Supplementary Information

for

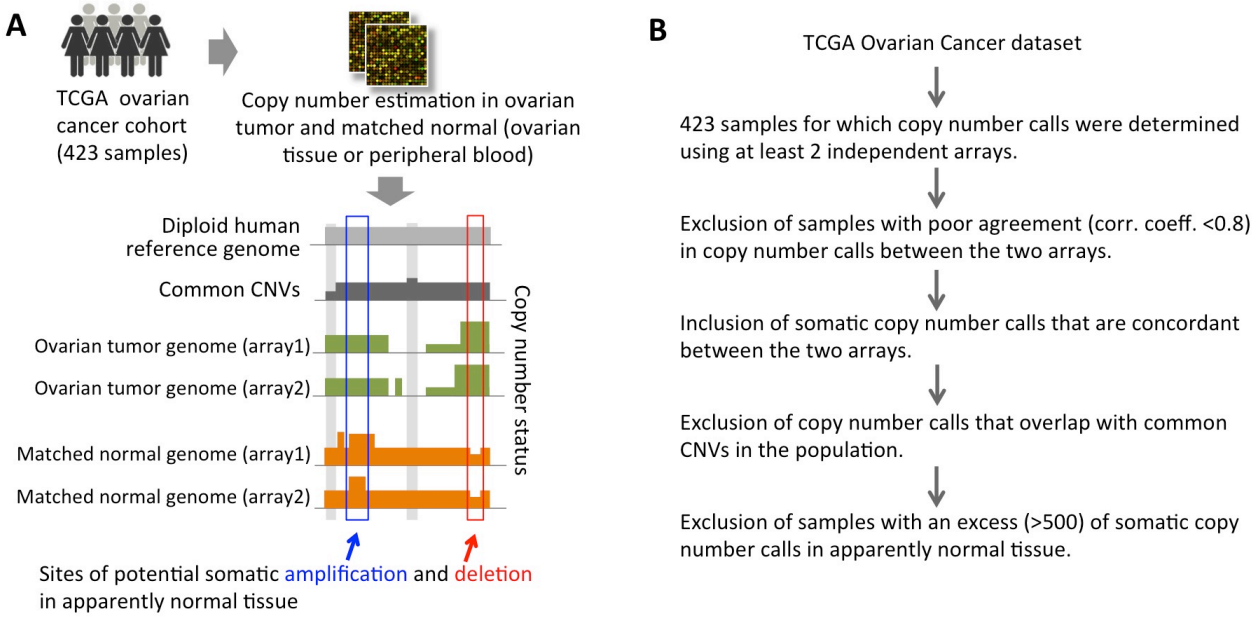
Patterns of somatically acquired amplifications and deletions in apparently normal tissues of ovarian cancer patients

Leila Aghili, Jasmine Foo, James DeGregori, Subhajyoti De

Index:

Supplementary Module 1: Preprocessing of the datasets

The Cancer Genome Atlas (TCGA, 2011) recruited ~570 ovarian cancer patients, a subset of which also had genomic data available. We obtained the aCGH-based copy number calls of 423 serous ovarian cancer samples and matched normal tissue from the Cancer Genome Atlas (TCGA, 2011), for which copy number calls were available using two independent arrays. Of them, 109 and 314 samples had normal ovarian tissue and peripheral blood as matched normals, respectively. Copy number status for ovarian tumor-normal pairs was determined using three aCGH arrays in two genome analysis centers: Agilent HG-CGH-415K_G4124A and HG-CGH-244A arrays at Harvard Medical School, and Agilent CGH-1x1M_G4447A array at MSKCC. In general, copy number calls were highly similar between pairs of arrays for a vast majority (~91%) of the samples at a base-pair resolution (correlation coefficient >0.9). We selected these samples for further analysis to minimize false positive calls.

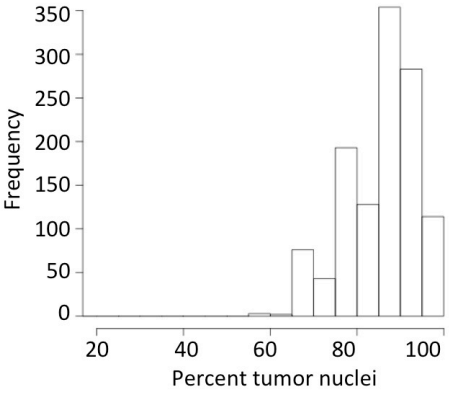


Supplementary Figure S1: A) Calling potential somatic amplifications and deletions in apparently normal tissue from copy number calls from tumor-normal pairs using two independent arrays. B) Filtering the dataset of somatic copy number alterations.

We applied several filters to refine our dataset (**Supplementary Figure S1A-B**). First, we excluded the tumor-normal pairs for which there were poor agreements in copy number calls between the pairs of aCGH arrays (correlation coefficient <0.8). Second, we considered only the copy number calls that were consistent between both the arrays, using guidelines as outlined in the 'Detection of somatic copy number status' sub-section of the Methods section of the main text. Third, we obtained the list of common CNVs present in the human population from the UCSC Genome Browser (*`Structural var`* track; data freeze 03/2013), which were curated by the Database of Genomic Variants (Macdonald et al., 2014) from published papers. Fourth, we excluded the samples with an excess of potential somatic copy number calls (>500). For instance, the sample TCGA-13-0797 had ~1000 potential somatic amplification and deletion calls in apparently normal tissue distributed through out the genome, which was 3 orders of magnitude more than the median for the dataset. Taken together, we have applied multiple filters to minimize potential biases arising from biological and technical artifacts; but since array-based approaches are error-prone, we cannot exclude the possibility of any false positives and false negatives in our dataset.

At the end, our filtered dataset had 607 potential somatic amplifications and deletions in 314 normal peripheral blood samples, and 494 potential somatic amplifications and deletions in 109 normal ovarian tissue samples.

Issues associated with tumor purity: Besides tumor cells, tumor samples also contain non-tumor cells (e.g. stromal and immune cells), which can affect copy number assessment. Different tumor samples can have different proportions of tumor and non-tumor cells, and therefore the aCGH log2 threshold for diploid (or diploid equivalent) copy number in tumor samples need to be corrected for tumor purity. However, purity is not a concern for matched normal samples, which have no tumor cell contamination, and thus the threshold for copy number alteration in normal sample is expected to remain unaffected. We recognize that there might be between-individual variations in normal tissue composition that arose during normal development or sample preparation – but such variations in normal tissue composition are expected to be minor, and there is no appropriate data to adjust for such variations. Hence, while comparing aCGH signals from normal and matched tumor samples to identify pSCNAs in normal samples, we only adjust the tumor aCGH threshold for purity.



