoma and hypoxic upregulation of *AHNAK2* supports epithelial-mesenchymal transition (EMT) and cancer cell stemness (47). Cancer cells acquire characteristics of self-renewal, motility, and invasiveness, traits that facilitate metastatic dissemination during EMT (48). That is, the driver gene mutation of *AHNAK2* may promote metastatic colonization of the lung to the brain by supporting EMT.

Our findings, depicting that chronic myeloid leukemia and the ErbB signaling pathway were mutated in the majority of LCBM patients, supports our hypothesis that mutations in these pathways may indeed provide a survival advantage to these cells and help them reach distant sites. Of note, glioma pathways have also been identified. The Ti/Tv profile showed our mutation profile was much closer to brain cancer mutation profile as brain cancer has high C to T transitions were more frequent, while LC had a higher frequency of C to A transversion. These evidences strongly indicate that brain tumor-related events are involved in the process of LCBM.

Multivariate analysis demonstrated that the high ratio of shared mutations were associated with better prognosis (49,50). This implies that patients might have had a preferable prognosis when the tendency of metastatic cancer mutation was more inclined to primary cancer, that is the BM sample might not have evolved from the primary cancer but rather they had a shared antecedent. Limited inter-tumor heterogeneity between LC and BM within the same patient results in effective postoperative chemotherapy and radiotherapy in this circumstance.

Overall, we revealed the genomic difference between metastatic and primary tissues, and identified several genes associated with LCBM. While further molecular biology studies to validate the role of identified candidates and a larger-scale of LCBM samples would be required to confirm our findings, this study provides a preliminary evidence of the genetic evolution of LC metastases and furthers our advantage in revealing the therapeutic vulnerabilities of LC metastatic tumors.

Page 10 of 12

Liu et al. WES identifies lung cancer metastasis to the brain

Footnote

Reporting Checklist: The authors have completed the STREGA reporting checklist. Available at http://dx.doi.org/10.21037/atm-21-1555

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-21-1555

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-21-1555). All authors thank Shanghai Tongshu Biotechnology Co., Ltd., for technical support and have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All procedures performed in this study involving human participants were in accordance with the Declaration of Helsinki (as revised in 2013). This study protocol was approved by the Ethics Committee of Sun Yat-sen Memorial Hospital and Tongji Medical College. All informed consents were taken from 12 LC patients.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7-34.
2. Mujoomdar A, Austin JH, Malhotra R, et al. Clinical predictors of metastatic disease to the brain from non-small cell lung carcinoma: primary tumor size, cell type, and lymph node metastases. Radiology 2007;242:882-8.
3. Guo S, Yan F, Xu J, et al. Identification and validation of the methylation biomarkers of non-small cell lung cancer (NSCLC). Clin Epigenetics 2015;7:3.
4. Langley RR, Fidler IJ. The seed and soil hypothesis revisited--the role of tumor-stroma interactions in metastasis to different organs. Int J Cancer 2011;128:2527-35.
5. Landau DA, Carter SL, Getz G, et al. Clonal evolution in hematological malignancies and therapeutic implications. Leukemia 2014;28:34-43.
6. Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012;481:306-13.
7. Naxerova K, Jain RK. Using tumour phylogenetics to identify the roots of metastasis in humans. Nat Rev Clin Oncol 2015;12:258-72.
8. Yousefi M, Bahrami T, Salmaninejad A, et al. Lung cancer-associated brain metastasis: Molecular mechanisms and therapeutic options. Cell Oncol (Dordr) 2017;40:419-41.
9. Ceresoli GL, Reni M, Chiesa G, et al. Brain metastases in locally advanced nonsmall cell lung carcinoma after multimodality treatment: risk factors analysis. Cancer 2002;95:605-12.
10. Ji Z, Bi N, Wang J, et al. Risk factors for brain metastases in locally advanced non-small cell lung cancer with definitive chest radiation. Int J Radiat Oncol Biol Phys 2014;89:330-7.
11. Wang SY, Ye X, Ou W, et al. Risk of cerebral metastases for postoperative locally advanced non-small-cell lung cancer. Lung Cancer 2009;64:238-43.
12. Cox JD, Scott CB, Byhardt RW, et al. Addition of chemotherapy to radiation therapy alters failure patterns by cell type within non-small cell carcinoma of lung (NSCCL): analysis of radiation therapy oncology group (RTOG) trials. Int J Radiat Oncol Biol Phys 1999;43:505-9.
13. Robnett TJ, Machtay M, Stevenson JP, et al. Factors affecting the risk of brain metastases after definitive chemoradiation for locally advanced non-small-cell lung carcinoma. J Clin Oncol 2001;19:1344-9.
14. Nguyen DX, Massague J. Genetic determinants of cancer metastasis. Nat Rev Genet 2007;8:341-52.
15. Yoo JY, Yang SH, Lee JE, et al. E-cadherin as a predictive marker of brain metastasis in non-small-cell lung cancer, and its regulation by pioglitazone in a preclinical model. J Neurooncol 2012;109:219-27.

16. Grinberg-Rashi H, Ofek E, Perelman M, et al. The expression of three genes in primary non-small cell lung cancer is associated with metastatic spread to the brain. Clin Cancer Res 2009;15:1755-61.

17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.

18. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.

19. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res 2016;44:e108.

20. Talevich E, Shain AH, Botton T, et al. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol 2016;12:e1004873.

21. Xia H, Beck AS, Gargili A, et al. Transstadial Transmission and Long-term Association of Crimean-Congo Hemorrhagic Fever Virus in Ticks Shapes Genome Plasticity. Sci Rep 2016;6:35819.

22. Mermel CH, Schumacher SE, Hill B, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12:R41.

23. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122.

24. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013;3:246-59.

25. Blokzijl F, Janssen R, van Boxtel R, et al. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med 2018;10:33.

26. Rosenthal R, McGranahan N, Herrero J, et al. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 2016;17:31.

27. Miller CA, White BS, Dees ND, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol 2014;10:e1003665.

28. Dang HX, White BS, Foltz SM, et al. ClonEvol: clonal ordering and visualization in cancer sequencing. Ann Oncol 2017;28:3076-82.

29. Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44-57.

30. Wang J, Vasaikar S, Shi Z, et al. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res 2017;45:W130-W137.

31. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016;32:2847-9.

32. Bailey MH, Tokheim C, Porta-Pardo E, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 2018;174:1034-5.

33. Castro A, Ozturk K, Pyke RM, et al. Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. BMC Med Genomics 2019;12:107.

34. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415-21. Erratum in: Nature 2013;502:258.

35. Paterlini-Bréchot P. Organ-specific markers in circulating tumor cell screening: an early indicator of metastasis-capable malignancy. Future Oncol 2011;7:849-71.

36. Popper HH. Progression and metastasis of lung cancer. Cancer Metastasis Rev 2016;35:75-91.

37. Turajlic S, Swanton C. Metastasis as an evolutionary process. Science 2016;352:169-75.

38. Tomasini P, Barlesi F, Gilles S, et al. Comparative genomic analysis of primary tumors and paired brain metastases in lung cancer patients by whole exome sequencing: a pilot study. Oncotarget 2020;11:4648-54.

39. Tang WF, Wu M, Bao H, et al. Timing and Origins of Local and Distant Metastases in Lung Cancer. J Thorac Oncol 2021. [Epub ahead of print]. doi: 10.1016/j.jtho.2021.02.023.

40. Preusser M, Berghoff AS, Koller R, et al. Spectrum of gene mutations detected by next generation exome sequencing in brain metastases of lung adenocarcinoma. Eur J Cancer 2015;51:1803-11.

41. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013;502:333-9.

42. Chen C, Liu Y, Rappaport AR, et al. MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. Cancer Cell 2014;25:652-65.

43. Ziemin-van der Poel S, McCabe NR, Gill HJ, et al. Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias. Proc Natl Acad Sci U S A 1991;88:10735-9.

44. Aljohani HM, Aittaleb M, Furgason JM, et al. Genetic mutations associated with lung cancer metastasis to the

Page 12 of 12

Liu et al. WES identifies lung cancer metastasis to the brain

brain. Mutagenesis 2018;33:137-45.

45. Matsumoto S, Takahashi K, Iwakawa R, et al. Frequent EGFR mutations in brain metastases of lung adenocarcinoma. Int J Cancer 2006;119:1491-4.

46. Liao L, Ji X, Ge M, et al. Characterization of genetic alterations in brain metastases from non-small cell lung cancer. FEBS Open Bio 2018;8:1544-52.

47. Wang M, Li X, Zhang J, et al. AHNAK2 is a Novel Prognostic Marker and Oncogenic Protein for Clear Cell Renal Cell Carcinoma. Theranostics 2017;7:1100-13.

48. Wan L, Pantel K, Kang Y. Tumor metastasis: moving

new biological insights into the clinic. Nat Med 2013;19:1450-64.

49. Gaidzik VI, Teleanu V, Papaemmanuil E, et al. RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. Leukemia 2016;30:2160-8.

