Original research

Genomic landscape and evolution of arm aneuploidy in lung adenocarcinoma [☆], [☆☆]

Beili Gao [a,1]; Fujun Yang [b,1]; Ming Han [c]; Hua Bao [c]; Yi Shen [c]; Ran Cao [c]; Xue Wu [c]; Yang Shao [c,d]; Changhong Liu [e]; Zhe Zhang [f,g,*]

[a] Department of Respiration, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China
[b] Department of Oncology, Weihai Municipal Hospital, Cheeloo College of Medicine, Shandong University, Weihai, Shandong, China
[c] Nanjing Geneseeq Technology Inc. Nanjing, Jiangsu, China
[d] School of Public Health, Nanjing Medical University, Nanjing, China
[e] Department of Thoracic Surgery, The Second Hospital of Dalian Medical University, Dalian, Liaoning, China
[f] Department of Medical Oncology, Fudan University Shanghai Cancer Center, Shanghai, China
[g] Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

Abstract

For lung adenocarcinoma, arm aneuploidy landscape among primary and metastatic sites, and among different driver and frequently mutated gene groups have not been previously studied. We collected the largest cohort of LUAD patients (n=3533) to date and analyzed the profiles of chromosome arm aneuploidy (CAA), and its association with different metastatic sites and mutated gene groups. Our results showed distant metastasis (bone, brain, liver) were characterized by high CAA burden and biased towards arm losses compared to regional metastasis (pleura, chest) and primary tumors. Moreover, EGFR, MET, PIK3CA, PKHD1 and RB1 mutant groups were found to have high CAA burden, while those with BRAF, ERBB2 and KRAS mutations belonged to the low CAA burden group. Comparing EGFR L858R and EGFR 19del mutants, distinct CAA co-occurrences were observed. Network-based stratification with population based genomic evolution analysis revealed two distinct subtypes of LUAD with different CAA signatures and unique CAA order of acquisition. In summary, our study presented a comprehensive characterization of arm aneuploidy landscape and evolutionary trajectories in lung adenocarcinoma, which could provide basis for both biological and clinical investigations in the future.

Neoplasia (2021) 23, 870–878

Keywords: Lung adenocarcinoma, Chromosome arm aneuploidy, Metastatic sites, EGFR, Mutations, Arm aneuploidy evolution

Introduction

Chromosome arm aneuploidy (CAA) is characterized by copy number gain or loss of chromosome arms [1]. CAA is the most prevalent type of genomic alteration in solid tumors, occurring in 90% of cases [1]. Both aneuploidy burden as well as specific CAAs are associated with clinical outcome in a number of cancers. For example, high levels of aneuploidy is associated with lung cancer progression [2], whereas 1p and 9p loss are associated with favorable prognosis in glioma [3]. CAA have been shown to outperform mutations and focal somatic CNAs in predicting drug response [4]. Tumors with distinct histological subtypes can also be clustered using arm loss and gain patterns [5].

[*] Corresponding author:
E-mail address: zhangzhe2021fduscc@gmail.com (Z. Zhang).

[☆] ☆ Funding: Nanjing Geneseeq Technology Inc. provided support in the form of salaries for authors MH, HB, YS, RC and XW, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

[☆☆] ☆☆ Conflicts of interest: MH, HB, YS, RC, XW and YS are employees and/or shareholders of Nanjing Geneseeq Technology, Inc. All other authors declared no conflicts of interest.

[1] Co-first authors: these authors contributed equally to this work, and should be considered joint first authors.

Received 30 March 2021; received in revised form 28 May 2021; accepted 3 June 2021

https://doi.org/10.1016/j.neo.2021.06.003

Lung cancer is the most commonly diagnosed cancer worldwide [6], with lung adenocarcinoma (LUAD) accounting for 40% of all lung cancers [7]. Early studies of CAA in lung cancer showed agreement for some arm gains (1q, 3q, 5p and 8q) and losses (3p, 8p, 9p, 9q and 13q), but with inconsistent prevalence, and disagreements for certain CAAs [8]. A larger cohort may address these inconsistencies. Evidence suggest metastatic spread depend on specific aneuploidies [9]. CAAs 6q-, 7p+, 9p- and 13q- were enriched in metastatic versus primary non-small cell lung cancer [4], however this study did not investigate different sites of metastasis, such as lymph, pleura, liver and brain. Certain mutant genes can be associated with specific CAAs. For example, 3p- with mutant BAP1, PBRM1 and VHL and 5p+ with mutant PBRM1 and VHL in kidney cancer [10]. In addition, evidence suggest solid tumors are characterized by initial arm gains, followed by numerous arm losses [4]. These have not been studied in LUAD.

A recent pan-cancer study analyzed CAA in 503 LUAD samples from The Cancer Genome Atlas (TCGA) whole exome sequencing (WES) dataset [5]. However, a larger cohort is required to reliably detect differences in CAA characteristics across different metastatic sites and mutant gene groups. Targeted panel sequencing data can often provide large cohorts, and have been shown to be reliable in making large scale genomic alteration calls such as arm aneuploidy. One study used a 110-gene kidney cancer panel and found aneuploidy call concordance with shallow whole genome sequencing (sWGS) data was 0.8723 [11]. Another study using a 100-gene colorectal cancer panel had concordance of 0.82128 with sWGS [12].

We retrospectively examined a large dataset from a 425-cancer gene targeted panel, with a large cohort of 3533 LUAD patients. This dataset consisted of samples from primary tumor and metastasis to 6 different sites, and well documented mutation status in driver and other frequently mutated genes. We aimed to uncover the association between CAA and site of metastasis, drivers and frequently mutated genes. And whether CAAs are acquired in any specific order in LUAD.

Materials and methods

Patients and samples

Custom 425-gene targeted sequencing panel data from 3533 patients with primary or metastatic lung adenocarcinoma were retrospectively examined. Written informed consent of sample usage for research was collected from each patient according to ethical regulations of respective hospitals. Tissue samples were FFPE. Paired normal samples were whole blood. All samples passed in-house QC procedures, including FFPE damage, contamination and matched normal control tests. An additional 20 LUAD samples underwent both targeted panel sequencing and WES for CAA concordance analysis to validate the use of the 425-gene targeted panel for making CAA calls.

