Thermodynamic energetics underlying genomic instability and whole-genome doubling in cancer

Francoise Remacle[a,b], Thomas G. Graeber[c,d,e,f,g,1], and R. D. Levine[a,c,e,f,h,1]

[a]Fritz Haber Center for Molecular Dynamics, Institute of Chemistry, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel; [b]Theoretical Physical Chemistry, UR MolSys B6c, University of Liège, B4000 Liège, Belgium; [c]Department of Molecular and Medical Pharmacology, University of California, Los Angeles, CA 90095; [d]Crump Institute for Molecular Imaging, University of California, Los Angeles, CA 90095; [e]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095; [f]California NanoSystems Institute, University of California, Los Angeles, CA 90095; [g]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, David Geffen School of Medicine, University of California, Los Angeles, CA 90095; and [h]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095

Contributed by R. D. Levine, May 20, 2020 (sent for review November 27, 2019; reviewed by Samuel Bakhoum, Ilya Nemenman, and Gustavo Stolovitzky)

Genomic instability contributes to tumorigenesis through the amplification and deletion of cancer driver genes. DNA copy number (CN) profiling of ensembles of tumors allows a thermodynamic analysis of the profile for each tumor. The free energy of the distribution of CNs is found to be a monotonically increasing function of the average chromosomal ploidy. The dependence is universal across several cancer types. Surprisal analysis distinguishes two main known subgroups: tumors with cells that have or have not undergone whole-genome duplication (WGD). The analysis uncovers that CN states having a narrower distribution are energetically more favorable toward the WGD transition. Surprisal analysis also determines the deviations from a fully stable-state distribution. These deviations reflect constraints imposed by tumor fitness selection pressures. The results point to CN changes that are more common in high-ploidy tumors and thus support altered selection pressures upon WGD.

aneuploid | genomic instability | whole-genome doubling | free energy | surprisal analysis

Tumorigenesis is a complex process with multiple alterations combining together to rewire the circuitry of the cell toward unrestrained growth and survival. Cellular loss of genomic stability programs can accelerate tumor evolution. Genomic instability at the chromosome and subchromosome level is a hallmark of aggressive and lethal cancers. Oncogene amplification and tumor suppressor gene deletion are classical examples of DNA copy number alterations (CNAs) contributing to tumor phenotypes. More recent high-resolution and high-throughput studies of tumor cohorts have revealed highly recurrent patterns of CNAs across the full genome, including both shared and tumor type-specific patterns (1, 2). These recurrent patterns point strongly to the shaping of the genome by conserved selection forces related to increases in tumor phenotype-linked fitness.

Many processes such as DNA damage repair defects, replication stress, breakage-fusion-bridge cycles, telomere attrition, lagging chromosomes, nuclear envelope defects, defective mitosis, and epigenetic-guided mechanisms can lead to CNA (2–6). Defective mitosis can lead to whole-genome duplication (WGD), which can also be observed in experimental systems (2, 7, 8). Analysis of somatic and germline single-nucleotide polymorphism (SNP) patterns supports WGD as typically an early event in the genomic instability components of tumorigenesis (2, 7, 9, 10). Both elevated CNA and WGD events are linked to poor prognosis (2, 9, 11, 12).

The fragmented but recurring patterns of tumor genomes are natural candidates for entropy-based analysis. Fragmentation of atomic nuclei (2, 13), molecules (2, 14), clusters (15, 16), aggregates (17), china plates, and so on has been discussed using different information theoretic points of view. In biology, the fragmentation of the nucleus of a dying cell, known as "karyorrhexis," has long been studied.

Here we seek to characterize CNAs—often summarized as the number of times a DNA locus is present in a cellular genome. Experimentally, it is easier to measure the variation per DNA mass or to measure on a relative basis, in part because sample input to microarray or sequencing assays is normalized based on DNA amount, and not on cell number (18–20). Analysis pipelines for inferring absolute copy numbers (CNs) from allelic frequencies have been established (21–23). Likewise, the ploidy, or average chromosome copy number of a cell, is also typically inferred. Such inferences are not free of error (10, 21), reflecting among other aspects the purity and the possible heterogeneity of the sample.

The availability of rather large datasets providing CNAs for many hundreds of patient tumors and the availability of such data for different cancer types makes an analysis a worthy challenge. Our aim here is to quantify the free energy changes that accompany CNs and to relate them to biological processes taking place in the tumors such as the WGD and the effects of fitness selection pressures on the genome.

For each tumor one now has the CN of the different genes as shown in *SI Appendix*, Fig. S1. It is a single peak distribution characterized by a mean CN for that tumor, the ploidy, and the fluctuations about the mean; see *SI Appendix*, Fig. S2 for several distinct cases. Somatic cells can have a ploidy different from two. One often represents the CN of a gene in a given tumor as the ploidy of the tumor times the CNA. In log space, the CNA is thus the additive deviation from the mean and it can be positive or negative. By definition the mean value of the additive deviation is zero, but the range of CNA can be large so that the histogram

Significance

Genomic instability is characteristic of the majority of cancers. It includes changes of copy numbers of genes and chromosomes during diverse cell processes, such as genome replication. Genomic instability is a strong biomarker of poor prognosis. An analysis of DNA copy numbers in human tumors across different cancers reveals an instability in an energetic sense: The free energy of the distribution of gene loci copy numbers in each tumor is found to be a monotonically increasing and universal function of the mean chromosome ploidy in that tumor. This relates the biological (genomic instability) with the physicochemical stability criterion. The analysis also shows that it is energetically more favorable for genome doubling to occur from a not-already-fragmented genome.

Author contributions: F.R., T.G.G., and R.D.L. designed research, performed research, analyzed data, and wrote the paper.

Reviewers: S.B., Memorial Sloan Kettering Cancer Center; I.N., Emory University; and G.S., IBM Research.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: tgraeber@mednet.ucla.edu or rafi@fh.huji.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1920870117/-/DCSupplemental.

First published July 21, 2020.

of the CNs of a given tumor can be broad, especially so when there are gains or losses of a subset of whole chromosomes. A number of histograms for tumors of a given cancer are centered about a ploidy of two, which is also the case for control (normal) tissue samples. Often the histogram is broader and centered about a ploidy of four. These distributions support that the genome is typically first doubled and then fragments, as has been reported based on allele polymorphism analysis (7, 9, 10). A small group of tumors have a very broad distribution without a clear most probable value and with a mean significantly above four. The width and the typical asymmetry of the histogram are determined by the second and higher moments of the distribution of the CNs. As is to be expected on general grounds, these higher moments are more sensitive to any errors in the absolute quantification as compared to the mean. The data of Carter et al. (21), and see also Zack et al. (10), suggest that the rms error of the mean CN, the ploidy, is roughly 0.6. While this error is large as compared to the two means of two and four it is not large enough to hide the difference between the histograms centered at these two values.

To analyze the energetics of the fragmentation we use the concept of free energy. The free energy is a measure along the path in the genome full energy landscape (24–26). Here we determine the free energy and its dependence on the ploidy by surprisal analysis (27) of the distribution of CNs of a particular tumor. This route has been previously used to analyze the distribution of transcripts as measured for different individual tumors (28), including both diseased and control (normal tissue) states. Transcripts are messenger RNA (mRNA) molecules and so are individual chemical species. Here the distribution is the number of copies of a particular gene locus on the genome. Such a distribution does not necessarily represent a system in equilibrium. It is therefore possible and indeed appropriate to distinguish two contributions that together make the free energy of the system. To do so we refer to an intermediate distribution, a distribution that in biology is often not known from experiment but it can be generated from the experimental data during the analytical process of surprisal analysis. This is the distribution of the system when it is in a balanced state where each possible forward process is matched by a process in the reversed direction. Such a balanced state is stable to small perturbations and so it is also referred to as the stable state. The free energy of the system in its stable state is the major contribution to the total free energy. The other and smaller part of the total is the minimal work that is necessary to bring the system from the stable state to the actual state it is in. Computing this free energy in terms of the distribution of the species is discussed in detail in ref. 29. Here we compute the free energy in terms of the distribution of CNs. A summary of the thermodynamic description of the stable state and the deviations from it is provided in *SI Appendix*, section S1. A technical discussion for the multispecies situation common in biology can be found in ref. 30.