50. Zhong Q, Peng HL, Zhao X, et al. Effects of BRCA1- and BRCA2-related mutations on ovarian and breast cancer survival: a meta-analysis. Clin Cancer Res 2015;21:211-20.

(English Language Editor: J. Jones)

Table S1 Clinicopathological information of the 12 NSCLC participants

Sample ID	Overall survival (months)	Gender	Histology	Survival Condition	Smoking	Drinking	Number of metastatic tumors	Maximum diameter of metastatic tumor (cm)	The number of mutations in LC	The number of mutations in BM	The number of mutations both in LC and BM	Share Ratio
P01	76	Female	Adenocarcinomas	Alive	No	No	Single	<3	63	206	31	0.492063
P02	10	Female	Large cell carcinoma	Dead	No	No	Single	>3	148	154	44	0.297297
P03	26	Female	Adenocarcinomas	Dead	No	No	Single	>3	104	173	50	0.480769
P05	15	Female	Adenocarcinomas	Dead	No	No	Multiple	>3	325	372	5	0.015385
P06	64	Male	Adenocarcinomas	Dead	Yes	Yes	Multiple	<3	62	110	38	0.612903
P07	40	Female	Adenocarcinomas	Dead	Yes	No	Multiple	<3	153	97	20	0.130719
P08	16	Male	Adenocarcinomas	Dead	Yes	No	Single	<3	582	345	191	0.328179
P09	56	Male	Clear cell carcinoma and tubular adenocarcinoma	Alive	Yes	Yes	Single	<3	463	556	241	0.520518
P10	38	Female	Adeno-squamous carcinoma	Dead	No	No	Single	>3	167	306	28	0.167665
P11	50	Male	Adenocarcinomas	Dead	Yes	Yes	Single	<3	187	490	22	0.117647
P13	19	Male	Squamous cell carcinomas	Dead	Yes	No	Single	<3	278	377	148	0.532374
P15	39	Male	Adenocarcinomas	Alive	No	No	Multiple	<3	308	212	73	0.237013

LC, lung cancer; BM, brain metastasis; NSCLC, non-small cell lung cancer.
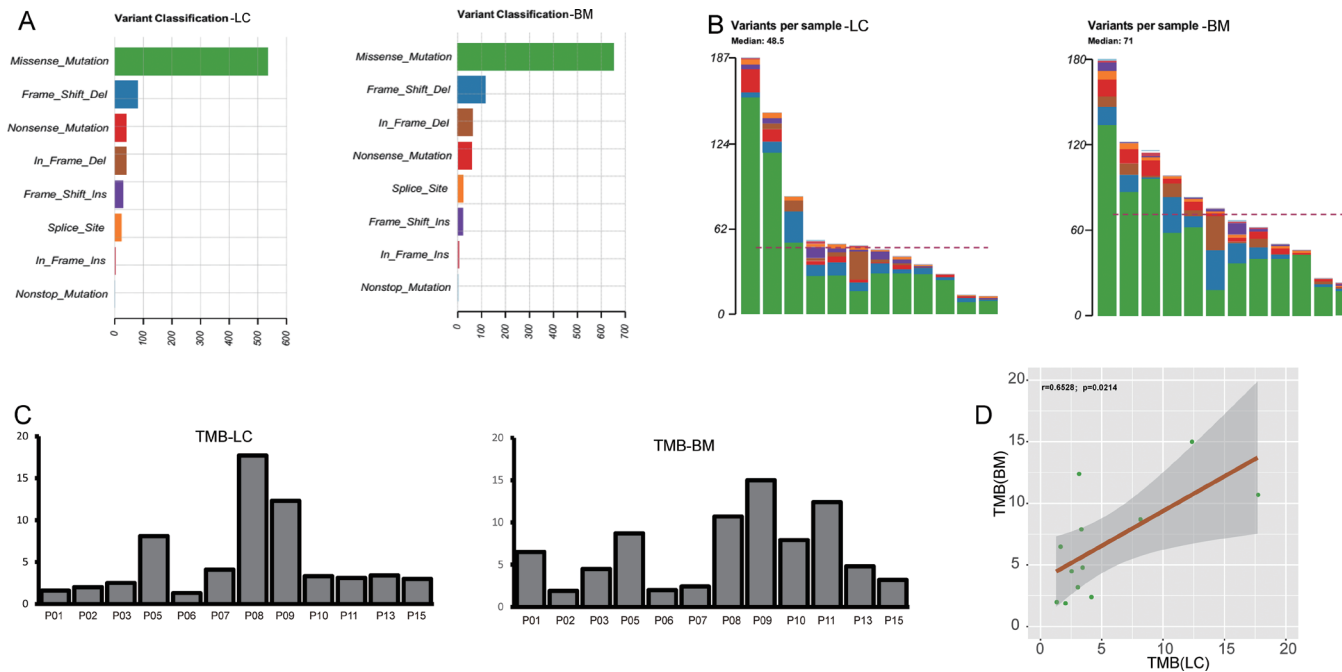
Figure S1 Landscape of somatic mutations of driver gene in LC and BM. (A) Number of variants, (B) classification of variant, (C) TMB of each patient and (D) the correlation between TMB of LC and BM. TMB, tumor mutation burden; LC, lung cancer; BM, brain metastasis.

Table S2 Enriched pathway of frequently mutated genes

Category	Term	Count	%	P value	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
KEGG_ PATHWAY	hsa05220:Chronic myeloid leukemia	18	1.393189	1.19E-05	BCR, PIK3CB, MAP2K2, STAT5A, STAT5B, CBL, TP53, MECOM, CBLC, CDKN2A, HDAC2, KRAS, SOS1, RUNX1, CRK, CHUK, SHC4, PIK3R2	359	75	5085	3.399443	0.001992
KEGG_ PATHWAY	hsa04012:ErbB signaling pathway	19	1.470588	2.51E-05	EGFR, NRG3, PIK3CB, MAP2K2, STAT5A, STAT5B, CBL, ELK1, CBLC, KRAS, SOS1, CAMK2B, MTOR, MAP2K7, CRK, NRG2, CAMK2A, PIK3R2, SHC4	359	87	5085	3.093363	0.001404
KEGG_ PATHWAY	hsa05214:Glioma	13	1.006192	0.001237	EGFR, PIK3CB, MAP2K2, TP53, CDKN2A, KRAS, SOS1, CAMK2B, CALML6, MTOR, CAMK2A, PIK3R2, SHC4	359	63	5085	2.922801	0.050663

