Published in final edited form as:
Nat Biotechnol. ; 29(12): 1109–1113. doi:10.1038/nbt.2049.

High-order chromatin architecture determines the landscape of chromosomal alterations in cancer

Geoff Fudenberg[1], **Gad Getz**[2], **Matthew Meyerson**[2,3,4,5], and **Leonid Mirny**[2,6,7]

[1]Harvard University, Program in Biophysics, Boston, Massachusetts

[2]The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[3]Harvard Medical School, Boston, MA 02115, USA

[4]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA

[5]Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02115, USA

[6]Harvard-MIT, Division of Health Sciences and Technology

[7]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA

The rapid growth of cancer genome structural information provides an opportunity for a better understanding of the mutational mechanisms of genomic alterations in cancer and the forces of selection that act upon them. Here we test the evidence for two major forces, spatial chromosome structure and purifying (or negative) selection, that shape the landscape of somatic copy-number alterations (SCNAs) in cancer[1]. Using a maximum likelihood framework we compare SCNA maps and three-dimensional genome architecture as determined by genome-wide chromosome conformation capture (HiC) and described by the proposed fractal-globule (FG) model[2,3]. This analysis provides evidence that the distribution of chromosomal alterations in cancer is spatially related to three-dimensional genomic architecture and additionally suggests that purifying selection as well as positive selection shapes the landscape of SCNAs during somatic evolution of cancer cells.

Somatic copy-number alterations (SCNAs) are among the most common genomic alterations observed in cancer, and recurrent alterations have been successfully used to implicate cancer-causing genes[1]. Effectively finding cancer-causing genes using a genome-wide approach relies on our understanding of how new genome alterations are generated during the somatic evolution of cancer[4–7]. As such, we test the hypothesis that three-dimensional chromatin organization and spatial co-localization influences the set of somatic copy-number alterations observed in cancer (Fig. 1A, recently suggested by cancer genomic data in a study of prostate cancer[8]. Spatial proximity and chromosomal rearrangements are discussed more generally[9–12]). Unequivocally establishing a genome-wide connection between SCNAs and three-dimensional chromatin organization in cancer has until now been limited by our ability to characterize three-dimensional chromatin architecture, and the resolution with which we are able to observe SCNAs in cancer. Here, we ask whether the "landscape" of SCNAs across cancers[1] can be understood with respect to spatial contacts in a 3D chromatin architecture as determined by the recently developed HiC method for high-throughput chromosome conformation capture[2] or described theoretically via the fractal globule (FG) model (theoretical concepts[13], review[3]). Specifically, we investigate the model presented in Figure 1A, and test whether distant genomic loci that are brought spatially close

Correspondence to: Leonid Mirny.

by 3D chromatin architecture during interphase are more likely to undergo structural alterations and become end-points for amplifications or deletions observed in cancer.

Towards this end, we examine the statistical properties of SCNAs in light of spatial chromatin contacts in the context of cancer as an evolutionary process. During the somatic evolution of cancer[14,15] as in other evolutionary processes, two forces determine the accumulation of genomic changes (Fig. 1A): generation of new mutations and fixation of these mutations in a population. The rate at which new SCNAs are generated may vary depending upon the genetic, epigenetic, and cellular context. After an SCNA occurs, it proceeds probabilistically towards fixation or loss according to its impact upon cellular fitness. The fixation probability of an SCNA in cancer depends upon the competition between positive selection if the SCNA provides the cancer cell with a fitness advantage, and purifying (i.e. negative) selection if the SCNA has a deleterious effect on the cell. The probability of observing a particular SCNA thus depends upon its rate of occurrence via mutation, and the selective advantage or disadvantage conferred by the alteration (Fig. 1A). Positive, neutral, and purifying selection are all evident in cancer genomes[16].

Our statistical analysis of SCNAs argues that both contact probability due to chromosomal organization at interphase and purifying selection contribute to the observed spectrum of SCNAs in cancer. From the full set of reported SCNAs across 3,131 cancer specimens in[1], we selected 39,568 intra-arm SCNAs (26,022 amplifications and 13,546 deletions) longer than a megabase for statistical analysis, excluding SCNAs which start or end in centromeres or telomeres. To establish that our results were robust to positive selection acting on cancer-associated genes, we analyzed a collection of 24,301 SCNAs that do not span highly-recurrent SCNA regions (16,521 amplifications and 7,789 deletions, respectively 63% and 58% of the full set[1], see Methods). We present results for the less-recurrent SCNAs, and note that our findings are robust to the subset of chosen SCNAs. We performed our analysis by considering various models of chromosomal organization and purifying selection, which were used to calculate the likelihood of the observed SCNA given the model. The likelihood framework was then used to discriminate between competing models. Statistical significance was further evaluated using permutation tests. The strong association we find between SCNAs and high-order chromosomal structure is not only consistent with the current understanding of the mechanisms of SCNA initiation[17], but provides insight into how spatial proximity may be arrived at via chromosomal architecture and the significance of chromosomal architecture for patterns of SCNAs observed at a genomic scale.

Results

Patterns of three-dimensional chromatin architecture are evident in the landscape of SCNAs

The initial motivation for our study was an observation that the length of focal SCNAs and the length of chromosomal loops (i.e. intra-chromosomal contacts) have similar distributions (Figs. 1B and 1C), both exhibiting ~ $1/L$ scaling. Analysis of HiC data for human cells showed that the mean contact probability over all pairs of loci a distance L apart on a chromosome goes as $P^{HiC}(L) \sim 1/L$ for a range of distances L= 0.5 to 7Mb[2]. This scaling for mean contact probability was shown to be consistent with a fractal globule (FG) model of chromatin architecture. Similarly, the mean probability to observe a focal SCNA of length L goes approximately as $P^{SCNA}(L) \sim 1/L$ for the same range of distances L= 0.5 to 10 Mb as noted in [1]. Mathematically, the observation that the mean probability to observe an SCNA decays with its length is quite significant. If two SCNA ends were chosen randomly within a chromosome arm, the mean probability to observe an SCNA of length L would remain constant. Positive selection, which tends to amplify oncogenes or delete tumor suppressors,

again does not give rise to a distribution whose mean decreases with length. Either purifying selection or a length-dependent mutational mechanism is required to observe this result.

