Supplementary Methods

Whole-genome doubling (WGD) estimation

Briefly, each sample was represented as a copy number profile of major and minor allele copy numbers at the level of each chromosome arm. Using these arm-level summaries, the total alterations (relative to diploid) and the probabilities of loss/gain for each allele at each chromosome arm were calculated. $10,000$ simulations were then run for each sample. In each simulation, a number of sequential aberrations, based on the probabilities of loss/gain for each allele already calculated, were applied to a diploid profile. A p-value for WGD was obtained by counting the percentage of simulations where the proportion of chromosome arms with a major allele copy number ≥ 2 was higher than that observed in the sample.

GISTIC2.0 peak definition with multi-sample tumours

The summary profiles described below reduce the data from the multiple samples analysed from an individual tumour to a single set of copy number segments across the genome. This allows each tumour to be represented by a single summary profile corresponding to either clonal or subclonal SCNA that may be used as input to GISTIC2.0.

In order to identify clonal and subclonal SCNA as well as to include LOH as loss, four separate summary profiles for each multi-sample tumour were generated:

- Profile 1: A clonal profile that did not take into account the presence of LOH.
- Profile 2: A clonal profile that takes into account the presence of clonal LOH.
- Profile 3: A subclonal profile that does not take into account the presence of LOH and investigates subclonal gains rather than losses when they occur at the same genomic loci in distinct samples from the same tumour.
- Profile 4: A subclonal profile that takes into account the presence of subclonal LOH and investigates subclonal losses/LOH rather than gains when they occur at the same genomic loci in distinct samples from the same tumour.

For each tumour sample and the most recent common ancestor (MRCA) (Online methods "Ancestral reconstruction and phylogeny inference") from an individual tumour, seg_CN values for each segment in the tumour's minimum consistent segmentation are generated as described (Online methods "GISTIC2.0 peak definition").

Clonal profiles 1 and 2

For the summary clonal SCNA profiles (Profiles 1 and 2) the allele-specific copy number profiles of the MRCA for each tumour were inferred using MEDICC[Schwarz, 2014].

- Profile 1 reflects gains, without considering the presence of LOH, and is the seg_CN profile of the inferred MRCA.
- Profile 2, taking into account the presence of LOH, sets any area affected by LOH in the MRCA to a seg_CN equivalent of a total integer copy number value of 1 in a diploid tumour leaving the value of seg_CN in areas of the genome unaffected by LOH unchanged.

Subclonal profiles 3 and 4

To determine the significantly recurrent focal peak regions of gain and loss in a subclonal context we generated two separate sets of summary SCNA profiles (profiles 3 and 4). For both of each tumour's summary subclonal SCNA copy number profiles, seg_CN data across all its individual samples were considered.

In the subclonal summary profiles for each tumour, in the areas of the minimum consensus segmentation across the genome that our analysis identified as a clonal loss, clonal gain, clonal amplification or a clonal neutral event across all samples (Online methods, "SCNA intratumour heterogeneity and clonality definitions") the sample segments with the total integer copy number with the smallest absolute difference to their sample's respective ploidy were selected. This is because these segments are likely to be the most similar to the tumour's MRCA and demonstrate the smallest subclonal change in copy number.

These summary segments corresponding to areas of the genome with clonal copy number changes are used in both Profile 3 and Profile 4. For the remaining minimum consensus segments across the genome that our analysis identified as subclonal (Online methods, "SCNA intratumour heterogeneity and clonality definitions") two separate procedures for Profile 3 and Profile 4 were followed that are detailed below.

- Profile 3 aims to investigate subclonal gains rather than losses if they occur at the same position and does not take into account LOH. In this profile, in segments classified as subclonal gains/amplifications, the seg_CN corresponding to the greatest positive total copy number difference from its sample's respective ploidy was selected. The remaining subclonal segments that were only identified as subclonal losses and not as subclonal gains/amplifications, had the segment with the greatest negative total copy number difference lower than its sample's respective ploidy chosen.
- Profile 4 aims to investigate subclonal losses and LOH rather than gains/amplifications if they occur at the same position. Across all samples from each tumour, any segment that was identified as subclonal LOH had its seg_CN value set to the equivalent of an integer copy number value of 1 in a diploid tumour. Then, in segments identified as subclonal losses by our analysis (Online methods, "SCNA intratumour heterogeneity and clonality definitions"), the seg_CN corresponding to the sample with the greatest negative total copy number difference to its sample's respective ploidy was selected and included in the summary profile. The remaining subclonal segments that were only identified as a subclonal gain/amplification and not as a subclonal loss had the segment with the greatest additional number of copies above its sample's respective ploidy chosen.

