Patterns of selection against centrosome amplification in human cell lines

Marco António Dias Louro[1,✉], **Mónica Bettencourt-Dias**[1], **and Claudia Bank**[1]

[1]Instituto Gulbenkian de Ciência, Oeiras 2780-156, Portugal

The presence of extra centrioles, or centrosome amplification, is a hallmark of cancer cells. Centriole numbers in cancer cell populations appear to be at an equilibrium maintained by centriole overproduction and selection, reminiscent of mutation-selection balance. It is not known if the interaction between centriole overproduction and selection can quantitatively explain the intra- and inter-population heterogeneity in centriole numbers.

Here, we define mutation-selection-like models and employ a model selection approach to infer patterns of centriole overproduction and selection in a diverse panel of cell lines. Surprisingly, we infer strong and uniform selection against any number of extra centrioles, in most cell lines. However, we do not detect significant differences in parameter estimates between different cell lines. Finally we assess the accuracy and precision of our inference method and find that it increases non-linearly as a function of the number of sampled cells. We discuss the biological implications of our results and how our methodology can inform future experiments.

modelling | centrioles | cancer | evolutionary theory | inference

Correspondence: ***mlouro@igc.gulbenkian.pt***

Introduction

Centrioles are microtubule-based structures that organize the centrosome and thereby orchestrate microtubule nucleation in vertebrate cells (1, 2). Centriole number abnormalities are a source of phenotypic heterogeneity in cancer cells. Indeed, centriole numbers show variability both within cancer cell populations and between different cancers (3–5). The causes and consequences of this heterogeneity are still poorly understood. Moreover, to the best of our knowledge, there exists no quantitative description of how centrioles numbers are distributed in cancer cells.

In a proliferating cell population, cells start the cell cycle with two centrioles, which duplicate once and only once during S-phase. After cytokinesis, both daughter cells inherit two centrioles each. This centriole duplication and segregation cycle ensures that centriole number is kept constant across generations (3, 6, 7).

In stark contrast with most proliferating cells, centriole numbers are often de-regulated during cancer development. In particular, cells with abnormally high numbers of centrioles are common in tumors and cancer-derived cell lines, and have been recently identified in pre-neoplastic tissues (8–12). Interestingly, within a single population of cancer cells, individual cells often carry different numbers of centrioles. However, the number of centrioles per cell in the population seems to display a specific distribution depending on the cell type (9–12). The source of this variability within and between cell populations is still poorly understood and calls for the development of quantitative approaches.

The occurrence of extra centrioles, termed centrosome amplification, tends to bear deleterious consequences for the cell by triggering multipolar divisions, cell cycle arrest, and/or by promoting chromosome missegregation (13–16). Thus, excess centrioles are typically counter-selected and rarely observed in healthy tissues. However, some mechanisms are known to provide protection against centrosome amplification. For example, centrosome clustering mechanisms allow cells to group extra centrioles in two spindle poles, thus improving the viability of daughter cells (13, 14, 16). Thus, cancer cell lines are generally regarded as being more tolerant to centrosome amplification than normal cells.

Recent data suggest that centriole numbers are maintained at an equilibrium in cell line populations. For instance, it has been observed that after transient centriole elimination, p53-deficient cell populations can seemingly recover their initial distribution of centriole numbers (17, 18). Similarly, there are reports of extra centrioles being lost over time in cell populations after induction of cytokinesis failure (19). Since centrosome amplification is typically deleterious for cells, it is likely that centriole numbers in these populations are maintained by a balance of centriole (over)production and negative selection. These dynamics are similar to an evolutionary mutation-selection process, where the *de novo* appearance of deleterious variants in a population is counteracted by natural selection, eventually converging to so-called mutation-selection balance (20).

How centrosome (over)production and selection behave quantitatively is poorly understood. For instance, extra centrioles may be gained "smoothly" in a dose-dependent fashion through overexpression of key centriole biogenesis regulators, such as Plk4, STIL, and SAS-6 (21–24). Alternatively, extra centrioles may be gained in sharp transitions - if an otherwise normal cell undergoes cytokinesis failure, it may restart the cell cycle with at least double the normal number of centrioles (25). Similarly, it is not known if selection strength varies with the number of centrioles. For example, it is possible that centriole clustering is less efficient in resolving multipolar spindles if the cell contains a high number of extra centrioles. In the absence of protective mechanisms, it is possible that the presence of extra centrioles is deleterious, regardless of absolute centriole numbers. Thus, it is not a trivial question how centriole (over)production and selection can generate equilibrium distribution of

centriole numbers, and if these two processes are sufficient to explain the observed centriole number heterogeneity within and between cell populations.

Here, we develop mathematical models of centriole overproduction and selection against centrosome amplification that are predicated on different assumptions on how supernumerary centrioles are produced and how selection operates. We use these models to analyze recently published data on representative cell lines of the progression from Barrett's esophagus to gastroesophageal adenocarcinoma (9) and the NCI-60 panel of cancer cell lines (10). These two data sets provide us with the opportunity to study how centriole number distributions vary along cancer progression, from pre-malignant to malignant stages, in the case of the Barrett's esophagus data set, and between different cancer types, in the NCI-60 data set.

Employing a model selection approach, we found that models featuring a constant cost of centrosome amplification, irrespective of the number of centrioles in a cell, best explain the empirical distributions for most of the cell lines. Moreover, our results suggest that the distribution of centriole numbers is generally super-exponential, which could be indicative of multi-step centriole number increments. We identified a general trend in the parameter estimates indicating strong selection against extra centrioles but we did not detect significant differences between cell lines. Using simulations, we show that our parameter estimation method is accurate and we predict that its precision increases non-linearly with the number of sampled cells. In summary, our work presents the first quantitative description of how centriole numbers evolve in proliferating cell populations with persistent centrosome amplification and provides a statistical tool for further dissecting the processes that generate within- and between-population variation in centriole numbers.

Results

A model of centriole number dynamics in proliferating cell populations. To study how centriole number distributions in proliferating cell populations are generated, we developed a general mathematical model grounded in mutation-selection theory. Our subject of focus is a population of proliferating cells subject to centriole overproduction and selection against extra centrioles. For the purpose of data analysis, we consider that individual cells are in mitosis and fully characterised by their number of extra centrioles, i, which can range from zero to an arbitrarily high upper bound, i_{max}. Thus, the population can be split into subpopulations by centriole numbers such that, for example, the zeroth subpopulation ($i = 0$) represents all cells containing wild-type centriole numbers (four centrioles). Each sub-population has an intrinsic growth rate r_i. Centriole overproduction occurs at a rate $\mu_{i,j}$ from subpopulation i to subpopulation j, where $j > i$. Thus, we assume that there is no loss of centrioles across cell division. Centriole overproduction events can be interpreted as gain of $j - i$ centrioles. Since in these cell lines there is a net increase in the number of centrioles compared to the wild-type situation, we make the simplifying assumption that there is no loss of centrioles. Cells that contain fewer than wild-type numbers are rarely observed in the analysed data sets; for simplification purposes, we disregard them in our model and in our analysis. Finally, our model is deterministic and all sub-population frequencies P_i are continuous variables; i.e., we assume an effectively infinite population size. Taken together, the temporal rate of change in the relative frequency of cells containing i centrioles is given by the following ordinary differential equation:

$$\frac{dP_i}{dt} = \Big(r_i - \sum_{j=0}^{i_{max}} r_j P_j(t)\Big) P_i(t) + \sum_{k=0}^{i} \mu_{k,i} P_k(t) - \sum_{l=i+1}^{i_{max}} \mu_{i,l} P_i(t) \quad \text{, for all } i \leq i_{max}. \tag{1}$$

The dynamics of the population are thus described by a system of i_{max} differential equations. As previously mentioned, experimental data suggest that centriole numbers in the population follow a stable equilibrium distribution when unperturbed (17, 18). We propose the following expression that describes a fully polymorphic locally stable equilibrium distribution (Figure S1) for an arbitrary value of i_{max} (see also Methods):

$$P_i^* = \frac{\eta_i}{\sum_{j=0}^{i_{max}} \eta_j}, P_0(0) > 0 \tag{2}$$

$$\eta_i = \prod_{j=i+1}^{i_{max}} f(j) \Bigg(\sum_{a_o \in A(i)} \Big(\prod_{k \in a_o} f(k) \prod_{m=1}^{|\Phi(i,k)|-1} \mu_{\phi_m, \phi_{m+1}} \Big) \Bigg) \tag{3}$$

where

$$f(i) = r_0 - \sum_{j=1}^{i_{max}} \mu_{0,j} - r_i + \sum_{n=i+1}^{i_{max}} \mu_{i,n} \text{for all } i <= i_{max}, \tag{4}$$

and

$$A(j) = (a_o)_{o \in \mathcal{P}(S^j)} \tag{5}$$

is the sequence containing all elements of the power set $\mathcal{P}(S^j)$. The set S^j is defined as

$$S^j = y \in \mathbb{N} : 0 < y \leq j - 1. \tag{6}$$

Finally, we define

$$\Phi(j,k) = (\phi_m)_m \in \{\{0\} \bigcup \{a_j \neg k\}\}, \tag{7}$$

with a_j corresponding to the elements of $A(j)$ from the definition in equation Eq. (5).
This set of equations determines the equilibrium balance of centriole overproduction and selection in a cell population, i.e., it determines the predicted proportion of cells with i extra centrioles in an unperturbed cell population, given an arbitrary set of overproduction rates and fitness functions. To our knowledge, this equilibrium solution has not been described in the rich population-genetic literature on mutation-selection models to date (20). Furthermore, this general solution allows for the computation of analytical expressions for the equilibrium distributions and their (log-)likelihood under more specific centriole overproduction and selection scenarios, as shown below.

Distributions of centriole numbers in cell populations tend to be heavy-tailed. Our goal is to infer the balance between selection and centriole overproduction from the shape of the distribution of centriole numbers in samples from 67 cell lines (9, 10). In these data sets, between 35 and 82 mitotic cells were sampled from a population of cultured cells, and centrioles were identified, and counted, by co-immunostaining of two centriolar markers.
As a first step, we characterised which type of distribution most likely underlies these data; this helps to identify more specific models for parameter inference. For example, consider a simple model where extra centrioles are produced at a constant rate μ and extra centrioles induce a uniform growth rate penalty r. This model is analogous to the classical mutation-selection model and can be obtained in our general framework by substituting $\mu_{i,j} = \mu$ for $j = i+1$ and 0 otherwise, and $r_i = r$. Under these assumptions, equation Eq. (2) can be written as

$$P_i^* = \frac{(1-r-\mu)\mu^i}{(1-r)^{i+1}}, P_0(0) > 0 \tag{8}$$

It can be readily seen by substitution that this equilibrium is formally equivalent to the geometric distribution:

$$P(X = i) = (1-p)^i p, \tag{9}$$

where p represents the probability of success in i trials (roughly, the probability of observing an extra centriole in our model). If an analogue to the classical mutation-selection model is sufficient to explain the data, then the data should be geometrically distributed. In contrast, we observed an overrepresentation of cells with high centriole numbers and an underrepresentation of wild-type-like cells (Figure S2) compared to a geometric distribution. This is a coarse indication that the distribution of centriole numbers in cell populations is heavy-tailed. If that is the case, then more complex centriole overproduction and/or selection dynamics are required to explain the data.
To test this at the level of individual empirical distributions for each of the 67 cell lines, we fitted geometric and Waring-Yule distributions to each of the empirical population-level distributions of centriole numbers. Here, the Waring-Yule distribution represents a generalised discrete distribution that can potentially account for heavy tails. Then, we calculated the value of the X^2 statistic as a measure of goodness-of-fit for both distributions. Finally, we determined the difference between the value of the X^2 statistic, ΔX^2, under the geometric and Waring-Yule models, such that positive values indicate a better fit of the Waring-Yule distribution. Conversely, if the geometric distribution is a better fit, we expect that both distributions should converge and yield a ΔX^2 of approximately 0.
Visual inspection of model fits suggests the Waring-Yule (heavy-tailed) distribution is a better fit to the represented empirical distributions (Figure 1A, B, and C). In addition, our results indicate positive values of ΔX^2 for the majority of cell lines, suggesting a better fit of the Waring-Yule (heavy-tailed) distribution (Figure 1D). For 16 out of 67 cell lines, we obtained values of $\Delta X^2 \leq 0$, indicating exponential-like and not heavy tails. Thus, our results suggest that a simple model reminiscent of classical mutation-selection balance, which results in a geometric distribution of centriole numbers in the population, fails to explain the data for most cell lines.
Ultimately, we are interested in relating the distribution of centriole numbers in the population to the processes that generate it. If the distributions are indeed heavy-tailed, this could be achieved either by weak selection against cells with high centriole numbers or by more complex centriole overproduction mechanisms. For example, since centrioles duplicate in most healthy cells, it is possible that centriole overproduction also occurs in multiples of two (which we will refer to as overduplication). Similarly, after cytokinesis failure, a cell may restart the cell cycle and reduplicate all four centrioles, gaining four extra (25). Intuitively, overduplication or cytokinesis failure could produce cells with multiples of two and/or four centrioles. Coherently, it can be seen in both individual and pooled distributions that cells with four, and eight, extra centrioles are particularly abundant in the data (Figure S2).