Supplementary Figure S2: Pathological tumor purity estimate for the TCGA ovarian cancer samples.

We obtained pathological tumor purity estimates for the TCGA ovarian cancer samples from the Cancer Genome Atlas (TCGA, 2011). A vast majority of the samples had high (80-100%) tumor purity (**Supplementary Figure S2**). We then calibrated the tumor aCGH threshold after adjusting for pathological tumor purity provided by the TCGA, and repeated the analyses described in the **Supplementary Figure S2**. For instance, for the tumor samples with purity >90%, 80-90%, 70-80%, 60-70%, 50-60%, and <50%, we used a copy number log2 cut-off of 0.1, 0.09, 0.08, 0.07, 0.06, and 0.05 respectively. The threshold in paired normal samples remained unchanged.

Adjustment for tumor purity imposes a stricter aCGH threshold in tumor samples, and thus we expected to detect fewer pSCNAs in the filtered analysis. Indeed, in the purity-adjusted analysis we detected 1084 pSCNAs (98.5% of 1101 pSCNAs in the original analysis). The remaining 17 cases (1.5% cases) were randomly distributed in 14 samples in the dataset. Excluding these small number cases does not affect any of our key conclusions. For instance, none of these pSCNAs overlapped with the loci discussed in Table 2; none of these samples had BRCA germ-line or somatic mutations (so our results in Figure 1D-E would be marginally stronger after excluding these cases), and it did not affect the survival analysis (p-value <0.05 in each case after excluding these cases). Therefore, tumor purity estimates are unlikely to affect our key conclusions.

Taken together, we have taken measures to minimize biases arising from biological and technical artifacts, but nevertheless, we cannot rule out the possibility of any false positives and false negatives in our dataset.

Supplementary Module 2: Mutation rate estimation

Let us assume that λ is the log2 signal intensity at a given locus in a given sample in the copy number microarray, and accordingly, C is the aggregated copy number call at the tissue-level resolution, such that $\lambda = log_2(\frac{C}{2})$.

Let α be the fraction of cells with δ copies of the locus, so that, $C = (\alpha\times\delta + (1 - \alpha) \times 2)$. Solving for α we get $\alpha = \frac{2^{\lambda+1}-2}{\delta-2}$. Since α is obtainable from λ and δ we will consider α as observable from here onwards. If we assume that genomes of normal somatic cells are close to diploid such that $\delta \in$ (Hoffman et al., 2012), then α is obtainable from λ alone.

Let N and L be the frequency and median length of the detectable somatic copy number alterations per sample, so that they collectively affect $N \times L$ bp of approximately 3×10^9 bp of the diploid genome.

Here we attempt to estimate, using two different models, the somatic copy number alteration rate per locus per generation in the hematopoietic stem cell (HSC) lineage since conception during development. Similar estimation for ovarian tissue was challenging, since the time of separation of the tumor and normal ovarian stem cell could not be ascertained with certainty, and relevant parameters were not available.

Model 1: We assume that since conception during the course of development the HSC pool follows a discrete time pure birth process (Galton-Watson process with zero death rate) (Karlin, 1966). In other words, at each generation each cell splits into two, and there is no cell death. Note that in this model if we observe α, we can estimate that the mutation occurred at generation $n = log_2(\frac{1}{\alpha})$. Next we calculate the number of cell divisions that occurred until the mutation happened (i.e. how many cell divisions in the first n generations). We can calculate that this number (call it D) is

$$D = 1 + 2 + 2^2 + \cdots + 2^n = 2^{n+1} - 1.$$

Then, the mutation rate per locus per symmetric division r is roughly estimated as:

$$r = \frac{NL}{3\times10^9}\frac{1}{D} = \frac{NL}{3\times10^9}\frac{1}{2^{n+1}-1}. \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where $n = log_2(\frac{1}{\alpha})$ and α is given in terms of λ and δ above.

Model 2: In this model, let us consider the possibility of cellular death and relax the assumption of simultaneous generations (the assumption that all cells divide at the same time in each generation). Let us use a continuous-time birth-death process. In order to use this process we must have two additional estimates – cell birth rate (b) and death rate (d).

In this model the population P grows as $e^{(b-d)t}$. After observing α, we can estimate that the mutation happened at a time when the population was of size $P = \frac{1}{\alpha}$.

Next we proceed to estimate D, the number of cell divisions that have occurred until the mutation happened, under this model. The mutation happened at time $t_m = log\left(\frac{P}{b-d}\right)$. Then the number of cell divisions happening until this time t_m is approximately

$$D = \int_0^{t_m} e^{(b-d)s} b\,ds = \frac{b}{(b-d)}(P-1) = \frac{b}{b-d}\left(\frac{1}{\alpha} - 1\right).$$

Then, the mutation rate per locus per symmetric division r is roughly estimated as:

$$r = \frac{NL}{3\times10^9}\frac{b-d}{b}\left(\frac{\alpha}{1-\alpha}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Application of the models, underlying assumptions, and rate estimation: In the aCGH copy number microarray, we chose to set the detection limit at λ > 0.2 or λ < -0.2. The average value of N in peripheral blood is approximately 2, and the median length of such events is 5.65×10^4 bp. We assume that the genomes of normal

somatic cells are close to diploid such that [$\delta \in \{1,2,3\}$]. Using these estimates in the ***model-1*** based on the *pure birth process*, the rate of somatic genomic alteration per locus per cell division is 6.58×10^{-6}.
We assume that the HSC birth rate $b = 0.02$ week^{-1}, death rate $d = 0.00\,2$ week^{-1} based on published reports(Abkowitz et al., 2002; DeGregori, 2013; Michor et al., 2005; Sehl et al., 2011). Accordingly, Using these estimates in the ***model-2***, where we consider the possibility of cellular death and relax the assumption of simultaneous generations, the rate of somatic genomic alteration per locus per cell division is 1.38×10^{-5}.

We then estimated the rates for the sample, TCGA-13-0757, which had 2.44 Mb of genomic region affected by $pSCNA^{norm}$, considerably more than others. The average rate of somatic genomic alteration per locus per cell division for this sample was 1.42×10^{-4} and 2.98×10^{-4}, according to the model-1 and model-2 respectively.