For each of the four profiles, GISTIC v2.0.23 was applied to the transformed copy number data using default settings with run_broad_analysis set to FALSE. In addition, we used a seg_CN total

copy number ratio cap of 1.5 as used by Zack et al. [Zack, 2013]. All peaks were mapped to the affected hg19 cytobands and subsequent analyses performed at the cytoband level.

- Peaks of clonal SCNA gain/amplification were taken from the GISTIC2.0 run with the above settings on Profile 1.
- Peaks of clonal SCNA loss/LOH were taken from the GISTIC2.0 run with the above settings on Profile 2.
- Peaks of subclonal SCNA gain/amplification were taken from the GISTIC2.0 run with above settings on Profile 3. In addition, in order to be considered as a subclonal peak of gain/amplification, the same area of the genome must also have been identified using a separate permutation test for recurrence of subclonal gain SCNAs across samples (Online methods "Permutation test for recurrence of SCNA across tumours").
- Peaks of subclonal SCNA loss/LOH were taken from the GISTIC2.0 run with the above settings on Profile 4. In order to be considered as a subclonal peak of loss/LOH, the same area of the genome must also have been identified using a separate permutation test for recurrence of subclonal loss/LOH SCNA across samples (Online methods "Permutation test for recurrence of SCNA across tumours").

Prevalence and clonality within consensus peak regions

See Online methods "GISTIC2.0 consensus peak definition" for details on the creation of consensus peaks.

The number of tumours with a SCNA event overlapping at least one cytoband within a consensus peak region were reported in Figure 3B. Across-genome plots at the single cytoband level showing the proportion of the cohort affected by SCNAs overlapping each cytoband are shown in Extended Data Fig. 6 and 7.

Description of the Markov chain model that incorporates arm-level events

We adapted a previously described Markov chain model [3] that keeps track of the distribution of the number of copies of a given chromosome arm. Below is an example for chromosome $1p$.

States of the Markov chain are triples of (a, f, σ), where

- a is the number of attached copies of $1p$ (either part of chromosome 1 or of a neo-chromosome that contains $1p$)
- f is the number of free copies of $1p$
- $\sigma = +1, -1, 0$ depending on whether there are any free p-arms $(+1)$, any free q-arms (-1), or no free arms (0) in the cell.

There is an additional state corresponding to dead cells. Cells are considered dead if $a + f$, which is the total number of copies of $1p$, goes below 1 or above N.

Each step of the Markov chain corresponds to one generation. The transition probabilities are computed from the following scenarios. At each step, each cell dies spontaneously with probability computed from its fitness score, as previously described [3] but with arm-specific rather than

whole chromosome scores. The contribution of a given chromosome arm (e.g. $1p$) to the survival probability of the cell is

$$\exp(d \cdot \text{score}_{1p} \cdot (a + f)),$$

where d controls the probability of death after each cell division [4]. The value of d was determined empirically based on apoptosis rates found in tumour samples as described [4].

Assuming the cell survives, it undergoes WGD with probability p_{GD}, otherwise it divides and the values a, f, σ are updated as follows:

1. Each copy of a whole chromosome or neo-chromosome containing $1p$ breaks at the centromere with a probability p_{split}.

2. Each copy of a whole chromosome or neo-chromosome containing $1p$ missegregates with probability p_{misseg}.

3. Each free arm $1p$ and $1q$ missegregates with probability 0.5, as these broken arms cannot form proper attachments to the mitotic spindle.

4. The value of σ is updated by estimating chromosome breakages and free arm missegregations in the whole cell.

5. Any remaining free $1p$ arms fuse with any existing free q-arms in the cell. In particular, after these fusions, the resulting state (a, f, σ) cannot have $f > 0$ and $\sigma < 0$ simultaneously, since in that case the free p-arms would fuse with the free q-arms.

Starting in state $(A, 0, 0)$ and running the Markov chain for g steps, we obtain the probability distribution on the states of the Markov chain after g generations when the founder cell has A copies of arm $1p$. The probability that a random cell in the colony after g generations has B total copies of arm $1p$ is then equal to the sum of the probabilities of states (a, f, σ) with $a + f = B$, divided by the sum of the probabilities of all states (a, f, σ) with $1 \leq a + f \leq N$, which correspond to live cells. Deviance scores were computed to take into account the square of the differences in copy number between the average predicted karyotype and the actual sample karyotype and normalized to the unweighted evolution model.

Markov chain model parameters

Experimental measurements of chromosome missegregation rates have revealed that in most chromosomally unstable cancer-derived cell lines, missegregation rates fall in the range between 0.001 and 0.00422 per chromosome copy per cell division, so this was the tested range for p_{misseg}. The tested range for g, between 0 and 300, was derived from both experimental and computational studies [5, 4, 3]. The tested range for p_{GD} was between 0 and 0.014, based on in silico modelling predictions. For samples for which no WGD was observed on the genomic level, p_{GD} values of 0 were chosen. Below is the full list of parameter values.