The connection between three-dimensional genomic architecture and SCNA structure goes beyond the similarity of their length distributions: loci that have higher probability of chromosomal contacts are also more likely to serve as SCNA end points (Fig. 2). To quantitatively determine the relationship between three-dimensional genomic architecture and SCNA, both data sets were converted into the same form. For each chromosome, we represent HiC data as a matrix of counts of spatial contacts between genomic locations i and j as determined in the GM06990 cell line using a fixed bin size of 1 Mb^2. Similarly, we construct SCNA matrices by counting the number of amplifications or deletions that start at genomic location i and end at location j of the same chromosomes across the 3,131 tumors. Figure 2 presents HiC and SCNA matrices (heatmaps) for chromosome 17. Away from centromeric and telomeric regions, which are not considered in this analysis, the SCNA heatmap appears similar to the HiC heatmap (Pearson's $r = .55$, $p < 0.001$, see Supplementary Table S1 for other chromosomes). In particular, regions enriched for 3D interactions also appear to experience frequent SCNA. Since the Pearson correlation coefficient is not suited for describing rare probabilistic events like SCNAs, for further analysis we employ the Poisson likelihood, a widely-used method to statistically analyze rare events[18].

Likelihood analysis demonstrates that observed SCNAs are fit best by fractal globular chromatin architecture, and all fits are improved when purifying selection is considered

To further test the role of chromosome organization for the generation of SCNA, we developed a series of statistical models of possible SCNA-generating processes, computed the Poisson likelihoods of the SCNA data given these models (see Eq. 6), and performed model selection using their Bayesian Information Criterion (BIC) values, which is the log-likelihood of a given model penalized by its number of fitting parameters (see Eq. 7). Considered models take into account different mechanisms of the generation of SCNA, with a mutation rate either: uniform in length (*Uniform*), derived from experimentally determined chromatin contact probabilities (*HiC*) or derived from contact probability in the fractal globular chromatin architecture (*FG*). We note that the *FG* model specifies a contact probability that depends on the distance between genomic loci, but does not include positional differences at a given distance.

We also took into account possible deleterious effects of SCNAs due to purifying selection, which can lead to a reduced probability of fixation (see Eq. 1). Deleterious effects of SCNAs on cellular fitness may arise from the disruption of genes or regulatory regions; as such, we expect longer SCNAs to be more deleterious. A relationship between SCNA length and its deleterious effect on cellular fitness is supported by the observation that whole-arm SCNAs are less likely for longer chromosomal arms[1], as well as an observation of linearly decreasing bacterial growth rate with longer amplifications[19]. If we assume that the deleterious effect of an SCNA increases linearly with its length L, and consider the somatic evolution of cancer as a Moran process[15,20], we find that the probability of fixation decays roughly exponentially with length at a rate that reflects the strength of purifying selection (see Eq. 4, Fig. 1B). Combining the effects of purifying selection on fixation probability with the mutational models leads to the following six models: *Uniform*, $Uniform^{+sel}$, *HiC*, HiC^{+sel}, *FG*, FG^{+sel}, with no fitting parameters for models without selection and a single fitting parameter for selection, where the additional parameter is penalized via BIC.

Model selection provides two major results (Fig. 3): First, among models of SCNA generation, a model that follows the chromosomal contact probability of the fractal globule ($\sim 1/L$) significantly outperforms other models. Second, since considering purifying

selection helps fit the observed roll-over in the number of SCNAs at longer distances (L >20Mb, Fig. 1B), every model is significantly improved when purifying selection is taken into account ($p < .001$ via bootstrapping), suggesting that SCNAs experience purifying selection. We note that the additional decline in the number of SCNA at long distances could possibly be due to alternative chromatin-independent mechanisms that further disfavor the formation of exceptionally long SCNAs. Figure 3 presents log-likelihood ratios of the models (with and without purifying selection) with respect to the uniform model. If models are fit on a chromosome-by-chromosome basis (Supplementary Fig. 2) we observe that for long chromosomes, the *FG* model fits better than purifying selection alone. We also find that the best-fit parameter describing purifying selection is proportional to chromosome length (Supplementary Fig. 2A). Since smaller values for the best-fit parameter correspond to stronger purifying selection, these two results suggest that short, gene-rich, chromosomes may experience greater purifying selection. However, we note that purifying selection proportional to the genomic length of an SCNA fits the data better than purifying selection proportional to the number of genes affected by an SCNA (Supplementary Fig. 3).

Permutation analysis supports the connection between SCNAs and experimentally determined three-dimensional chromatin architecture

We next tested whether the position-specific structure of chromosomal contacts observed in experimental HiC data, and absent for the *FG*, was evident in the SCNA landscape. The test was performed using permutation analysis (Fig. 4). Since both the probability of observing an SCNA with a given length and intra-chromosomal contact probability in HiC depend strongly on distance L, we permuted SCNAs in a way that preserves this dependence but destroys the remaining fine structure. This is achieved by randomly reassigning SCNA starting locations within the same chromosomal arm, while keeping their lengths fixed. We find that HiC fits the observed SCNAs much better than it fits permuted SCNAs (Fig. 4, $p<.001$). Similar analysis within individual chromosomes shows that the fit is better for 18 of the 22 autosomal chromosomes, except for chromosomes 10, 11, 16, and 19, and is significantly better ($p<.01$) for nine chromosomes 1, 2, 4, 5, 7, 8, 13, 14, and 17 (Fig. 4B). While the observed amplification and deletions each separately fit better on average than their permuted counterparts (Supplementary Fig. 5), deletions fit considerably better than amplifications ($p<0.001$ vs. $p<0.05$).