Typically, a system in a state of nonequilibrium, if left on its own, will change. To maintain a state of nonequilibrium one needs to impose constraints on the system. It is these constraints that are identified by surprisal analysis. It is these constraints that prevent the free energy from spontaneously going down toward its minimal value that is reached at the state of equilibrium. If we allow the constraints to relax the free energy can go down in value. The free energy acts as a potential that directs the motion toward the stable state where the free energy is a minimum. This means that in a spontaneous time evolution of the system it moves to a more thermodynamically stable lower-energy state. Explicit biological examples are provided, for example, in refs. 31, 32.

We conclude the paper with a free energy analysis of the energetics underlying WGD. WGD occurs due to a failed mitosis following the S phase in the cell cycle. Our analysis identifies the biological meaning of the dominant constraint on the genome as governing the extent of fluctuations of the CNs about their value in the stable state. The distribution can be broader or narrower and we show that this corresponds to two sides of a biological transition. Our detailed result is that the transition from a diploid case to a whole-genome doubling favors cells with a narrower distribution. We show that it takes additional energy to double the genome in cells with a broad distribution of CNs. Qualitatively and quantitatively we find the very same transition to WGD in all of the seven cancer types that we analyzed.

Results

Surprisal Analysis.

Data. The data in the literature are often given as the CN of a gene with respect to the mean CN, the ploidy, of that tumor. The actual CN requires scaling by the ploidy, CN $= \langle p \rangle$CNA. The ploidy is determined, for example, from the allelic imbalances of SNPs in the genome (21, 22). For a given tumor many genes have a CN of approximately two or four. A value of around four supports that the tumor's history includes a WGD. In some cases, the CN is different from the mean only for rather specific genes. An example is shown for a particular diploid tumor with breast cancer in *SI Appendix*, Fig. S1*A*. In this case, only few genes deviate from a CN of two. These deviant genes belong to specific affected chromosomal regions. Two more examples of increasingly fragmented genomes are shown in *SI Appendix*, Fig. S1 *B* and *C*. Both reflect WGD. One example is a duplication or two whole genomes, with a mean ploidy of 3.8, and one is a fourfold duplication, four whole genomes with a mean ploidy of 6.9. The histogram of CNs as shown is peaked about the ploidy (*SI Appendix*, Fig. S2) with a width smaller than the mean. In a stable state of the cell we therefore expect a priori that the CNs are about equal for all genes but possibly fluctuate about a mean equal to the ploidy. One can also say that the most probable CN is the ploidy. We do not, however, impose these expectations but recover them from the surprisal analysis. The results of the stable state as determined by surprisal analysis for the CNs in the breast cancer tumors are also shown in *SI Appendix*, Fig. S3*A*. It is the same tumor whose distribution is shown in *SI Appendix*, Figs. S1*C* and S2*C*. As can be seen in *SI Appendix*, Fig. S3*A* the stable state is not the same as a strictly flat genome where all genes have the very same CNs. Rather, there are fluctuations of the CNs about the ploidy. The fluctuations of the CNs in the stable state are often quite small for a tumor with a ploidy of about 2. However, the fluctuations tend to be higher for higher ploidy (*SI Appendix*, Fig. S3*B*). This is to be expected from the results for the stable state as determined by surprisal analysis and the correlation between more fluctuations and higher ploidy is derived in *SI Appendix*, section S2.

The surprisal. Surprisal analysis is used to quantitatively represent the distribution of measured CN for each gene locus i, $X_i(n)$ in a given tumor, labeled by tumor index n. In our analysis we use gene loci as a representative set of DNA loci spanning the genome (1). We use the expression for a distribution of minimal free energy subject to constraints. This is a standard approach in statistical thermodynamics and leads to the generic expression $X_i(n) = X_i^o(n)\exp((\mu_i(n) - \mu_i^o(n))/kT)$. $X_i^o(n)$ is the value of the CN of gene i of tumor n in the stable state. The exponential factor is a multiplicative term representing the deviation from the stable state. $\mu_i(n)$ is the chemical potential of locus i in tumor n and $\mu_i^o(n)$ is the potential in the stable state. The thermodynamic potential has the units of energy and so it is scaled by the thermal energy $kT$to make the fold change a dimensionless quantity. k is Boltzmann's constant. When $\mu_i(n) > \mu_i^o(n)$ the CN of locus i is higher than in the stable state. (For applications of similar expressions to the distribution of transcript expression levels see, e.g., refs. 27, 28. See ref. 30 for a detailed statistical

mechanics derivation. See *SI Appendix*, section S1 for a brief review of the thermodynamics.)

The primary mathematical output of minimizing the free energy is an expression for the CNs as fold deviation from the stable state, $\ln(X_i(n)/X_i^o(n))$. It can be written as an additive deviation from the stable state:

$$\underbrace{\ln X_i(n)}_{\substack{\text{logarithm of CN}\\ \text{of locus } i \text{ in tumor } n}} = \underbrace{\ln X_i^o(n)}_{\substack{\text{logarithm of CN}\\ \text{of locus } i \text{ in tumor } n\\ \text{in the stable state}}} + \underbrace{\left(\mu_i(n)-\mu_i^o(n)\right)/kT}_{\substack{\text{difference in chemical potentials}\\ \text{of locus } i \text{ of tumor } n \text{ between}\\ \text{the observed and the stable state}}} . \quad [1]$$

The free energy of tumor n with respect to its stable state is $\Delta G(n) = \sum_i X_i(n)(\mu_i(n) - \mu_i^o(n))$. We determine the stable state and the chemical potentials by fitting the right-hand side of Eq. **1** to the data for many tumors. We do so by relating the deviation of the actual expression level from its stable value to the operation of constraints on the system. We do not use the value of $X_i^o(n)$ that we expect but use surprisal analysis to extract it from the data for each gene and tumor. Eq. **1** is derived at a fixed temperature and pressure. The temperature T is a given constant and not a variable.

The constraints that confine the system to its present state are labeled by the index $\alpha = 1,2,..$ We impose the very same constraints on all of the tumors of a given cancer type, supported by the highly recurrent pattern of CNs seen across tumors from multiple patients. Biologically, these recurrent patterns reflect shared tumor fitness selection pressures related to the hallmarks of cancer (e.g., increased proliferation, decreased growth, and angiogenesis). $G_{i\alpha}$, the value of constraint α, is only a function of the gene index i but not of the tumor index n. The constrained quantity is the mean value of the constraint over the CNs $\langle G_\alpha\rangle(n) = \sum_i G_{i\alpha}X_i(n)$. We minimize the free energy of the system relative to its stable state. Using the well-established technique of Lagrange multipliers to minimize the free energy subject to the values of the constraints leads to an explicit expression for the CN $X_i(n)$ of locus i in tumor n. Written in terms of the chemical potentials of the constraints this reads

$$\ln X_i(n) = \ln X_i^o(n) + \sum_{\alpha=1,2,..} G_{i\alpha}\left(\mu_\alpha(n)/kT\right). \quad [2]$$

A derivation of this equation is outlined in *SI Appendix*, section S1. It is this equation that is directly fitted to the data by surprisal analysis. Eq. **2** determines the chemical potential of locus i in terms of the constraints is $\mu_i(n) = \sum_{\alpha=0,1,..} G_{i\alpha}\mu_\alpha(n)$. The chemical potentials are related to the Lagrange multipliers used in the process of minimizing the free energy as $\mu_\alpha(n) = -kT\lambda_\alpha(n)$. What we aim to do is a thermodynamic analysis of a system of many components (a grand canonical ensemble). Therefore, we analyze the CNs of each gene locus, and not the CNAs.

The problem in fitting Eq. **2** is that in biology one often does not know beforehand the values $G_{i\alpha}$ of the constraints for each locus. We use an approximation that is often good enough, where by "good enough" we mean that the inferred values of the Lagrange multipliers are within the inevitable experimental error. The error is estimated by the procedure given in *SI Appendix*, section S3 from the errors reported in the data and is quite large for the currently available measured CNs. To implement surprisal analysis, we use the mathematical technique of matrix diagonalization known as singular value decomposition, SVD (*Methods*).

The two terms in the free energy. Surprisal analysis provides the two components of the free energy that we need to characterize the system. First, surprisal analysis as in Eq. **2** determines the distribution of CNs at the stable equilibrium state. From that distribution we have the free energy at equilibrium for tumor n, $\langle \ln X_i^o(n)\rangle$, where the mean is over all loci (indexed by i). Thereby the free energy is computed for each tumor n. It is convenient to define a chemical potential $\mu_0(n)$ that characterizes the distribution of CNs at equilibrium by the definition $\ln X_i^o(n) \equiv G_{i0}(\mu_0(n)/kT)$. This leads to an alternate form of Eq. **2**,

$$\ln X_i(n) = \sum_{\alpha=0,1,2,..} G_{i\alpha}\left(\mu_\alpha(n)/kT\right). \quad [3]$$

In regard to nomenclature, the zeroth constraint, $\alpha = 0$, is the stable state or balanced state, and the first constraint is the dominant constraint. The chemical potential of locus i in the stable state, cf. Eq. **3**, is $\mu_i^o(n) = G_{i0}\mu_0(n)$.