Table S3 Enrichment analysis of frequently mutated genes

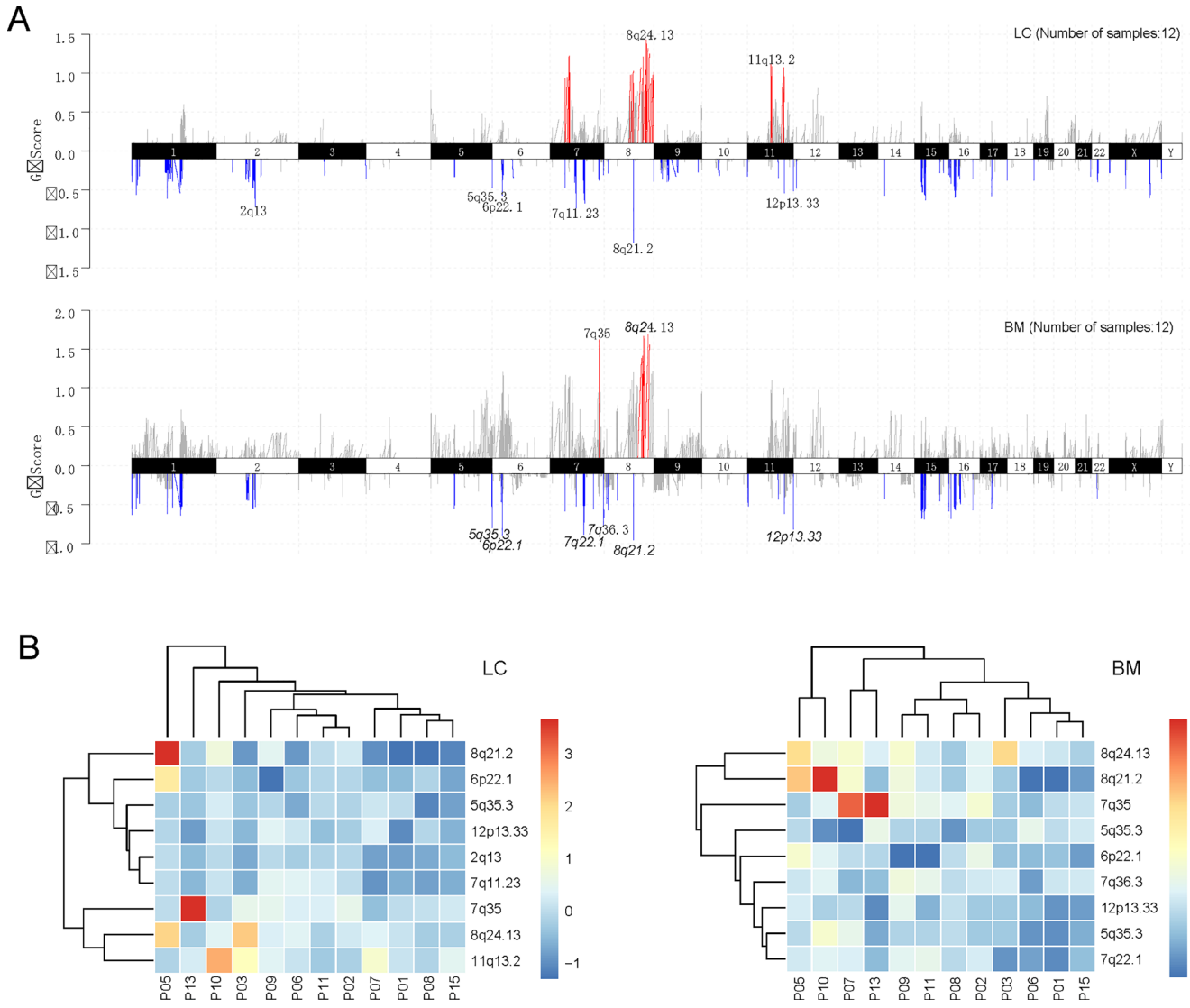| Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
SP_PIR_ KEYWORDS	egf-like domain	46	3.560372	1.17E-11	NRG3, CRELD2, LDLR, ADAM21P1, ASTN1, MMRN1, ITGBL1, HMCN2, HMCN1, CD93, FAT1, IMPG2, ACAN, CNTNAP2, TPO, FBN3, TNN, FBN2, NRG2, MUC12, CUBN, NRXN2, NRXN3, ATRN, NOTCH2NL, CELSR3, CELSR2, SSPO, MUC4, SLIT3, NOTCH2, MUC3A, CLEC18A, THBD, LRP1, SNED1, FBLN2, OTOG, RELN, ADAM22, ADAM19, LRP2, MEGF6, ADAM12, SELE, LRP5	1227	230	19235	3.135289	2.21E-09
SP_PIR_ KEYWORDS	tyrosine-specific phosphatase	11	0.851393	1.38E-05	PTPRC, PTPRD, PTPN3, PTPRZ1, PTPRH, PTPRN2, PTPN4, PTPN21, PTPRO, CDC25A, PTPN12	1227	31	19235	5.56261	9.78E-04
SP_PIR_ KEYWORDS	actin-binding	34	2.631579	4.49E-05	HIP1R, MYO7B, SSFA2, MYBPC3, SSH2, WASF2, PXK, VILL, VCL, ESPN, MACF1, DMD, WIPF3, INF2, DBNL, PHACTR2, SPTBN5, MYH1, SPTBN4, MYH7, ACTN2, EVL, MYH6, VASP, MYO18B, SYNE2, TNS1, MIB2, FHOD3, MYH13, WASL, MYH14, MYH7B, PARVA	1227	247	19235	2.157891	0.002833
SP_PIR_ KEYWORDS	sh3 domain	30	2.321981	6.99E-05	MYO7B, BCAR1, NCF1C, CACNB2, ASAP1, ITSN2, CACNB4, ARHGAP4, MIA2, MACF1, PACSIN3, PACSIN2, PSTPIP1, DLG5, RASA1, PIK3R2, CASKIN1, SH3PXD2B, DBNL, SASH1, NOXO1, EFS, SH3BP4, ARHGAP33, FCHSD2, SH3RF3, CRK, TJP2, SASH3, SRGAP2	1227	210	19235	2.239492	0.003611
SP_PIR_ KEYWORDS	atp-binding	119	9.210526	1.15E-04	POTEKP, STK35, TTK, ITPKA, KIF13A, NLRC5, ATP2B2, DDX17, OPLAH, DHX35, MAP2K7, MAP2K5, EGFR, MAGI3, CARS, BCR, MYH1, MAGI1, PIK3CB, MYH7, MYH6, HNRNPU, MYO18B, NME3, NAV2, SMARCA5, BMP2K, FLAD1, MYO18A, MYH7B, NEK6, NEK7, DNAH11, FGFR2, DNAH17, TDRD9, CHEK1, CHEK2, RIOK1, ABCA4, ABCA3, TTBK1, DDX60L, MKI67, MAP2K2, MAP2K3, ATP11A, BRIP1, SMG1, ABCB1, ATP11C, DDX4, RPS6KL1, ATP13A4, EPHA6, EPHA8, ATP2A1, MYH13, HSPD1, MTOR, MYH14, FPGT, GUK1, UBE2E1, PC, KIF25, CDK17, HELQ, MYO7B, ATP10A, HLCS, MOV10L1, HELZ, PIP5K1A, DDR2, MCM8, ATAD3B, CHUK, ABCE1, SGK1, LIMK1, MINK1, CFTR, TOP1MT, CBWD3, PANK3, CBWD1, CDK11A, RRM1, ERN2, TNNI3K, CDK11B, SMARCAD1, DNAH9, MKNK2, BRSK2, AKAP9, BRSK1, DNAH7, BMS1, STK32C, MAP3K3, SPEG, DGKG, ABCD2, HSPA7, CAMK2B, PRKAA1, CAMK2A, AATK, ABCA12, ACSL5, FLT1, FLT4, NLK, UBE2Q2, RPS6KA5, ROCK1P1, ABCC3, SLFN13, ABCC5	1227	1326	19235	1.406861	0.005455
SP_PIR_ KEYWORDS	extracellular matrix	32	2.47678	1.53E-04	ADAMTSL1, ADAMTS14, CALR, LAMB4, COL9A3, HMCN1, COL27A1, COL6A3, ACAN, FBN3, FBN2, COL11A1, SPP1, COL18A1, WNT10A, HAPLN3, ZP3, COL13A1, LGALS1, MEPE, COL5A3, COL5A1, COL4A6, COL4A5, LAMA2, ADAMTS6, LAMA5, FBLN2, COL1A2, RELN, COL1A1, CHL1	1227	241	19235	2.08152	0.00666
SP_PIR_ KEYWORDS	metal-binding	236	18.26625	1.67E-04	ALAD, PDLIM5, RNF213, RNF212, ZNF251, ATP2B2, PLOD2, PRIM2, BIRC8, MAP2K7, FTL, MAP2K5, ZNF43, UBR4, ZNF503, MECOM, ADAMTS6, UHRF1, NME3, NEK6, NEK7, FUS, ZNF519, TRIM64C, ZNF611, CYP2B6, ZNF76, ASAP1, CHEK2, CALR, ADAP1, TCF20, DAGLA, RPS29, DMD, CASZ1, COL18A1, DNMT3A, ZNF620, KLF10, KLF11, KLF17, ATP11A, ATP11C, DAGLB, ZNF525, ATP13A4, PXDNL, ZBTB42, CBLC, NR1I2, ATP2A1, ZFHX3, RERE, DPP3, ATP10A, ZMAT2, ZNF530, ARSI, HELZ, ZEB2, ZNF347, ZNF346, CANX, GLI3, STAC3, ASH2L, ZNF491, ZNF735, NOS2, AGAP2, NR2F1, KCNMA1, TP53, ZNF141, HERC2, ZSWIM2, MMP14, VPS8, ZDHHC14, ZDHHC17, ZDHHC11, PPEF2, ASH1L, ADAM19, ADAM12, KSR1, NSD1, REV3L, CDRT1, THAP7, LIMS1, ZNF469, FHL3, MKNK2, POLA1, CBLL1, ALPP, LOC440434, CYP2A13, CYP27B1, GNPTAB, BCL11B, TYW1, DGKG, DTNB, GATAD2B, PRKAA1, PPP3CA, RFPL3, RASA3, MLLT6, ZNF267, NLK, ZNF770, PDZRN3, SOD2, RPS6KA5, LRP1, PHF14, JMJD6, ROCK1P1, COG8, SH3RF3, ZBTB4, PHF21A, NLN, DPYD, ADH1C, ZXDB, RPS27L, APOBEC3H, ENPEP, ZIC2, APOBEC3D, TRIM46, TRIM48, AGAP9, TRIM43, TRIM44, ZNF578, RBM10, SAR1B, RNF145, CARS, ZNF813, NPEPPS, SUZ12, KDM2B, MIB2, RELN, ZNF436, ZNF845, TSHZ2, ZC3H4, ADAM21P1, ZNF846, PRUNE2, PEG10, FGD6, ADSSL1, LPO, EHMT1, LPP, ZMYM5, NR4A2, SMG1, NR4A3, CACNA2D2, FOXP1, MAN2C1, TET3, CADPS2, POLD1, ANTXR2, HIVEP1, ADAM22, FPGT, PC, ZNF85, RAI1, KDM6A, NPEPL1, CTCF, ZKSCAN3, ZZEF1, MAZ, ZNF727, TPO, RANBP2, MAN2B1, IMPDH1, PHC2, NFX1, ZCCHC4, ZNF280C, NRXN2, LIMK1, NRXN3, COQ7, MYT1L, ZNF718, ZMIZ2, ZMIZ1, UBC, CYP2A6, ERN2, TNNI3K, ZFPM2, CHFR, CPSF3, ZNF99, ADAMTS14, ABLIM3, PRSS1, BRSK2, ADH5, BRSK1, PPM1B, STK32C, NPTX1, PRDM15, MAP3K3, TGM3, XAF1, ING1, DUS3L, ADARB1, ADARB2, CBL, CREB5, ZNF705G, CADPS, RNF112, ADI1, RNF44, SULF2, SUMF2	1227	2972	19235	1.244832	0.006765