Finally, we examined the possible influence of chromosomal compartments (domains, as determined in[2]) on the landscape of SCNAs by fitting models where SCNA formation is favored if both ends are in the same type of domain (see Methods). Maximizing the likelihood of this two-parameter $FG^{+\text{domains}}$ model demonstrated a marginal increase in the BIC-corrected likelihood above the *FG* model for deletions, and not for amplifications (Supplemental Fig. 8 and 9). The best-fitting domain strength parameter values favored small (10–20%) increases in the relative probability of intra-domain SCNAs. Additionally, the best-fitting $FG^{+\text{domains}}$ model shows a smaller amount of position-specific information than HiC, as determined by permutation tests (Supplemental Fig. 8).

Discussion

Our genome-wide analysis of HiC measurements and cancer SCNA finds multiple connections between higher-order genome architecture and re-arrangements in cancer. Using an incisive likelihood-based BIC framework, we found that: (1) probability of a 3D contact between two loci based on the *FG* model explains the length distribution of SCNA better than other mechanistic models or than a model of purifying selection alone; (2) comparisons with permuted data demonstrate the significant connection between megabase-level position-specific 3D chromatin structure observed in HiC and SCNA; (3) a multiplicative model favoring intra-domain SCNAs provides little improvement beyond the

FG model and has less position-specific information than HiC; (4) SCNA data reflect mutational mechanisms and purifying selection, in addition to commonly considered positive selection.

These results argue strongly for the importance of 3D chromatin organization in the formation of chromosomal alterations. While the distribution of SCNAs could conceivably depend on a complicated mutation and selection landscape, which is merely correlated with 3D genomic structure, a direct explanation via 3D genomic contacts is more parsimonious. Along these lines, two recent experimental studies of translocations suggest that physical proximity is among the key determinants of genomic rearrangements[21,22]. Additionally, a genome-wide analysis of translocations across cancers demonstrates an enrichment of translocations among chromosomal loci with greater numbers of experimentally determined chromosomal contacts[23].

Genomic architecture may vary with cancer cell type of origin and the specific chromatin states of these cells[24,25], thus influencing the set of observed SCNAs in each cancer type; for example, re-arrangement breakpoints in prostate cancer were found to correlate with loci in specific chromatin states of prostate epithelial cells[8]. In fact, if HiC data matching the tumor cell-types of origin for the set of observed SCNAs becomes available, we may find that the cell-type specific experimental 3D contacts fit the observed distribution of SCNAs better than the fractal globule model. Despite this limitation, when we perform a permutation analysis on SCNAs grouped by cancer lineage (epithelial, hematopoietic, sarcomas and neural), we still find that HiC fits the observed SCNAs significantly better than it fits permuted SCNAs consistently across cancer lineages for deletions, but not for amplifications (Supplementary Fig. 6).

Differences between amplifications and deletions (Supplementary Figs. 4, 5, 6) may reflect differences in the strength of selection and mechanisms of genomic alteration: conceivably a simple loss of a chromosomal loop could lead to a deletion, while amplifications may occur through more complicated processes[17] and may require interactions with homologous and non-homologous chromosomes that are not necessarily directly related to intra-chromosomal spatial proximity during interphase.

Our results suggest that a comprehensive understanding of mutational and selective forces acting on the cancer genome, not limited to positive selection of cancer-associated genes, is important for explaining the observed distribution of SCNAs. Furthermore, comparing model goodness-of-fits for the distribution of SCNAs argues that purifying selection is a common phenomenon, and that many SCNAs in cancer may be mildly deleterious "passenger mutations" (reviewed in[26,27]). We note that while we find evidence for both chromatin organization and purifying selection in the length distribution of SCNAs, in our best-fitting model, 3D chromatin architecture explains a factor of ~100 in relative frequencies of SCNAs, whereas purifying selection contributes an additional factor of ~3 for long SCNAs (L > 20–100Mb) and has little effect on the frequency of shorter SCNAs (L < 20Mb). Presumably, mechanisms other than purifying selection could lead to additional suppression of excessively long SCNAs. However, the observed exponential rollover in the number of SCNAs at long distances is unlikely to be caused by limitations arising from SCNA mapping, since whole-arm SCNAs are successfully detected at high frequencies.

The sensitivity and relevance of comparative genomic approaches to chromosome rearrangements can only increase as additional HiC-type datasets become available. Future studies will be able to address the importance of different 3D structures to the observed chromosomal rearrangements across cell types and cell states. Perhaps even more importantly, cancer genomic sequencing data will allow for significantly more detailed

analyses than the current array-based approaches, allowing for greater mechanistic insight into SCNA formation. In particular, high-throughput whole-genome sequencing data will allow for both a high-resolution analysis of interchromosomal rearrangements and yield insight into the interplay between sequence features, chromatin modifications, and 3D genomic structure.

Methods

Constructing heatmaps

We generated SCNA heatmaps from the data of Beroukhim et al.[1] who reported a total of 75,700 amplification and 55,101 deletion events across 3,131 cancer specimens; reported events are those with inferred copy number changes >.1 or <!.1, due to experimental limitations. We restricted our analysis to intra-arm SCNAs which do not start/end near telomeric/centromeric regions separated by more than one megabase bin, giving a set of 39,568 SCNAs (26,022 amplifications and 13,546 deletions). We note that SCNAs starting/ending in centromeres/telomeres (which include full-arm gain/loss) display a very different pattern of occurrence from other focal SCNAs, particularly in terms of their length distribution, which may indicate a different mutational mechanism. Requiring a separation of greater than one megabase bin is due to resolution limits of both SCNA and HiC data (see Supplementary Fig. 1 for details). SCNA matrices are constructed by counting the number of amplifications or deletions starting at Mb i and ending at Mb j of the same chromosomes. Similarly, HiC heatmaps were generated by counting the number of reported interactions[2] between Mb i and j of the same chromosome in human cell line GM06690.