Nucleic acid isolation, library construction, targeted panel sequencing and WES

Genomic DNAs from FFPE samples and the whole blood control samples were extracted using Qiagen QIAamp DNA FFPE Tissue Kit and DNeasy Blood and tissue kits (Qiagen, USA)), respectively, and quantified using Qubit 3.0 with dsDNA HS Assay Kit (ThermoFisher Scientific, USA). Sequencing library preparation was performed with KAPA Hyper Prep Kit (KAPA Biosystems, USA). DNA libraries were pooled and captured with a custom 425 cancer-gene panel. The capture reaction was performed with Dynabeads M- 270 (Life Technologies, CA, USA) and xGen Lockdown hybridization and wash kit (Integrated DNA Technologies) according to manufacturers' protocols. Captured libraries were PCR amplified with KAPA HiFi HotStartReadyMix (KAPA Biosystems), followed by purification using AgencourtAMPure XP beads. Libraries were quantified by qPCR using KAPA Library Quantification kit (KAPA Biosystems). Library fragment size was determined by Bioanalyzer 2100 (Agilent Technologies). The target-enriched library was then sequenced on HiSeq4000 NGS platforms (Illumina) to a minimum coverage depth of 100X and 600X for blood and FFPE, respectively. Exome capture was performed using the IDT xGen Exome Research Panel V1.0 (Integrated DNA Technologies) and sequenced using HiSeq4000 to a mean coverage depth of ~60X for the normal control (white blood cells samples) and ~150X for the tumor FFPE samples.

Sequencing data processing

FASTQ file quality control was performed using Trimmomatic [13], where N bases and low quality (score <20) bases were removed. Pair-end reads were aligned to the human reference genome (hg19) using Burrows-Wheeler Aligner (BWA), then) with default parameters, followed by PCR deduplication with Picard V2.9.4 (Broad Institute, MA, USA). Local realignment around indels and base quality score recalibration was performed with the Genome Analysis Toolkit (GATK 3.4.0). Somatic single-nucleotide variants (SNVs) were identified using MuTect2. Final list of mutations was annotated using vcf2maf (available on GitHub). Resulting mutation list was filtered through an internally collected list of recurrent sequencing errors on the same sequencing platform, summarized from the sequencing results of ~500 normal samples. Mutations occurring within repeat masked regions were also removed. For additional filtering, mutations were called when VAF is above 1% with a minimum of 3 mutant reads for hotspot COSMIC mutations (>20 recurrences); or have VAF above 2% with a minimum of 5 mutant reads for all other mutations. Gene fusions were identified using DELLY [14] using default parameters and manually inspected using IGV.

CNV analysis and CAA call

Sequenza v2.1.2 [15] was used to call somatic segment level copy number. Sequenza file is compressed by default binning window size of 50. Tumor versus normal depth ratio was normalized against GC content, followed by allele-specific segmentation. Tumor sample purity and ploidy was inferred from B allele frequency and depth ratio of segments. Final purity and ploidy adjusted total copy number (CNt) was calculated for each segment. A segment was neutral if CNt equals ploidy, loss if CNt was less than ploidy, and gain if CNt was more than sample ploidy. Similar to a previous arm aneuploidy study,[11] arm gain or loss was called when >50% of the chromosome arm length have segment gain or loss, respectively.

CAA pattern, co-occurrence and evolution analysis

An individual was defined as being a mutant for a driver or frequently mutated non-driver gene if he/she has at least one non-silent mutation in that gene. Frequently mutated non-driver genes are defined as genes with non-silent mutation frequency > 5% of entire study cohort. CAA burden was calculated as the total number of aneuploid arms in an individual. For gain-loss difference, an individual is in the gain>loss category if total number of arm gains in that individual was greater than the total number of arm losses, vice versa for gain<loss category, and gain=loss if number of arm gains and losses were equal.

R package 'cooccur' [16] was used to study co-occurrence or exclusion of CAA pairs and CAA with mutant gene pairs. Cooccur compares the observed versus expected co-occurrence frequencies under a probabilistic model. For each event pair, cooccur produces an FDR adjusted q-value as a measure of significance, in addition to the magnitude of co-occurrence (positive) or exclusivity (negative) (log2(O/E), centered on zero (neutral). Q-value < 0.1 was considered significant.

CAA evolution graph was generated using TRONCO (TRanslational ONCOlogy) [17], which infers CAA acquisition using population level data from a cohort of patients. TRONCO input was a binary matrix with

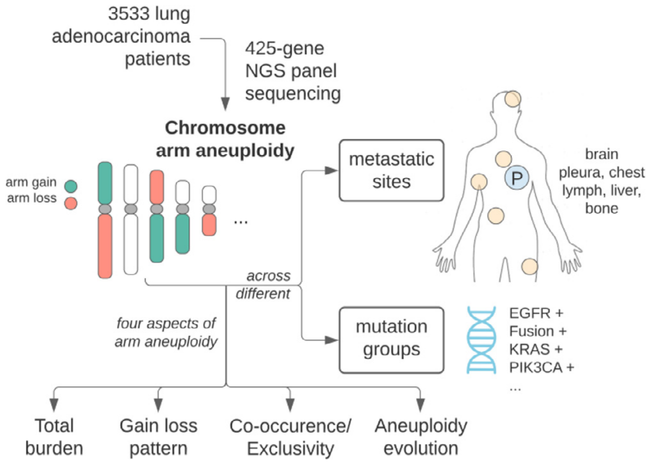

Fig. 1. CAA study overview.

rows as samples and columns as CAAs. Results using both AIC and BIC regularizations were displayed, and only connections with $P < 0.05$ were displayed. Python implementation of Network-based stratification (NBS) [18] was used for LUAD subtype clustering. NBS input was a binary matrix with rows as samples and columns as CAAs. NBS is based on non-negative matrix factorization (NNMF) algorithm, which decomposes a large matrix into initial cluster membership and signatures with reduced dimensionality. Patient co-clustering map is generated by hierarchical clustering of above-mentioned signatures, producing clusters of patients of specific LUAD subtypes. Clustering was performed with default parameters (network propagation coefficient = 0.7, number of iterations = 100). Silhouette score was calculated using the silhouette_score function from Scikit-learn python package [19]. Silhouette score is a measure of both intra-cluster and inter-cluster distance, and ranges from -1.0 (poor) to 1.0 (perfect separation of clusters). NBS clustering with clusters ranging from 2 to 10 were performed, and number of clusters with the highest overall silhouette score was chosen.

Statistical analysis

CAA call concordance between targeted panel sequencing and WES was assessed using Cohen's Kappa and Pearson's R. Numerical differences between categorical variables were assessed using Kruskal-Wallis test. A two-sided p value of less than 0.05 was considered statistically significant. Pairwise Wilcoxon rank sum test was used to test numerical differences between every category pair. Pairwise Chi-square or Fisher's exact test was used to test proportional differences between every category pair. *P*-values were FDR adjusted for multiple hypothesis testing correction, adjusted *P*-values (q-values) of less than 0.1 was considered statistically significant. All statistical analyses were performed in R (v.3.6.1).