Surprisal analysis also determines the free energy needed to bring the distribution of CNs from equilibrium to its present state, a state under constraints and therefore a state of higher free energy, $\langle \ln(X_i(n)/X_i^o(n))\rangle$.

In terms of the Lagrange multipliers we write the free energy of the genome in its stable state as $kT\langle \ln X_i^o(n)\rangle = \mu_0(n)\langle G_0\rangle(n)$. The values of $\mu_0(n)$ and G_{i0} are determined by the SVD of the matrix of the logarithm of the data, $\ln X_i(n)$, where the different columns are different tumors n. Our notation emphasizes that the value of the constraint does depend on the tumor index. The chemical potential, dimensions work, is related to the Lagrange multiplier as $\mu_0(n) = -kT\lambda_0(n)$. $\lambda_0(n)$ corresponds to the largest eigenvalue of the SVD matrix of the data $\{\ln X_i(n)\}$. The free energy of the stable state, $\ln X_i^o(n) = G_{i0}(\mu_0(n)/kT)$, is therefore the dominant term in Eq. **3**.

The free energy needed to bring the distribution of CNs from the stable to its actual state is $\sum_{\alpha=1,2,..}\mu_\alpha(n)\langle G_\alpha\rangle(n)$. Again, the values $\mu_\alpha(n)$ and $G_{i\alpha}$, $\alpha = 1,2,..$ are determined by the singular-value decomposition. The free energy of the genome in its actual state is the sum, $\sum_{\alpha=0,1,..}\mu_\alpha(n)\langle G_\alpha\rangle(n)$.

A histogram of the measured CNs can be quite tight about its peak (see *SI Appendix*, Fig. S2), implying that the deviations from the stable state are limited. Therefore, the work required to rearrange the system from its stable state, $\sum_{\alpha=1,2,..}\mu_\alpha(n)\langle G_\alpha\rangle(n)$, a free energy that can also be written as $\sum_i X_i(n)\ln(X_i(n)/X_i^o(n))$ meaning the average of the deviations from the stable state, will not be nearly large as compared to the free energy of the stable state itself (further discussed below).

Analysis of the CN distribution. The typical dataset contains the CNs of about 20,000 genes in many hundreds of patient tumors for each cancer (33). Performing surprisal analysis on data for a particular cancer determines the stable balanced state and the deviations as spelled out in Eq. **1** for all of the tumors of that cancer. The complete set of results of the analysis of a particular cancer are the chemical potentials of each constraint for each tumor, $\mu_\alpha(n)$, $\alpha = 0,1,2,..$, and the values of the constraints, $G_{i\alpha}$, for each location i.

Fig. 1*A* is the histogram of the ploidy in a cohort of breast cancer patients. Fig. 1*B* shows the value of the chemical potential of the stable state $\mu_0(n)/kT$ for each tumor n in breast cancer vs. the ploidy of that tumor. Shown also are the relevant error bars. These reflect two quite distinct sources. One is the error in the estimate of $\mu_0(n)$ which reflects the error in the experimental reading of the CN. Shown is an upper bound on the error of $\mu_0(n)$ determined as for other applications of surprisal analysis (34) and summarized in *SI Appendix*, section S3. As is evident from Fig. 1 the upper bound on this error, while not negligible, is quite small as compared to $\mu_0(n)$ itself. The convex trend of $\mu_0(n)$with the ploidy is therefore secure. Indeed, as shown in *SI Appendix*, Fig. S4 the very same trend is rather exactly the same for all of the seven cancer types that were examined individually. The other error in $\mu_0(n)$ is the error implied by the procedure of estimating the ploidy. It is an error in placing the particular tumor along the abscissa. This error plays an important role in plotting the tumor dependencies because the ploidy of a particular tumor is the prime descriptor of that tumor. We use an rms error of 0.6 as

discussed in ref. 21. As is clear from Fig. 1, even this high value does not mask the clear convex trend of with the ploidy.

Together with $\mu_0(n)$ surprisal analysis also yields the values of the constraint of equilibrium, G_{0i}, in the different loci i. Intuitively we expect that the actual CNs of different loci do not differ much from the ploidy of that tumor. Therefore, the CN of any specific locus should be about the mean except for a very few highly deviant ones. We verify this expectation from the results of the analysis. A typical example is shown in *SI Appendix*, Fig. S3.

The chemical potential for the dominant constraint, $\mu_\alpha(n)$, $\alpha = 1$ is shown in Fig. 1*C*. The effect of errors is now more evident. Yet, it is clear that even with the larger error bars on the ploidy there are three groups of tumors. There is the diploid set centered at a ploidy of about 2, the WGD cluster centered not quite about 4 but clearly centered nearby, and a third set with the most extreme ploidy. The theoretic significance of this identification is to be judged from the error bars on the values of $\mu_1(n)$. The potentials are determined as Lagrange multipliers. They arise in imposing the constraints. If a multiplier has the value zero it is the same as saying this constraint need not be imposed as this constraint does not constrain the value of the free energy. If an error bar spans a value of zero, the constraint is therefore not warranted by the data because of the possible errors. The magnitude of the error in $\mu_1(n)$ for this group of breast cancer tumors is ± 7. As can be seen in the plot for many tumors the error bar spans the value of zero. For those tumors one cannot say that the constraint is needed. However, for many others, and in all three groups of tumors there are cases for which the constraint is required.

Tumors can have both positive and negative values of $\mu_1(n)$ showing that deviations of the CN from the stable state can be in both directions. In thermodynamics a change in sign of a chemical potential is a transition in the state of the system so we need to identify this transition. Indeed, we will further argue below that the tumors where the deviations are negative, $\mu_1(n) < 0$, have a special role.

Pan-cancer results. Surprisal analysis was implemented separately on data files for seven different cancer types, breast invasive carcinoma (BRCA), ovarian serous cystadenocarcinoma (OV), colon adenocarcinoma (COAD), glioblastoma (GBM), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and bladder urothelial carcinoma (BLCA). For all seven cases the free energy vs. ploidy is the same monotonically rising plot, Fig. 2. The monotonic increase of the free energy is evidence for increasing thermodynamic instability as the ploidy deviation from normal becomes more extensive. The results shown are the free energies of the stable states, $(\mu_0(n)/kT)\langle G_0\rangle(n)$, plotted vs. the ploidy of tumor n for all seven cancers, each one identified in the legend by the color used. This commonality is already anticipated in that $\mu_0(n)$(see *SI Appendix*, Fig. S4) is itself already the same function of the ploidy for all cancers studied. The density of symbols as a function of ploidy in *SI Appendix*, Fig. S4 and in Fig. 2 is not uniform. Since tumors of different cancers are plotted in different colors this shows that some cancers have a few more tumors with particularly low or high ploidy. However, to within the error bars on the ploidy (Fig. 1) we cannot draw definite and secure conclusions about intracancer variability.

The plot of the free energy $(\mu_0(n)/kT)\langle G_0\rangle(n)$ vs. ploidy shows a high density of symbols about ploidies of 2 and 4 because of the higher number of tumors about these two values (compare Figs. 1*A* and 2). One can also plot the free energy of the actual distribution of CNs. The stable state is so stable that adding the work needed to bring the system to its actual ploidy is not very clearly seen on such a plot (*SI Appendix*, Fig. S5). It is a noticeable bump about ploidies of 2 and 4. We will, however, have more to say about this energy but first to the main component, the work required to alter the genome. This is shown for tumors of two distinct cancer types, breast and GBM in Fig. 3. For clarity the error bars are not shown but are shown in *SI Appendix*, Fig. S6. For both cancers one clearly sees three groups of tumors. Diploid, more common for GBM, the spread about ploidy of 4 presumably reflecting WGD, and some rarer tumors with higher ploidy where it is not possible to categorically say if it reflects a fourfold process.