These models make certain assumptions such as (i) the cell division, cell death, and mutation rates do not change during development and across the HSC population, (ii) there is no clonal selection during normal HSC development, (iii) HSC divisions are always symmetric, and (iv) all mutations occur in the same clone. The biology of hematopoietic development is more complex than that assumed in the model, but in absence of information regarding relevant parameters, these assumptions were rational choices. Since we adopted a simplistic study design, these estimates likely represent a parsimonious estimate of the lower bound of the rate of somatic genomic alterations in peripheral blood.

Single cell genome sequencing analysis: Voet et al. (Voet et al., 2013) identified copy number alterations in (i) individual cells and also (ii) in daughter cells during cell division using single cell paired end genome sequencing data. They reported de novo events arising in the daughter cells in one cell cycle (**Supplementary table S1**). Accordingly, using the equation (1), we calculate the rate of somatic genomic alteration per locus per cell division.

Supplementary Table S1: Summary statistics showing the rate of somatic genomic alteration per locus per cell division. A broken branch in the phylogenetic tree indicates multiple cell divisions between the ancestral and progenitor cells. Amplifications (blue) and deletions (red) are as reported by Voet et al. (Voet et al., 2013)

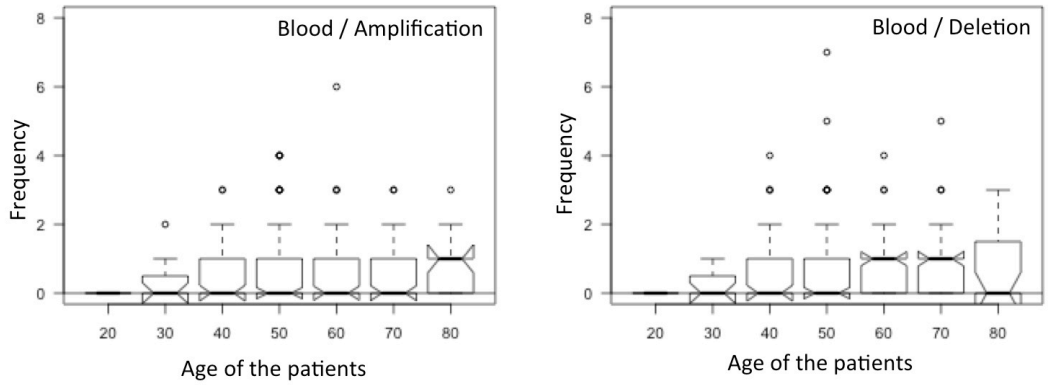
Cells	De novo aberrations in cell cycle	Estimated rate (relative to parent cell) $r = \frac{N \times L}{3 \times 10^{-9}}$
mda-sc82	Chr1:87.32 Mb–121.45 Mb	1.14E-02
mda-sc83	-	0.00E-00
mda-sc1	Chr1:66.68 Mb–114.15 Mb	1.58E-02
mda-sc2	Chr4:118.95 Mb–191.03 Mb; Chr12:45.61 Mb–133.83 Mb;	5.34E-02
HCC38 breast cancer cell line		
PicoPlex-sc1	-	0.00E-00
PicoPlex-sc2	-	0.00E-00
PicoPlex-sc9	Chr2:95.52 Mb–119.2Mb; Chr5:106.79 Mb–153.13 Mb; Chr5:164.47 Mb–180.71 Mb; Chr8:125.38 Mb–131.96 Mb; Chr18:20.23 Mb–46.66 Mb; Chr18:55.30 Mb–78.01 Mb;	4.73E-02
PicoPlex-sc10	Chr5:106.79 Mb–153.13 Mb; Chr5:164.47 Mb–180.71 Mb; Chr8:125.38 Mb–131.96 Mb; Chr18:20.23 Mb–46.66 Mb; Chr18:55.30 Mb–78.01 Mb;	3.94E-02

We also acknowledge the potential caveats of using single cell sequencing data in this context. First, HCC38 is a highly unstable breast cancer cell line and thus has relatively high number of genomic alterations; second, the de novo alterations detected in individual cells are subject to stochasticity, especially when the sample size is small, and only a single cell division is considered; third, small events could not be detected in single cell genomic analysis, which also has a high error rate; fourth, the rates of de novo alterations in a tissue and in cell culture are likely to differ; and finally, purifying selection operates on these genomic changes in a tissue, excluding many of these events from reaching a high frequency in the population. Anyhow, we decided to provide single cell analysis as a complementary investigation to our tissue-level analysis.

Supplementary Module 3: Control for potential covariates

Effects of age on pSCNA^{nbl}, after adjusting for amplification/deletion status

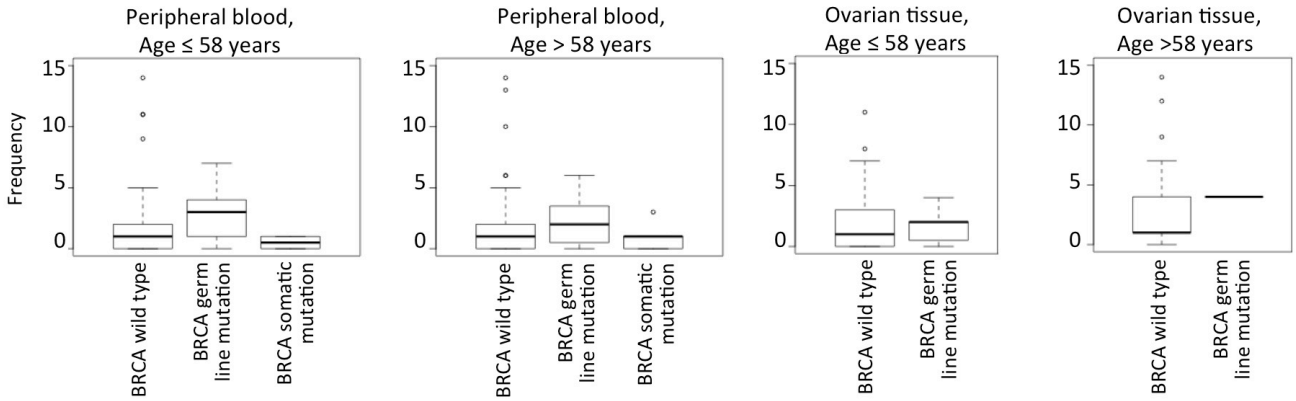
In main text Fig 1, we reported that older individuals had more pSCNA^{norm}. We repeat the analyses after analyzing amplifications and deletions separately. We find consistent results for both (**Supplementary Figure S3**).