Results

LUAD patients and CAA evaluation using panel sequencing

Data from 3533 LUAD patient samples sequenced using a 425-cancer gene panel were retrospectively examined in this study. This is the largest genomic dataset assembled for LUAD to date, and offered us a unique opportunity to assess the association of arm level aneuploidy with site of metastasis and various driver as well as frequently mutated non-driver genes, as well as probable order of CAA acquisition in this LUAD cohort (Figure 1). Patients in this study consisted of roughly equal proportions in terms of gender and age group (Table 1).

Table 1

Patient demographic and clinical information.

Group	Patients	
Total	3533	
Gender	n	%
Male	1877	53.1
Female	1551	43.9
Gender Unknown	105	3
Age group	n	%
Age<=60	1737	49.2
Age>60	1692	47.9
Age Unknown	104	2.9
Tumor status	n	%
Primary	2735	77.4
Metastatic	798	22.6
Metastatic site	n	%
Lymph	451	12.8
Bone	96	2.7
Liver	79	2.2
Pleura	54	1.5
Brain	48	1.4
Chest	17	0.5
Other	53	1.5

Two previous aneuploidy studies demonstrated the validity of using panel sequencing to evaluate CAA status in tumor samples [11,12]. We also conducted our own concordance study using 20 samples sequenced with 425-cancer gene targeted panel and whole exome sequencing (WES). Comparison of CAA calls between panel and WES data showed high level of concordance with a Pearson's r mean score of 0.85 (Figure S1A).

Overall, the top 5 most frequent arm aneuploidies in the entire LUAD cohort (N = 3533) were 7p gain (61%), 5p gain (52%), 8q gain (51%), 1q gain (47%) and 19p loss (46%) (Figure S2). Previous lung cancer aneuploidy studies only assessed primary tumor samples. To facilitate comparisons, we also assessed frequent arm aneuploidies in primary samples. Among primary samples (N = 2735), the top arm gains were in 1q (47%), 3q (28%), 5p (52%), 7p (59%) and 8q (50%) (Figure S3). Out of 14 previous studies surveyed [8,20,29,30,21–28], 1q, 3q, 5p and 8q gains were found in 7 or more studies, while 7p gain was found in 4 previous studies. Three other arm gains occurred in high proportion of primary samples: 7q (34%), 14q (32%) and 20q (30%), which were only seen in 2– 3 previous studies (see supplementary data "LUAD_CAA_papers_compilation.xlsx" for details). Top arm losses were 3p (31%), 8p (36%), 9p (33%), 9q (33%), 10q (39%), 13q (36%), 15q (39%), 17p (35%), 18q (32%), 19p (45%), 19q (34%) and 22q (35%). Loss of 3p, 8p, 9p, 9q, 13q, 17p, 18q and 19p were seen in 5 or more previous studies (see supplementary data "LUAD_CAA_papers_compilation.xlsx" for details). Most previous studies only reported presence or absence of arm aneuploidy. Our results not only confirmed prior findings, but also establishes here the prevalence of each arm aneuploidy in LUAD.

CAA differences between primary and metastatic sites

CAA associations with primary and metastatic sites were examined from 3 perspectives: total CAA burden, arm gain-versus-loss and whether individual CAAs are enriched at any particular metastatic site. Sites of metastasis included pleura, chest, lymph, bone, brain, liver and other (Table 1).

CAA burden was defined as the 'sum of aneuploid arms' in an individual, as previously described [5]. This method has been shown to be reliable

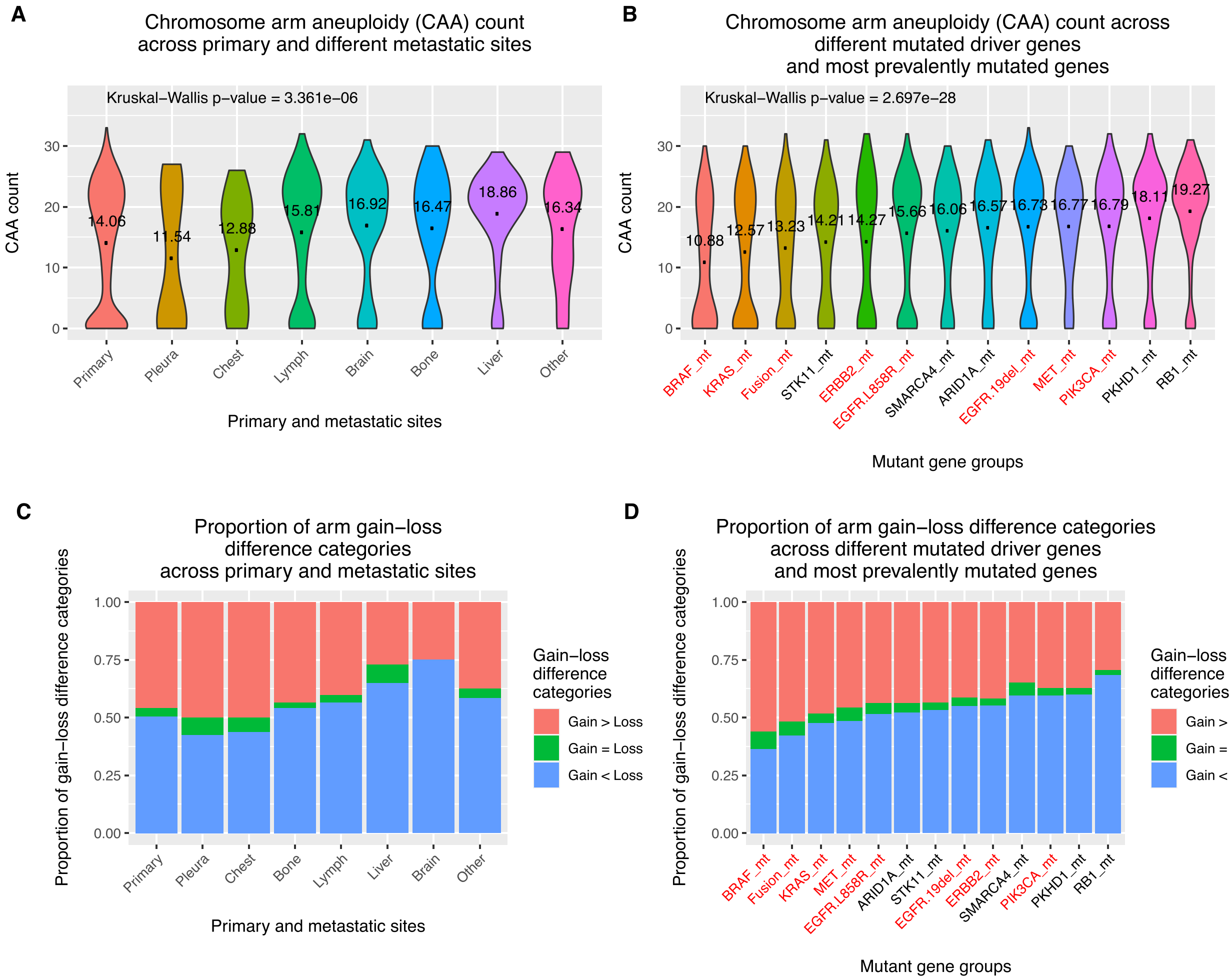

Fig. 2. Distribution of CAA counts and proportion of arm gain and loss difference.
(A) Distribution of CAA count across primary and metastatic sites. (B) Distribution of CAA count across different mutant gene groups. CAA counts are based on the sum of aneuploid arms in a patient. For pairwise Wilcoxon tests, please see Table S1 and S2. (C) Arm gain-loss difference categories across primary and metastatic sites. (D) Arm gain-loss difference categories across different mutant gene groups. A patient's gain-loss difference category is based on whether a patient has more, less, or equal arm gains and arm losses. For pairwise Fisher's exact tests, please see Table S3 and S4. For figure B and D, red x-axis label indicates driver mutants; black x-axis label indicates prevalent non-driver mutant genes.