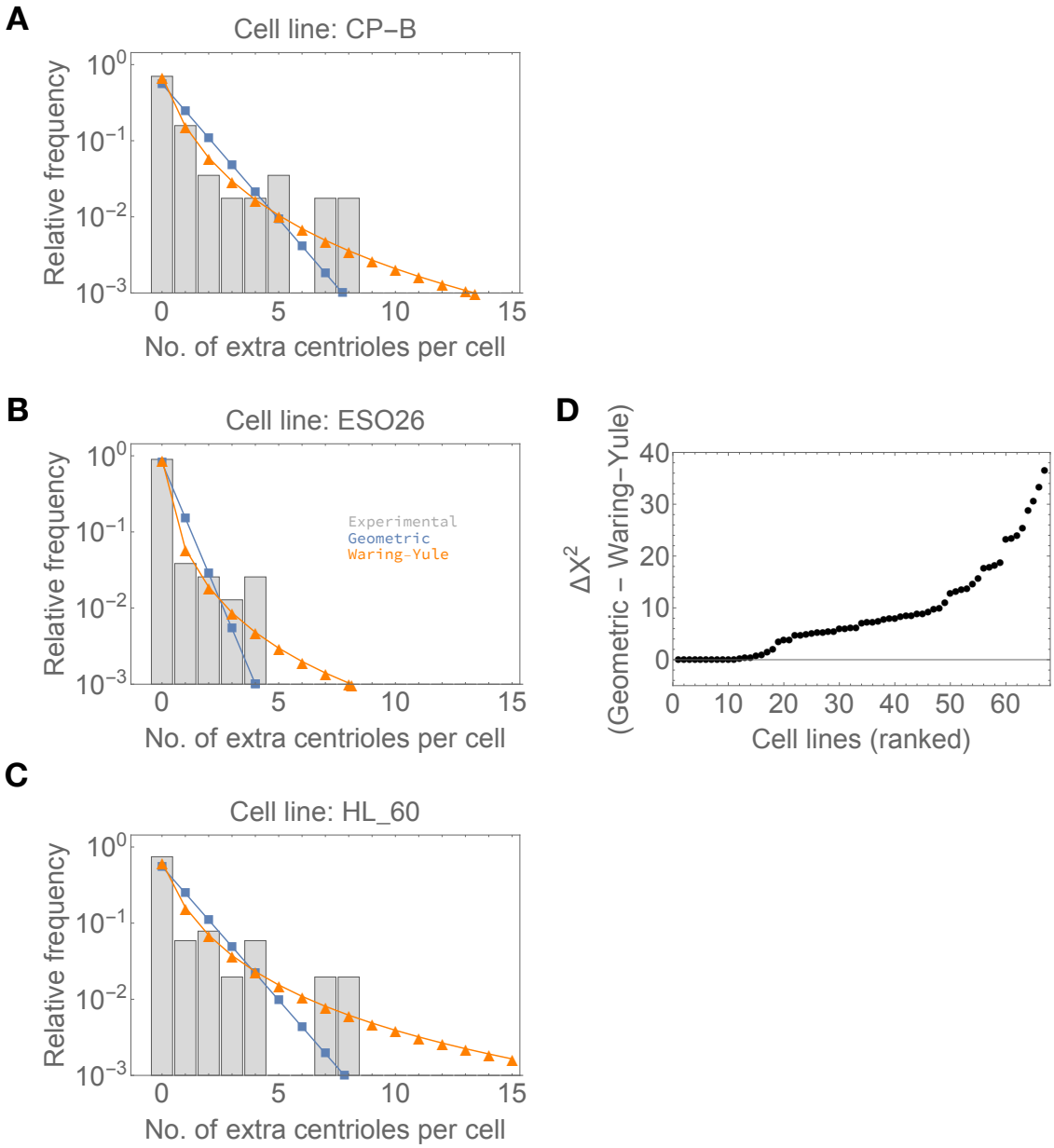

Fig. 1. Distributions of centriole numbers in cell populations tend to be heavy-tailed.. A-C - Examples of empirical distributions for three cell lines (grey bars) and the predicted relative frequencies under geometric (blue) and Waring-Yule distributions (orange), which are representative of distributions with exponential-like and heavy tails, respectively. Number of sampled cells: (A) n=57; (B) n=78; (C) n=51. D - Difference between the calculated X^2 value under geometric and Waring-Yule distributions. Cell lines were ranked in ascending order according to the ΔX^2 value. Higher positive values indicate a better fit of the Waring-Yule distribution to the corresponding empirical distribution, suggesting heavier-than exponential tails.

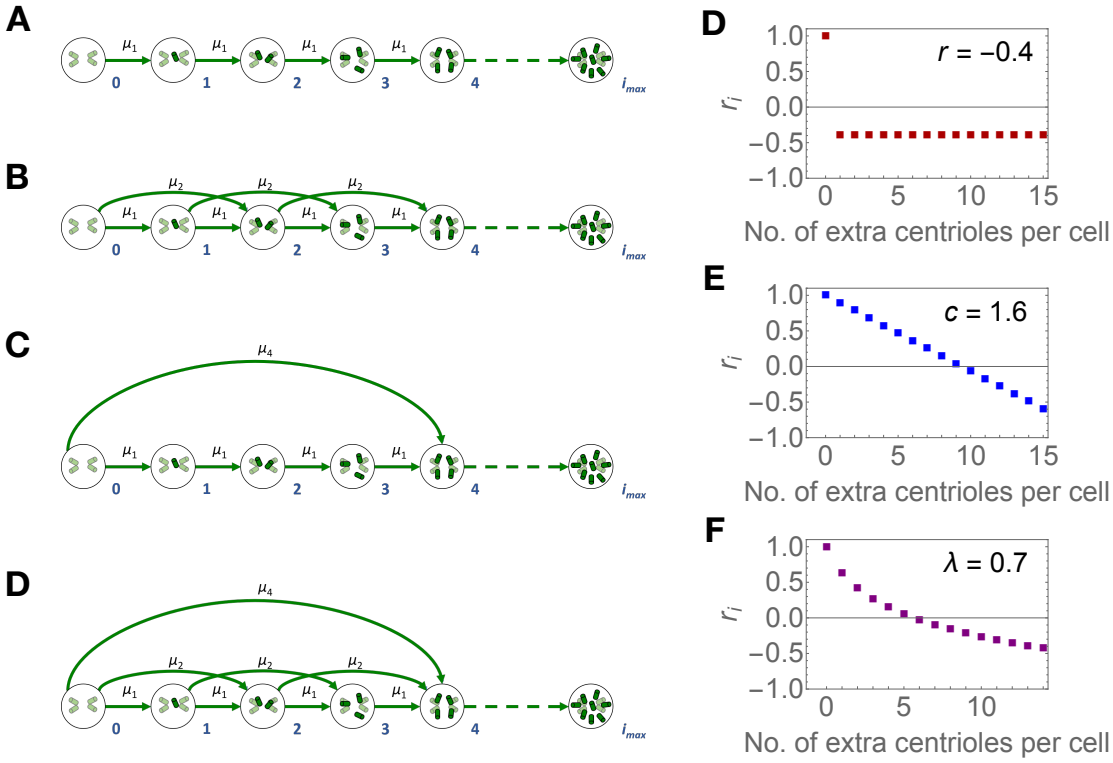

Fig. 2. Centriole overproduction and fitness functions in the candidate models. (a) Single-step centriole overproduction, at a rate μ_1, (b) single- and double-step centriole overproduction events, at rates μ_1 and μ_2, (c) single- and quadruple-step centriole overproduction events, at rates μ_1 and μ_4, (d) and all three centriole overproduction events. Circles represent the subpopulation of cells containing i extra centrioles (green cylinders, i indicated by the numbers in blue). Green arrows represent transitions between subpopulations, which correspond to centriole overproduction, occurring at a rate μ_1, μ_2 or μ_4. Overproduction events can occur for all i up to i_{max}. (d-f) The value of r_i , for cells with i centrioles, under the (d) flat, (e) linear, and (f) power-law fitness functions, evaluated at the indicated parameter values.

In summary, a model parameterised by a single centriole overproduction and instrinsic growth rate is not sufficient to explain the data. By visually inspecting the empirical distributions, we hypothesised that including more complex centriole overproduction events in the models, such as literal centriole overduplication, may provide better fits.

A candidate set of models based on hypotheses on centriole biology. The general model described in equation Eq. (1) provides a powerful starting point for inferring the distribution of centriole numbers in proliferating cell populations but it is overparameterised with respect to the data under consideration. Moreover, we are interested in comparing different hypotheses regarding specific fitness functions and overproduction parameters that could generate said centriole number distributions. To avoid overfitting and to inspect relevant biological scenarios, we generated 12 candidate models by imposing a set of constraints on centriole overproduction and on selection, based on the previous analysis on the tail of the distributions.

Several cellular processes are known to yield supernumerary centrioles but we still lack a quantitative description of their contribution to the generation of cells with extra centrioles. We reasoned that these processes may be parameterised as the rate of gain of a given number of centrioles. As a universal scenario across all models, we considered that extra centrioles can be gained one at a time, at rate μ_1. Since centrioles typically duplicate in number and since we observed an excess of multiples of two and four centrioles in the data, we reasoned that centriole overproduction could also be thought of in terms of extra duplication events. Thus, we considered two additional overproduction "rules", which state two and four extra centrioles can be gained at a rate μ_2 and μ_4, respectively.

Similarly, how strongly selection acts depending on the number of extra centrioles remains unknown. We focused on two main possibilities: first, that any abnormal number of centrioles is equally deleterious (resulting in a flat fitness function for all $i > 0$), and second, that the deleterious effect of extra centrioles increases with their number in a cell. Regarding the latter, we assume either an additive or power-law relationship between the number of extra centrioles and intrinsic growth. In all models, we set the intrinsic growth rate for cells with wild-type centriole numbers, r_0, to be maximal and equal to 1; i.e., cells that contain no excess centrioles always have maximum fitness. Otherwise, a fully polymorphic equilibrium (i.e. where cells with any number of extra centrioles could, in theory, be observed) would not be reached under these models. The intrinsic growth rates for cells with abnormal centriole numbers ($i > 0$) are defined as (1) $r_i = r$ for all i>0 ("flat" model), (2) $r_i = 1 - \frac{ci}{i_{max}}$ for all $i > 0$ ("linear" model), and (3) $r_i = 1 - \frac{\log(i+1)}{\lambda \log(i_{max}+1)}$ for all $i > 1$ ("power-law" model).