Table S3 (*continued*)

http://dx.doi.org/10.21037/atm-21-1555

| Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
SP_PIR_ KEYWORDS	repressor	48	3.71517	2.79E-04	ZNF85, SBNO2, ETV7, BCLAF1, CBX4, ZEB2, CTCF, TCERG1, BCL11B, ANP32A, GATAD2B, LRRFIP1, NFX1, NFATC1, NKRF, ATF7IP, DNMT3A, ERF, SATB2, RCOR2, JARID2, KLF10, KLF11, MECP2, MED12, KLF17, ZNF141, TLE4, ZNF503, BANP, MED13, MED13L, FOXP1, FOXN3, SUZ12, CHMP1A, KDM2B, MSX1, HDAC2, PHF21A, CELF2, ZFPM2, PBX1, CUX1, NSD1, NCOR1, HDAC7, RERE	1227	435	19235	1.729815	0.009885
SP_PIR_ KEYWORDS	isopeptide bond	38	2.941176	3.13E-04	HIST1H2AC, TNRC18, GRIK2, RAD23A, CBX4, ELK1, SYNCRIP, CHEK1, ZEB2, GON4L, CTCF, PIP5K1A, HIST2H2AB, CUL2, HSF1, RANBP2, HNRNPC, ACIN1, KPNB1, POTEE, EGFR, SATB2, TP53, MYH7, MYH6, CCDC138, HIST2H3D, PSMA1, UHRF1, MSX1, EVPL, TRAF3IP3, ASH1L, UBC, TFAP2A, CUL4B, UBXN7, CPSF3	1227	319	19235	1.867414	0.009845
SP_PIR_ KEYWORDS	Endocytosis	17	1.315789	3.53E-04	DBNL, SYNRG, STON1, CUBN, CALY, LDLR, HIP1R, GTF2A1L, EPS15L1, ITSN2, CLEC10A, SH3BP4, LRP1, THBD, PACSIN3, PACSIN2, LRP2, LRP5	1227	96	19235	2.776037	0.00999
SP_PIR_ KEYWORDS	transcription regulation	166	12.8483	4.94E-04	STAT5A, STAT5B, CNOT3, HOXD12, CBX4, CNOT2, ZXDB, ZNF251, CTNNB1, CDKN2A, CGGBP1, SND1, ZNF578, SSX9, ZNF43, SATB2, RCOR2, MED12, MECP2, ZNF813, ZNF503, ARID1A, MED13, GRHL1, MECOM, FOXN3, SUZ12, UHRF1, KDM2B, MSX1, TFAP2A, TFAP2D, ZNF436, ZNF519, CRTC2, MEAF6, TSHZ2, ZNF611, HOXA13, ZNF76, SETD1A, ELK1, KEAP1, TRRAP, ZNF846, TCF21, TCF20, AHRR, TCERG1, CASZ1, TFDP3, RUNX1, RUNX2, TCF3, SSX1, TAF2, NKRF, ZNF620, KLF10, KLF11, NR4A2, KLF17, EN1, TEAD2, GTF2A1L, NR4A3, MED13L, ZNF525, FOXP1, ATF6, NOTCH2, NR1I2, HDAC2, SMARCC2, HIVEP1, NCOR1, ZFHX3, HDAC7, RERE, ZNF85, TAF1C, ZNF530, ZEB2, CTCF, ZNF347, ZKSCAN3, GLI3, PAX1, FUBP1, ARHGAP22, MCM8, MAZ, HSF1, ASH2L, ZNF727, ZNF491, ANP32A, ZNF735, LRRFIP1, NFX1, NR2F1, ATF7IP, SNAPC5, SSBP3, CCNK, ZNF280C, ARID5B, SNAPC3, SOX12, TP53, ZNF141, TLE4, BANP, MYT1L, ZNF718, CHMP1A, BRWD1, ZMIZ2, ZMIZ1, ASH1L, KHSRP, BRDT, FOXC2, ERN2, ZFPM2, WASL, CUX1, NSD1, ZNF99, FOXD4L5, THAP2, SBNO2, BCLAF1, ETV7, ZNF469, STON1, NPAS1, PRDM15, NPAS2, NPAS3, RB1CC1, BCL11B, POU2F2, POU2F1, GATAD2B, BCL3, PER3, FOXD4, NFATC1, ZNF267, SIM2, ERF, JARID2, NLK, ZNF770, AFF3, CREB5, WWTR1, POU5F1B, ZNF705G, JMJD6, ATXN7, FOXE1, ZBTB4, PHF21A, PBX1	1227	2026	19235	1.284447	0.013298
SP_PIR_ KEYWORDS	triple helix	9	0.696594	5.61E-04	COL13A1, DMD, COL6A3, COL1A2, COL1A1, COL11A1, COL5A1, COL4A6, COL4A5	1227	31	19235	4.551226	0.014407
SP_PIR_ KEYWORDS	nucleus	321	24.8452	6.04E-04	STAT5A, PPP2R5C, STAT5B, HOXD12, INTS1, SYNCRIP, INTS2, WTAP, ZNF251, CTNNB1, DDX17, INTS5, INTS4, RPP30, MAP2K7, ZNF43, MAGI3, RCOR2, PTBP1, MED12, MECP2, UBR4, ZNF503, NAV3, MED13, MECOM, ERGIC2, MYO18B, UHRF1, MSX1, MAD2L1, NAV2, CMIP, PDE4DIP, NEK6, FUS, TP53TG3, CRTC2, ZNF519, ZNF611, HOXA13, RAD23A, ZNF76, CHEK1, CHEK2, XAB2, ADAP1, TCF21, TCF20, AHRR, FRG2, FRG1, CASZ1, AHNAK2, PABPC1, TCF3, NKRF, DNMT3A, NOC4L, MKI67, ZNF620, KLF10, KLF11, KLF17, EN1, DUSP21, ZNF525, ZBTB42, NOTCH2, CBLC, NR1I2, SYNE2, DCP1B, ETS2, NCOR1, ZFHX3, RERE, TAF1C, UTP18, GAR1, ZNF530, TCOF1, ZMAT2, HELZ, ZEB2, ZNF347, ZNF346, GLI3, HIST2H2AB, DIP2A, ASH2L, HSF1, ANP32A, ZNF491, SPANXD, QKI, ZNF735, SPANXC, BRD4, TWISTNB, AGAP2, FANCA, NR2F1, ATF7IP, SNAPC5, RBM42, HIST1H1B, SNAPC3, TP53, ZNF141, TLE4, CHMP1A, BRWD1, DACT1, CDK11A, ASH1L, BRDT, NOL11, FOXC2, CDK11B, NSD1, ARL4C, FOXD4L5, REV3L, HIST1H2AC, ZNF469, ETV7, POLA1, BMS1, LOC440434, SPEG, BCL11B, RB1CC1, GATAD2B, BCL3, PPP3CA, PER3, SNRNP70, TBC1D1, MLLT6, FOXD4, SIM2, NMD3, ZNF267, RRP12, PDS5A, JARID2, NUCKS1, NLK, ZNF770, AFF3, WWTR1, RPS6KA5, LRP1, JMJD6, ZBTB4, FOXE1, PHF21A, PBX1, USP48, TJP2, RALY, STK35, CBX4, CNOT3, ZXDB, UBQLN4, CNOT2, ZIC2, CDKN2A, CGGBP1, SND1, ZNF578, SRRM1, PMS2, ACIN1, ITCH, RBM10, PABPN1, SATB2, ZNF813, ARID1A, NPEPPS, GRHL1, TACC1, FOXN3, HNRNPU, SUZ12, PNPLA7, PSMA1, KDM2B, BMP2K, SMARCA5, TFAP2A, CELF2, TFAP2D, ZNF436, ZNF845, MEAF6, TSHZ2, SETD1A, ELK1, KEAP1, TRRAP, HEXDC, DAZAP1, ZNF846, PEG10, TCERG1, NIPBL, POLE3, TFDP3, HNRNPC, RUNX1, RUNX2, TAF2, EHMT1, C17ORF49, LPP, ZMYM5, RFX7, NR4A2, SMG1, BRIP1, SYCE1, GTF2A1L, HEATR1, TEAD2, NR4A3, MED13L, DPPA3, RBMX, AIM2, FOXP1, HIST2H3D, ATF6, HDAC2, POLD1, SMARCC2, HIVEP1, HNRNPH1, FPGT, HDAC7, ZNF85, RAI1, KDM6A, CTCF, TIMM50, DEK, EPS15L1, ZKSCAN3, PTMA, PAX1, FUBP1, ARHGAP22, MCM8, MAZ, PLRG1, ZNF727, HNRNPCL1, LRRFIP1, RANBP2, KPNB1, CHUK, AHNAK, PHC2, NFX1, NOL6, SSBP4, SGK1, SSBP3, ZNF280C, ARID5B, SOX12, TP53TG3B, BANP, MYT1L, ZNF718, XPC, ZMIZ2, EIF5AL1, ZMIZ1, TPPP, CPSF7, KHSRP, UBC, TNNI3K, ZFPM2, WASL, CPSF3, CHFR, CUX1, ZNF99, SMARCAD1, BCLAF1, STON1, IRX5, GON4L, HCFC2, WBP11, BRSK1, NPAS1, NPAS2, PRDM15, NPAS3, POM121, POU2F2, MSI1, POU2F1, CALML6, XAF1, MYOF, ING1, GEMIN5, NFATC1, ERF, DGCR6L, ADARB2, NFRKB, SCOC, CREB5, POU5F1B, ZNF705G, ADI1, SH3BP4, ATXN7	1227	4283	19235	1.17491	0.014827

| Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
SP_PIR_ KEYWORDS	Transcription	168	13.0031	7.29E-04	STAT5A, STAT5B, CNOT3, HOXD12, CBX4, CNOT2, ZXDB, ZNF251, CTNNB1, CDKN2A, CGGBP1, SND1, PRIM2, ZNF578, SSX9, ZNF43, SATB2, RCOR2, MED12, MECP2, ZNF813, ZNF503, ARID1A, MED13, GRHL1, MECOM, FOXN3, SUZ12, UHRF1, KDM2B, MSX1, TFAP2A, TFAP2D, ZNF436, ZNF845, ZNF519, CRTC2, MEAF6, TSHZ2, ZNF611, HOXA13, ZNF76, SETD1A, ELK1, KEAP1, TRRAP, XAB2, ZNF846, TCF21, TCF20, AHRR, TCERG1, CASZ1, TFDP3, RUNX1, RUNX2, TCF3, SSX1, TAF2, NKRF, ZNF620, KLF10, KLF11, NR4A2, KLF17, TEAD2, GTF2A1L, NR4A3, MED13L, ZNF525, FOXP1, ATF6, NOTCH2, NR1I2, HDAC2, SMARCC2, HIVEP1, NCOR1, ZFHX3, HDAC7, RERE, ZNF85, TAF1C, ZNF530, ZEB2, CTCF, ZNF347, ZKSCAN3, GLI3, PAX1, FUBP1, ARHGAP22, MCM8, MAZ, HSF1, ASH2L, ZNF727, ZNF491, ANP32A, ZNF735, LRRFIP1, TWISTNB, NFX1, NR2F1, ATF7IP, SNAPC5, SSBP3, CCNK, ZNF280C, ARID5B, SNAPC3, SOX12, TP53, ZNF141, TLE4, BANP, MYT1L, ZNF718, CHMP1A, BRWD1, ZMIZ2, ZMIZ1, ASH1L, KHSRP, BRDT, FOXC2, ERN2, ZFPM2, WASL, CUX1, NSD1, ZNF99, FOXD4L5, THAP7, SBNO2, BCLAF1, ETV7, ZNF469, STON1, NPAS1, PRDM15, NPAS2, NPAS3, RB1CC1, BCL11B, POU2F2, POU2F1, GATAD2B, BCL3, PER3, FOXD4, NFATC1, ZNF267, SIM2, ERF, JARID2, NLK, ZNF770, AFF3, CREB5, WWTR1, POU5F1B, ZNF705G, JMJD6, ATXN7, FOXE1, ZBTB4, PHF21A, PBX1	1227	2071	19235	1.271677	0.016461
SP_PIR_ KEYWORDS	trimer	8	0.619195	9.44E-04	COL13A1, COL6A3, COL1A2, COL1A1, COL11A1, COL5A1, COL4A6, COL4A5	1227	26	19235	4.823522	0.019712
SP_PIR_ KEYWORDS	phosphoric monoester hydrolase	13	1.006192	0.001275	PTPRC, PTPRD, PTPN3, PTPRZ1, PTPRH, PTPRN2, PTPN4, PTPN21, PPP3CA, ALPP, PTPRO, CDC25A, PTPN12	1227	69	19235	2.953533	0.025592
SP_PIR_ KEYWORDS	calcium channel	12	0.928793	0.001479	RYR1, CACNA1G, CACNA1H, RYR2, CACNB2, CACNA1E, CACNB4, CACNA1C, CACNA2D2, ITPR1, CACNA1A, ITPR2	1227	61	19235	3.083891	0.028623
SP_PIR_ KEYWORDS	hydroxylation	13	1.006192	0.001647	COL18A1, COL9A3, THBD, COL6A3, COL1A2, CELSR3, CELSR2, COL1A1, COL5A3, COL11A1, COL5A1, COL4A6, COL4A5	1227	71	19235	2.870335	0.03078
SP_PIR_ KEYWORDS	collagen	15	1.160991	0.002715	COL18A1, COL23A1, COL13A1, COL5A3, COL5A1, COL4A6, COL4A5, COL9A3, PLOD2, COL27A1, COL6A3, COL1A2, COL1A1, EDA, COL11A1	1227	95	19235	2.475228	0.048681