Supplementary Figure S3: The number of somatic genomic alterations (amplifications and deletions) per normal tissue sample (peripheral blood and ovarian tissue), grouped according to amplification/deletion status and age.

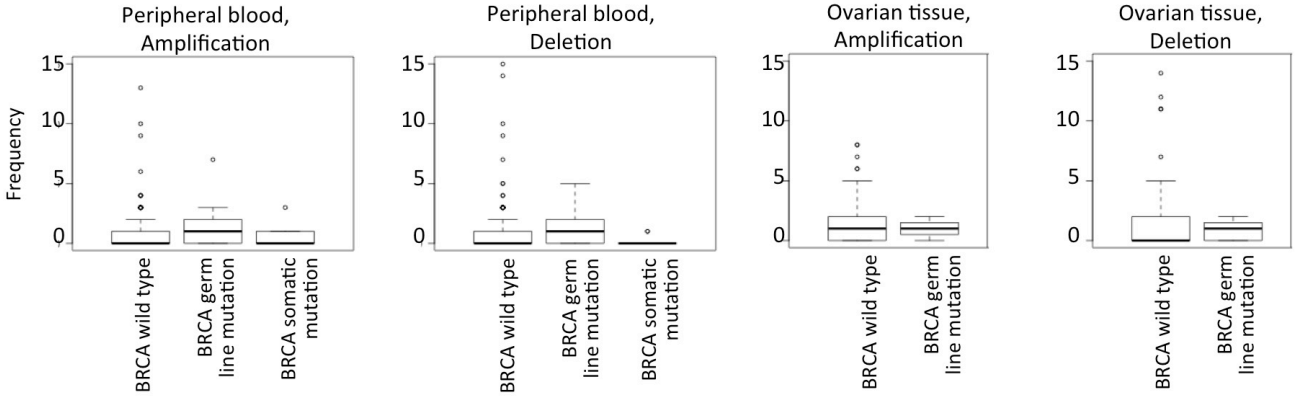
Effects of BRCA mutations on pSCNA^{norm}, after adjusting for age and amplification/deletion status

In Figure 1 of the main text, we reported that the BRCA mutation carriers harbor more potential somatic amplifications and deletions than those with no BRCA mutations. The number of samples with BRCA1 or BRCA2 germ line mutation was too small to warrant separate analysis. Anyhow, we investigated whether the results were confounded by the age of the samples. Thus, we reanalyzed the data, after grouping the samples by their age, and analyzing the samples with age ≤58 years (below median) and >58 years (above median) separately (**Supplementary Figure S4**). For each age group, we classify the samples according to their BRCA mutation status and compare the number of pSCNAs between BRCA germ-line and somatic mutation carriers with that for the wild-type samples within respective age group (**Supplementary Figure S4**). Within each age group we found consistent results: BRCA germ line mutation carriers had higher pSCNAs compared to those with wild-type copy BRCA genes. Therefore, excess of pSCNAs in BRCA germlne mutation carriers was not due to age differences.



Supplementary Figure S4: The number of somatic genomic alterations (amplifications and deletions) per normal tissue sample (peripheral blood and ovarian tissue), grouped according to BRCA mutation status and age.

We then investigated whether the results were similar for both amplifications and deletions. We reanalyzed the data, after grouping the analyzing the somatic amplifications and deletions separately (**Supplementary Figure S5**). We found that for the BRCA mutation carriers typically had higher frequency for both somatic amplifications and deletions relative to those with no BRCA mutations.



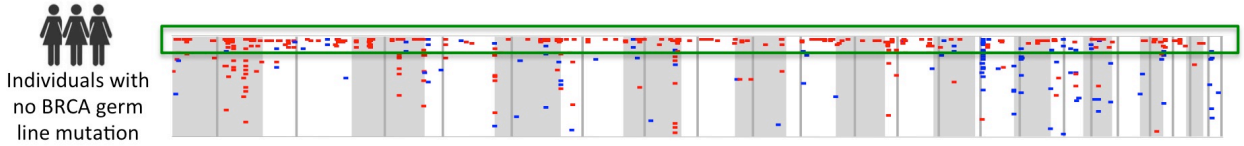
Supplementary Figure S5: The number of somatic amplifications and deletions per normal tissue sample (peripheral blood and ovarian tissue), grouped according to BRCA mutation and amplification/deletion status.

Taken together, our findings suggest that the BRCA mutation carriers harbor more potential somatic amplifications and deletions than those with no BRCA mutations, and that these results are independent of age.

Supplementary Module 4: Context of genomic alterations in apparently normal tissue

Individuals with excess of potential somatic genomic alterations in apparently normal ovarian tissue

Some individuals had an excess of potential somatic amplifications and deletions. We excluded the samples (e.g. TCGA-13-0797) which had more than 1000 $pSCNA^{norm}$ calls in apparently normal tissue distributed through out the genome, which was 3 orders of magnitude more than the median for the dataset.



Supplementary Figure S6: Personal genomes with an excess of $pSCNAs^{nov}$ in apparently normal ovarian tissue. Two individuals (TCGA-57-1584 and TCGA-57-1993) with an excess of somatic amplifications are marked with green box.

Nonetheless, we decided to survey the top ranking samples in terms of $pSCNA^{norm}$ frequency. We found that a considerable proportion of focal deletions in ovarian tissue were clustered in the genomes of two individuals (TCGA-57-1584 and TCGA-57-1993, **Supplementary Figure S6**). We found no evidence for technical artifacts; instead both were stage IIIC, grade-3 early age (47 and 56 years) ovarian cancer patients with considerable genomic abnormality also in their tumor genomes. It is possible that genomic instability in their normal and malignant somatic tissue stemmed from extensive DNA damage and/or impaired repair, but limited amount of functional, clinical, and cancer gene mutation data precluded any detailed investigation into the origin of genomic instability in these patients.