The full combinatorial set of the above-fitness functions and "overproduction" rules yields 12 different models, in the following

named using the initial of the fitness function and the overproduction steps (Figure 2). For example, F1-- refers to the model featuring a flat fitness function (parameterised by r) and single centriole overproduction events (parameterised by μ_1). Note that the models with fewer parameters are nested within the more parameter-heavy models and can be obtained by setting excluded centriole overproduction rates to zero. For instance, F124 yields identical equilibria to F1-- if $\mu_2 = 0$ and $\mu_4 = 0$.
In addition, the 12 models in our candidate set can be obtained by modifying equation Eq. (2) according to the specified centriole overproduction and intrinsic growth rate "rules". In the case of model F124, and by extension, all models nested in it, equation Eq. (2) simplifies to:

$$P_i^*\Big|_{r,\mu_1,\mu_2,\mu_4} = \frac{(1-r-\mu_1-\mu_2-\mu_4)(\sum_{j=0}^{\lfloor i/4 \rfloor}\sum_{k=0}^{\lfloor (i-4j)/2 \rfloor}\binom{i-3j-k}{\max(j,k)}\binom{i-4j-2k-\min(j,k)}{\min(j,k)})\mu_1^{i-4j-2k}\left((1-r)\mu_2\right)^k\left((1-r)^3\mu_4\right)^j}{r^{i+1}}. \quad \textbf{(10)}$$

This form is convenient because it is independent of i_{max}, therefore bypassing the need to assume a potentially artificial upper bound for the number of centrioles per cell.
We note that our focal set of models is neither an exhaustive nor systematic exploration of all possibilities. However, it includes models of different complexity, depending on the number of centriole overproduction parameters. Moreover, it incorporates different biological hypotheses with respect to fitness, and results in both exponential and heavy-tailed equilibrium distributions as described above (see also Figure S3).

Models assuming a flat fitness function best explain the data for most cell lines. As a first approach we tested if the most complex models in our candidate set (i.e. the models assuming all three overproduction events) are a good fit to the data. First, we fitted the models to each empirical distribution. Then, we performed a Monte Carlo multinomial test to distinguish between predicted and empirical distributions (Figure 3). We considered that the models were a poor fit if the test yielded a significant p-value (p-value $\leq 0.05/67$, adjusted according to a Bonferroni correction). Our models showed a good fit for a majority (56 for the "flat" model, 52 for the "linear" model, 57 for the "power-law model) of cell lines.
Thus, we concluded that the most complex models are a good fit to the data. For 10 out of 67 cell lines, all three model are a poor fit to the corresponding empirical distributions. These cell lines tend to display some proportion of cells with extremely high centriole numbers (≥ 15), which are very rare under all our models. We took a conservative approach and excluded these cell lines from further analysis.
Next, we asked which of the 12 models best explains the observed data. We generated 200 bootstrap distributions for each cell line by drawing a random sample with replacement from each empirical distribution, and fitted each of the 12 models. Then, we calculated the Bayesian Information Criterion (BIC) score from the resulting maximum log-likelihood value. The model that minimised the BIC for *the largest number* of bootstrap distributions was selected as the best for each cell line (Fig. 4). Strikingly, the best models for 41 out of 57 cell lines assumed the flat fitness function. 16 cell lines are best explained by models assuming the linear fitness function; these cell lines are not associated to any specific tissue types or developmental stages (Figure S4. No cell line is best explained by models with a power-law fitness function. In contrast, there is more variability with respect to the best set of centriole overproduction parameters.
Next we analysed the number of bootstrap distributions were selected for each cell line (Figure S5. We observed that the best models for each cell line are selected for a maximum of 198 (99%) and a minimum of 40 (20%) bootstrap distributions, with a median of 110.5 (55.25%) bootstrap distributions (Figure S5B). In addition, in some cases, the BIC score is equal for models assuming the flat and power-law fitness functions (see supporting code). This means that the decision for the most appropriate model is sometimes not clear, which can be either due to the models yielding indistinguishable distributions, or the data not being sufficiently informative to distinguish between models (see also below).
Then we examined models sharing the same fitness function as the best model. In total, the same fitness function was selected for a maximum of 199 (99.5%) and a minimum of 78 bootstrap distributions (39%), with a median of 150 (75%) (Figure S5C). Therefore, our results are more consistent regarding the selected fitness function, when compared to individual models.
Despite some uncertainty in model selection, we concluded that most empirical distributions are best explained by models assuming a flat fitness function, with models assuming a linear fitness function best explaining a smaller subset.

Parameter estimates indicate strong selection against excess centrioles. For the following parameter estimation and in order to further reduce the complexity of choices, we thus focused on the model F124, which contains the models with a constant fitness function but simpler centriole overproduction rules.
We next tested if we could distinguish between different cell lines based on their parameter estimates. First, we tested the sensitivity of the fitting as a function of the parameters, by evaluating the log-likelihood expression at values around the maximum, and looked for correlations between pairs of parameters. We observed that for model F124, all parameters are uncorrelated, and thus can be independently estimated from the data (Figure S6).
Second, we analysed the maximum-likelihood r and μ_1 estimates obtained for the previously generated non-parametric bootstrap distributions, under model F124 (Figure 5). We observed globally negative median estimates of the intrinsic growth rate, r, indicating strong selection against cells with extra centrioles. The median estimates for the single-step overproduction rate,

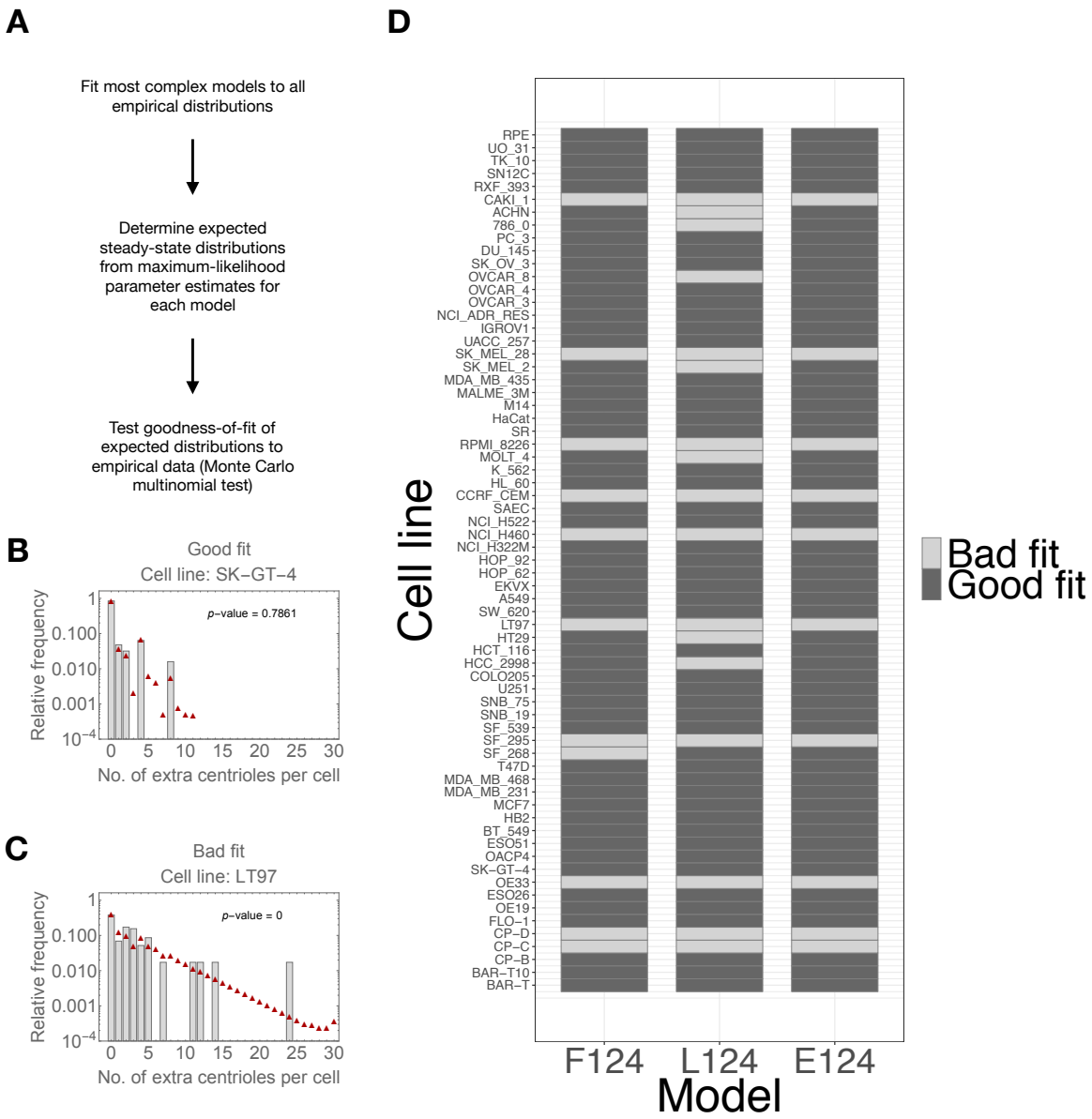

Fig. 3. The most complex models are a good fit to the majority of empirical distributions. (a) Procedure for determining goodness-of-fit. The significance value was set to $\alpha = 0.05$, and adjusted according to the Bonferroni correction for 67 tests. Note that under the null hypothesis, the predicted distribution under a given model is identical to the empirical distribution. (b) Experimentally observed frequencies of centriole numbers per cell (grey bars) and the predicted frequencies under model F124 - flat fitness function, single-, double-, and quadruple-step overproduction parameters (red triangles). The two examples include cases where we obtained non-significant (good fit) and significant (bad fit) p-values. Number of sampled cells: (b) n=82 and (c) n=63. (d) Goodness-of-fit of the three most complexes models to each cell line.

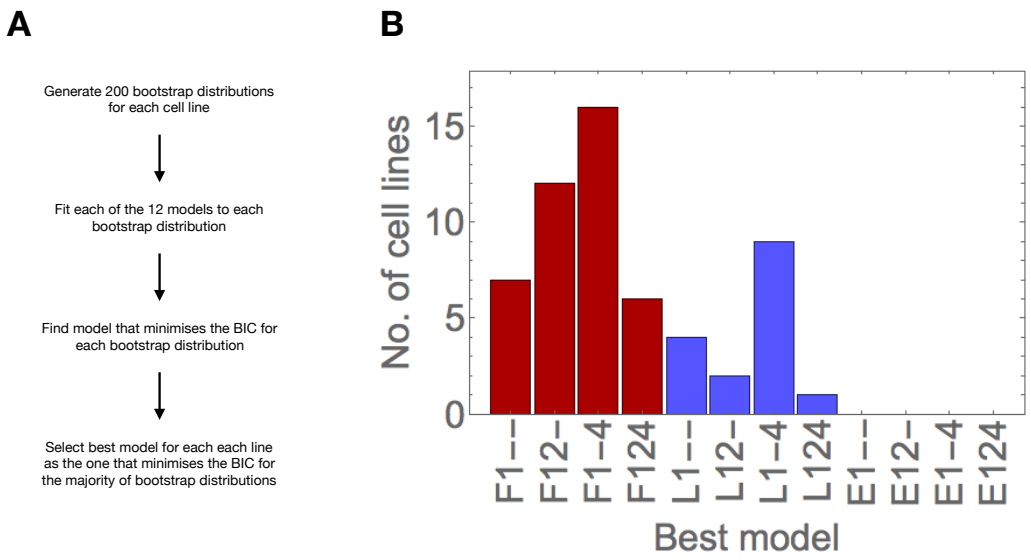

Fig. 4. Models assuming a flat fitness function best explain the data for most cell lines.(a) Procedure for model selection. (b) Models with a flat fitness function explain the data for most cell lines. Number of cell lines for which the corresponding model was selected as the best. The fitness function of the models is indicated in red - flat; blue - linear.