Mutational and Evolutionary Models of SCNA

To test the respective contributions of mutational and selective forces on the distribution of SCNAs, we consider the probability of observing at SCNA that starts and end at i and j

$$P_{ij}=\mu_{ij}\cdot\pi(L) \quad (1)$$

as the product of the probability of a mutation, i.e. an SCNA to occur in a single cell μ_{ij}, and the probability to have this mutation fixed in the population of cancer cells $\pi(L)$, where $L=|i-j|$ is the SCNA length. The mutation probability μ_{ij} depends on the model that describes the process leading to chromosomal alterations: (*Uniform*) two ends of an alteration are drawn randomly from the same chromosomal arm, giving $\mu_{ij}^{Uniform}=const$; (*HiC*) the probability of an alteration depends on the probability of a 3D contact between the ends as given by HiC data, $\mu_{ij}^{HiC}\sim P_{ij}^{HiC}$; (*FG*) the probability of alteration depends upon the probability of 3D contact according to the fractal globule model, i.e. on SCNA length L: $\mu_{ij}^{FG}=\mu^{FG}(L)\sim 1/L$. The probability of fixation depends on the fitness of a mutated cell as compared to non-mutated cells (see below). Each mutational model is considered by itself and in combination with purifying selection, giving six models: *Uniform, HiC, FG, Uniform*$^{+sel}$, *HiC*$^{+sel}$, and *FG*$^{+sel}$. For example, $P_{ij}^{FG}=\mu^{FG}(L)$, and $P_{ij}^{FG^{+SEL}}=\mu^{FG}(L)\cdot\pi(L)$. The additional parameter describing selection is accounted for using BIC (described below).

We also examined a mutational model which combines the effects of chromosomal compartments as determined by HiC[2] with the *FG* model (*FG*$^{+\text{domains}}$). Domains are brought into our models by assuming different likelihoods of SCNA ends to be located active-active, active-inactive and inactive-inactive domains (two independent parameters). This domain structure is then multiplied by the fractal globule contact probability, $P_{ij}^{FG^{+domains}}=\mu^{FG}(L)\cdot D_{ij}$, where $D_{ij}=1$ if i and j are in different domains, $D_{ij}=\kappa$ if i and j are both in an open

domain, and $D_{ij} = \nu$ if i and j are both in a closed domain. We exclude chromosomes 4, 5, and 15 for the domain analysis, as these chromosomes have a poor correspondence between HiC domains (as determined in the original analysis of HiC[2]) and the HiC contact map.

Effects of Selection on the Probability of Fixation

Two major selective forces act on SCNAs: positive selection on SCNAs that amplify an oncogene or delete a tumor suppressor, and purifying selection that acts on all alterations. Purifying selection results from the deleterious effects of an SCNA that deletes or amplifies genes and regulatory regions of the genome that are not related to tumor progression. We assume that deleterious effect of an SCNA, and the resulting reduction in cells fitness ΔF, is proportional to SCNA length: $|\Delta F| \propto L$.

The probability of fixation is calculated using the Moran process as model of cancer evolution[15,20]:

$$\pi(\Delta F)=\frac{1-1/(1+\Delta F)}{1-1/(1+\Delta F)^{N}}, \tag{2}$$

where ΔF is a relative fitness difference (selection coefficient), N is the effective population size. For weakly deleterious mutations ($\Delta F < 0$, $N|\Delta F| \gg 1$, $|\Delta F| \ll 1$)

$$\pi(\Delta F) \approx \frac{\Delta F}{1-\exp(-\Delta F N)} \tag{3}$$

Note that for sufficiently deleterious mutations this leads to an exponentially suppressed probability of fixation: $\pi(\Delta F) \propto \exp(\Delta F N)$ ($\Delta F < 0$), a useful intuitive notion. Assuming a deleterious effect linear in SCNA length, $\Delta F = -L/\lambda$, we obtain the probability of fixation for purifying selection acting on an SCNA

$$\pi(L)=C\frac{L}{\exp(L/\alpha)-1} \tag{4}$$

where C is an arbitrary constant obtained from normalization of $P(L)$, and $\alpha = \lambda/N$ is a fitting parameter which quantifies the strength of purifying selection. For gene-based purifying selection, L is simply replaced by the number of genes altered. Mutations that are selectively neutral have no length dependence, so $\pi(L) = C$, and thus $P_{ij} \sim \mu_{ij}$.

Controlling for Positive Selection

Positive selection acting on cancer-associated genes (eg. oncogenes and tumor suppressors) presents a possible confounding factor to our analysis. To establish that our results were robust to positive selection acting on cancer-associated genes, we analyzed the subset of the 39,568 SCNAs (26,022 amplifications and 13,546 deletions) that do not span highly-recurrent SCNA regions identified by GISTIC with a false-discovery rate q-value for alteration of <.25 as listed in Beroukhim et al.[1], a collection of 24,310 SCNAs (16,521 amplifications and 7,789 deletions, respectively 63% and 58% of the full set). After SCNAs spanning highly-recurrent regions are removed, permutations are performed under the constraint that permuted SCNAs do not cross any of the highly-recurrent regions. Positive selection can also be somewhat controlled for by setting a threshold on the inferred change in copy number, to filter SCNAs that may have experienced strong positive selection in individual cancers. We note that our findings are robust to the subset of chosen SCNAs, most likely because there are many fewer driver SCNAs than passenger SCNAs (Supplementary Fig. 7).