Enrichment for genomic features:

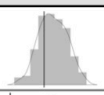
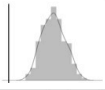
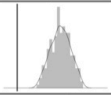
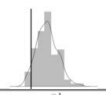
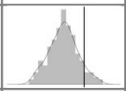
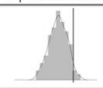
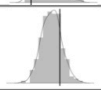
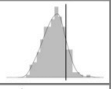
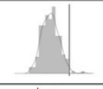
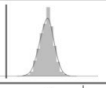
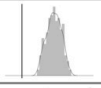
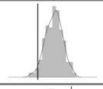
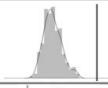
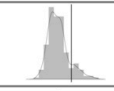
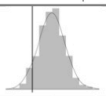
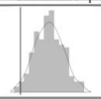
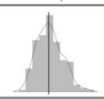
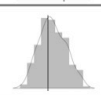
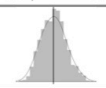
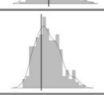
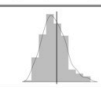
We analyzed genomic context of the $pSCNA^{norm}$ after segregating amplifications and deletions in blood and ovarian tissue separately, using an approach as described in the **Method**.

For each normal sample, we first calculated the extent of overlap using intersectbed after masking selected regions: 1Mb centering centromeres, 500kb from the tip of the telomeres, and also the genomic regions that underwent copy number changes in its matched tumor genome (and thus was not assessed for copy number status in the paired normal sample). We masked the centromere and telomere regions since the aCGH arrays (and also some genomic and epigenomic features sucha s replication timing) have poor representation there. We then permuted the $pSCNA^{norm}$ within respective chromosomes using shufflebed, while keeping the location and higher order organization of genomic features unchanged, and after masking the same selected regions in each sample. Since the regions that underwent copy number changes in its matched tumor genome would be specific to a tumor-normal pair, we ran the permutation analysis separately for each sample, and then combined the results. We reported q-value for statistical significance (**Supplementary Figure S7**). We analyzed potential somatic amplifications and deletions in blood and ovarian tissue separately, and chose to highlight those that were deemed significant in at least 3 out of 4 scenarios.

We also note the challenge for estimating statistical significance (Bilke and Gindin, 2012; De et al., 2013) and the limitations of combining heterogeneous data types from different sources(Sima and Gilbert, 2014; Sugihara et al., 2012). Permutation is undoubtedly the preferred method in this scenario as (i) for a large number of data-points (e.g. the number of $pSCNA^{norm}$ is ~10^3), classical statistical analyses (e.g. t-test) are expected to return extremely significant p-value even for very minor differences, (ii) permutation allowed us to make necessary adjustments for genomic context (e.g. masking centromeres and telomeres), and (iii) provides realistic p-values when correctly implemented (Bickel et al., 2010; Bilke and Gindin, 2012; De et al., 2013). However, one can perform permutation in many different ways (e.g. permuting within respective chromosomes, or over all chromosomes; controlling for minimum distance between two events in the same sample). Choice of permutation constraints has the potential to affect the null model and hence the statistical inferences drawn (De et al., 2013). In our analysis, we chose a simple and yet biologically relevant null model, and the permutation results for different functional elements including G4, Alu, L1, and evolutionarily conserved elements were generally consistent for different categories in **Supplementary Figure S7.** We found similar results when the $pSCNA^{norm}$ were permuted across all autosomes.

Nevertheless, inferring causality from association is non-trivial, especially when heterogeneous data types from different sources are integrated. Hence we cautiously interpret the data. In any case, our findings are consistent with the function of these elements.

Supplementary Figure S7: Enrichment for genomic contexts for all pSCNAs, and also when the somatic amplifications and deletions in blood and ovarian tissues were analyzed separately.

Overlapping genomic feature	Total		Blood				Ovary			
			Amplifications		Deletions		Amplifications		Deletions	
	q-value	Distribution	q-value	Distribution	q-value	Distribution	q-value	Distribution	q-value	Distribution
Protein coding genes	>0.05		>0.05		>0.05		>0.05		>0.05	
28 way conserved elements	<0.05		<0.05		<0.05		<0.05		<0.05	
Alu elements	<0.05		>0.05		<0.05		>0.05		<0.05	
L1 elements	>0.05		>0.05		<0.05		>0.05		>0.05	
L2 elements	<0.05		<0.05		<0.05		<0.05		<0.05	
G quadruplex motifs	<0.05		<0.05		>0.05		>0.05		<0.05	
Constant early replicating regions	>0.05		<0.05		>0.05		>0.05		>0.05	
Constant late replicating regions	>0.05		>0.05		>0.05		>0.05		>0.05	
Early replicating fragile sites	>0.05		>0.05		>0.05		>0.05		>0.05	
Common fragile sites	>0.05		>0.05		>0.05		>0.05		>0.05	

Supplementary Module 5: Correlation with cancer gene mutation and survival patterns

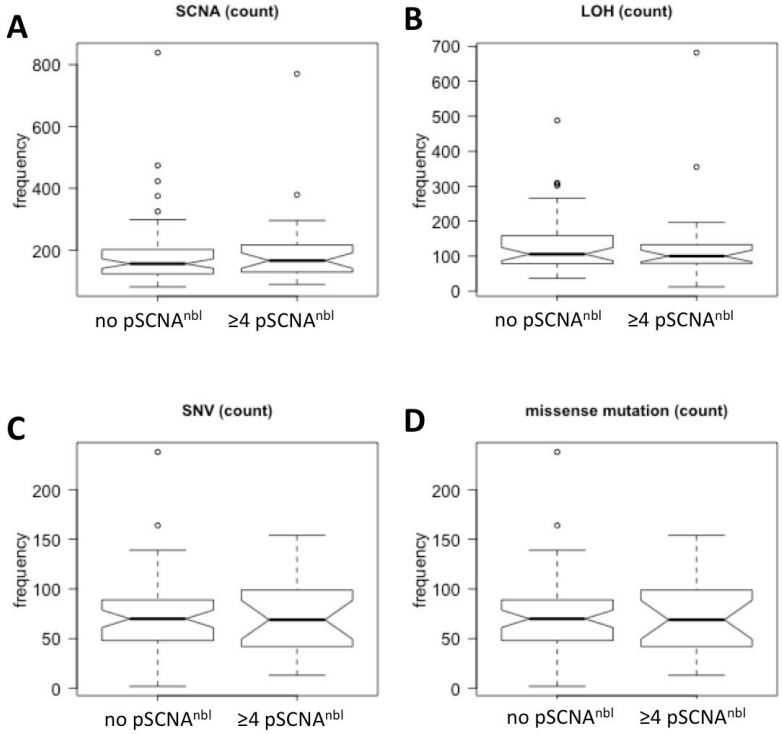
Cancer gene mutation analysis: We obtained data on somatic point mutations in protein coding genes for the ovarian cancer samples from the TCGA(TCGA, 2011). The variants were identified using Illumina GaII and ABI SOLiD sequencing, and then comparing tumor and matched normal samples as a part of the TCGA initiative. We analyzed only the missense mutations that occurred in the set of 121 classic cancer genes, as recorded in the COSMIC database(Forbes et al., 2011). We calculated the frequency of cancer gene mutations in the samples that had no detectable $pSCNA^{nbl}$ ($pSCNA^{nbl}=0$), and compared that with the samples that had more detectable $pSCNA^{nbl}$ ($pSCNA^{nbl} \geq 3$; ≥ 4; ≥ 5; **Supplementary Table S2**). Mutations in RB1, MLL3, CREBBP were present in >5% of the samples with no detectable $pSCNA^{nbl}$, but rarely occurred in the samples with an excess of $pSCNAs^{nbl}$, while some other genes (e.g. TP53, NF1) were mutated in proportionally less samples in the former group. The results were not biased by age, stage, and BRCA mutation status (data not shown).