Fig. 3 shows the work needed to deviate the system from its stable state, $\langle \ln(X_i(n)/X_i^o(n))\rangle$ (discussed above), where the averaging is over the locus index i for each tumor n. Each tumor is labeled by its ploidy and the plot is vs. this ploidy. Qualitatively the plot looks the same if, say, just the contribution of the one or two most important constraints are plotted.

The Landscape and Energetics of the Stable State. The next sections investigate the CN patterns linked to the surprisal analysis results, for both the stable state and the primary constraints, as well as the thermodynamic implications for the energetics of DNA instability. From Eq. **1**, the log of the DNA profile of a tumor n is its stable state value $G_{i0}\mu_0(n)/kT$ plus the sum over all constraints $\sum_{\alpha=1,2,..} G_{i\alpha}\mu_\alpha(n)/kT$. First and foremost, $\mu_0(n)$ scales with the ploidy of tumor n, and this contribution to the DNA profile reflects that a higher ploidy sample has more DNA on average across its full genome (Fig. 1). This pattern typically well reflects the "dominant CNA pattern" for a particular tumor type. The dominant CNA pattern for any tumor type is readily visible in a genome-view heat map of a cohort of tumor samples, as well as in a genome-oriented view of G_{i0} (both shown in Fig. 4*A*). As expected, the most extreme weights in the G_{i0} profile include known oncogene amplifications (e.g., *MYC* and *TERC*) and known tumor suppressor gene deletions (*NF1* and *TP53*) on the opposite side (Fig. 4*B*).

We have identified a dominant zeroth eigenvalue throughout our work in biology (e.g., ref. 27). This reflects the stability of the equilibrium state against small pertrubations. In a previous paper dedicated to the analysis of mRNA expression levels we showed that the processes present in the stable state are common to cells across different types of cancers and even across of organisms (35). Here, we determine a dominant constraint because of the stability of the stable state and as in gene expression levels, a predominant portion of the stable state is common across different cancers. The results of a pan-cancer dominant constraint support the hypothesis that the human genome cannot be too disrupted by available genomic instability mechanisms and still support cellular life.

The free energy of the CN distribution has two contributions, the equilibrium free energy of the stable state and the free energy resulting from the work done by the constraints. The by far larger contribution is the equilibrium free energy of the stable state of the gene CNs, as shown in Fig. 2. Comparing the stable state free energy in units of the thermal energy kT, (Fig. 2; $\sim 10^4$ to 10^5 for ploidy 2 to 4) to the free energy contribution from the dominant constraint (see Fig. 7; $\sim 10^2$), we see that the contribution from the dominant constraint is two to three orders of magnitude smaller. The work needed to bring the genome from the stable state to its actual fragmented state, the contribution from all constraints ($\alpha = 0, 1, 2, \ldots$), is of order $\sim 10^3$ (Fig. 3). The stable state free energy function is common for all of the seven cancers that we examined. It is a monotonic, convex function of the ploidy. The result that the free energy of the stable state is an increasing function of the ploidy has potential, nonexclusive, physical interpretations. For example, the free energy would be expected to increase with ploidy since after a whole-genome doubling there are twice as many nucleotides arranged into a specific, ordered DNA code. Additionally, the increasing free energy could reflect an aspect analogous to pressure, as the nucleus becomes crowded with DNA it takes extra energy to add even more. The increase in equilibrium free energy is potentially provided by the cell machinery during each

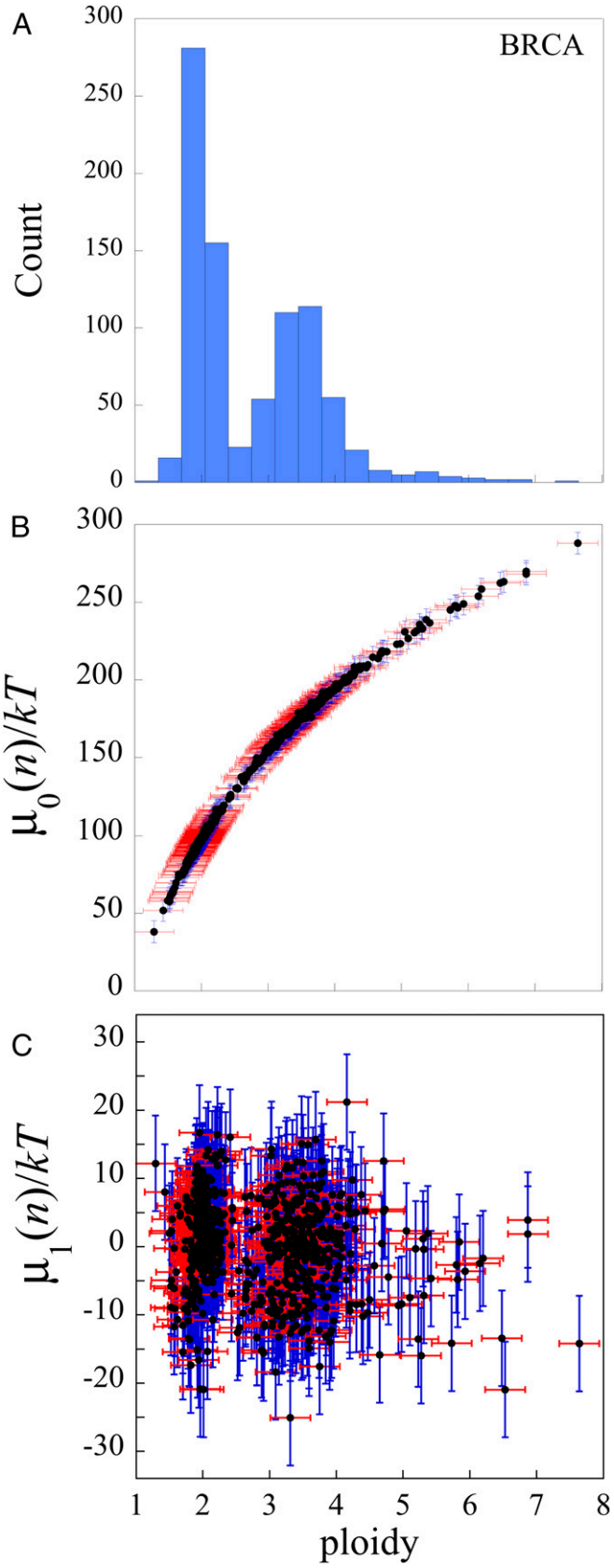

Fig. 1. The chemical potential of the stable state and of the dominant constraint as a function of ploidy. (*A*) Histogram of ploidy for a cohort of 862 BRCA tumors. (*B*) The chemical potential of the stable state $\mu_0(n)$ over kT for each tumor n in the BRCA cohort vs. the ploidy of that tumor. kT is the unit of thermal energy where k is Boltzmann's constant. (*C*) The chemical potential of the first constraint $\mu_1(n)$ over kT for each tumor n vs. the ploidy of that tumor. Two error bars are shown for each tumor: the error in the determination of the ploidy, an error in the abscissa, and the error in the chemical potential due to the error of measuring the CN of the genes. Estimation of the errors is discussed in *SI Appendix*, section S3. [See also *SI Appendix*, Figs. S1–S4. It is shown in *SI Appendix*, Fig. S5 that the plot of the chemical potential of the stable state $\mu_0(n)$ against the ploidy is very much the same for all of the seven types of tumor analyzed.]

cell cycle, with the machinery, for example, bringing in the nucleotide building blocks and linking them in the correct DNA sequence order. In the case of a failed mitosis, this approximate doubling of free energy remains with the resultant WGD cell, rather than being divided among two daughter cells.

The Landscape of the Dominant Constraint. Next in our analysis of the surprisal results, we see that the dominant constraint, the μ_1 term in Eq. **2**, plays two roles. Primarily μ_1 refines the deviation from the ploidy-proportional dominant-pattern scaling defined by μ_0 (the balanced state pattern). Specific examples of how negative and positive μ_1 values contribute to overall CNA patterns are shown in Fig. 5 *A* and *B*. The sign of the first constraint also plays a key role in the understanding of the energetics of the WGD. We turn next to a discussion of this role. We will further continue with this theme later.