since arm aneuploidy calls has been shown to not be biased by arm length. Also, the 'sum of aneuploid arms' is closely correlated with aneuploidy at a genome wide level [5]. Primary tumors had significantly lower CAA burden compared to metastasis (14.06 vs 15.94, *P*-value = 2.35e-06). A detailed look at each metastatic site showed pleura metastasis had the lowest CAA burden (11.54), whereas liver had the highest at 18.86 (Figure 2A). CAA burden of primary tumor was significantly different from lymph and liver metastasis ($q < 0.1$). Pleura metastasis had significantly different CAA burden from all other metastatic sites, and chest metastasis was significantly lower than liver ($q < 0.1$) (for results of all pairwise tests, see Table S1).

The next CAA metric we examined was patients' bias towards CAA gain or loss. This CAA gain or loss bias was found by a previous study to distinguish solid tumor from haematological tumors [4]. We wanted to assess whether this gain-loss difference varied among primary and metastatic sites. Three gain-loss difference categories were defined: patients with more arm gains than losses (gain>loss), equal number of arm gains and losses (gain=loss), or more arm losses than gains (gain<loss). Primary and regional metastasis had similar proportion of patients with more arm gains than loss and more arm losses than gains (Figure 2C). Distant metastasis however, had far greater proportion of patients with more arm losses than gains (liver: gain<loss = 65%, gain>loss = 27%; brain: gain<loss = 75%, gain>loss = 25%,). (For results of all pairwise tests, see Table S3).

Tumor purity was not significantly different across primary and metastatic groups (Figure S1D, Table S7), therefore was unlikely to have contributed to differences in CAA burden. A previous study showed TP53 mutations to be strongly correlated with amount of arm level aneuploidy [5], which was also seen in our dataset. We performed stratified analysis based on TP53 mutation status. Results showed only in TP53 wild type patients were CAA burden and arm gain-versus-loss significantly different between primary and metastatic sites (Figure S4A and Figure S4C). CAA metrics was also not affected by disease stage, as differences in CAA burden and arm gain-versus-loss patterns remained in primary and metastatic samples of stage IV patients (Figure S5A and Figure S5C).

At the individual CAA level, brain metastasis was enriched for specific arm losses compared to other sites, including losses of 1p, 3p, 6q, 7q, 9p and 9q (Figure S6A). Distant metastasis (bone, brain and liver) when compared to regional metastasis (pleura and chest), were enriched in gain of 1q and losses of 8p, 9p, 9q, 10q and 13q (Figure S6B).

Table 2

Frequency of mutations in driver and frequently mutated non-driver genes.

Driver genes	Number	Frequency	Non-driver	Number	Frequency
EGFR total	1953	0.55	RB1	214	0.061
EGFR.19del	1005	0.28	PKHD1	192	0.054
EGFR.L858R	838	0.24	ARID1A	184	0.052
KRAS	421	0.12	SMARCA4	181	0.051
PIK3CA	247	0.070	STK11	178	0.051
ERBB2	198	0.056			
BRAF	137	0.039			
MET	73	0.021			
Fusion total	307	0.085			
ALK	189	0.053			
ROS1	85	0.024			
RET	33	0.0093			
TP53	1867	0.53			

These results indicate for LUAD, primary and metastatic sites had varying levels of CAA burden. Distant metastasis such as liver and brain were biased towards arm losses. And specific arm aneuploidies were enriched in distant metastasis, especially the brain.

CAA differences between driver and frequently mutated non-driver gene groups

Next, we examined whether CAA burden and arm gain-versus-loss was associated with specific mutant driver genes or frequently mutated non-driver gene groups. Several LUAD driver mutations were selected based on existing literature [31]. These included BRAF, ERBB2, KRAS, EGFR L858R, EGFR 19del, MET and PIK3CA, as well as prevalent fusions including ALK, RET, and ROS1 fusions. EGFR was the most frequently mutated driver gene in this cohort of Chinese patients (Table 2). Frequently mutated non-driver genes are defined as genes with non-silent mutation frequency > 5%, which were RB1, PKHD1, ARID1A, SMARCA4 and STK11 (Table 2). Genes were considered mutated if it carried one or more non-silent mutations.

CAA burden was significantly different among mutant gene groups (*P*-value = 2.697e-28) (Figure 2B). BRAF mutant group had the lowest CAA burden at 10.88, while EGFR 19del, MET, PIK3CA, PKHD1 and RB1 had high CAA burden, at 16.73, 16.77, 16.79, 18.11 and 19.27 respectively. (For results of all pairwise tests, see Table S2). In terms of arm gain-versus-loss proportions, BRAF mutation group was the most biased towards arm gain, with 56% of patients having more arm gains than losses, while only 36% of patients had more arm losses than gains (8% of patients had equal number of arm losses and gains). While RB1 mutation group was the most biased towards arm losses, with 67% of patients having more arm losses than gains (Figure 2D). (For results of all pairwise tests, see Table S4).

Tumor purity was significantly different across mutant gene groups. However, mutant gene groups with high tumor purity such as BRAF mutants had low CAA burden (Figure S1E, Table S8). Since normal cells are unlikely to be aneuploid, it is unlikely that high tumor purity is driving low CAA burden, and vice versa. Therefore, purity was unlikely a contributing factor in CAA burden differences seen between mutant gene groups. In stratified TP53 mutation analysis, difference in CAA burden between mutant gene groups were much more pronounced in TP53 wild type patients compared to TP53 mutant patients (Figure S4E and Figure S4F). For gain-versus-loss bias, differences between mutant gene groups remained significant for TP53 wild type (Figure S4G), while no longer significant among TP53 mutants (Figure S4H). Disease stage did not affect CAA metrics, as differences in CAA burden and arm gain-versus-loss patterns remained significant between mutant gene groups in stage IV patients (Figure S5B and Figure S5D).