Supplementary Table S2: Table showing the frequency of missense mutations in classic cancer genes.

Gene	Samples with pSCNA=0	Samples with pSCNA≥3	Samples with pSCNA≥4	Samples with pSCNA≥5
Number of sample	*82*	*63*	*27*	*15*
TP53	65	53	25	14
BRCA1	6	2	1	1
CREBBP	5	1	0	0
RB1	5	0	0	0
MLL3	4	0	0	0
APC	2	1	0	0
NF1	2	6	2	1
ABL1	1	0	0	0
ASXL1	1	0	0	0
ATM	1	1	1	0
ATRX	1	0	0	0
BAP1	1	0	0	0
BCOR	1	1	0	0
BRCA2	1	1	0	0
CSF1R	1	0	0	0
DAXX	1	0	0	0
EGFR	1	0	0	0
EP300	1	1	0	0
ERBB2	1	2	0	0
FAM123B	1	0	0	0
GATA3	1	0	0	0
GNA11	1	0	0	0
GNAS	1	1	0	0
HNF1A	1	1	0	0
IL7R	1	2	0	0
KDR	1	2	2	2
KIT	1	0	0	0
KRAS	1	0	0	0
NF2	1	1	0	0
NOTCH2	1	0	0	0
NRAS	1	1	0	0
PTCH1	1	0	0	0
RET	1	2	1	0
SETD2	1	1	0	0
TET2	1	0	0	0
TNFAIP3	1	1	1	1
CDC73	0	1	1	1
CDH1	0	1	0	0
DNMT3A	0	1	0	0
FGFR2	0	1	0	0
FLT3	0	2	2	0
NTRK3	0	1	1	0
PIK3CA	0	1	0	0
PRDM1	0	1	1	1

Genomic alterations in the matched tumor genomes

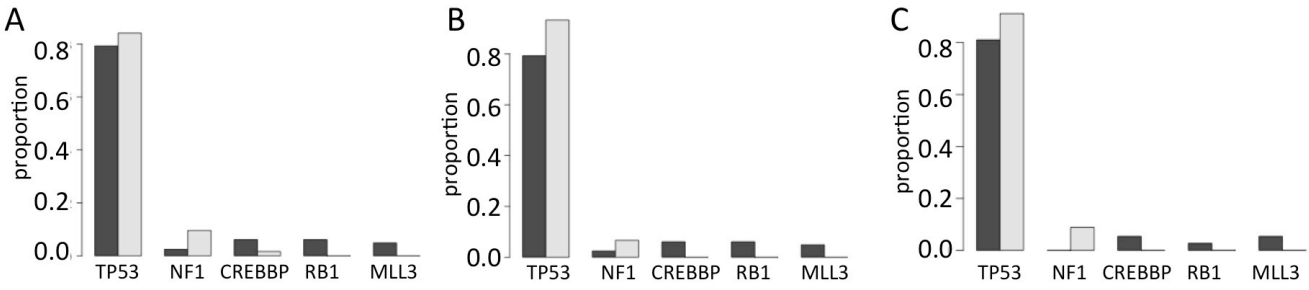
We compared the burden of genomic alterations (*i.e.* point mutations, copy number alterations, and LOH events) between the tumor genomes of the patients who had no pSCNA^{nbl} and the patients who had considerable excess of such events in apparently normal blood (pSCNA^{nbl}≥4). In particular, we compared the number of (i) somatic copy number alterations, (ii) LOH events, (iii) single nucleotide mutations in protein coding regions, and (iv) missense mutations in their tumor samples, and in all the four categories, there were no significant differences between the two patient groups (**Supplementary Figure S8)**.



Supplementary Figure S8: Boxplots showing differences in the frequencies of (A) somatic copy number alterations (B) loss of heterozygosity events, (C) single nucleotide mutations in protein coding regions, and (D) missense mutations in the tumor genomes of the patients who had no pSCNA^{nbl} (left box) and those who have ≥4 pSCNA^{nbl} (right box) in their peripheral blood.

Cancer gene mutation analysis after adjusting for potential covariates

In the main text and Fig-3B, we reported that the patients with no pSCNA^{nbl} had different driver mutations in their ovarian tumor compared to those who had more pSCNAs^{nbl}. To ensure that our results are not due to potential covariates, we repeat the analyses after choosing alternate pSCNA^{nbl} threshold (**Supplementary Figure S9A-B**). We also repeated the analyzes after grouping the samples by age and stage, and only considering the patients who are of age between 50 and 69 years, and had tumor of stage II or III; there were 71 such cases in the filtered dataset (**Supplementary Figure S9C**). The results were consistent in all the three cases.