μ_1 are relatively low but more variable than those of the intrinsic growth rate r. Indeed, for most cell lines, the median estimate of μ_1 falls within the range of 0-0.423, with two cell lines scoring over 0.7 (SE268 and HT19). However, the confidence intervals for both parameters are considerably wide, spanning almost the whole parameter range in the case of the intrinsic growth rate r, such that we cannot identify significant differences between cell lines. Thus, our results indicate a pattern of low to moderate centriole overproduction rates and intrinsic growth rates for cells with extra centrioles.

Next, we addressed the accuracy and precision of our inference method. If our method is accurate, we expect that parameter estimates from model simulations converge to the input parameter values. In addition, we expect that errors in parameter estimation should be similar between model simulations and the ones obtained in the non-parametric bootstrap, for samples of the same size. To test this, we used a parametric bootstrap, where we resampled the expected distributions under model F124 instead of resampling the empirical distributions as in the non-parametric bootstrap. As input values, we used the previously calculated median r, μ_1, μ_2, and μ_4, values from the 200 non-parametric bootstrap distributions for each cell line. We assumed a sample size, i.e. number of sampled cells, equal to that obtained in the corresponding data set and generated 200 parametric bootstrap distributions. Then, we fitted model F124 to each parametric distribution and analysed the estimated parameter values. Our results show that the median parameter estimate obtained from the parametric mostly agrees with the input value (Figure S7), indicating that our method is accurate. Moreover, we obtained confidence intervals of similar length to those obtained from the non-parametric bootstrap distributions. For some cell lines, the errors obtained for the intrinsic growth rate r differed between the non-parametric and parametric bootstrap distributions. This is likely due to the lower sensitivity of the maximum likelihood values to changes in r compared to μ_1 (Figure S6).

We concluded that parameter estimation accuracy and precision are similar when data is simulated either from the empirical or predicted distributions. Since confidence interval length is influenced by sample size and we did not detect systematic biases in our inference method, it is likely that the precision of our parameter estimates mainly depended on the number of sample cells per cell line.

Accuracy and precision of the inference method increase non-linearly with the number of sampled cells. To provide statistical guidance for future experiments, we asked how much the precision of the parameter estimates and accuracy of model selection could be improved by increasing the sample size, within experimentally feasible limits.

To address this question, we reasoned that the estimated values from the median expected distribution would provide a useful test case. We identified the median expected distribution by measuring the Euclidean distance between the vector whose elements are the four estimated parameter values for a given bootstrap distribution to the vector of the lowest possible values for each parameter.

We used the parameter values corresponding to the median Euclidean distance as inputs for model F124 and simulated 200 parametric bootstrap distributions in a range of sample sizes. For comparison purposes, we assumed a sample size of 35 and 83, which correspond respectively to the lowest and the highest number of cells obtained experimentally in our data sets. In addition, we considered realistic sample sizes of 50, 100, 150, 200, and 250 cells. Finally, we simulated parametric bootstrap distributions with a sample size of 1000 to analyse the properties of our inference method if larger sample sizes were attainable.

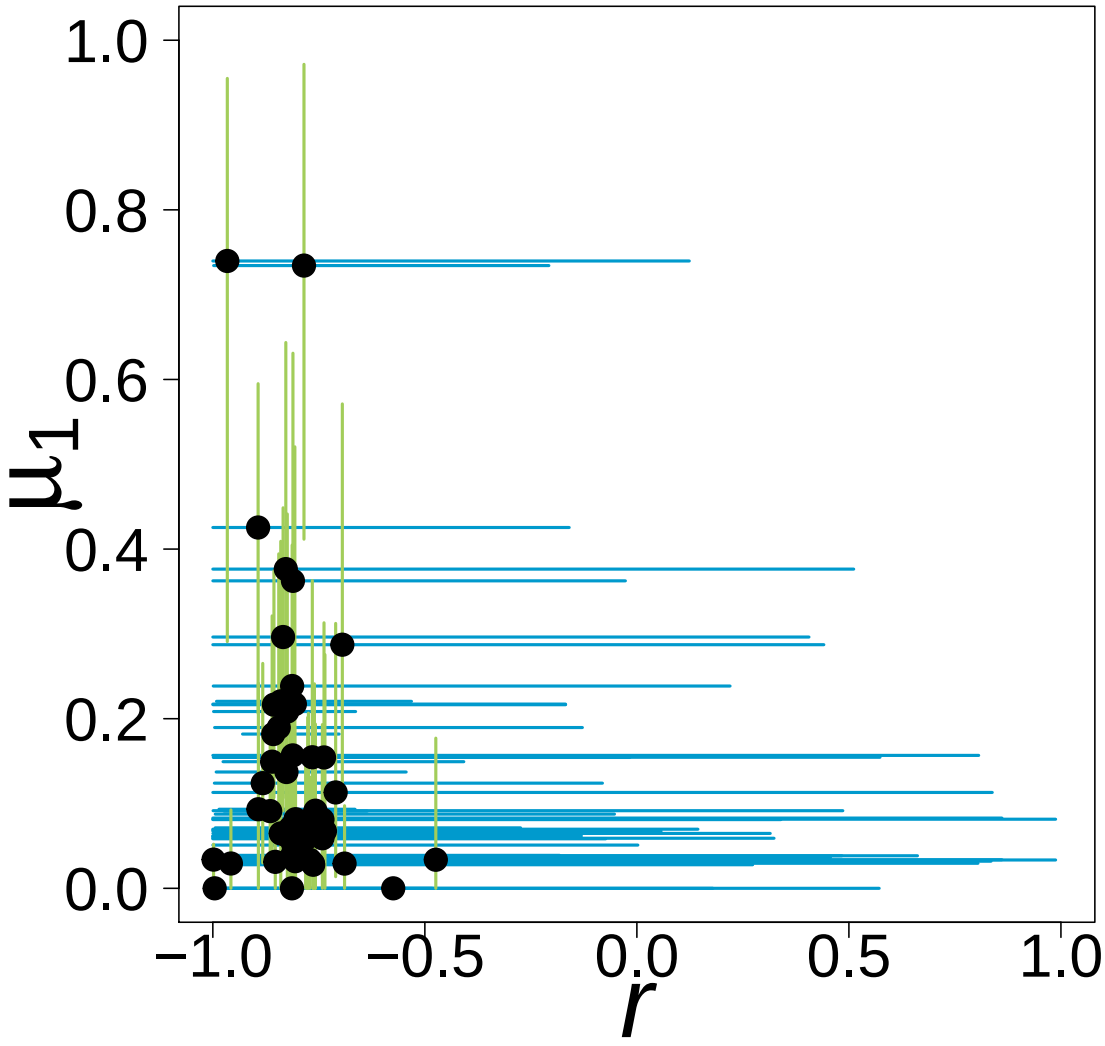

Fig. 5. Parameter estimates indicate strong selection against excess centrioles but considerable errors in parameter estimation. (a) Median estimates (black dots) of r and μ_1 obtained from 200 non-parametric bootstrap distributions for each empirical distribution. Lines indicate 95% confidence intervals (0.025 to 0.975 inter-quantile distance) for r (blue) and μ_1 (light green).

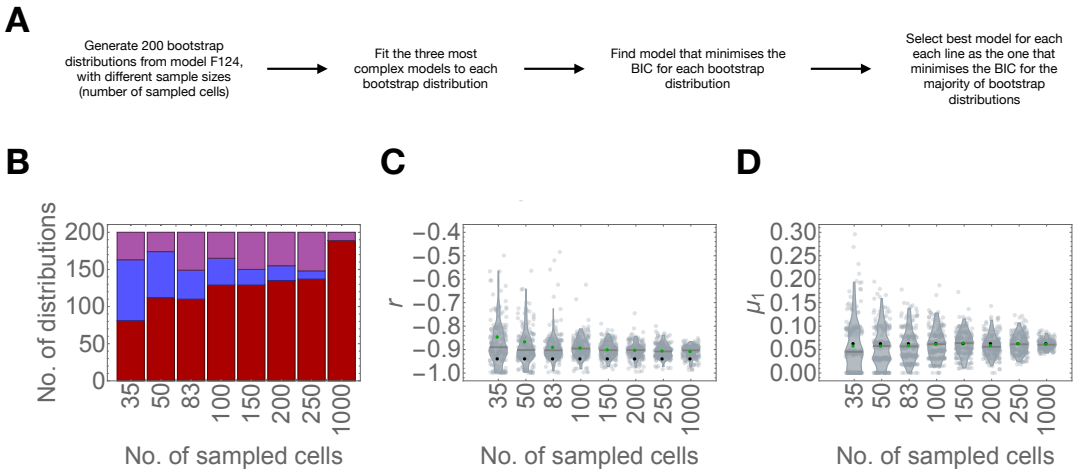

Fig. 6. Accuracy and precision of the inference method increase non-linearly with the number of sampled cells. (a) Number of bootstrap distributions for which model F124 (red), L124 (blue), and E124 (purple) fitness functions were selected as the best in simulations of model F124 as a function of bootstrap sample size. Input parameter values for the simulations: $r = -0.940$, $\mu_1 = 0.062$, $\mu_2 = 0.096$, $\mu_4 = 0.288$. (b-c) Bootstrap distribution of parameter estimates for r (b) and μ_1 (c) as a function of bootstrap sample size. Black dots indicate the median estimated value and the green dot indicates the input value used in the simulations.

Subsequently, we fitted the three most complex models. We counted the number of times each model maximized the log-likelihood function out of the 200 bootstrap distributions for each simulated sample size. Note that the BIC is unnecessary because the three models have the same number of parameters. In addition, we analysed the parameter estimates of model F124 from the simulated distributions.

Since we performed simulations under model F124, then this model should fit best most of the bootstrap distributions. It is also trivial that errors in parameter estimation will become smaller for larger sample sizes. Regardless, we were interested in *quantifying* how often model F124 is identified as the best model and how confidence interval length varies with the number of sampled cells.

Figure 6 shows the results for model selection from simulated data. We observed that within the range of experimentally obtained sample sizes, model F124 is correctly identified in 81 (40.5%), 112 (56%), and 110 (55%), out of 200 parametric bootstrap distributions, for a sample size of 35, 50, and 83 simulated cells, respectively. Model F124 is only marginally outperformed by model L124 for a sample size of 35 (selected for 81 and 82 bootstrap distributions, respectively). Nevertheless, we observed that the number of bootstrap distributions for which the true model is selected increases to 129 (59.5%) for a sample size of 100, and to 137 for a sample size of 250 (67.5%). For the maximum sample size tested, the true model was selected for 189 (94.5%) bootstrap distributions.

Next, we inspected how the distributions of parameter estimates vary with sample size. We observed that the input value always falls within the confidence interval but we noted that the median r value is consistently overestimated with respect to the input value (Fig. 6C). Nevertheless, parameter estimation is fairly accurate for higher sample sizes. We also observed a near two-fold decrease in confidence interval length between the minimum (35) and maximum (83) experimentally obtained sample sizes, and a further 1.55, 1.91, 2.02 and 2.01-fold decrease between the maximum experimentally obtained sample size (83) and examples containing 100, 150, 200, and 250 simulated cells, respectively. For 1000 simulated cells, we obtained confidence intervals with a length of 0.109, corresponding to a 2.43-fold decrease compared to the maximum experimentally obtained sample size.