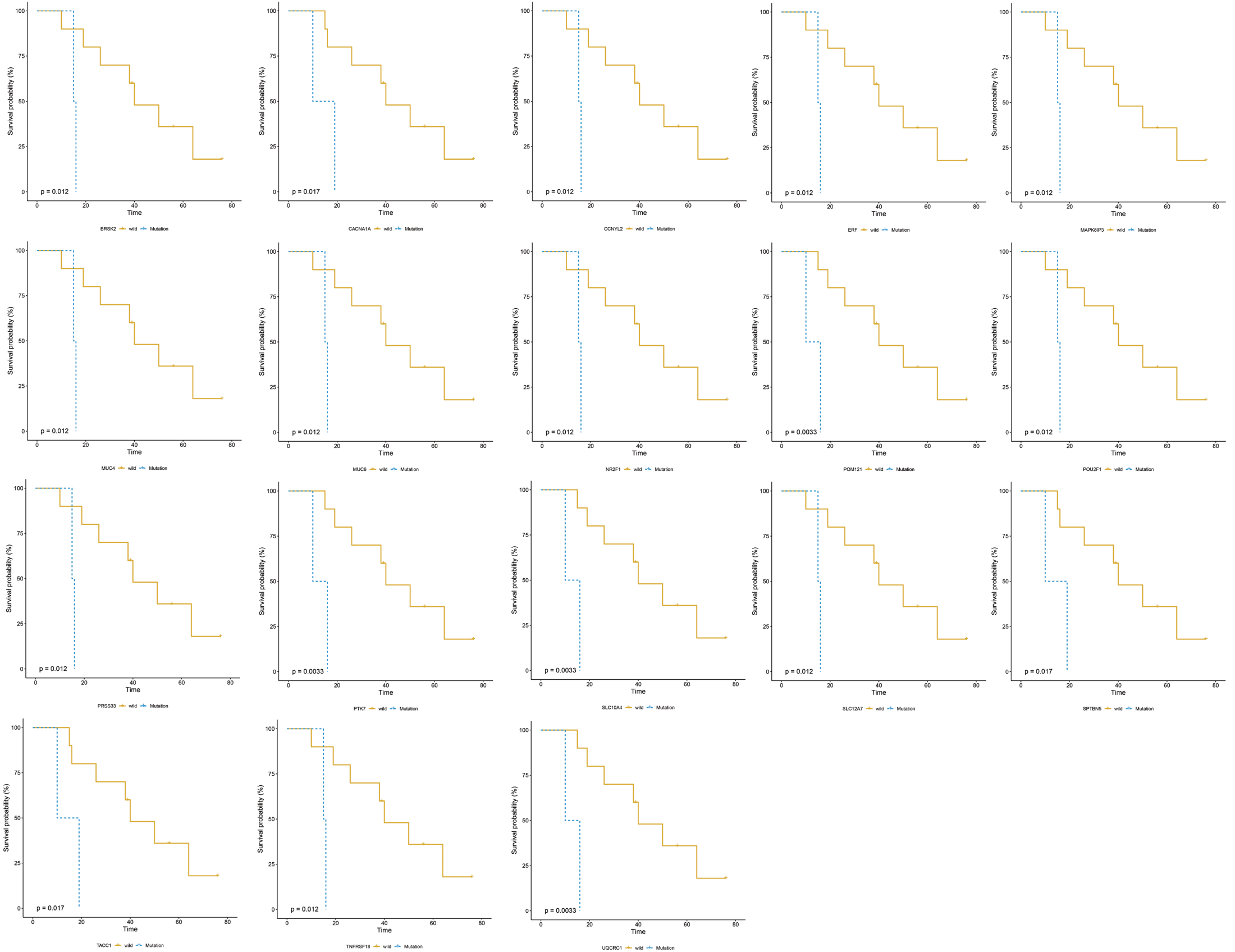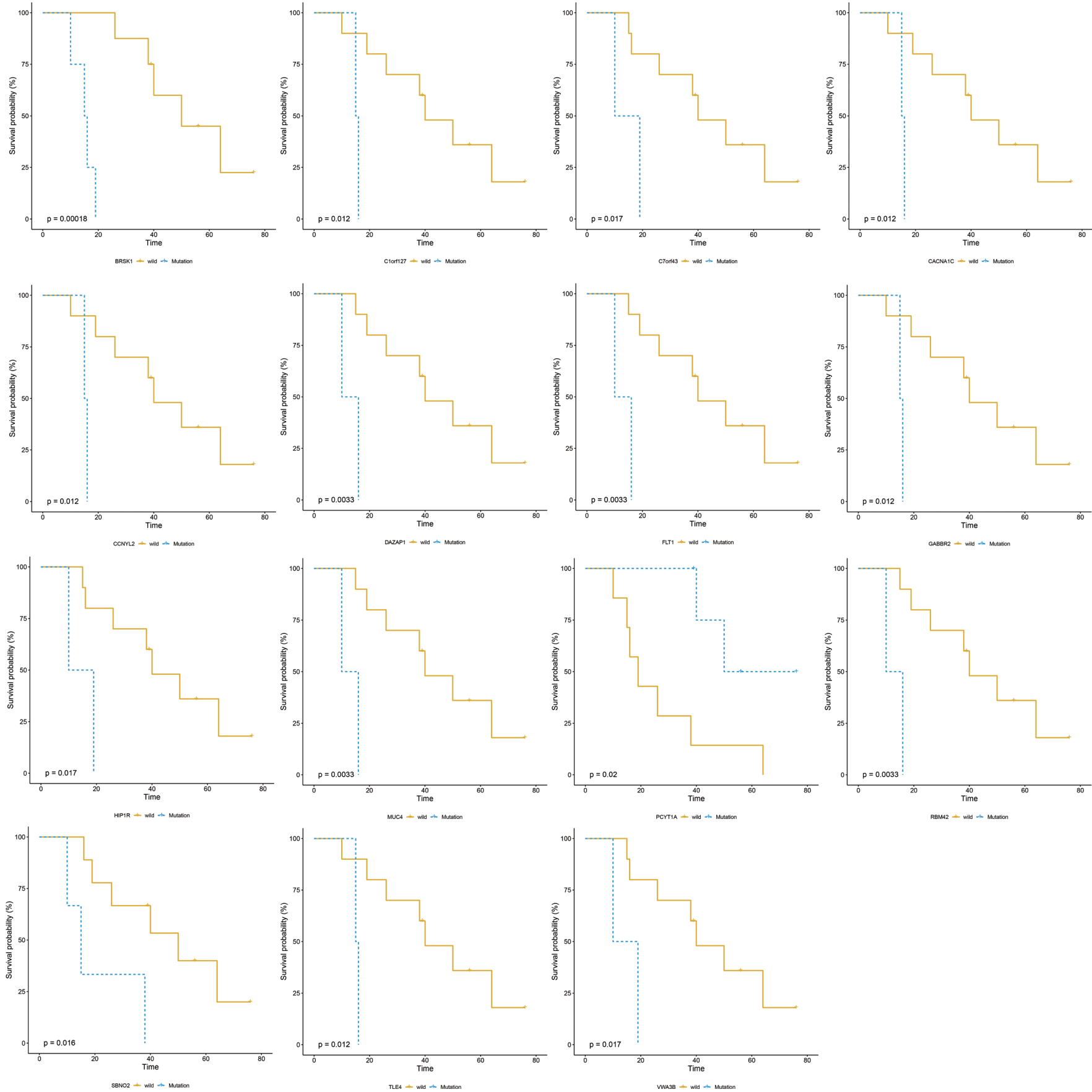**Figure S2** CNVs in lung tumor and brain tumor samples. CNVs, copy number variants.
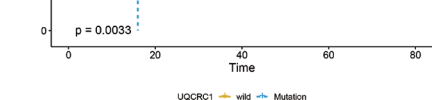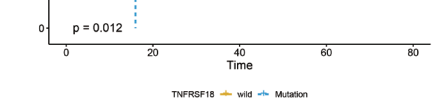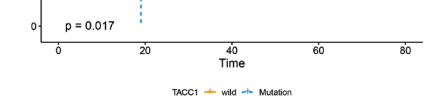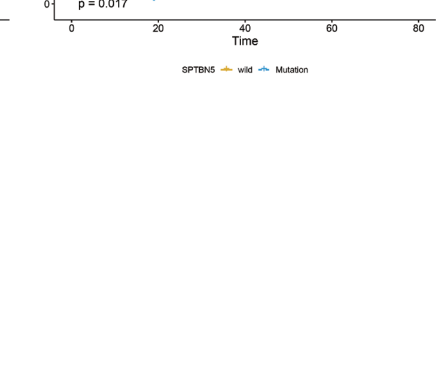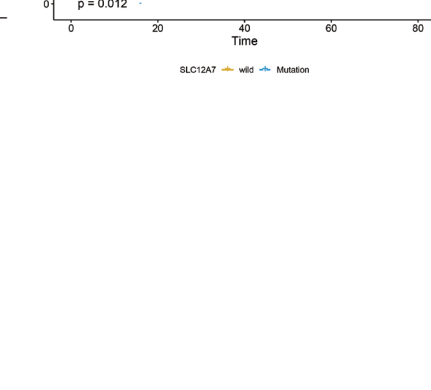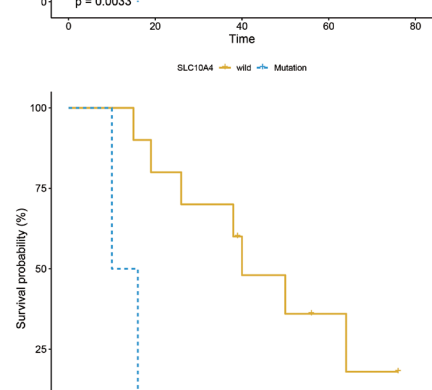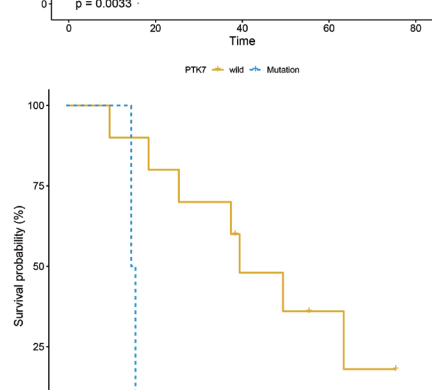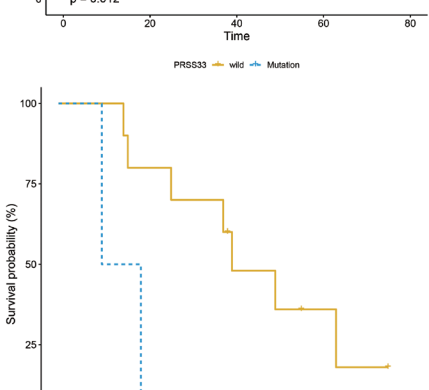