On a locus-by-locus basis, how are the fluctuations of the CNs about their mean reduced when μ_1 is negative? This has to do with a property of the CN distribution $[\ln(X_i)]$ which is not commonly observed for other distributions such as those of transcript levels (28, 32). In surprisal analysis, the weights of the stable state for locus i satisfy $\sum_{i=1}^{N} G_{i0}^2 = 1$ so thatG_{i0} has an average value of $1/\sqrt{N}$ with typically limited fluctuations about the average. We designate these fluctuations as δG_{i0}. The unexpected result, not commonly seen in other biological profiles, is that for most/all tumor types the fluctuations $\delta G_{i0}$are more or less the variation of the values of G_{i1}s about their mean value of zero. In short, $\delta G_{i0} \sim G_{1i}$, and both generally reflect the "dominant CNA pattern" for that tumor type (Fig. 4 *A* and *B*). Thus, assuming that the first constraint dominates the other constraints (so that they can be, to first order, neglected), the change in CNs in tumor n along its genome is $\ln(\mu_0(n)(N^{-1/2} + \delta G_{i0}) + \mu_1(n)G_{1i}) \cong \ln(\mu_0(n)N^{-1/2} + (\mu_0(n) + \mu_1(n))\delta G_{i0})$. Since $\mu_0(n)$ is positive (see Fig. 1), if $\mu_1(n) < 0$ the fluctuations along the genome are reduced, while if $\mu_1(n) > 0$ the fluctuations are enhanced. Again, illustrations of these two cases are shown in Fig. 4*A*. The change of state corresponding to the change in sign of μ_1 (negative to positive) is the change from a more flat to a more fluctuating genome. For the subsets of tumors where the mean CN (the ploidy) is about the same, the variance of the CN increases as μ_1 increases (*SI Appendix*, Fig. S7). For brevity we use WG1 (whole genome 1) to indicate the subset of tumors with ploidy close to 2 and WG2 for tumors with ploidy close to 4 (*Methods*). When additional constraints beyond the first significantly contribute, the trend can be graphically somewhat less clear but it remains.

μ_1 and higher constraints can play a second role, adding into the DNA profile description (Eq. **3**) any high ploidy- or low ploidy-biased amplification or deletion events. Deviations from the overall correlation of the G_{i0} and G_{i1} profiles provide a guide to these regions (Fig. 4*B*). These ploidy-biased amplifications and deletions can be on the chromosome scale (Fig. 6 *A* and *B*), or more focally defined (Fig. 6*C* and *SI Appendix*, Fig. S8), and are interpreted to reflect selection for amplifications and deletions that particularly benefit the fitness of non-WGD or WGD tumors. In OV and LUSC cancers, these aspects of the surprisal results point to focal amplification of CCNE1 (Cyclin E) (Fig. 6*C*) (9, 10) and KRAS (*SI Appendix*, Fig. S8*A*) and a Chr 22

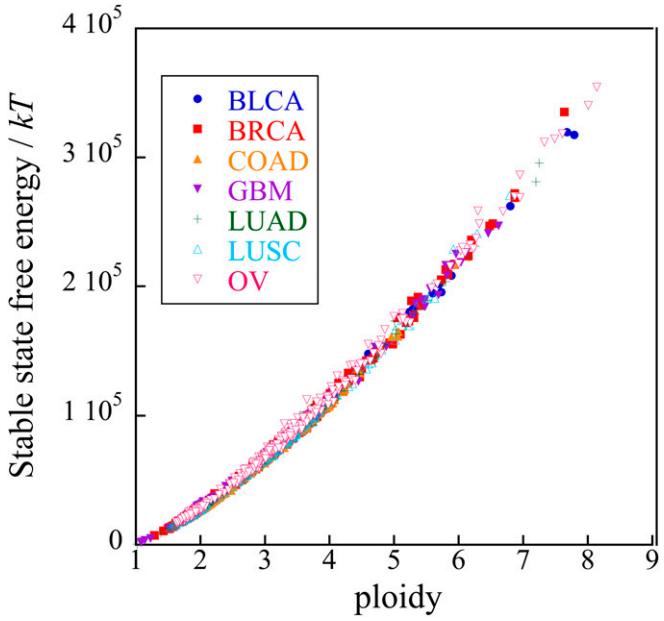

Fig. 2. Free energy of the stable state. The free energy of the stable state vs. ploidy for the seven patient tumor types investigated (BCLA, BRCA, COAD, GBM, LUAD, LUSC, and OV). The plot shows 3,140 patients color-coded for the particular cancer type. The free energy is determined by surprisal analysis. The free energy of the stable state is the product of the stable state (zeroth constraint) chemical potential $\mu_0$times the mean value of the zeroth constraint as derived in *SI Appendix*, section S1. In units of the thermal energy kT the expression of the free energy for tumor index n is $(\mu_0(n)/kT)\langle G_0\rangle(n)$ (see also *SI Appendix*, Fig. S5).

sub q-arm telomere-proximal deletion (*SI Appendix*, Fig. S8*B*) being more prominent in WG2 tumors as compared to WG1 tumors. We turn to show that the transition from a flatter to a more disordered genome is reflected in the propensity of the genome to double.

Free Energy Analysis of WGD. Thus far, we used a tumor-centric point of view. The quantitative statement is that of Eq. **1**. It characterizes the CNAs of different loci$\{i\}$ in a particular tumor in a particular cancer. Here we turn to an ensemble of all tumors of a particular cancer. If one had a dynamical theory or data over time it might have been possible to argue that the ensemble of all tumors at a given point in time is equivalent to a time history of the CNAs of one (or a few) tumors over time. This is consistent with the results shown in Figs. 1, 2, or 3. Many tumors are centered about a ploidy of 2, close to diploid (WG1), or 4, the results of a WGD (WG2). Few tumors are in the intervening space, with the density of tumors being minimal in the transition region between WG1 and WG2, say in the range 2.3 < ploidy < 2.7. Below we analyze the differences between the WG1 and WG2 subensembles.

Our first step is to determine the distribution of the work needed to deviate the CNAs of a tumor from its stable state. For any particular tumor n this is the mean value of the surprisal $\sum_i X_i(n)\ \ln\ (X_i(n)/X_i^o(n))$. The contribution of the first constraint, $\alpha = 1$, to this sum is from Eq. **2**, $(\mu_1(n)/kT)\langle G_1\rangle(n)$. The two terms in the product are not independent. As discussed in *SI Appendix*, section S4, surprisal analysis determines$\mu_1(n)$ as a function of $\langle G_1\rangle(n)$. The connection is one-to-one, so the opposite route is equally possible, to determine $\langle G_1\rangle(n)$ as a function of $\mu_1(n)$. (That a Lagrange multiplier and its associated constraint are conjugate variables is the case throughout thermodynamics, with pressure and volume being perhaps the most familiar.) The deviations from the stable state are quite small so to determine analytically $\langle G_1\rangle(n)$ as a function of $\mu_1(n)$ we resort to the lowest-order contribution. It is shown in *SI Appendix*, section S5 that to leading order in $\mu_1(n)$: $\langle G_1\rangle(n) \simeq \langle G_1\rangle^0(n) + \mu_1(n)(\partial\langle G_1\rangle(n)/\partial\mu_1(n))_{\mu_1(n)=0}$. The superscript on the first term means that it is the value at the stable state for which, by construction, $\mu_1(n) = 0$. The derivative in the second term is the ploidy of the ensemble (*SI Appendix*, section S5). Therefore, we have for the distribution of the free energy due to the first constraint

$$\mu_1(n)\langle G_1\rangle(n) = \langle G_1\rangle^0(n)\mu_1(n) + \overline{\text{ploidy}}\mu_1^2(n). \qquad [4]$$

As a function of $\mu_1(n)$ the work needed to be done by the first constraint is a parabola in $\mu_1(n)$(Fig. 7). The origin of the parabola is shifted from the plot origin due to the mean value $\langle G_1\rangle^0(n)$ typically not being equal to zero. This occurs when the mean