Unlike intrinsic growth rates, we obtained extremely accurate median estimates for μ_1 regardless of sample size (Figure 6D). Thus, it is possible that the overestimation of r is because the estimated values lie close to the lower bound (-1), and not an inconsistency generated by the model. We observed that the confidence interval for μ_1 decreased from 0.194 to 0.138 (1.41-fold decrease), for sample sizes of 35 and 83, respectively, and further to 0.114 for a sample size of 100 (1.21-fold decrease compared to a sample size of 83) and 0.083 for a sample size of 250 (1.66-fold decrease compared to a sample size of 83). For 1000 simulated cells, confidence interval lengths were as low as 0.0412.

In conclusion, both model selection accuracy and parameter estimation precision increased non-linearly with sample size. Importantly, we predict that increasing the number of sampled cells within a feasible range (between 100 and 200) can greatly improve our inference power. However, it should be stated that these results may change depending on the number of cells with centrosome amplification. For example, if there are few abnormal cells in the population, it is probably harder to distinguish between models. Conversely, if cells with extra centrioles are more frequent, it is expected that models become easier to distinguish. In addition, the range of parameter values should be taken into account to avoid estimation biases, such that estimated values do not fall close to the bounds.

Discussion

We here combined analytical and statistical methods to characterise abnormal centriole number distributions in populations of human cell lines. Adopting classical mutation-selection balance theory from population genetics, we developed a set of mathematical models for analysing a broad panel of cell lines, which are representative of the diversity along cancer development and between different cancer types. Using a model selection approach, we found that a constant and heavy cost of excess centriole numbers is a common feature of the best approximating models for the majority of cell lines. In addition, we quantified how uncertainty in the model selection and parameter estimation procedures can be ameliorated by obtaining larger sample sizes in the future. We show that integrating statistical information into experimental setups could reveal potential differences between cell lines in the mechanisms that cause abnormal centriole number distributions. Importantly, our population-level approach recognizes and quantifies the variation in centriole numbers that has recently been observed in experimental data.

Dynamics of centriole numbers in proliferating cell populations. To the best of our knowledge, we provide the first quantitative description of centriole number dynamics in populations of proliferating cells. Past studies, both experimental and theoretical, on the population-level response to supernumerary centrioles investigated how wild-type numbers are recovered after perturbation (17, 18)), and also highlighted the role of negative selection in driving this process. These previous studies have mainly focused on distinguishing cells with wild-type centriole/centrosome numbers from cells with abnormal numbers. Here, we explored the full centriole number variation that has been observed in experimental studies. We described the heterogeneity in centriole numbers per cell within and between cell populations, and we evaluated which type of underlying fitness function is most likely to generate the observed variation within the population. Interestingly, our analysis suggests that selection acts strongly against any number of excess centrioles in most cell lines. This means that deleterious effects arise as soon as excess centrioles are produced, whereas the actual number does not seem to matter for selection.

Implications for the biology of centrosomes. Our results show that the centriole number distribution within a population carries important biological information. First, as we argued above, we inferred a constant and heavy cost of abnormal centriole numbers. Whereas understanding variation has long been a staple of evolutionary studies, it has often been overlooked in cell biology. However, on a similar subject, it has been reported that mechanisms different organelle biosynthesis, such as *de novo* assembly, or fusion display specific signatures that can be identified by relating the mean and the variance in their distributions (26). Thus, valuable insight can be gained from a broader quantitative description of the data.

Second, simple models incorporating single-centriole overproduction events and a constant fitness function (i.e. akin to the classical formulation of mutation-selection balance in population genetics) were sufficient to explain the shape of the centriole distribution in a few cell lines, whereas in others a more complex relationship between selection and overproduction improved the fitting. In the latter case, our analysis indicates that the shape of the distribution depends more on how supernumerary centrioles are acquired rather than on how they are eliminated - i.e. model fits are less sensitive to changes in intrinsic growth rates than to the mode of overproduction. This raises the hypothesis that various cellular mechanisms might lead to overproduction whereas selection 'punishes' the presence of any number of excess centrioles.

The biological processes associated with our postulated single-, double- and quadruple-step centriole overproduction events may be entirely different. For example, overproduction of two centrioles could occur due to centriole re-duplication after premature disengagement (27). Quadruple overproduction could be a consequence of cytokinesis failure followed by reduplication of all four centrioles (25).

We identified models with a flat fitness function as the best for most cell lines, suggesting centrosome amplification *per se* is deleterious, regardless of the number of extra centrioles. Intuitively, one could expect that higher number of centrioles would induce stronger selection because the mechanisms that eliminate cells with centrosome amplification seem to involve some form of "counting". For example, one of the main sources of cell death in cells with extra centrioles is multipolar divisions (13, 14). It is likely that it is harder for cells to cluster extra centrioles if there are more of them. Thus, one could expect that the probability that a cell undergoes a multipolar division increases with the number of extra centrioles. Likewise, it has been recently proposed that a molecular complex called the PIDDosome triggers p53-dependent cell cycle arrest by "counting" excess mother centrioles (15). Thus, it is probably easier for the PIDDosome to detect extra centrioles if they occur in larger numbers. Ultimately, it is still not clear how these or other mechanisms respond to the number of extra centrioles and somewhat surprisingly, our results indicate this response should not affect the strength of selection in various cell lines.

We anticipate that future experimental work will address these mechanisms in greater detail, upon which our models can be refined to integrate mechanistic details of overproduction rather than the current general overproduction rates. That could, in turn, allow for a more specific statistical inference of when and how centriole overproduction occurs in different cell lines. Ideally, our modeling and inference approach will eventually link experimental information about centriole distributions with genomic inference of cancer-line and -stage specific molecular alterations.

Limitations of this study. The observation that cells with extra centrioles are relatively rare in the analysed data sets is a major determinant of the uncertainty of our model selection and parameter estimation procedures. However, we showed that a

modest increase in the number of analysed cells can potentially mitigate these issues. In addition, mapping model parameters to the biological system may not be trivial. Importantly, one of the aspects we simplified is centriole segregation, i.e. how many centrioles are inherited by each daughter cell after cell division. Centriole segregation is negligible if selection is sufficiently strong, as cells with extra centrioles would not be expected to produce viable offspring. If that is the case, given that the cells in the analysed data sets are mitotic, all extra centrioles would likely have originated in the preceding interphase. However, we cannot rule out that centriole segregation may have an effect on how centriole number distributions are shaped.
Furthermore, the fact that cells in the data sets are mitotic is convenient for the inference of centriole overproduction and selection parameters it because allowed us to eliminate confounding variables related to the cell cycle. First, it means that centriole number variations along the cell cycle, which would be expected in an asynchronously dividing population, need not be considered. Second, it allowed us to disregard differences in cell cycle progression within and between cell lines. Altogether, characterising centriole number distributions along cell cycle progression is an entirely different problem. From a purely theoretical standpoint, it requires a more detailed implementation of the timing of centriole duplication and the length of each cell cycle phase. This would be desirable in the broader context of characterising centriole number distributions in proliferating cell populations but is beyond the scope of our current analysis.

Towards unraveling the causes and consequences of centriole number changes in cancer. Numeric aberrations of the centrosome and their putative link to cancer formation have long been described (3), although accurate quantifications of centriole numbers in tumor biopsies and cancer-derived cell lines have emerged only recently (9–12) . To this day, the contribution of centrosomal anomalies to cancer development remain controversial, with some studies showing that higher numbers, via Plk4 overexpression, can initiate or aggravate tumorigenesis (28, 29), and others showing that it is not sufficient and may even slow down progression (30, 31). On the other hand, extra centrioles are associated with other cancer hallmarks, such as aneuploidy (14, 32) and invasion (33–35), and often correlate with a more aggressive cancer phenotype (3, 7). The widespread occurrence of centriole number abnormalities in a cancer setting makes them an attractive as prospective biomarkers and as therapeutic targets (36). Thus, understanding the underlying causes of these abnormalities is important for biomedical research.
Here, we adopted evolutionary theory to quantify the variation in centriole numbers within cancer cell populations. That is because ultimately, centriole number abnormalities are highly heterogeneous both within and between cancer cell populations. In order to predict how these abnormalities evolve during cancer development, and how they may interweave with other cancer hallmarks, it is crucial to have a quantitative understanding of how extra centrioles emerge in these cells, and how the cells cope with them. For example, in the case of the Barrett's esophagus progression model, the increase in centriole numbers from the metaplasia to the dysplasia stages can be explained by loss of p53 (9). This can be interpreted as a reduction in the strength of negative selection, since p53 can lead to cell-cycle arrest or cell death in the presence of extra centrioles. If this is true, it would be interesting to quantify how strong the decrease in selective pressure and if it is sufficient, by itself, to account for the shift in centriole number distributions.
We showed that, pending an increase in statistical power, our modelling framework can be used to infer these changes and further our understanding of the relationship between extra centriole numbers and cancer development.

Methods

Experimental data. For our analysis, we considered two recently published data sets. The first data set corresponds to 13 cell lines derived from different tissues in the Barrett's esophagus cancer progression model. These include pre-malignant (metaplasia and dysplasia) and malignant stages (adenocarcinoma and lymph node metastasis). The second data set corresponds to 53 cell lines from the NCI-60 panel, a group of cell lines that spans multiple cancer types (leukemia, melanoma and lung, colon, brain, ovary, breast, prostate, and kidney, cancers), as well as five non-cancerous cell lines. In both cases, the data correspond to centriole number counts in mitotic cells. Out of the 71 cell lines, four were discarded from the analysis: for three of the cell lines, we could not compute the equilibrium expression, due to high i_{max}); for the remaining cell line, no cells with centrosome amplification were recorded, under which conditions the models can be trivially solved by setting μ_1, and/or μ_2, and/or μ_4 to zero. As previously stated, we do not take into account any cell with less than four centrioles (which represent approximately 1.9% of the total). Additional experimental details can be found in the corresponding publications (9, 10).

Mathematical analysis. Equation Eq. (1) describes the rate of change in the equilibrium frequency of the subpopulation of cells containing i centrioles. To obtain the equilibrium solution in equation Eq. (2), we solved $\frac{dP_i}{dt} = 0$ for all i, for $i_{max} = \{3, 4, 5\}$, using Wolfram Mathematica, and proposed a general expression for increasing i_{max}. Then, we verified if the expression was correct by comparing it to the steady-state obtained from numerical integration of equation Eq. (1) (Figure S1). To ensure equation Eq. (2) yields exclusively non-negative values for all i, we added the following constraint:

$$\mu_{i,j} > 0 \quad \wedge \quad r_0 > \sum_{j=0}^{i_{max}} \mu_{0,j} + r_i - \sum_{k=i}^{i_{max}} \mu_{i,k}. \tag{11}$$

Parameter	Range
r	$]-1,1[$
c	$]0,2[$
λ	$]0,2[$
μ_1	$[0,1[$
μ_2	$[0,1[$
μ_4	$[0,1[$

Table 1. Model parameters. The indicated ranges were used as constraints when estimating best fitting parameters

The first term indicates that centriole overproduction rates $\mu_{i,j}$ are strictly positive, otherwise higher centriole numbers would not be reachable. The second term indicates that the intrinsic growth rate of wild-type cells must exceed the sum of the intrinsic growth rates of cells with i extra centrioles and the rate at which their frequency increases as a function of centriole overproduction. Breaking this constraint would lead to the depletion of cells with wild-type centriole numbers, which is not observed in any of the data in our analysis.