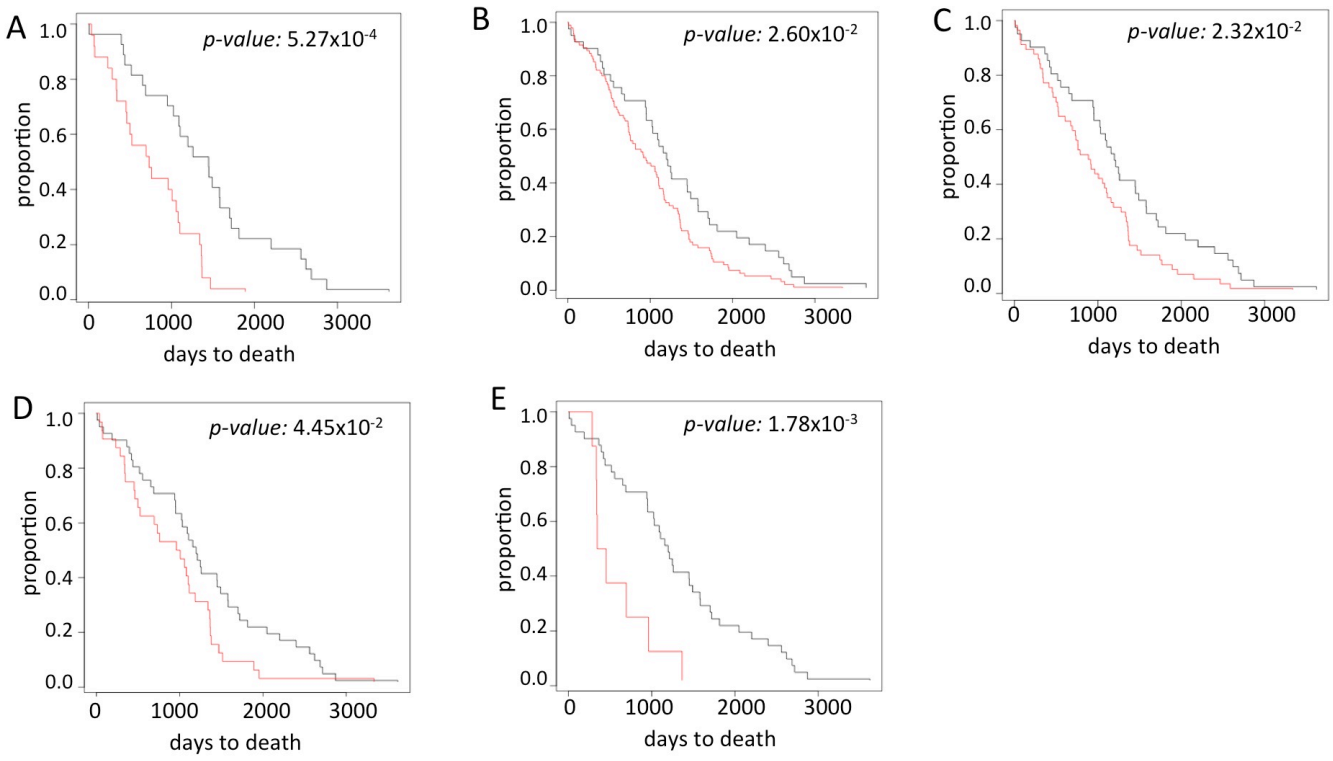
Supplementary Figure S9: Frequency of the common cancer gene mutations in ovarian cancer patients (A) who had no pSCNA^{nbl} (black) and those who have ≥3 pSCNA^{nbl} (grey), (B) who had no pSCNA^{nbl} (black) and those who have ≥5 pSCNA^{nbl} (grey), and (C) who were of age between 50 and 69 years, had stage II or III cancer, and had no pSCNA^{nbl} (black) and those who have ≥3 pSCNA^{nbl} (grey),.

Survival analysis after controlling for potential covariates

In the main text and Fig-3B, we reported that the pSCNA^{nbl} predicted survival. To ensure that our results are not due to potential covariates, we adjusted for age, stage, and tumor purity using a cox-proportional hazards regression model (survival R Package, method=breslow). Pathological tumor purity for these samples was obtained from the TCGA ovarian cancer clinical datasets (TCGA, 2011). We found that, even adjusting for these covariates the patients with ≥4 pSCNAs^{norm} in peripheral blood had significantly shorter survival (p-value $4.3x10^{-3}$).

In a complementary analysis, we focused on the patients who are of age between 50 and 69 years, and had tumor of stage II or III, there were 108 such cases in the filtered dataset. Using Kaplan Meier survival analysis, we found that the patients with ≥4 pSCNAs^{norm} in peripheral blood had significantly shorter survival (log rank test; p-value: $5.27x10^{-4}$; **Supplementary Figure S10A**) compared to those with no pSCNAs detected in blood. The results were consistent in other age- or stage groups, but the sample sizes were too small for statistically meaningful comparison.

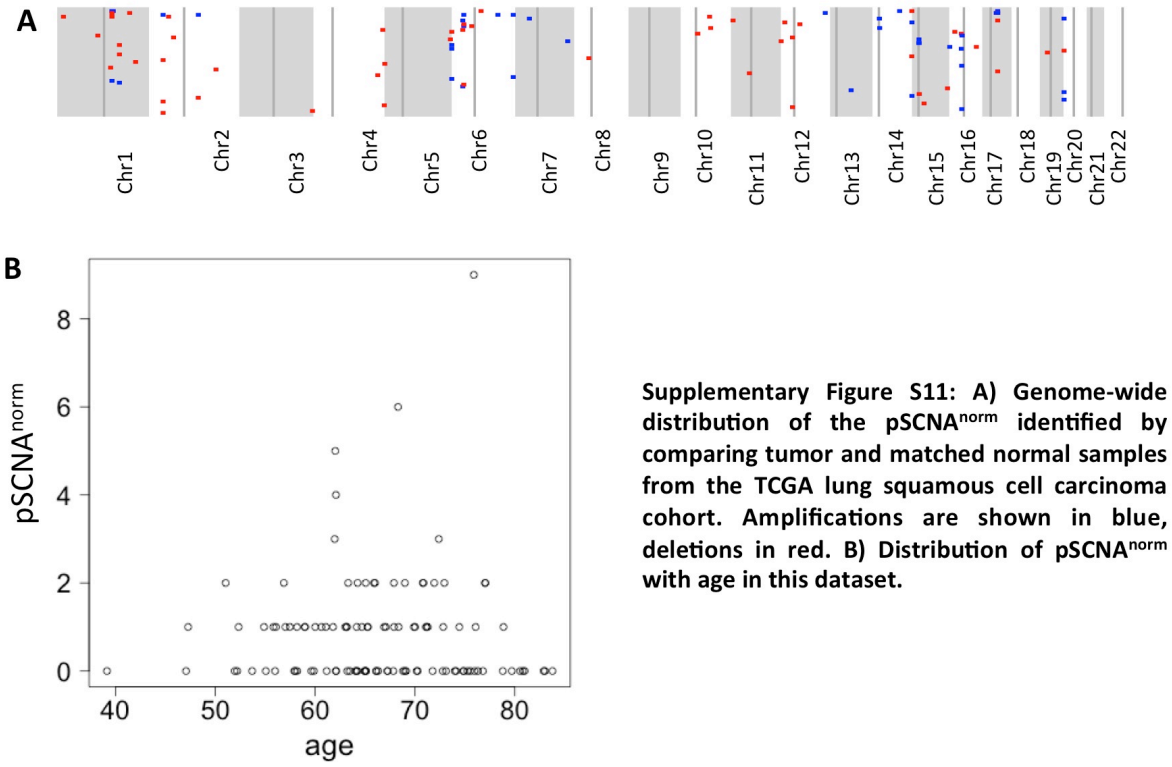
We then repeated the analyses described main text and Fig-3, after changing the pSCNA^{norm} threshold. Using Kaplan Meier survival analysis, we found that compared to those with no pSCNAs detected in blood the patients with ≥1, ≥2, ≥3 or ≥5 pSCNAs^{norm} in peripheral blood had significantly shorter survival (**Supplementary Figure S10B-E**).