These results show there are clear differences in total CAA burden and arm gain-versus-loss bias among mutant gene groups in LUAD, which are influenced by TP53 mutation status. BRAF, KRAS, ERBB2 and those with fusions belong to a low CAA burden group with bias towards arm gain; whereas EGFR 19del, MET, PIK3CA, PKHD1 and RB1 mutants belong to a high CAA burden group biased towards arm loss.

CAA pair and mutant genes co-occurrence and exclusivity

Co-occurrence and exclusivity of specific CAA and mutant genes was assessed using ‘cooccur’ R package (see Methods for details). KRAS mutant group was most significantly in exclusion with gains of 1p, 7p 14q and 16p, and losses of 2q and 10q. Fusion group was significantly in exclusion with losses of 4q, 5q and 11p (Figure 3A). Next, EGFR 19del and L858R mutant groups were compared directly. We used stringent comparison criteria and found several CAAs that significantly co-occurred/exclusive with one EGFR mutation ($q < 0.1$), but was very non-significant ($q > 0.5$) with the other EGFR mutation group (Figure 3D). Nine CAAs were only co-occurring with EGFR 19del and not with EGFR L858R (losses: 4q, 5p, 5q, 17q, 22q; gains: 8q, 9p, 16q, 17p). In terms of exclusions, 14q loss was in exclusion with EGFR 19del, but not EGFR L858R. 1p loss, 8p gain and 16p loss were in exclusion with EGFR L858R only. In terms of similarities between EGFR L858R and EGFR 19del mutation groups, we found both EGFR mutations were significantly co-occurring with eleven CAAs (losses: 8p, 9q, 10q, 12q, 15q, 19p; gains: 1p, 7p, 7q, 14q, 16p), and in exclusion with 7p loss ($q < 0.1$) (Figure 3E). Lastly, among mutant gene groups with high CAA burden, we found MET mutation co-occurred with 1q loss and PIK3CA mutation co-occurred with 2p gain ($q < 0.1$), while these were not found to be co-occurring with EGFR mutations ($q > 0.5$), which had much larger sample sizes (Figure 3F). (For all CAA and mutant driver gene group co-occurrences and exclusions, see Figure S7). Here we successfully showed various CAAs co-occurred or were in exclusion with certain LUAD driver genes. We also found distinct arm aneuploidy co-occurrence between two EGFR mutation groups.

We also took the opportunity to examine which CAA pairs were significantly co-occurring or mutually exclusive. The top 10 most significantly co-occurring or exclusive CAA pairs are shown in Figure 3A. Overall, significant CAA co-occurrence was far more prevalent than exclusion (Figure 3A). In this cohort of lung adenocarcinoma patients, we found that arm gains tend to co-occur with other arm gains, while arm losses tend to co-occur with other arm losses. Among the top 10 co-occurring or exclusive

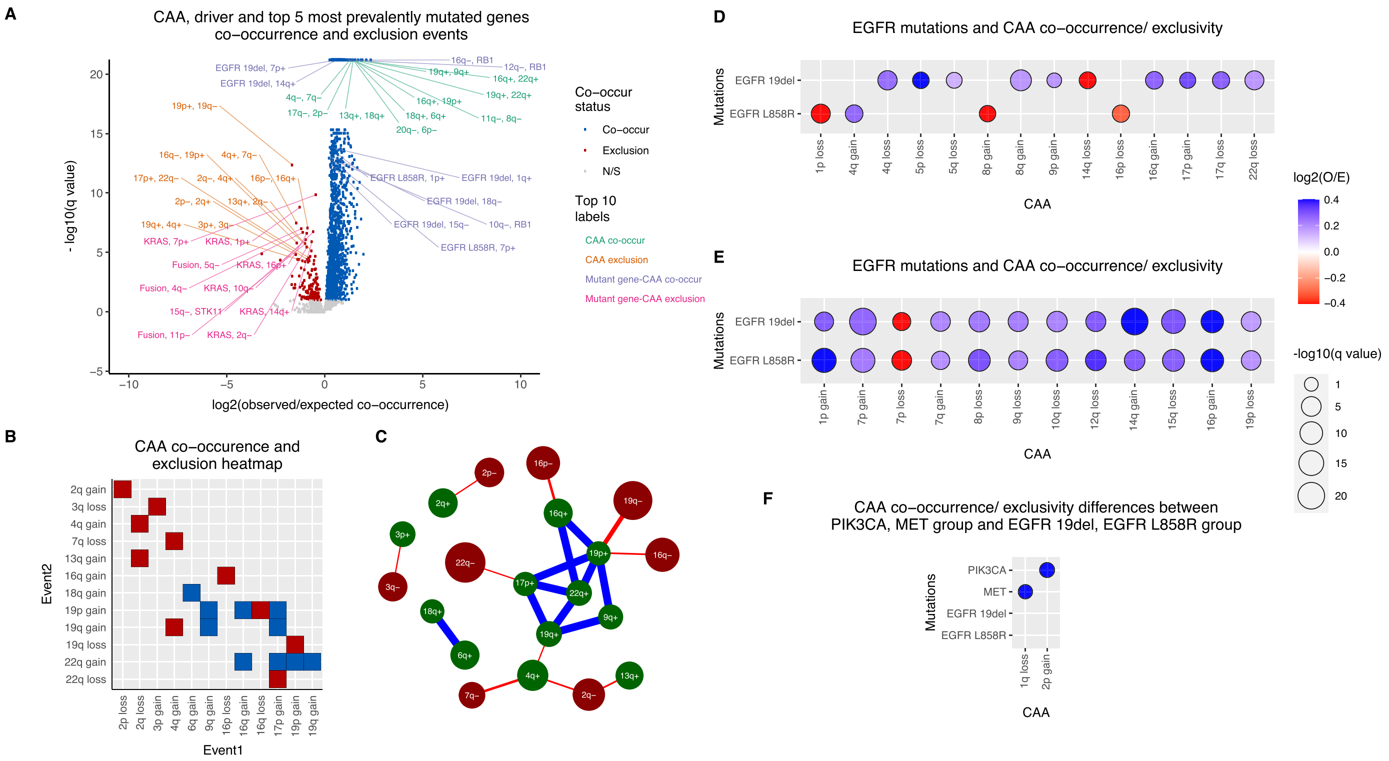

Fig. 3. CAA and mutant gene group co-occurrence and exclusivity.
(A) All CAA and mutant gene co-occurrence and exclusion events. FDR adjusted q−values < 0.1 was considered significant. Only the top 10 co-occurring and top 10 mutually exclusive CAA pairs and CAA mutation gene pairs were labeled. These were selected based firstly on pairs with the lowest q values, then those with the highest absolute value of log2 ratio between observed and expected co−occurrence. (B) CAA co-occurrence and exclusivity heatmap. Only top 10 co-occurring and exclusive arms are shown. (C) CAA co-occurrence and exclusivity network map. Only top 10 co-occurring and exclusive arms are shown. Node size represent CAA frequencies. Edge thickness represent inverse of q values. For (D, E and F), FDR adjusted q−values < 0.1 was considered significant. Log2 ratios were scaled to within -0.4 and 0.4. Circle sizes reflects -log10 of q-values. (D) EGFR mutations and CAA co-occurrence and exclusivity. Aneuploid arms significant in one EGFR mutation group but not the other are shown. Non-significant arms had q-value > 0.5. (E) EGFR mutations and CAA co-occurrence and exclusivity. Aneuploid arms significant in both EGFR mutation groups are shown. (F) CAA co-occurrence and exclusivity difference between PIK3CA and MET group and EGFR mutation groups. Non-significant arms had q-value > 0.5.