- Maximum number of allowed copies of any given arm: $N = 8$, which is the bound used in [4, 3].

- Missegregation rate: $p_{\text{misseg}} = 0.00422$.

- Probability of a whole chromosome splitting into 2 arms at a given cell division: $p_{\text{split}} = 0.4 \cdot p_{\text{misseg}}$.
- Missegregation probability for free broken arms: 0.5.
- Parameter used to translate chromosome scores into survival probabilities: $d = 0.00039047$, as computed previously [3, 4].
- Arm OG-TSG scores: as derived from [Davoli, 2013].
- Probability of WGD at each cell division: $p_{\text{GD}} = 0$ for non-WGD data, $p_{\text{GD}} = 0.005$ for WGD data, $p_{\text{GD}} = 0.012$ for subclonal WGD data.
- Number of generations: $g = 75$ for non-WGD and WGD data, $g = 150$ for subclonal WGD data.

Incorporation of OG-TSG scores in Markov chain modelling

The formula to compute the survival probability of each cell is similar to Equation (1) in [3], which depends on the chromosome scores (score_k = score of chromosome k) and the number of copies of each chromosome (num_copies_k = number of copies of chromosome k), plus two fixed parameters c and d that are used to translate cell scores into probabilities. The formula has been refined to keep track of the number of copies of each arm.

Specifically, the term $\text{score}_k \cdot \text{num\_copies}_k$ in the exponent in Equation (1) in [3] is now replaced with

$$(\text{score}_{kp} \cdot \text{num\_copies}_{kp}) + (\text{score}_{kq} \cdot \text{num\_copies}_{kq}),$$

where kp and kq denote the p-arm and the q-arm of chromosome k, respectively.

The model without OG-TSG scores corresponds to setting all the scores to 0, which makes the survival probability of each cell be independent of its karyotype.

The model with scrambled OG-TSG scores selects a permutation of the chromosome arms uniformly at random, and then assigns the scores to the chromosome arms according to that permutation.

In our figures involving the model with scrambled scores, an average of the model run for 1000 random permutations of OG-TSG scores is displayed. However, this is with the exception of panels h–j of Extended Data Fig 5, where we generate 10 random permutations to illustrate how the behavior of the model varies slightly depending on the random permutation that was generated. In addition, in Extended Data Fig 5 k–q a single random permutation is shown.

Investigating results in Markov chain modelling

A positive deviance score difference signifies that a predicted karyotype deviates further from the actual karyotype in the first model run more than in the second model run.

If selection shapes the clonal/subclonal SCNA landscape, chromosome arms with a higher density of OGs are more likely to be subject to clonal or subclonal gains while those with a higher density of TSGs are more likely to exhibit clonal or subclonal losses.

Characterisation of mitotic index and anisonucleosis

Although the prognostic value of nuclear atypia has been demonstrated [Von der Thusen, 2013; Kadota, 2014], no histopathological grading system is universally recognised in non-small cell lung carcinoma, aside from the description of architectural patterns in LUAD.

All NSCLC grading systems which have been proposed are specific to either LUAD or LUSC, owing to their inherent differences in cytological and architectural appearance. In this dataset, which incorporates both tumour types, the assessment of anisonucleosis was selected because it is the most robust characteristic of nuclear atypia and pleomorphism that could be graded uniformly across the histological tumour types.

References

[1] Schwarz, R.F. et al. Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLoS Comput Biol.* **10**, e1003535 (2014).

[2] Zack, T.I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

[3] Elizalde, S., Laughney, A.M. & Bakhoum, S.F. A Markov chain for numerical chromosomal instability in clonally expanding populations. *PLOS Comput. Biol.* **14**, e1006447 (2018).

[4] Laughney, A.M., Elizalde, S., Genovese, G. & Bakhoum, S.F. Dynamics of tumor heterogeneity derived from clonal karyotypic evolution. *Cell Rep.* **12**, 809–820 (2015).

[5] Dewhurst, S.M. et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).

[6] Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).

[7] von der Thüsen, J.H. et al. Prognostic significance of predominant histologic pattern and nuclear grade in resected adenocarcinoma of the lung: potential parameters for a grading system. *J. Thorac. Oncol.* **8**, 37–44 (2013).

[8] Kadota, K. et al. Comprehensive pathological analyses in lung squamous cell carcinoma: single cell invasion, nuclear diameter, and tumor budding are independent prognostic factors for worse outcomes. *J. Thorac. Oncol.* **9**, 1126–1139 (2014).