Model Selection using Poisson Log-likelihood, Bayesian Information Criterion

Since the occurrence of a particular SCNA starting at *i* and ending at *j* is a rare event, we evaluate the relative ability of a model to predict the observed distribution of SCNA by calculating the Poisson Log-likelihood of the data given the model:

$$\log L(\text{SCNA}|\text{Model}) = \log\left(\prod_{(i-j)>1} \frac{\exp(-P_{ij}^{\text{Model}})(P_{ij}^{\text{Model}})^{SCNA_{ij}}}{(SCNA_{ij})!}\right) = \sum_{(i-j)>1} -P_{ij}^{\text{Model}} + SCNA_{ij}\log(P_{ij}^{\text{Model}}) + const. \quad (6)$$

where P_{ij}^{Model} is dictated by the model as explained above, and $SCNA_{ij}$ is the number of SCNAs that start and end at i and j. Since recurrent regions of amplification and deletion are different, we calculate the log-likelihood separately for amplifications and deletions, and then aggregate across these two classes of SCNAs. After the log-likelihood is calculated, models are ranked and model selection is performed using Bayesian Information Criterion (BIC). BIC penalizes models based upon their complexity, namely their number of parameters. Penalizing *k* additional parameters for *n* observed SCNAs using Bayesian Information Criterion (BIC) is straightforward:

$$BIC = \log L(\text{SCNA}|\text{Model}) - \frac{1}{2}k\log(n) \quad (7)$$

where models with higher BIC are preferred[28]. For the permutation analysis, log-likelihood is calculated in the same way, first for the observed SCNAs, and then for permuted sets of SCNAs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Mirny Lab for helpful conversations, in particular with Christopher McFarland regarding purifying selection and Maxim Imakaev regarding fractal globules. We thank Craig Mermel for an introduction to SCNA data. We thank Vineeta Agarwala, Jesse Engrietz, Rachel McCord, and Job Dekker for helpful comments and suggestions. This work was supported by the NIH/NCI Physical Sciences Oncology Center at MIT (U54CA143874)

References

1. Beroukhim R, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899–905. [PubMed: 20164920]
2. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]
3. Mirny LA. The fractal globule as a model of chromatin architecture in the cell. Chromosome Research. 2011; 19:37–51. [PubMed: 21274616]
4. Greenman C, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446:153–158. [PubMed: 17344846]
5. Wood LD, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007; 318:1108–1113. [PubMed: 17932254]
6. Beroukhim R, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:20007–20012. [PubMed: 18077431]
7. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]

8. Berger MF, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220. [PubMed: 21307934]
9. Wijchers PJ, de Laat W. Genome organization influences partner selection for chromosomal rearrangements. Trends in genetics : TIG. 2011; 27:63–71. [PubMed: 21144612]
10. Meaburn KJ, Misteli T, Soutoglou E. Spatial genome organization in the formation of chromosomal translocations. Seminars in cancer biology. 2007; 17:80–90. [PubMed: 17137790]
11. Nikiforova MN, et al. Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. Science. 2000; 290:138–141. [PubMed: 11021799]
12. Branco MR, Pombo A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS biology. 2006; 4:e138. [PubMed: 16623600]
13. Grosberg AY, Nechaev SK, Shakhnovich EI. The Role of Topological Constraints in the Kinetics of Collapse of Macromolecules. Journal De Physique. 1988; 49:2095–2100.
14. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976; 194:23–28. [PubMed: 959840]
15. Merlo LM, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nature reviews. Cancer. 2006; 6:924–935.
16. Lee W, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010; 465:473–477. [PubMed: 20505728]
17. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nature reviews. Genetics. 2009; 10:551–564.
18. Armitage, P.; Berry, G.; Matthews, JNS. Statistical methods in medical research. Vol. xi. Malden, MA: Blackwell Science; 2001. p. 817
19. Hastings PJ, Bull HJ, Klump JR, Rosenberg SM. Adaptive amplification: an inducible chromosomal instability mechanism. Cell. 2000; 103:723–731. [PubMed: 11114329]
20. Moran, PAP. The statistical processes of evolutionary theory. Oxford: Clarendon Press; 1962. p. 200
21. Chiarle R, et al. Genome-wide Translocation Sequencing Reveals Mechanisms of Chromosome Breaks and Rearrangements in B Cells. Cell. 2011; 147:107–119. [PubMed: 21962511]
22. Klein IA, et al. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. Cell. 2011; 147:95–106. [PubMed: 21962510]
23. Jesse M, Engreitz VA, Leonid A Mirny. Three-dimensional genome structure influences partner selection for chromosomal translocations in human disease. 2011 submitted.
24. Mayer R, et al. Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. BMC cell biology. 2005; 6:44. [PubMed: 16336643]
25. Roix JJ, McQueen PG, Munson PJ, Parada LA, Misteli T. Spatial proximity of translocation-prone gene loci in human lymphomas. Nature genetics. 2003; 34:287–291. [PubMed: 12808455]
26. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–724. [PubMed: 19360079]
27. Haber DA, Settleman J. Cancer: drivers and passengers. Nature. 2007; 446:145–146. [PubMed: 17344839]
28. Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978; 6:461–464.