CAA pairs, three CAA triads co-occurred with the highest significance: 17p+, 19q+, 22q+; 17p+, 19p+, 22q+; 16q+, 19p+, 22q+ (Figure 3C).

CAA order of acquisition

A recent study showed solid tumors tend to gain chromosome arms initially, but subsequently suffer many arm losses [4]. We assessed whether this applies to LUAD. Probable order of CAA acquisition was inferred from population level data using TRONCO (see Methods for details). First, we observed our LUAD cohort could be separated into two subtypes using arm aneuploidy and a network-based stratification clustering technique (Figure 4A). Subtype 2 appeared to share order of CAA acquisition observed in solid tumors, with initial arm gains, followed by arm losses (8p, 17p, 18q and 19p) (Figure 4F). Subtype 1 however, showed a distinct order of CAA acquisition. It shared an initial 7p gain with subtype 2, but is followed by numerous arm losses, before resuming arm gains (1q, 5p, 7q and 8p) (Figure 4E). Subtype 1 is dominated by patients with brain and liver metastases (Figure 4C, for all pairwise tests, see Table S5), and those with EGFR, PIK3CA, SMARCA4, PKHD1 and RB1 mutations (Figure 4D, for all pairwise tests, see Table S6), while subtype 2 is had more patients with metastases to chest and pleura (Figure 4C), and those with BRAF and KRAS mutations and fusions (Figure 4D). These results show two LUAD subtypes with distinct CAA signatures and evolution trajectories, associated with sites of metastasis and mutant gene groups.

EGFR mutation status stratified analysis

High prevalence of EGFR mutation in this cohort allowed stratified analysis of CAA burden, gain-loss bias, CAA co-occurrence and CAA evolution under EGFR wild type and mutant conditions.

EGFR wild type and mutant subgroups showed similar patterns of CAA burden and gain-loss bias, with regional metastasis having lower CAA burden and bias toward arm gain compared to distant metastasis (Figure S8A, C, H and J), and BRAF, KRAS and Fusion having low CAA burden with bias towards arm gain, while PIK3CA, MET, ARID1A, PKHD1 and RB1 have high CAA burden with bias towards arm loss (Figure S8B, D, I, K). Therefore, EGFR mutation status did not have a modifying effect on CAA burden and gain-loss bias across primary and metastatic sites, or across mutant gene groups.

We also observed few similarities and many differences in CAA co-occurrence between EGFR wild type and mutant subgroups (Figure S8F, M). The large amount of differences in CAA co-occurrence suggest modifying effect of EGFR mutation on CAA co-occurrence and exclusion. Lastly, we observed distinct order of CAA acquisition between EGFR wild type and mutant subgroups. EGFR wild type begins with 8q gain to 7p gain, whereas

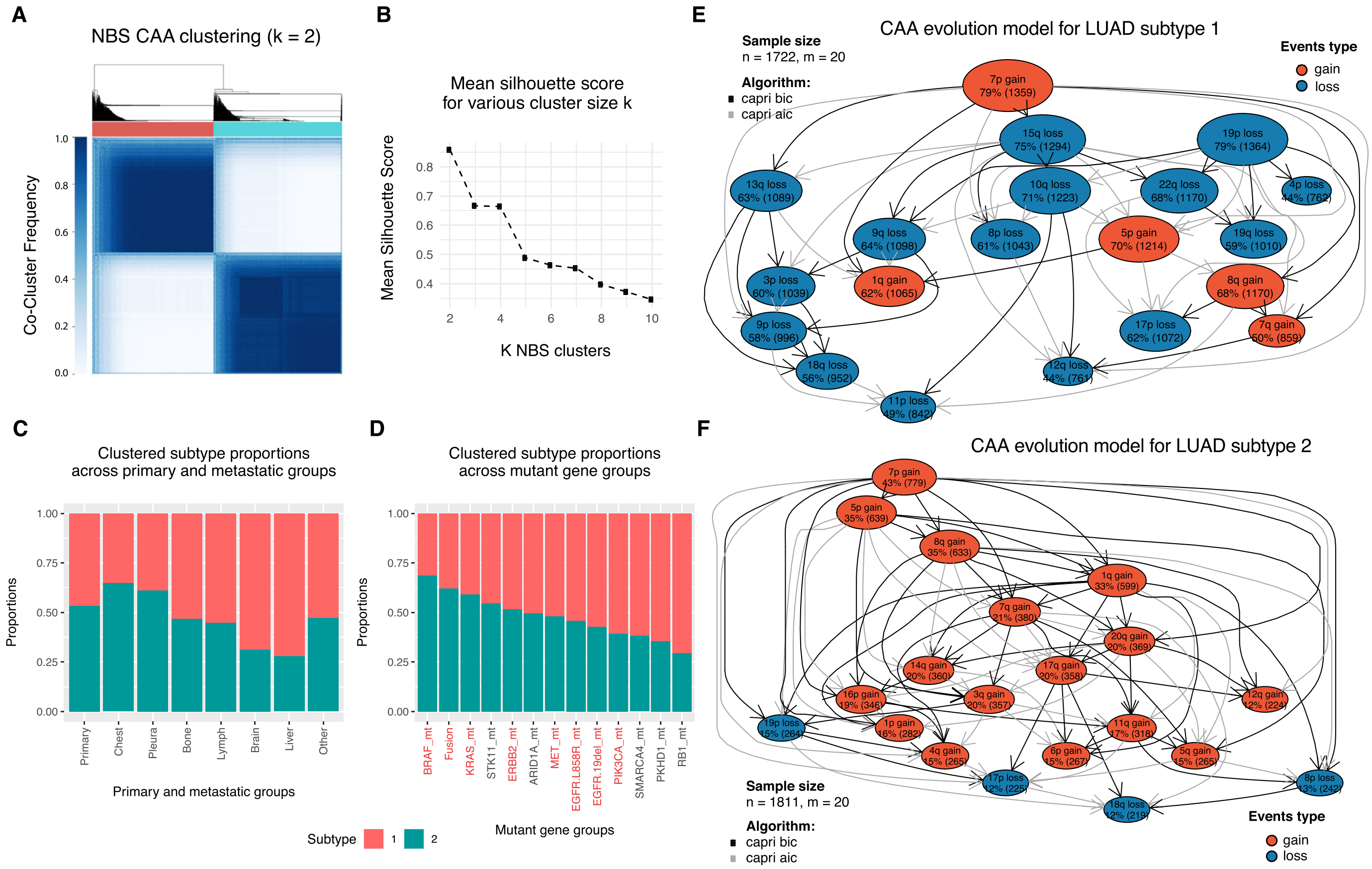

Fig. 4. CAA evolution.