Model fitting and selection. To fit the models, we first derived a general (log-)likelihood expression:

$$\ln \mathcal{L}(\theta_M|p_i) = \sum_{i=0}^{i_{max}} p_i \ln \left(P_i^*|\theta_M\right) \tag{12}$$

where θ is the tuple of parameters in model M, P_i is the equilibrium solution derived in equation Eq. (2), p_i is the observed relative frequency of cells containing i centrioles, and i_{max} indicates the subpopulation of cells harboring the maximum observed number of centrioles. For ease of comparison, we assumed i_{max} to be the maximum overall number of extra centrioles per cell in the data (30). We fitted the models by numerical maximization of the log-likelihood function, according to model parameters and the indicated range of values (Table 1).
The Bayesian Information Criterion (BIC) was calculated according to:

$$BIC = \ln(n)\kappa - 2\ln \hat{\mathcal{L}}(\theta_M|p_i), \tag{13}$$

where $\hat{\mathcal{L}}$ is the maximum likelihood estimator for a given model, κ is the number of parameters in the model and n is the sample size (number of cells observed in a given cell line). The BIC accounts for sample size and the number of model parameters, such that more complex models are penalized. When compared to another frequently used model selection criterion, the Akaike Information Criterion (AIC), the added sample size penalty is useful given that the number of sampled cells per cell line is limited. We selected the best model for each cell line by finding the one that minimizes the BIC score. We performed model fitting and selection using Wolfram Mathematica™. The results for model selection based on the empirical distributions alone (as opposed to the bootstrap samples) were confirmed in Python to check for numerical inconsistencies in log-likelihood values.

Bootstrapping. Non-parametric bootstrap samples were generated by drawing data points, with replacement, from the empirical distribution for each cell line. The sample size for each bootstrap distribution was equal to that obtained experimentally. Parametric bootstrap samples were generated by drawing, with replacement, from each predicted distribution. In other words, we fitted the models to the empirical distributions in question, to obtain an expected distribution under the respective models, and sampled thereof.

Statistical analyses and data visualization. After obtaining expected distributions under geometric and Waring-Yule models, using Wolfram Mathematica built-in functions, we calculated the value of the X^2 statistic for each pair of empirical and expected distributions. It should be noted that the X^2 statistic has limited for analysing the data at hand, such that we used it only for comparative purposes.
When testing goodness-of-fit of the most complex candidate models, we first determined the expected distribution under each of the models, for each cell line, and then performed Monte Carlo multinomial tests between the expected distribution and the corresponding empirical distribution. The Monte Carlo step consisted of 10,000,000 simulations under the expected distribution. We determined (approximated) p-values based on the proportions of simulations that yielded more extreme results than the data. The significance level was set to 0.05, and adjusted according to the Bonferroni correction for multiple testing (one for each of the 67 cell lines).
All plots were produced in Wolfram Mathematica™or R.

ACKNOWLEDGEMENTS
This work was supported by EMBO Installation Grant IG4152 and by ERC Starting Grant 804569 – FIT2GO, and FCT (Fundação para a Ciência e Tecnologia) grant PTDC/BIA-BID/32225/2017. M. A. Dias Louro was funded by FCT research fellowship PD/BD/139217/2018. We would like to acknowledge Carla A. M. Lopes and Gaëlle

Marteil for in-depth discussions on their work with the Barrett's esophagus and NCI-60 cell lines, and Telmo Cunha for his work on a preliminary version of the models. We thank all members of the Bank and Bettencourt-Dias labs for their insights and critical reading of this manuscript. We would like to acknowledge various participants of the "From Molecular Basis to Predictability and Control of Evolution" workshop at the Nordita Institute in Sweden. We would like to thank Ricardo Henriques for distributing a nice Overleaf template, which we adapted for this manuscript.

Bibliography

1. Juliette Azimzadeh and Wallace F. Marshall. Building the centriole. *Current Biology*, 20(18):R816–R825, 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.08.010.
2. Daniela A. Brito, Susana Montenegro Gouveia, and Mónica Bettencourt-Dias. Deconstructing the centriole: Structure and number control. *Current Opinion in Cell Biology*, 24(1):4–13, 2012. ISSN 09550674. doi: 10.1016/j.ceb.2012.01.003.
3. S. A. Godinho and D. Pellman. Causes and consequences of centrosome abnormalities in cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1650), 2014. ISSN 14712970. doi: 10.1098/rstb.2013.0467.
4. Mónica Bettencourt-Dias, Friedhelm Hildebrandt, David Pellman, Geoff Woods, and Susana A. Godinho. Centrosomes and cilia in human disease. *Trends in Genetics*, 27(8):307–315, 2011. ISSN 01689525. doi: 10.1016/j.tig.2011.05.004.
5. Pierre Gönczy. Centrosomes and cancer: Revisiting a long-standing relationship. *Nature Reviews Cancer*, 15(11):639–652, 2015. ISSN 14741768. doi: 10.1038/nrc3995.
6. Niccolò Banterle and Pierre Gönczy. Centriole Biogenesis: From Identifying the Characters to Understanding the Plot. *Annual Review of Cell and Developmental Biology*, 33(1):23–49, 2017. ISSN 1081-0706. doi: 10.1146/annurev-cellbio-100616-060454.
7. Erich A. Nigg and Andrew J. Holland. Once and only once: Mechanisms of centriole duplication and their deregulation in diseases. *Nature Reviews Molecular Cell Biology*, 19(5):297–312, 2018. ISSN 14710080. doi: 10.1038/nrm.2017.127.
8. J.Y. Chan. A Clinical Overview of Centrosome Amplification in Human Cancers. *International journal of biological sciences*, 7(8):1122, 2011.
9. Carla A.M. Lopes, Marta Mesquita, Ana Isabel Cunha, Joana Cardoso, Sara Carapeta, Cátia Laranjeira, António E. Pinto, José B. Pereira-Leal, António Dias-Pereira, Mónica Bettencourt-Dias, and Paula Chaves. Centrosome amplification arises before neoplasia and increases upon p53 loss in tumorigenesis. *Journal of Cell Biology*, 217(7):2353–2363, 2018. ISSN 15408140. doi: 10.1083/jcb.201711191.
10. Gaëlle Marteil, Adan Guerrero, André F. Vieira, Bernardo P. De Almeida, Pedro Machado, Susana Mendonça, Marta Mesquita, Beth Villarreal, Irina Fonseca, Maria E. Francia, Katharina Dores, Nuno P. Martins, Swadhin C. Jana, Erin M. Tranfield, Nuno L. Barbosa-Morais, Joana Paredes, David Pellman, Susana A. Godinho, and Mónica Bettencourt-Dias. Over-elongation of centrioles in cancer promotes centriole amplification and chromosome missegregation. *Nature Communications*, 9(1), 2018. ISSN 20411723. doi: 10.1038/s41467-018-03641-x.
11. Mengdie Wang, Beatrice S. Knudsen, Raymond B. Nagle, Gregory C. Rogers, and Anne E. Cress. A method of quantifying centrosomes at the single-cell level in human normal and cancer tissue. *Molecular Biology of the Cell*, 30(7):811–819, 2019. ISSN 19394586. doi: 10.1091/mbc.E18-10-0651.
12. David K. Breslow and Andrew J. Holland. Mechanism and Regulation of Centriole and Cilium Biogenesis. *Annual Review of Biochemistry*, 88(1):691–724, 2019. ISSN 0066-4154. doi: 10.1146/annurev-biochem-013118-111153.
13. William T. Silkworth, Isaac K. Nardi, Lindsey M. Scholl, and Daniela Cimini. Multipolar spindle pole coalescence is a major source of kinetochore mis-attachment and chromosome mis-segregation in cancer cells. *PLoS ONE*, 4(8), aug 2009. ISSN 19326203. doi: 10.1371/journal.pone.0006564.
14. Neil J. Ganem, Susana A. Godinho, and David Pellman. A mechanism linking extra centrosomes to chromosomal instability. *Nature*, 460(7252):278–282, 2009. ISSN 00280836. doi: 10.1038/nature08136.
15. Luca L. Fava, Fabian Schuler, Valentina Sladky, Manuel D. Haschka, Claudia Soratroi, Lisa Eiterer, Egon Demetz, Guenter Weiss, Stephan Geley, Erich A. Nigg, and Andreas Villunger. The PIDDosome activates p53 in response to supernumerary centrosomes. *Genes and Development*, 31(1):34–45, 2017. ISSN 15495477. doi: 10.1101/gad.289728.116.
16. Marco Raffaele Cosenza and Alwin Krämer. Centrosome amplification, chromosomal instability and cancer: mechanistic, clinical and therapeutic issues. *Chromosome Research*, 24(1):105–126, 2016. ISSN 15736849. doi: 10.1007/s10577-015-9505-5.
17. Bramwell G. Lambrus, Yumi Uetake, Kevin M. Clutario, Vikas Daggubati, Michael Snyder, Greenfield Sluder, and Andrew J. Holland. P53 protects against genome instability following centriole duplication failure. *Journal of Cell Biology*, 210(1):63–77, 2015. ISSN 15408140. doi: 10.1083/jcb.201502089.
18. Yao Liang Wong, John V. Anzola, Robert L. Davis, Michelle Yoon, Amir Motamedi, Ashley Kroll, Chanmee P. Seo, Judy E. Hsia, Sun K. Kim, Jennifer W. Mitchell, Brian J. Mitchell, Arshad Desai, Timothy C. Gahman, Andrew K. Shiau, and Karen Oegema. Reversible centriole depletion with an inhibitor of Polo-like kinase 4. *Science*, 348(6239):1155–1160, 2015. ISSN 10959203. doi: 10.1126/science.aaa5111.
19. Nicolaas C Baudoin, Kimberly Soto, Olga Martin, Joshua M Nicholson, Jing Chen, and Daniela Cimini. Asymmetric clustering of centrosomes defines the early evolution of tetraploid cells. *bioRxiv*, page 526731, 2019. doi: 10.1101/526731.
20. W. Hill. *The Mathematical Theory of Selection, Recombination and Mutation. R. Burger*, volume 79. Wiley, Chichester, 2002.
21. M. Bettencourt-Dias, A. Rodrigues-Martins, L. Carpenter, M. Riparbelli, L. Lehmann, M. K. Gatt, N. Carmo, F. Balloux, G. Callaini, and D. M. Glover. SAK/PLK4 is required for centriole duplication and flagella development. *Current Biology*, 15(24):2199–2207, 2005. ISSN 09609822. doi: 10.1016/j.cub.2005.11.042.
22. Carla A M Lopes, Swadhin Chandra Jana, Inês Cunha-Ferreira, Sihem Zitouni, Inês Bento, Paulo Duarte, Samuel Gilberto, Francisco Freixo, Adán Guerrero, Maria Francia, Mariana Lince-Faria, Jorge Carneiro, and Mónica Bettencourt-Dias. PLK4 trans-Autoactivation Controls Centriole Biogenesis in Space. *Developmental Cell*, 35(2):222–235, 2015. ISSN 18781551. doi: 10.1016/j.devcel.2015.09.020.
23. C. Arquint, K. F. Sonnen, Y.-D. Stierhof, and E. A. Nigg. Cell-cycle-regulated expression of STIL controls centriole number in human cells. *Journal of Cell Science*, 125(5):1342–1352, feb 2012. ISSN 0021-9533. doi: 10.1242/jcs.099887.
24. Petr Strnad, Sebastian Leidel, Tatiana Vinogradova, Ursula Euteneuer, Alexey Khodjakov, and Pierre Gönczy. Regulated HsSAS-6 Levels Ensure Formation of a Single Procentriole per Centriole during the Centrosome Duplication Cycle. *Developmental Cell*, 13(2):203–213, aug 2007. ISSN 15345807. doi: 10.1016/j.devcel.2007.07.004.
25. Susanne M.A. Lens and René H. Medema. Cytokinesis defects and cancer. *Nature Reviews Cancer*, 19(1):32–45, 2019. ISSN 14741768. doi: 10.1038/s41568-018-0084-6.
26. Shankar Mukherji and Erin K. O'Shea. Mechanisms of organelle biogenesis govern stochastic fluctuations in organelle abundance. *eLife*, 2014(3), jun 2014. ISSN 2050084X. doi: 10.7554/eLife.02678.001.
27. Jadranka Lončarek, Polla Hergert, and Alexey Khodjakov. Centriole reduplication during prolonged interphase requires procentriole maturation governed by plk1. *Current Biology*, 20(14):1277–1282, 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.05.050.
28. Özdemirhan Serçin, Jean Christophe Larsimont, Andrea E. Karambelas, Veronique Marthiens, Virginie Moers, Bram Boeckx, Marie Le Mercier, Diether Lambrechts, Renata Basto, and Cédric Blanpain. Transient PLK4 overexpression accelerates tumorigenesis in p53-deficient epidermis. *Nature Cell Biology*, 18(1):100–110, 2016. ISSN 14764679. doi: 10.1038/ncb3270.
29. Michelle S. Levine, Bjorn Bakker, Bram Boeckx, Julia Moyett, James Lu, Benjamin Vitre, Diana C. Spierings, Peter M. Lansdorp, Don W. Cleveland, Diether Lambrechts, Floris Foijer, and Andrew J. Holland. Centrosome Amplification Is Sufficient to Promote Spontaneous Tumorigenesis in Mammals. *Developmental Cell*, 40(3):313–322.e5, 2017. ISSN 18781551. doi: 10.1016/j.devcel.2016.12.022.
30. Benjamin Vitrea, Andrew J. Holland, Anita Kulukian, Ofer Shoshani, Maretoshi Hirai, Yin Wanga, Marcus Maldonado, Thomas Cho, Jihane Boubaker, Deborah A. Swing, Lino Tessarollo, Sylvia M. Evans, Elaine Fuchs, and Don W. Cleveland. Chronic centrosome amplification without tumorigenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 112(46):E6321–E6330, nov 2015. ISSN 10916490. doi: 10.1073/pnas.1519388112.
31. Jean-Philippe Morretton, Aurélie Herbette, Camille Cosson, Bassirou Mboup, Aurélien Latouche, Pierre Gestraud, Tatiana Popova, Marc-Henri Stern, Fariba Nemati, Didier Decaudin, Guillaume Bataillon, Véronique Becette, Didier Meseure, André Nicolas, Odette Mariani, Claire Bonneau, Jorge Barbazan, Anne Vincent-Salomon, Fatima Mechta-Grigoriou, Sergio Roman Roman, Roman Rouzier, Xavier Sastre-Garau, Oumou Goundiam, and Renata Basto. Centrosome amplification favours survival and impairs ovarian cancer progression. *bioRxiv*, page 623983, 2019. doi: 10.1101/623983.
32. Mijung Kwon, Susana A. Godinho, Namrata S. Chandhok, Neil J. Ganem, Ammar Azioune, Manuel Thery, and David Pellman. Mechanisms to suppress multipolar divisions in cancer cells with extra centrosomes. *Genes and Development*, 22(16):2189–2203, 2008. ISSN 08909369. doi: 10.1101/gad.1700908.
33. Susana A. Godinho, Remigio Picone, Mithila Burute, Regina Dagher, Ying Su, Cheuk T. Leung, Kornelia Polyak, Joan S. Brugge, Manuel Théry, and David Pellman. Oncogene-like induction of cellular invasion from centrosome amplification. *Nature*, 510(7503):167–171, 2014. ISSN 14764687. doi: 10.1038/nature13277.
34. Teresa Arnandis, Pedro Monteiro, Sophie D. Adams, Victoria Louise Bridgeman, Vinothini Rajeeve, Emanuela Gadaleta, Jacek Marzec, Claude Chelala, Ilaria Malanchi, Pedro R. Cutillas, and Susana A. Godinho. Oxidative Stress in Cells with Extra Centrosomes Drives Non-Cell-Autonomous Invasion. *Developmental Cell*, 47(4):409–424.e9, 2018. ISSN 18781551. doi: 10.1016/j.devcel.2018.10.026.
35. Gina M. LoMastro and Andrew J. Holland. The Emerging Link between Centrosome Aberrations and Metastasis. *Developmental Cell*, 49(3):325–331, 2019. ISSN 18781551. doi: 10.1016/j.devcel.2019.04.002.
36. Ryan A. Denu, Lauren M. Zasadil, Craig Kanugh, Jennifer Laffin, Beth A. Weaver, and Mark E. Burkard. Centrosome amplification induces high grade features and is prognostic of worse outcomes in breast cancer. *BMC Cancer*, 16(1), 2016. ISSN 14712407. doi: 10.1186/s12885-016-2083-x.