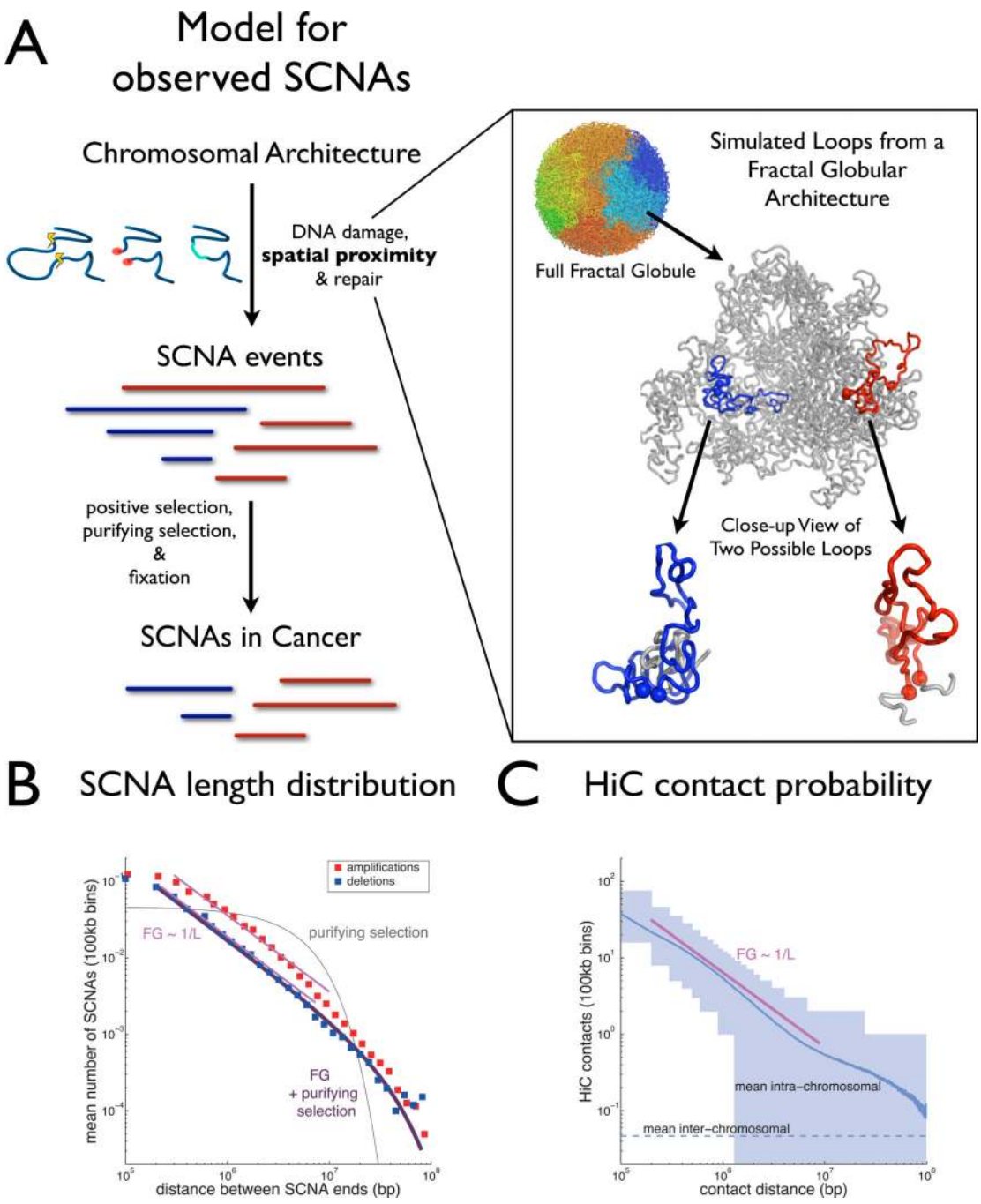

Figure 1. 3D proximity as mechanism for SCNA formation

A: Model of how chromosomal architecture and selection can influence observed patterns of somatic copy-number alterations (SCNAs). First spatial proximity of the loop ends makes an SCNA more likely to occur after DNA damage and repair. Next, forces of positive selection and purifying selection act on SCNAs which have arisen, leading to their ultimate fixation or loss. Observed SCNAs in cancer thus reflect both mutational and selective forces. Inset illustrates looping in a simulated fractal globule architecture (coordinates from M. Imakaev). Two contact points are highlighted by spheres and represent potential end-points of SCNAs.

B. SCNA length distribution for 60,580 less-recurrent SCNAs (39,071 amplifications, 21,509 deletions) mapped in 3,131 cancer specimens from 26 histological types[1]. Squares show mean number of amplification (red) or deletion (blue) SCNAs after binning at 100 kb resolution (and then averaged over logarithmic intervals). Light magenta lines show ~1/L distributions. Grey line shows the best fit for purifying selection (Eq 4) with a uniform mutation rate. Dark purple line shows best fit for deletions for FG^{+sel}.
C: Probability of a contact between two loci distance L apart on a chromosome at 100 kb resolution. The probability is obtained from intra-chromosomal interactions of 22 human chromosomes characterized by the HiC method (human cell line GM06690)[2]. Shaded area shows range from 5th and 95th percentiles for number of counts in a 100kb bin at a given distance. The mean contact probability is shown by blue line. Light magenta line shows ~1/L scaling also observed in the fractal globule model of chromatin architecture. Blue dashed line provides a baseline for contact frequency obtained as inter-chromosomal contacts in the same dataset.