(A) NBS CAA clustering. (B) Silhouette scores of NBS clustering. (C) Clustered subtype proportion across primary and metastatic groups. For pairwise proportion tests, please see Table S5. (D) Clustered subtype proportion across mutant gene groups. Red x-axis label indicates driver mutants; black x-axis label indicates prevalent non-driver mutant genes.For pairwise proportion tests, please see Table S6. (E, F) CAA evolution models. Node size represent CAA frequencies.

n = number of samples. m = number of arms displayed. Only top 20 arms with the highest CAA frequencies in each subtype are shown.

the reverse is seen in EGFR mutant – beginning with 7p gain to 8q gain (Figure S9A and Figure S9B). EGFR wild type has 3q gain and EGFR mutant has 14q gain early in their CAA evolution, whereas they are acquired much later in EGFR mutant and wild type subgroups respectively (later than top 20 CAA acquisitions).

Taken together, EGFR mutation does not interact with metastatic sites or other mutant gene groups in modifying their CAA burden or gain-loss bias. However, having mutations in EGFR affects CAA co-occurrence as well as CAA evolution.

Discussion

In this study, we performed a comprehensive analysis of arm aneuploidy landscape in a large LUAD cohort using target panel sequencing data. CAA burden, gain-loss difference and individual CAAs were associated with sites of metastasis and different mutant gene groups. Distinct orders of CAA acquisition were found in two LUAD subtypes, also associated with metastatic sites and mutant gene groups.

In terms of metastatic sites, regional metastasis (pleura, chest) had significantly lower CAA burden and biased towards arm gains compared to distant metastasis (brain, liver). Regional metastasis such as the pleural space can be readily invaded as it does not require metastatic seed to travel and survive in the circulatory or lymphatic system [32]. Metastasis to distant sites might require additional genomic alterations brought about not only by high levels of CAA, but specific CAAs such as gain of 1q and losses of several arms, including 8p, 9p, 9q, 10q and 13q. In terms of mutant gene groups, some are associated with a general increase in CAA burden. However, we also observed specific CAAs co-occurred with certain mutant genes far above what would be expected by chance. Co-occurrence analysis showed EGFR L858R and EGFR 19del mutant groups shared a number of CAA partners, but also had many distinct CAA associations. Questions still remain whether some arm aneuploidy events provide permissive environment for mutations, or if certain mutations are driving specific arm aneuploidies. RB1 mutant group was shown to have the highest CAA burden, and also shown to co-occur with specific arm losses (10q, 12q and 16q). Studies have shown RB1 inactivation indeed promotes aneuploidy, however the exact mechanism remains unclear [33]. Certain pairs of CAAs also co-occurred or were in exclusion above chance, suggesting that certain arm aneuploidies are not acquired at random. Furthermore, our results show that CAAs not only preferentially co-occur, but appear to be acquired in a sequential manner. A prior study demonstrated differences in CAA order of acquisition between haematological cancers and solid tumors [4]. Whereas we showed that within a single cancer type, there can be differing orders of CAA acquisition. In addition, EGFR wild type and mutant stratified analysis showed distinct order of CAA acquisition. Although mutant EGFR did not modify CAA burden or gain-loss bias, results from CAA co-occurrence and evolution analysis showed presence of EGFR mutation has modifying effects on the emergence of specific CAAs.

Several confounding factors were accounted for in this study, including purity, TP53 mutation status and disease stage. In terms of purity, total copy number estimated by Sequenza [15] are already purity adjusted. The eventual copy number estimates reflect tumor portion of the sample. A sample's tumor purity should have negligible impact on analysis. Lastly, our results along with other studies [11,12] have shown the reliability of using targeted panel sequencing data to study large scale genomic alterations. Methods used here can be applied in other disease settings.

In conclusion, findings of distinct arm aneuploidy differences across metastasis and mutant gene groups may hold biological insight for lung adenocarcinoma. Future functional studies may help validate and provide mechanistic explanation for these results.

Acknowledgments

Not applicable

Ethics approval and patient consent

Study conforms to the Declaration of Helsinki. Written informed consent was collected from each patient upon sample collection according to the protocols approved by the ethical committee of their respective hospitals.

CRediT author statement

Beili Gao: Methodology, Formal analysis, Writing – Original draft, Resources; Fujun Yang: Methodology, Formal analysis, Writing – Original draft, Resources; Ming Han: Methodology, Formal analysis, Visualization, Writing – Original draft; Hua Bao: Supervision, Methodology, Formal analysis, Writing – Review and editing, Yi Shen: Methodology, Formal analysis; Ran Cao: Writing – Review and editing; Xue Wu: Supervision, Writing – Review and editing; Yang Shao: Supervision, Writing – Review and editing; Changhong Liu: Supervision, Conceptualization, Resources; Zhe Zhang: Supervision, Conceptualization, Resources.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neo.2021.06.003.

References

1 Ben-David U, Amon A. Context is everything: aneuploidy in cancer. *Nat Rev Genet* 2020;**21**(1):44–62. doi:10.1038/s41576-019-0171-x.

2 Danielsen HE, Pradhan M, Novelli M. Revisiting tumor aneuploidy-the place of ploidy assessment in the molecular era. *Nat Rev Clin Oncol* 2016;**13**(5):291–304. doi:10.1038/nrclinonc.2015.208.

3 Roy DM, Walsh LA, Desrichard A, Huse JT, Wu W, Gao J, Bose P, Lee W, Chan TA. Integrated genomics for pinpointing survival loci within arm-level somatic copy number alterations. *Cancer Cell* 2016;**29**(5):737–50. doi:10.1016/j.ccell.2016.03.025.Integrated.

4 Shukla A, Nguyen THM, Moka SB, Ellis JJ, Grady JP, Oey H, Cristino AS, Khanna KK, Kroese DP, Krause L, et al. Chromosome arm aneuploidies shape tumor evolution and drug response. *Nat Commun* 2020;**11**(1):1–14. doi:10.1038/s41467-020-14286-0.