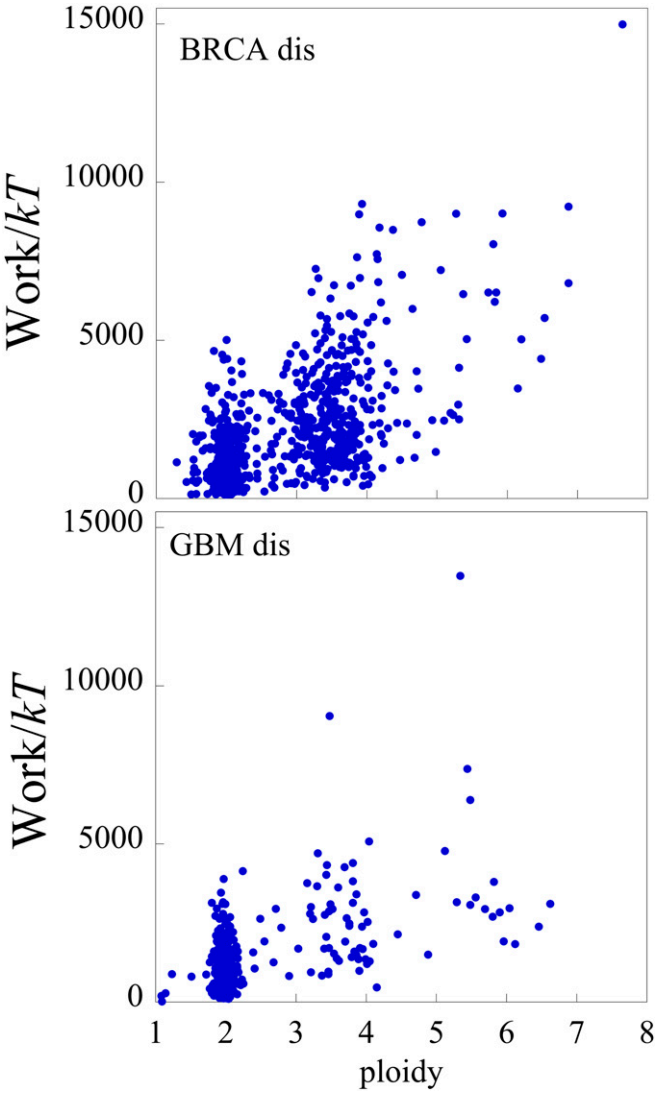

Fig. 3. The work needed to bring the genome from the stable state to its actual fragmented state. The work needed for two cancers, breast (*Upper*) and GBM (*Lower*), plotted vs. the ploidy. Here the three classes of patients with distinct ranges of ploidy are again evident. The GBM tumor genomes show far less genome-wide CNA than seen in breast tumors. The work needed to bring the genome from the stable state to its actual fragmented state is the sum over constraints $\alpha = 1,2,$ ..of the product of the chemical potential μ_α times mean value of the αthconstraint, as derived in *SI Appendix*, section S1 (see also *SI Appendix*, Fig. S6).

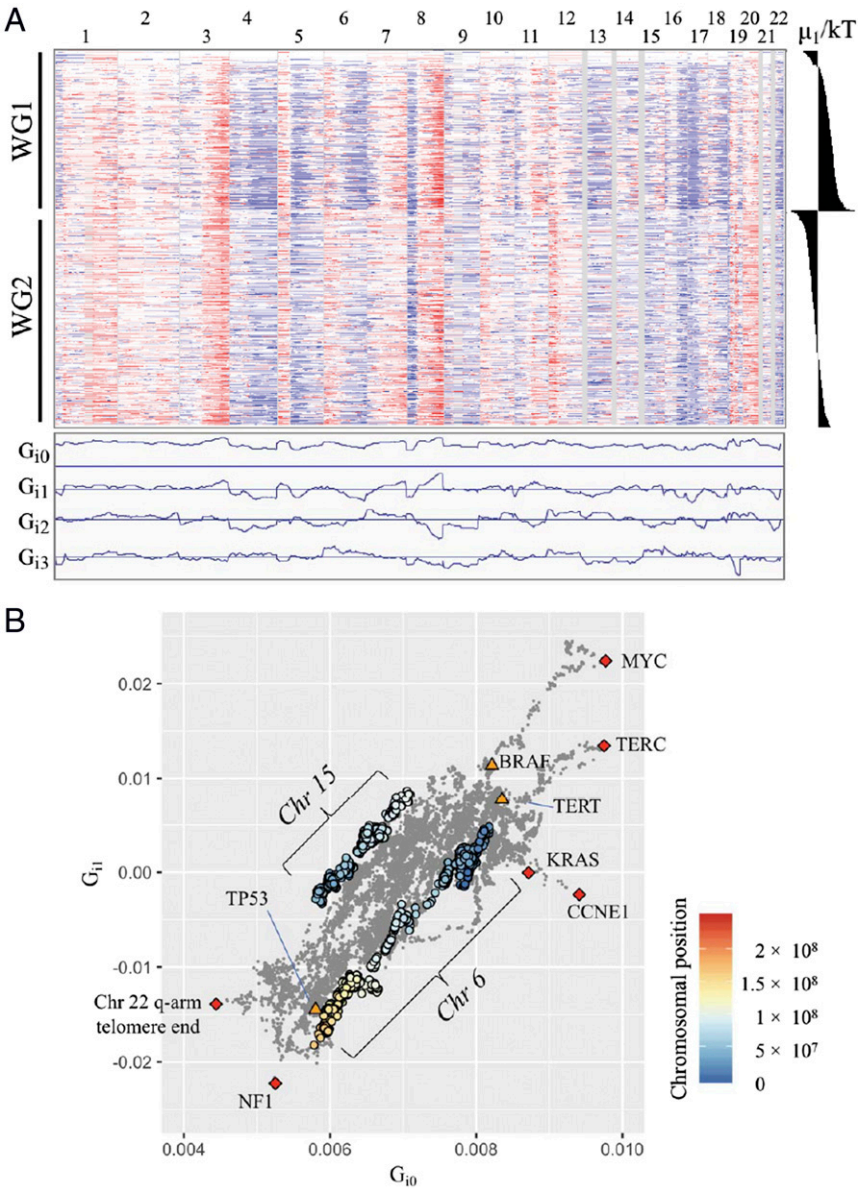

Fig. 4. The thermodynamic landscape of genomic instability. (*A*) Heat-map representation of CNA patterns in OV cancer. The DNA CN deviations shown in the heat maps are log-normalized relative to ploidy levels for visualization purposes. (*Lower*) Plots of the $G_{i\alpha}$profiles for the stable state ($\alpha = 0$) and the top three constraints. (*B*) Plot of G_{i0} versus G_{i1}based on gene index *i*, for OV cancer. Regions with strong amplifications (MYC, TERC, TERT, and BRAF), strong deletions (NF1, Chr 22 q-arm telomere end, and TP53) are indicated. Chromosomes with the most deviation from the central trend, and thus ploidy-biased, are indicated (Chr 6 and Chr 15). Amplifications that have high ploidy- or low ploidy-biased occurrence are indicated (CCNE1 and KRAS) (see also Fig. 6).

value of G_{1i} over all of the loci in the balanced state,$\langle G_1\rangle^0(n)$, imposed as a constraint does not increase the free energy. This is unlike the mean value $\langle G_1\rangle(n)$ in the actual state that does lead to an increase in free energy that equals the work done by the constraint. For all seven cancers that we examined, the exact numerical results fully validate the assumptions and approximations used in the above approach.

The theory derivations in *SI Appendix*, section S5 indicate that the work done by the first constraint is a parabola in the chemical potential $\mu_1(n)$ (Eq. **4**). We next ask, at a given ploidy [or a given value of $\mu_0(n)$] what additional work needs to be done to bring the system to its actual off-equilibrium state. This is the work done by the constraints. When $\mu_1(n) = 0$, then the tumor typically has the degree of "dominant CNA pattern" expected based on ploidy ($\mu_0(n)$) alone. From the landscape section, we saw that when $\mu_1(n)$ is positive (negative), then the tumor has more (less) of the dominant pattern than that typical for its ploidy level. From the plot of the work done by the dominant constraint (Fig. 7), we find that in the negative $\mu_1(n)$ realm, the free energy of the WG1 and WG2 ensembles is fairly overlapping. In contrast, in the positive $\mu_1(n)$ realm, when tumors have higher-than-average loads of CN variation, the energetic difference between the WG1 state and the WG2 state diverges, indicating that work is required to go from WG1 to WG2 (arrow in Fig. 7). Analytically, since the free energy has a $\overline{\text{ploidy}}\cdot\mu_1^2(n)$ term, the high ploidy WG2 tumor ensemble has a more convex free energy parabola. Also, due to the direction of the $\langle G_1\rangle^0(n)$-based shifts of the parabola origins described above, directions that are opposite for the differentially convex WG1 and WG2 parabolas, the resulting divergence of the parabolas is more pronounced in the positive $\mu_1(n)$realm.