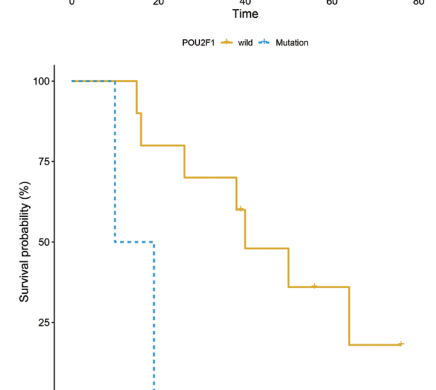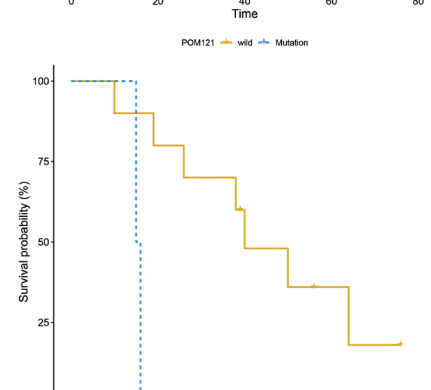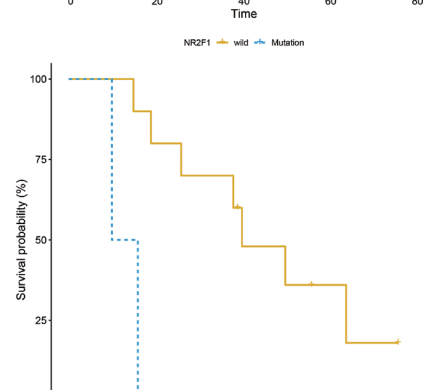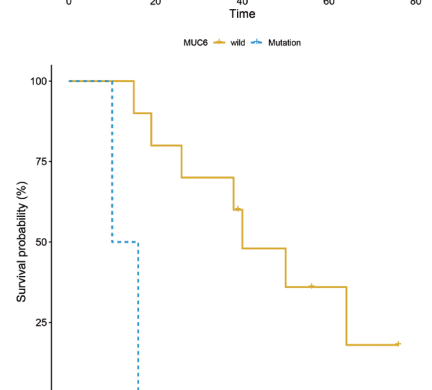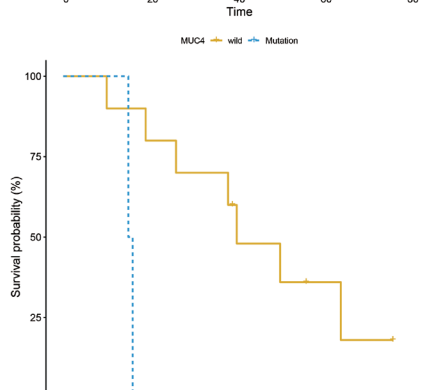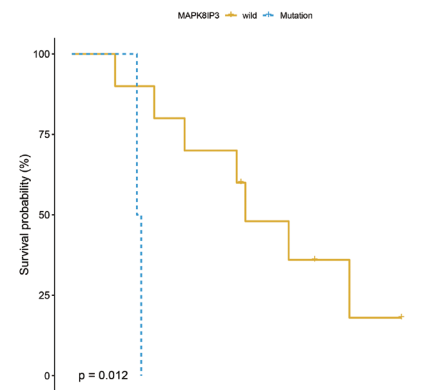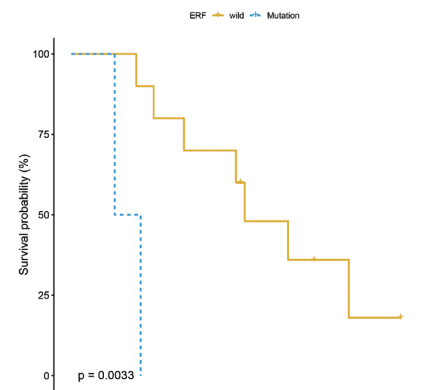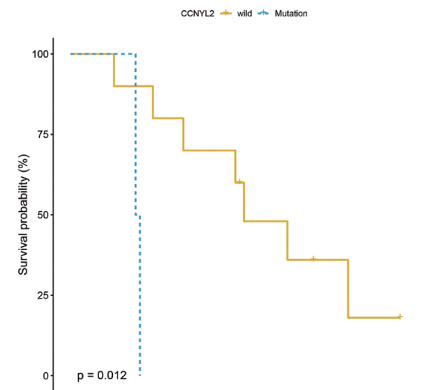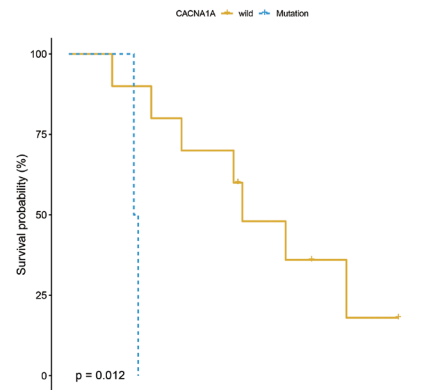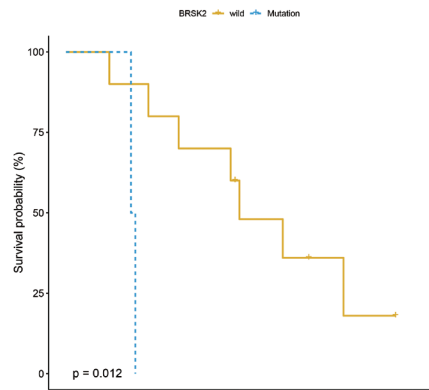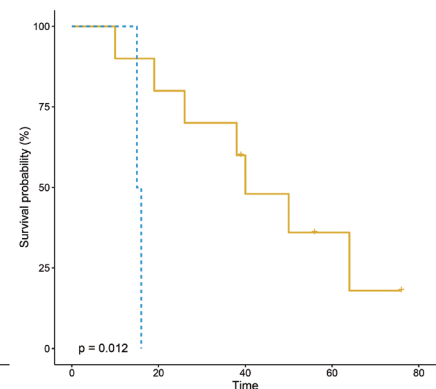