5 Taylor AM, Shih J, Ha G, Cherniack AD, Beroukhim R, Meyerson M. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* 2018;**33**:676–689.e3. doi:10.1016/j.ccell.2018.03.007.

6 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;**68**(6):394–424. doi:10.3322/caac.21492.

[7] Myers DJ, Wallen JM. Lung Adenocarcinoma. [Updated 2020 Jun 26]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan. Retrieved July 10, 2013, from: https://www.ncbi.nlm.nih.gov/books/NBK519578/. doi:10.1016/j.mayocp.2018.05.012

8 Danner BC, Hellms T, Jung K, Gunawan B, Didilis V, Fzesi L, Schöndube FA. Prognostic value of chromosomal imbalances in squamous cell carcinoma and adenocarcinoma of the lung. *Ann Thorac Surg* 2011;**92**(3):1038–43. doi:10.1016/j.athoracsur.2011.04.052.

9 Bloomfield M, Duesberg P. Inherent variability of cancer-specific aneuploidy generates metastases. *Mol Cytogenet* 2016;**9**(1):1–22. doi:10.1186/s13039-016-0297-x.

10 Benstead-Hume G, Wooller SK, Downs JA, Pearl FMG. Defining signatures of arm-wise copy number change and their associated drivers in kidney cancers. *Int J Mol Sci* 2019;**20**(22). doi:10.3390/ijms20225762.

11 Turajlic S, Xu H, Litchfield K, Rowan A, Horswell S, Chambers T, O'Brien T, Lopez JI, Watkins TBK, Nicol D, et al. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* 2018;**173**(3):595–610 e11. doi:10.1016/j.cell.2018.03.043.

12 Mamlouk S, Childs LH, Aust D, Heim D, Melching F, Oliveira C, Wolf T, Durek P, Schumacher D, Bläker H, et al. DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nat Commun* 2017;**8**. doi:10.1038/ncomms14093.

13 Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**(15):2114–20. doi:10.1093/bioinformatics/btu170.

14 Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**(18):333–9. doi:10.1093/bioinformatics/bts378.

15 Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, Szallasi Z, Eklund AC. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015;**26**(1):64–70. doi:10.1093/annonc/mdu479.

16 Griffith DM, Veech JA, Marsh CJ. Cooccur: Probabilistic species co-occurrence analysis in R. *J Stat Softw* 2016;**69**(1):1–17. doi:10.18637/jss.v069.c02.

17 De Sano L, Caravagna G, Ramazzotti D, Graudenzi A, Mauri G, Mishra B, Antoniotii M. TRONCO: An R package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics* 2016;**32**(12):1911–13. doi:10.1093/bioinformatics/btw035.

18 Huang JK, Jia T, Carlin DE, Ideker T. pyNBS: a Python implementation for network-based stratification of tumor mutations. *Bioinformatics* 2018;**34**(16):2859–61. doi:10.1093/bioinformatics/bty186.

19 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. *J ofMachine Learn Res* 2011;**12**(1):2825–30. doi:10.1145/2786984.2786995.

20 Petersen I, Bujard M, Petersen S, Wolf G, Goeze A, Schwendel A, Langreck H, Gellert K, Reichel M, Just K, et al. Patterns of chromosomal imbalances in adenocarcinoma and squamous cell carcinoma of the lung. *Cancer Res* 1997;**57**(12):2331–5. doi:10.1016/s0959-8049(97)85798-x.

21 Björkqvist AM, Tammilehto L, Nordling S, Nurminen M, Anttila S, Mattson K, Knuutila S. Comparison of DNA copy number changes in malignant mesothelioma, adenocarcinoma and large-cell anaplastic carcinoma of the lung. *Br J Cancer* 1998;**77**(2):260–9. doi:10.1038/bjc.1998.42.

22 Pei J, Balsara BR, Li W, Litwin S, Gabrielson E, Feder M, Jen J, Testa JR. Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas. *Genes Chromosom Cancer* 2001;**31**(3):282–7. doi:10.1002/gcc.1145.

23 Luk C, Tsao MS, Bayani J, Shepherd F, Squire JA. Molecular cytogenetic analysis of non-small cell lung carcinoma by spectral karyotyping and comparative genomic hybridization. *Cancer Genet Cytogenet* 2001;**125**(2):87–99. doi:10.1016/S0165-4608(00)00363-0.

24 Balsara BR, Testa JR. Chromosomal imbalances in human lung cancer. *Oncogene* 2002;**21**:6877–83 (45 REV. ISS. 5). doi:10.1038/sj.onc.1205836.

25 Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, et al. Characterizing the cancer

genome in lung adenocarcinoma. *Nature* 2007;**450**(7171):893–8. doi:10.1038/nature06358.

26 Berrieman HK, Ashman J, Cowen ME, Greenman J, Lind MJ, Cawkwell L. Chromosomal analysis of non-small-cell lung cancer by multicolour fluorescent in situ hybridisation. *Br J Cancer* 2004;**90**(4):900–5. doi:10.1038/sj.bjc.6601569.

27 Staaf J, Isaksson S, Karlsson A, Jönsson M, Johansson L, Jönsson P, Botling J, Micke P, Baldertorp B, Planck M. Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. *Int J Cancer* 2013;**132**(9):2020–31. doi:10.1002/ijc.27879.

28 Rotolo F, Zhu CQ, Brambilla E, Graziano SL, Olaussen K, Le-Chevalier T, Pignon JP, Kratzke R, Soria JC, Shepherd FA, et al. Genome-wide copy number analyses of samples from LACE-Bio project identify novel prognostic and predictive markers in early stage non-small cell lung cancer. *Transl Lung Cancer Res* 2018;**7**(3):416–27. doi:10.21037/tlcr.2018.05.01.

29 Zhao X, Weir BA, LaFramboise T, Lin M, Beroukhim R, Garraway L, Beheshti J, Lee JC, Naoki K, Richards WG, et al. Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res* 2005;**65**(13):5561–70. doi:10.1158/0008-5472.CAN-04-4603.

30 Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, Golas B, Wang L, Motoi N, Szoke J, et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. *Oncogene* 2009;**28**(31):2773–83. doi:10.1038/onc.2009.135.

31 Pao W, Girard N. New driver mutations in non-small-cell lung cancer. *Lancet Oncol* 2011;**12**(2):175–80. doi:10.1016/S1470-2045(10)70087-5.

32 Agalioti T, Giannou AD, Stathopoulos GT. Pleural involvement in lung cancer. *J Thorac Dis* 2015;**7**(6):1021–30. doi:10.3978/j.issn.2072-1439.2015.04.23.

33 Manning AL, Dyson NJ. RB: mitotic implications of a tumour suppressor. *Nat Rev Cancer* 2012;**12**(3):220–6. doi:10.1038/nrc3216.