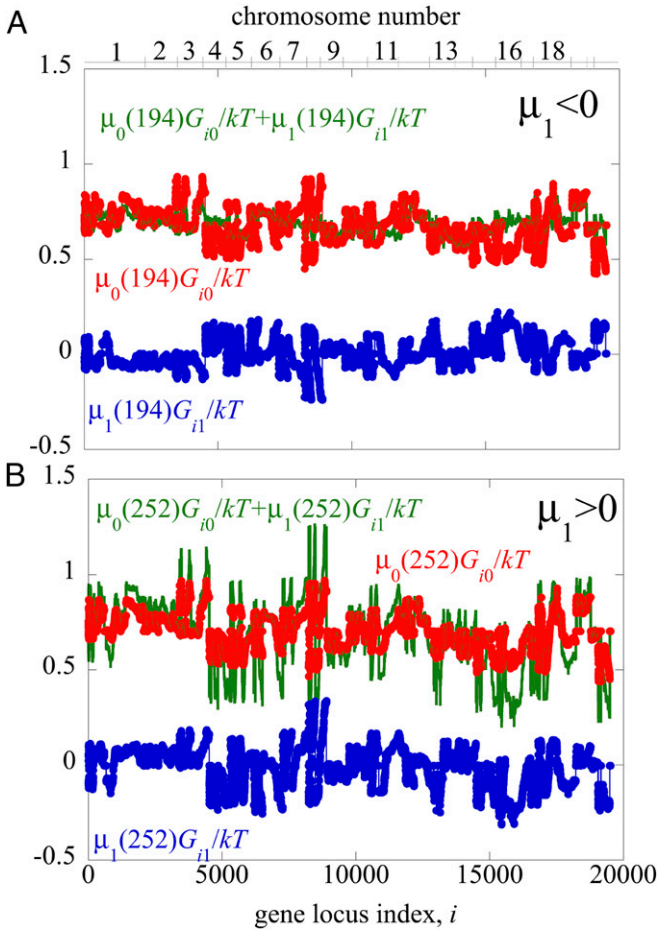

Fig. 5. The sign of the Lagrange multiplier of the dominant state subtracts or adds variance to the CNA pattern. (*A* and *B*) The dominant constraint decreases [when $\mu_1(n)<0$, *A*] or increases [when$\mu_1(n)>0$, *B*] the fluctuations of the CNs [$\ln(X_i)$] as compared to the balanced state. Shown for two OV WG1 tumors (n = 194, tumor TCGA-20-0996-01A-03D, *A*; n = 252, TCGA-24-1423-01A-01D, *B*). (*C*) Plot of variance of $\ln(X_i)$versus $\mu_1(n)$for OV, with WG1 and WG2 ensembles indicated (see also *SI Appendix*, Fig. S7).

These results have a clear implication for a thermodynamic description of a whole-genome doubling. These energetic considerations predict that a single genome cell with $\mu_1(n)>0$ (and the associated high CN variance) has a lower probability to undergo a genome duplication event, since additional work (energy input) would be required, compared to a $\mu_1(n)<0$ cell (with low CN variance) where no additional work other than that provided by the DNA replication machinery during S phase, is required (Fig. 6). Supporting evidence is also found in the histograms of μ_1 values, where WG2 tumors are biased to the negative μ_1, lower CN variation, side (Fig. 7*B*). In all seven cancers studied, the WG1 and WG2 nested parabolas are more overlapping in the low CN variance ($\mu_1(n)<0$) realm, and when CN variance is high ($\mu_1(n)>0$) the free energy levels of WG2 are elevated compared to WG1. Thus, in these latter cases, a WGD event would require an input of energy and thus would have a reduced probability to occur. This means that it is preferentially the lesser fluctuating genomes that are doubled.

Discussion

The theoretical analysis findings provide thermodynamic energetic underpinnings for several previously described attributes of genomic instability and whole-genome doubling. The thermodynamically determined stable state free energy scales monotonically with ploidy universally across all cancer types investigated. The monotonic increase of the free energy is evidence for increasing thermodynamic instability as the ploidy deviation from normal becomes more extensive. This result is in contrast to the stable-state free energy of other biological omic profiling data, such as the transcriptome, where the free energy is not strongly tied to any single parameter and does not show consistency across different tumor types (28). We found that the free energy scales with the average ploidy of the cell, thus tying the free energy to the overall DNA content of the cell. This supports the possibility that the free energy required to assemble nucleotides into a nonrandom sequence is a dominant aspect of the energetics influencing genomic stability.

We suggest that there are three different reasons why we identify a universal behavior. Two reasons are biological and one has to do with the inherent advantages of a thermodynamic approach. One reason is the dominant equilibrium constraint, the so-called stable state. There are two additional reasons and each one is just as important. First, the results of a pan-cancer dominant constraint support that the human genome cannot be too disrupted by available genomic instability mechanisms and still support cellular life. Technically we ask to emphasize that we analyze the CN of gene loci and not the gene loci CNA. CNA are the additive deviation with respect to the mean. Our choice is dictated by the thermodynamic approach to many component systems, what we called above a grand canonical approach. It is a technical point but without it we would not get the free energy to be a convex function of the ploidy, because in a canonical approach we would have normalized out the ploidy.

In addition to the free energy of the stable state, the free energy of the dominant constraint also points to biological insight. In any system, constraints must be imposed to maintain a state of nonequilibrium. In cancer biology, these constraints often reflect tumor fitness selection pressures. The unbiased theoretical analysis of CN data revealed that tumor cells that had undergone a WGD event had a distinct energy landscape. The resulting nested energy parabolas indicate that once a cell has gained sufficient DNA CNAs it is less likely to undergo a WGD event and is characterized by amplification and deletions that favor the fitness of non-WGD tumors. These thermodynamically revealed energy landscapes thus may explain past conclusions that when WGD occurs it tends to be an early event in tumor evolution, that is, prior to the accumulation of chromosome and subchromosome specific amplifications and deletions (accumulation of CNAs) (7, 9, 10). This result too was seen in all seven cancers investigated.

Methods

DNA CNA Data Availability. The Cancer Genome Atlas (TCGA) samples were downloaded from the TCGA portal (https://tcga-data.nci.nih.gov/tcga/), as reported in Graham et al. (1). CN profiles obtained were preprocessed level-3 data based on human genome 19, with CN variations removed. Prior to surprisal analysis, the CNA data were converted from $\log_2$ data to ln (natural log), more commonly used in thermodynamics.

Surprisal Analysis. SVD (see ref. 36 for the matrix algebra details) is a method for "pseudo diagonalization" of a rectangular matrix. In surprisal analysis the absolute quantities are important and thus the data are not centered. SVD is performed on the ln (natural log) data (= DNA CN levels) in rectangular matrix form where the columns index n is the tumors and the rows i are indexed by the gene locus. SVD is used to determine the constraints and the Lagrange multipliers (see refs. 27, 30 for more detailed description of the application and motivation). Very large datasets may require special handling but otherwise Fortran, MATLAB, or R SVD, or similar programs, are quite useful. The input matrix has gene loci as rows and samples as columns and the $G_{i\alpha}$ values are obtained from the SVD left eigenvectors, where i is the locus index and α is the SVD component index (α = 0, 1, . . .) that is used to index the constraints. Recall that n is the tumor index. The $\mu_\alpha(n)/kT$ values are obtained from the right singular vector indexed by α multiplied by the