Supplementary Figure S10: Kaplan Meier curve showing difference in the survival patterns between the ovarian cancer patients who have (A) no pSCNA^{nbl} (black) and those who have ≥4 pSCNA^{nbl} (red), after considering only the patients of age between 50 and 69 years, and stage II or III, (B) no pSCNA^{nbl} (black) and those who have ≥1 pSCNA^{nbl} (red), (C) no pSCNA^{nbl} (black) and those who have ≥2 pSCNA^{nbl} (red), (D) no pSCNA^{nbl} (black) and those who have ≥3 pSCNA^{nbl} (red), and (E) no pSCNA^{nbl} (black) and those who have ≥5 pSCNA^{nbl} (red).

Supplementary Module 6: Prevalence and significance of pSCNA^{norm} in lung cancer

It was challenging to find large cohorts where tumor and matched normal genomes were analyzed using more than one aCGH arrays/centers independently, and where aCGH calls were reasonably concordant between the experiments. The Cancer Genome Atlas lung squamous cell carcinoma dataset(TCGA, 2012) met both criteria, and we repeated some of the key analyses for lung squamous cell carcinoma. We obtained the aCGH-based copy number calls for tumor and matched normal tissue (apparently normal lung or peripheral blood) from the Cancer Genome Atlas. 110 samples had copy number analyzed using two independent centers: Agilent Human Genome CGH Custom Microarray 2x415K at Harvard medical School, and Agilent SurePrint G3 Human CGH Microarray at MSKCC. We processed the datasets as described in the Methods section of the Main-text (for ovarian cancer dataset). 18 and 92 of these samples had peripheral blood and lung tissue as the matched normals, respectively.



Supplementary Figure S11: A) Genome-wide distribution of the pSCNA^{norm} identified by comparing tumor and matched normal samples from the TCGA lung squamous cell carcinoma cohort. Amplifications are shown in blue, deletions in red. B) Distribution of pSCNA^{norm} with age in this dataset.

We detected a total of 95 amplifications and deletions (pSCNA^{norm}, per sample average: 0.8; **Supplementary Figure S11A**). In general, the relatively low number of pSCNA^{norm} could be due to tissue heterogeneity, or relatively weaker concordance in aCGH calls between the arrays. Nevertheless, most of the individuals with 2 or more pSCNA^{norm} had age ≥60 years, while those younger had on average fewer pSCNA^{norm} (**Supplementary Figure S11B**), supporting our original finding that the number of potential somatic genomic alterations detectable at a tissue-level resolution increases with age.

Analyzing genome-wide distributions of pSCNA^{norm}, we again found that the genomic neighborhoods of chr1q32, chr15q11, and chr17q21 had clusters of pSCNA^{norm}. Even though small size of the dataset precluded any rigorous statistical analysis, but this analysis suggested that these chromosomal regions might experience recurrent instability in apparently normal tissue types. Very few lung squamous cell carcinoma samples in our dataset had survival data (e.g. only one patient with pSCNA^{norm}≥4), so that survival analysis was not statistically meaningful.

References:

Abkowitz, J.L., Catlin, S.N., McCallie, M.T., and Guttorp, P. (2002). Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. Blood *100*, 2665-2667.

Bickel, P.J., Boley, N., Brown, J.B., Huang, H.Y., and Zhang, N.R. (2010). Subsampling Methods for Genomic Inference. Ann Appl Stat *4*, 1660-1697.

Bilke, S., and Gindin, Y. (2012). Analyzing the association of SCNA boundaries with replication timing. Nature biotechnology *30*, 1043-1045; author reply 1045-1046.

De, S., Pedersen, B.S., and Kechris, K. (2013). The dilemma of choosing the ideal permutation strategy while estimating statistical significance of genome-wide enrichment. Briefings in bioinformatics.

DeGregori, J. (2013). Challenging the axiom: does the occurrence of oncogenic mutations truly limit cancer development with age? Oncogene *32*, 1869-1875.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., *et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research *39*, D945-950.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z.P., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods *9*, 473-U488.

Karlin, S. (1966). Chapter 5: renewal processes. A First Course in Stochastic Process, 180.

Macdonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic acids research *42*, D986-992.

Michor, F., Hughes, T.P., Iwasa, Y., Branford, S., Shah, N.P., Sawyers, C.L., and Nowak, M.A. (2005). Dynamics of chronic myeloid leukaemia. Nature *435*, 1267-1270.

Sehl, M., Zhou, H., Sinsheimer, J.S., and Lange, K.L. (2011). Extinction models for cancer stem cell therapy. Mathematical biosciences *234*, 132-146.

Sima, J., and Gilbert, D.M. (2014). Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. Current opinion in genetics & development *25C*, 93-100.

Sugihara, G., May, R., Ye, H., Hsieh, C.H., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. Science *338*, 496-500.

TCGA (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

TCGA (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.

Voet, T., Kumar, P., Van Loo, P., Cooke, S.L., Marshall, J., Lin, M.L., Zamani Esteki, M., Van der Aa, N., Mateiu, L., McBride, D.J., *et al.* (2013). Single-cell paired-end genome sequencing reveals structural variation per cell cycle. Nucleic acids research *41*, 6119-6138.