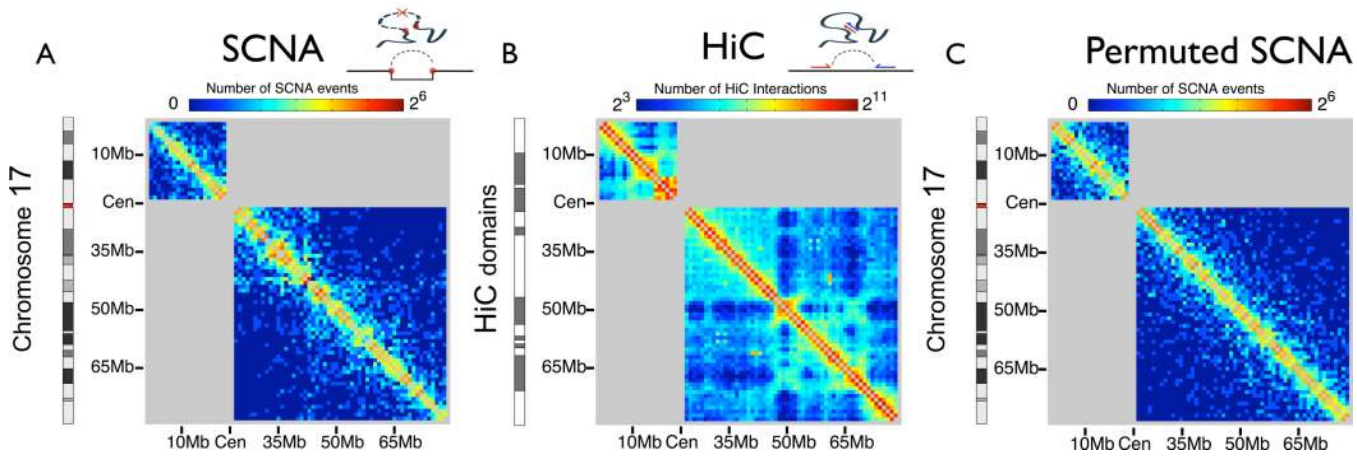

Figure 2. Heatmaps for chromosome 17 at 1 Mb resolution

A. SCNA heatmap: the value for site (*i*,*j*) is the number of SCNAs starting at genomic location *i* and ending at location *j* on the same chromosome. Chromosome band structure from UCSC browser shown on the left side with centromeric bands in red.

B. HiC heatmap: site (*i*,*j*) has the number of reported interactions between genomic locations *i* and *j* at Mb resolution. HiC domain structure is shown on the left side. Domains were determined by thresholding the HiC eigenvector (as in[2], white represents open domains, dark gray represents closed domains).

C. Permuted SCNA heatmap: as in **A**, but after randomly permuting SCNA locations while keeping SCNA lengths fixed.

Visually, the true SCNA heatmap is similar to HiC (Pearson's r = .55, p <.001, see Supplementary Table S1 for other chromosomes), displaying a "domain" style organization. Cartoons above the heatmaps illustrate how mapped HiC fragments and SCNA end-points can be converted into interactions between genomic locations *i* and *j*. Since inter-arm SCNAs, SCNAs with end-points near centromeres or telomeres, and SCNAs < 1Mb were not considered in our statistical analysis, these areas of the heatmaps are grayed out.