Supplementary Note 1: Figures

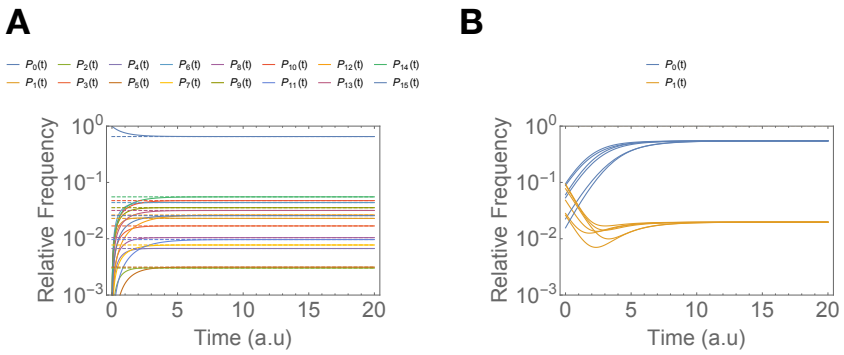

Fig. S1. The solutions to the system of ordinary differential equations converge to the expected equilibrium value. A - Comparison between numeric integration of the general model and the corresponding equilibrium expression Eq. (2), evaluated at the same parameter values. We assumed $i_{max} = 15$ and generated pseudo-random parameter values for all r_i and $\mu_{i,j}$. B - Comparison between numeric integration of the general model, from different initial conditions and the corresponding equilibrium expression. We generated a set of pseudo-random parameter values for all r_i and $\mu_{i,j}$ and initial conditions. Note that the y-axis is in log-scale.

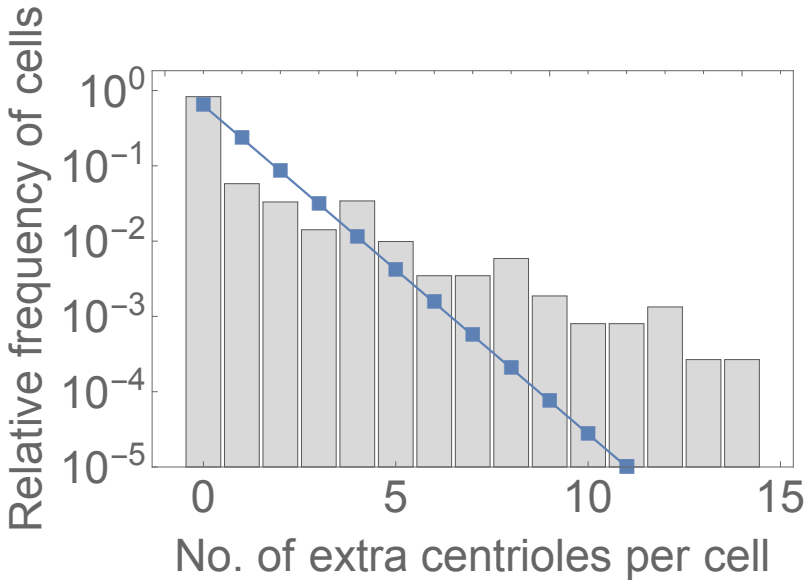

Fig. S2. Cells with high number of centrioles occur frequently across all data sets. Best fitting geometric distribution (blue) to the pooled distribution of centriole numbers in all sampled populations. Number of sampled cells: n=3746.

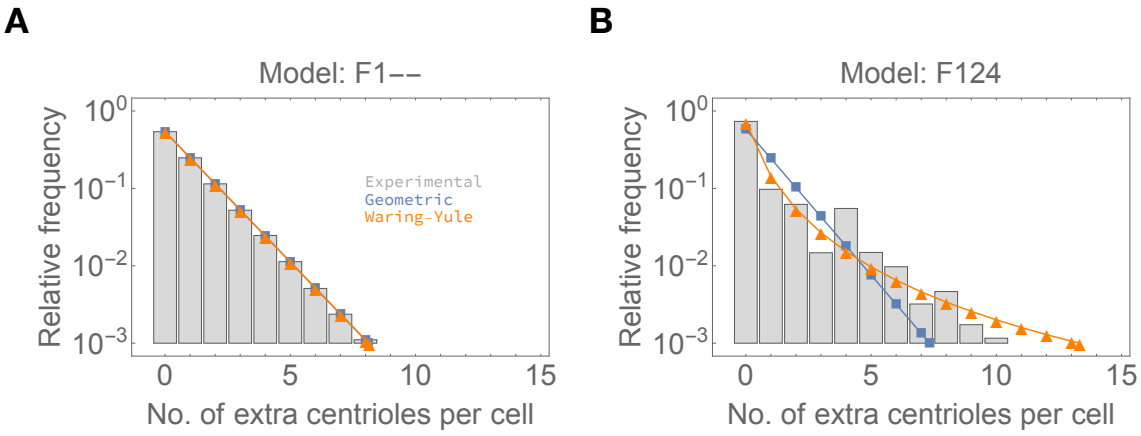

Fig. S3. Models in the candidate set can produce both geometric- and heavy-tailed distributions.. Best fitting geometric (blue) and Waring-Yule (orange) distributions to 1,000,000 simulated data points (relative frequencies indicated as grey bars) from (a) model F1-- ($r = -0.3$, $\mu_1 = 0.6$) and (b) model F124 ($r = -0.5$, $\mu_1 = 0.2$, $\mu_2 = 0.1$, $\mu_4 = 0.1$). Note that the y-axis is in log-scale

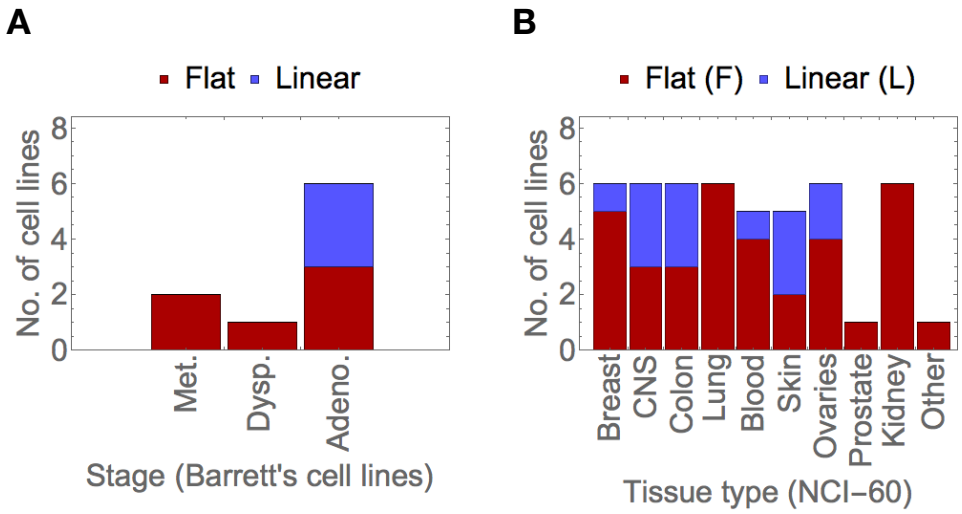

Fig. S4. Best fitness functions per tissue type. Models sharing the fitness function of the best model for each cell line in the data sets. A - Barrett's esophagus data set, grouped by developmental stage. B - NCI-60 data set, grouped by tissue type (including cancer and non-cancer cell lines). "Other" refers to a RPE cell line that was used as a control.