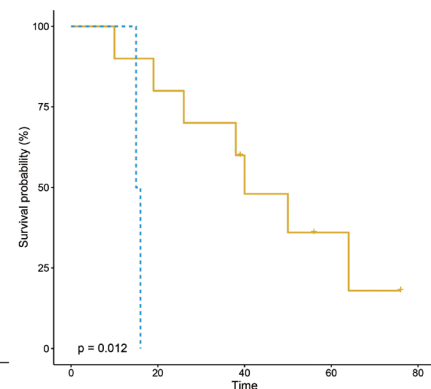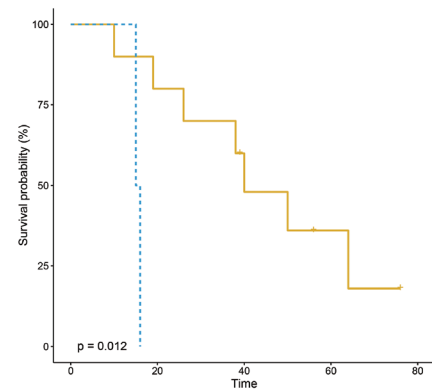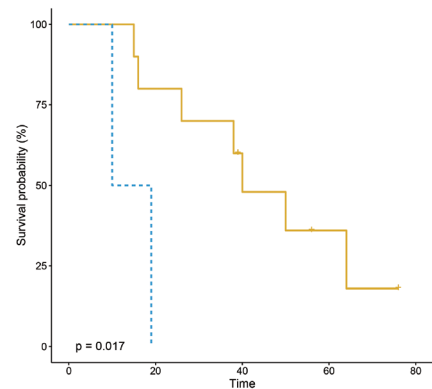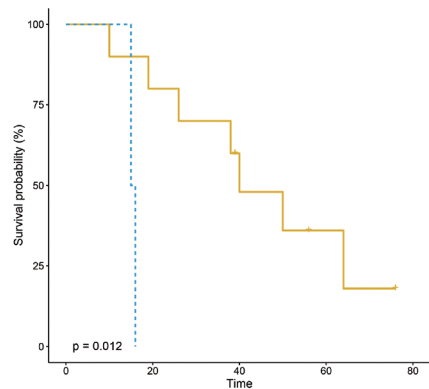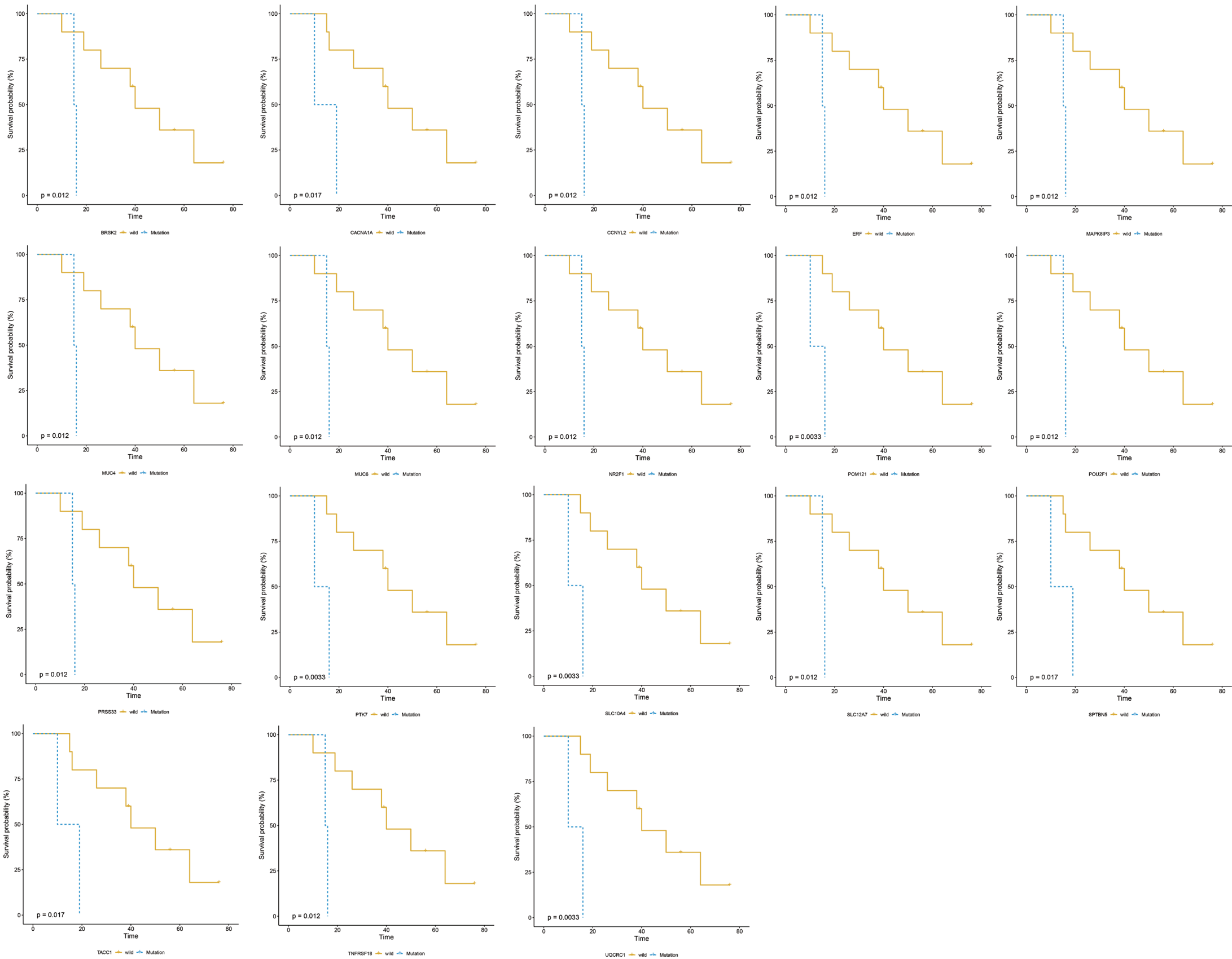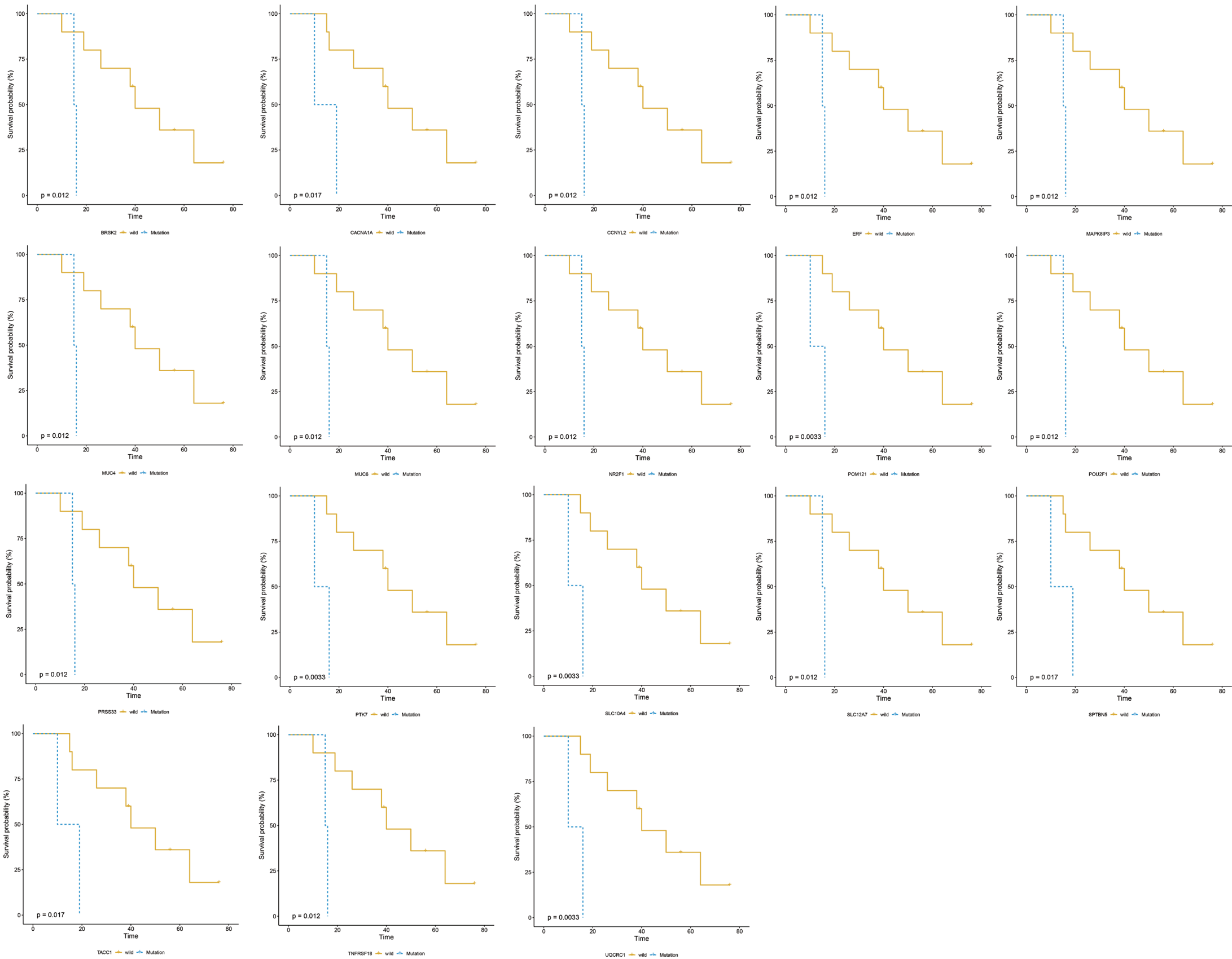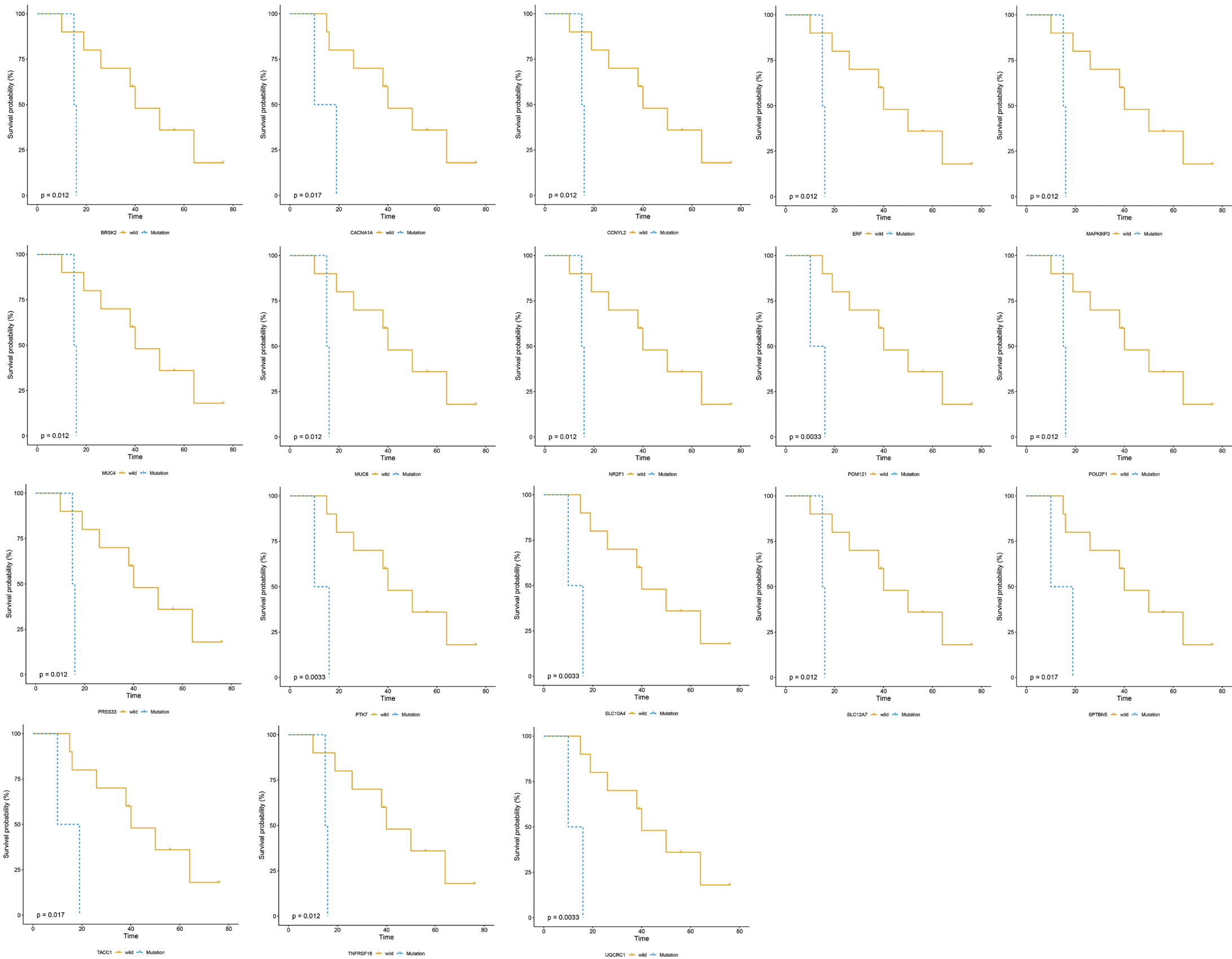
Figure S3 Survival analysis between mutations in LC and OS times. LC, lung cancer; OS, overall survival.

Figure S4 Survival analysis between mutations in BM and times. OS, overall survival; BM, brain metastasis.