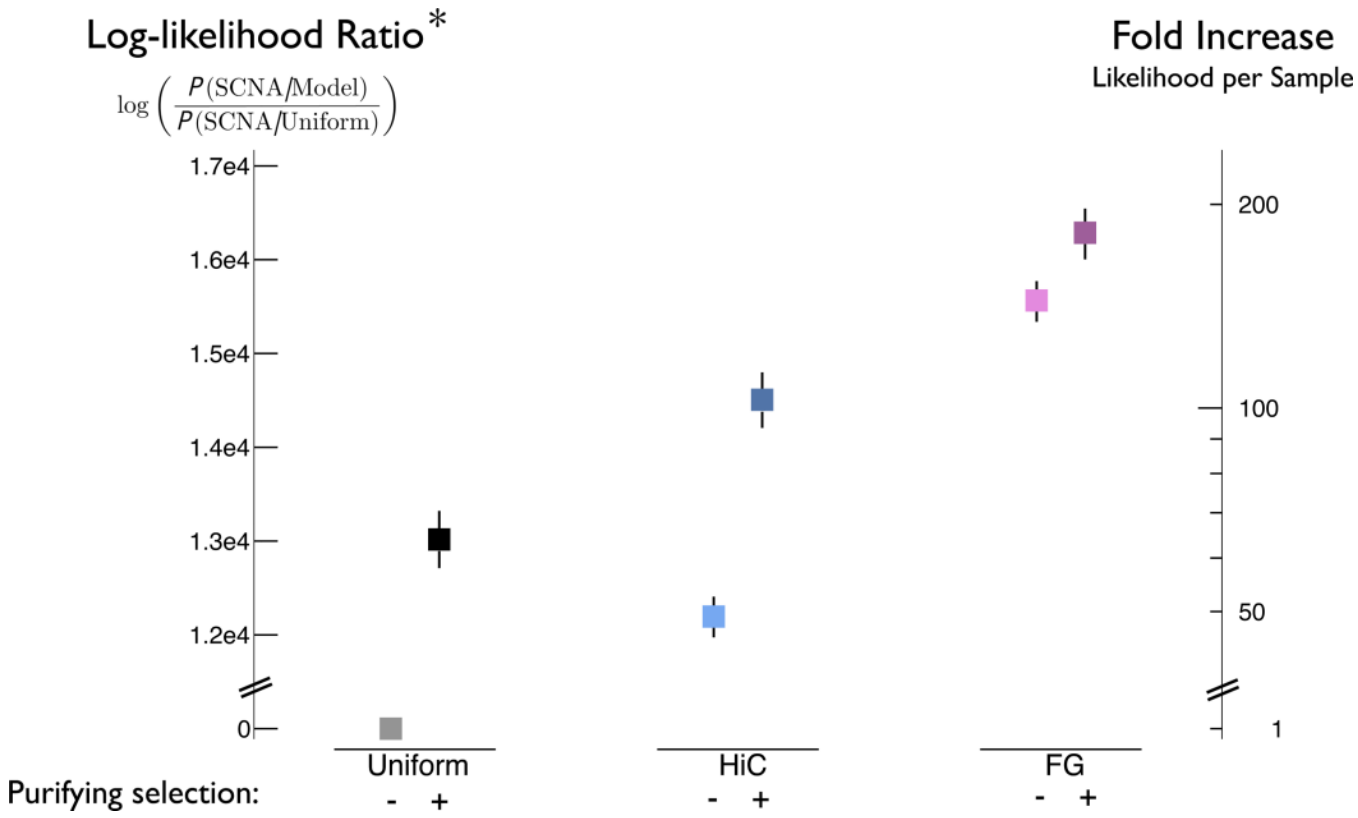

Figure 3. Selecting a model of SCNA formation

For each model, the log-likelihood ratio (*BIC-corrected log-likelihood ratio) is shown for the 24,310 observed SCNAs that do not span highly-recurrent SCNA regions listed in[1]. The following six models are considered: *Uniform, Uniform*$^{+sel}$*, HiC, HiC*$^{+sel}$*, FG, FG*$^{+sel}$. *HiC* model assumes mutation rates proportional to experimentally measured contact probabilities, while *FG* model assumes mutation rates proportional to mean contact probability in a fractal globule architecture (~1/L). Left y-axis presents BIC-corrected log-likelihood ratio for each model vs. *Uniform* model. Each model was considered with (+) and without (−) purifying selection. Right y-axis shows the same data as a fold difference in likelihood per cancer specimen (sample) vs. *Uniform*. Error bars were obtained via bootstrapping: squares represent the median values, bar ends represent the 5th and 95th percentiles. The FG model significantly outperforms other mutational models of SCNA formation, and every model is significantly improved when purifying selection is taken into account.

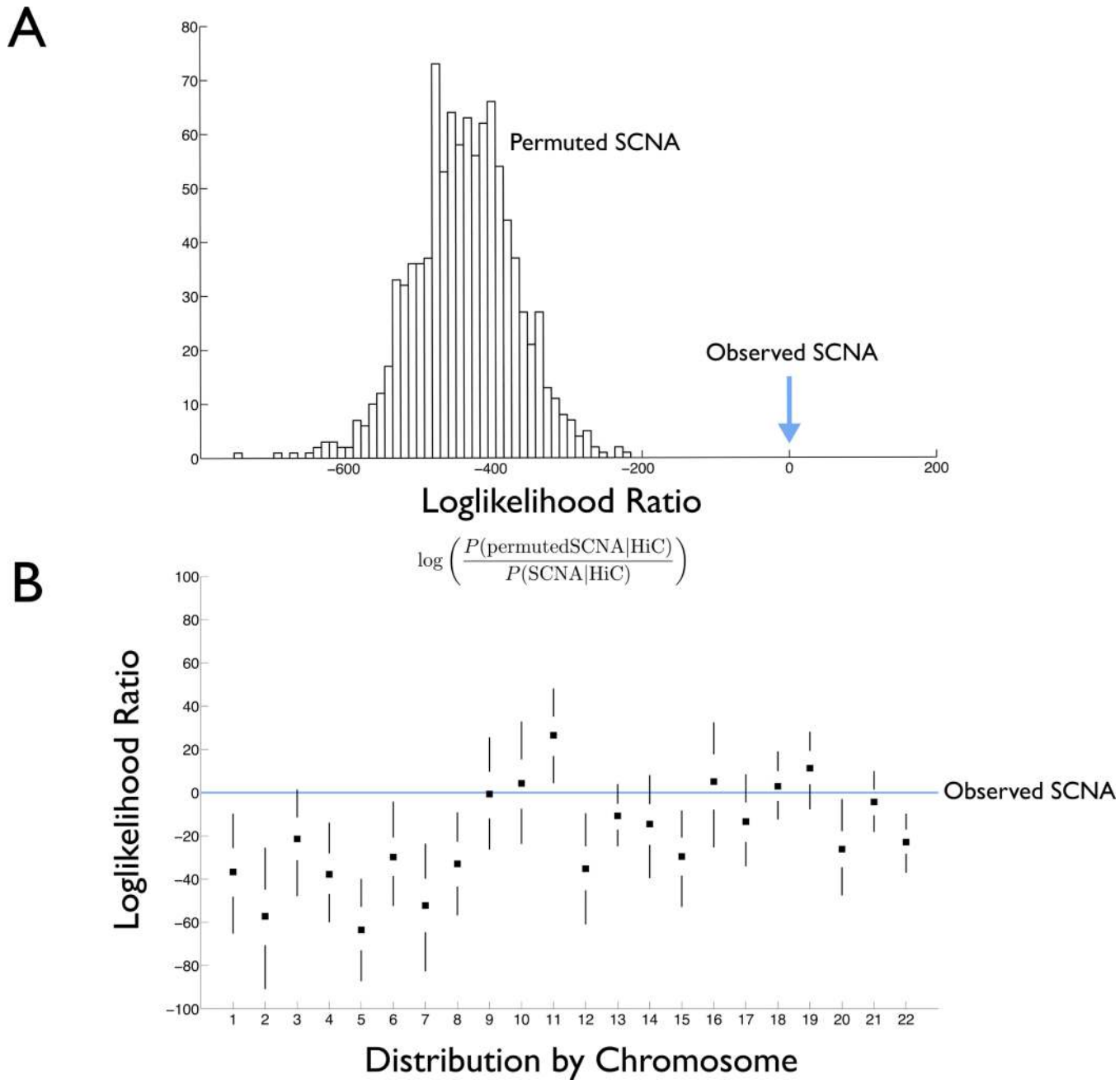

Figure 4. Permutation analysis of the relationship between SCNAs and megabase-level structure of HiC chromosomal interactions

A. Distribution of log-likelihood ratios for randomly permuted SCNAs given HiC vs. observed SCNAs given HiC over all 22 autosomes. Observed SCNAs (blue arrow) are fit better by HiC contact probability with $p<.001$. Permutations are performed by shuffling SCNA locations while keeping SCNA lengths fixed. **B:** Distributions of the same log-likelihood ratios for individual chromosomes (vs their corresponding observed SCNA, blue line). Squares represent median values, error bars respective represent the range from 5th to 25th percentile and 75th to 95th percentile.