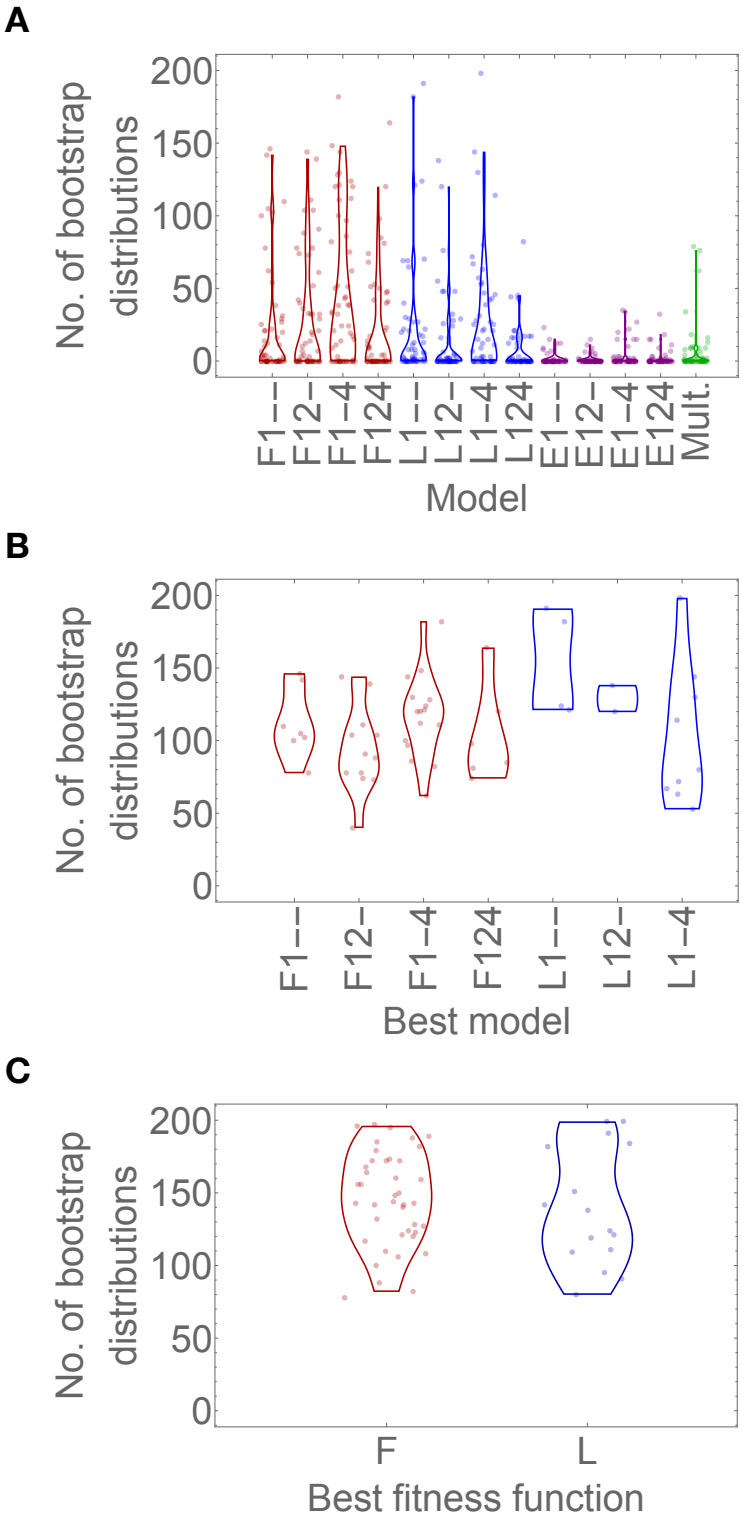

Fig. S5. Bootstrap support of the candidate models. A - Number of bootstrap distributions (out of the 200 generated for each cell line) explained by each model. We obtained indistinguishable BIC scores for multiple models in 381 bootstrap distributions spread across 26 different cell lines ("Mult."). C - Number of bootstrap distributions explained by models sharing the same fitness function as the best model for each cell line. The fitness function of the models is indicated in red - flat; blue - linear; purple - power-law; green - multiple models.

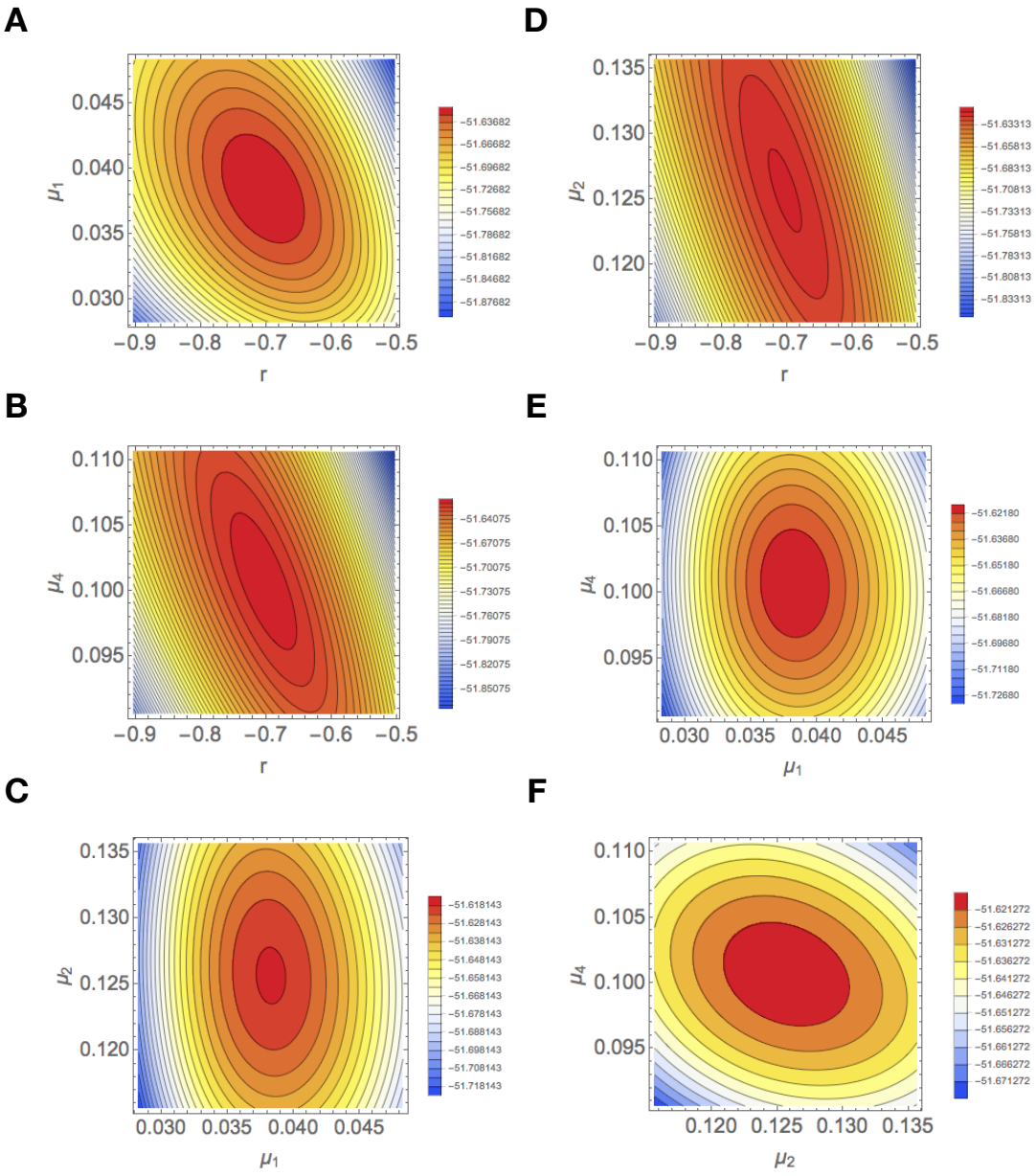

Fig. S6. Model parameters are not correlated and the likelihood value is less sensitive to r. We fitted the model to a random empirical distribution from the analysed data sets and calculated the likelihood values centered around the maximum as a function pairwise combinations of parameter values. A - r and μ_1; B - *r* and μ_2; C - *r* and μ_4; D - μ_1 and μ_2; E - μ_1 and μ_4; F - μ_2 and μ_4. Note scales for r is different the centriole overproduction parameters.

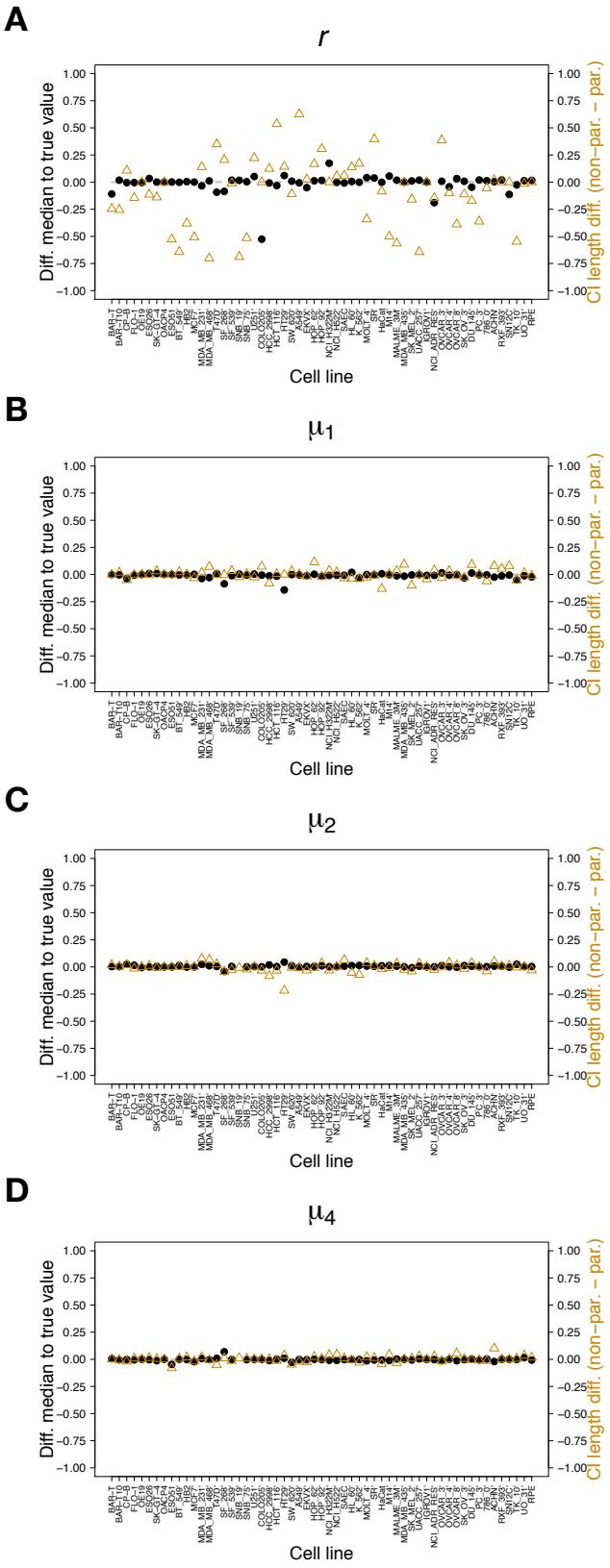

Fig. S7. Parameter estimation is accurate and estimation errors are likely due to small sample sizes. Difference between the median of the parametric bootstrap distribution for (A) r, (B) μ_1, (C) μ_2, and (D) μ_4, and the input value for the simulated data (in black) and difference between the confidence interval length of the non-parametric and parametric bootstrap distributions (in yellow).