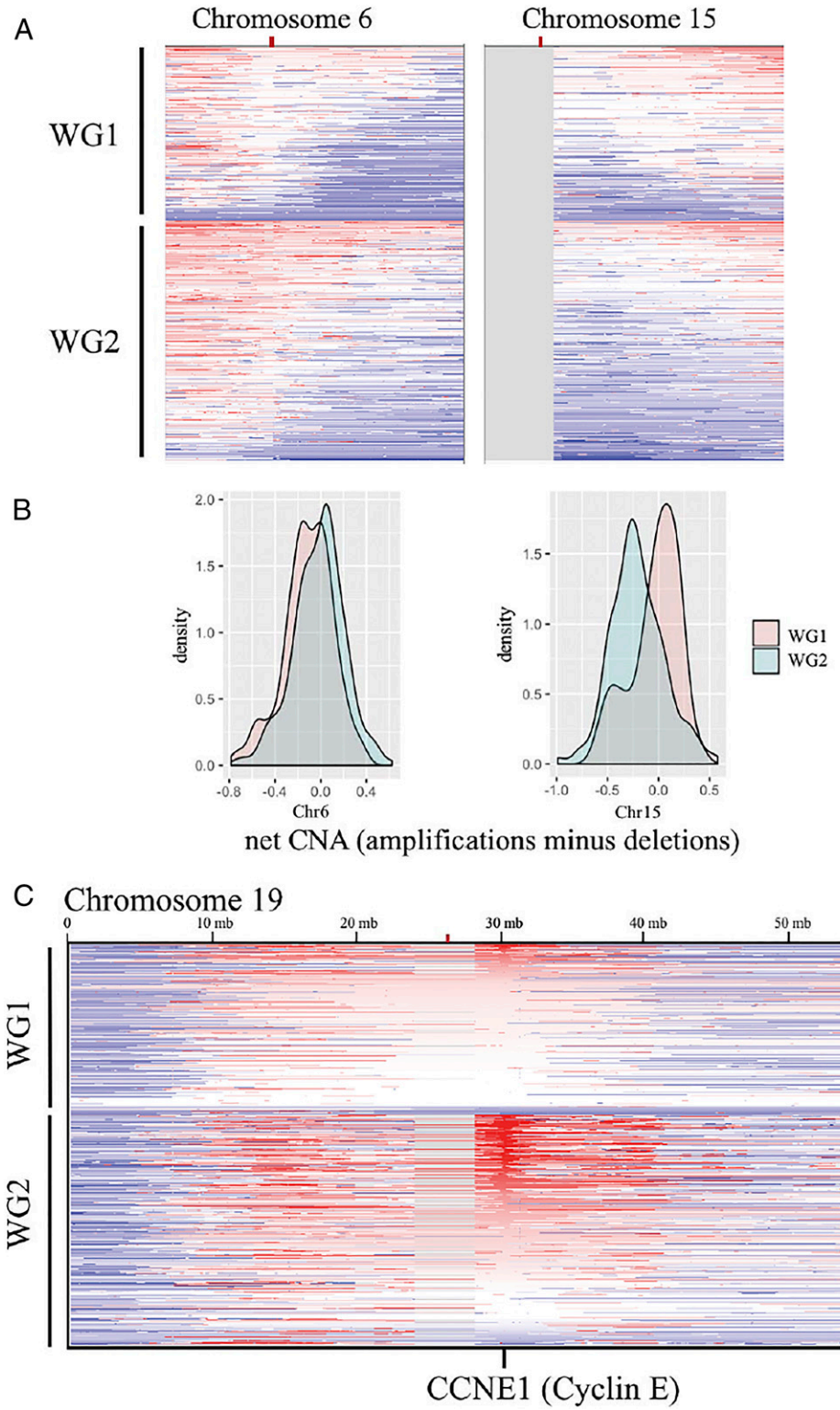

Fig. 6. Ploidy-biased trends revealed by surprisal analysis. (*A*) Deletion of Chromosome 6q is more prominent in WG1 tumors. Deletion of Chromosome 15q is more prominent in WG2 tumors. Samples are grouped as WG1 and WG2 ensembles then sorted by the degree of chromosome deletion. Centromere locations indicated by a red tick. (*B*) Histograms of the net CN alterations (amplifications minus deletions, summed over the bases in the chromosome), corresponding to *A*. (*C*) CCNE1 amplification is biased toward WG2 tumors. Samples are grouped as WG1 and WG2 ensembles then sorted by the degree of CCNE1 amplification (see also *SI Appendix*, Fig. S8).

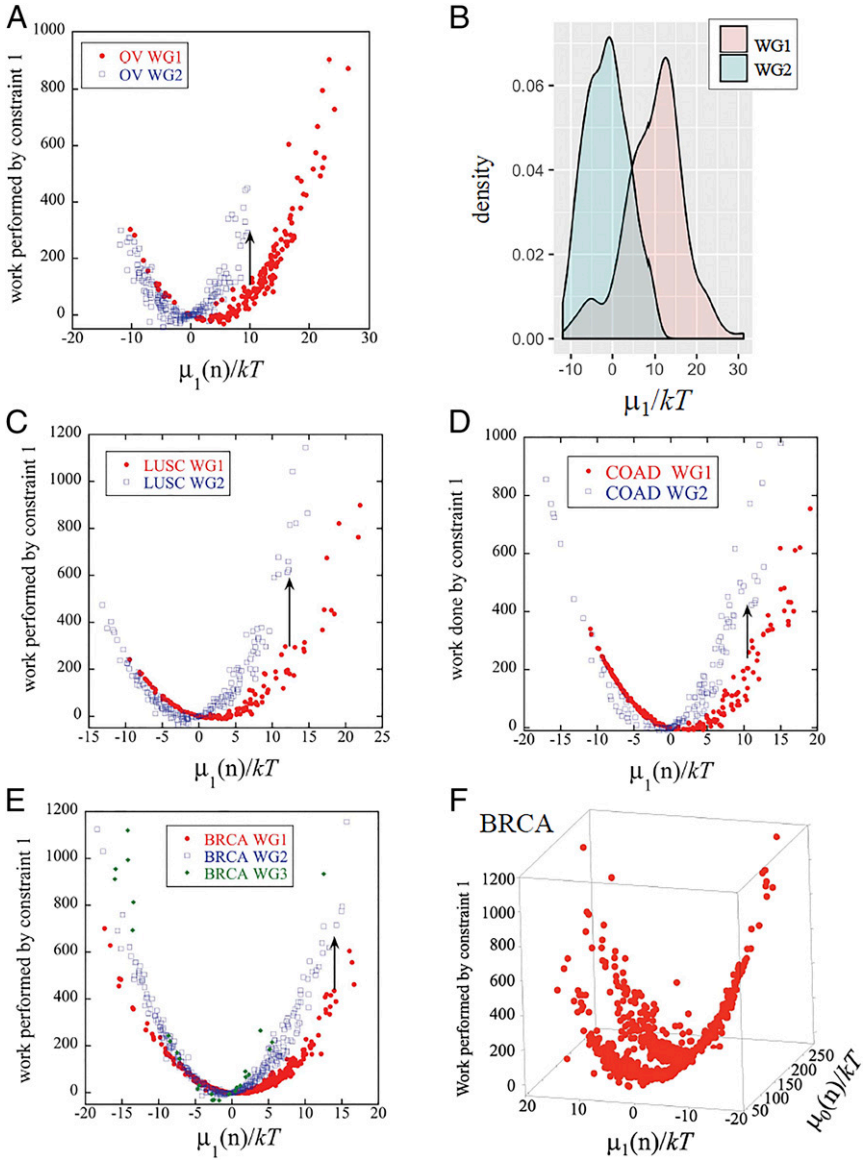

Fig. 7. Energetic considerations of whole-genome doubling (WGD). (*A*) Nested parabolas of the work done by the dominant constraint $\mu_1(n)$ for the WG1 and WG2 ensembles in ovarian cancer (OV). The vertical black arrow indicates the additional work needed to undergo a WGD (transition from WG1 to WG2) for tumor cells with $\mu_1(n) > 0$. The work done by the dominant constraint is the product of the stable-state chemical potential $\mu_1$times the mean value of the dominant (first) constraint as derived in *SI Appendix*, section S1. The energy plotted is in addition to the change in free energy of the stable state, as indicated in Fig. 2. (*B*) Histogram of the number of tumors with the indicated μ_1 values for the data in *A*. (*C*–*E*) Nested parabolas as in *A* for LUSC (*C*), COAD (*D*), and BRCA (*E*). (*F*) A three-dimensional plot of the work done by the dominant constraint vs. the two chemical potentials $\mu_0(n)$ and $\mu_1(n)$. Each point is a WG1 or WG2 BRCA tumor. The two branches correspond to WG1 and WG2 respectively. Note that for clarity of visualization the μ_1 axis is reversed in this panel compared to *E*, here running from positive to negative. Thus, *E* is a two-dimensional projection where the work is shown vs. $\mu_1(n)$ irrespective of the value of$\mu_0(n)$ (see also *SI Appendix*, Fig. S9 for ploidy consistent and purity corrected versions of the LUSC analysis and *SI Appendix*, Fig. S11 for an analysis of tumors with average ploidy 3).

corresponding singular eigenvalue and kT is the unit of thermal energy, with k being Boltzmann's constant.

Ploidy and Purity. Tumor ploidy estimates were inferred from the DNA CAN data, based on publicly available calculations using either the Absolute algorithm (https://www.synapse.org/#!Synapse:syn1710466, using the file named pancan12.sample_info.txt and https://gdc.cancer.gov/about-data/publications/pancanatlas, using the ABSOLUTE purity/ploidy file named TCGA_mastercalls.abs_tables_JSedit.fixed.txt, whole-exome-sequencing-based) (21) or the ASCAT algorithm (https://cancer.sanger.ac.uk/cosmic/sftp_file_info?data=%2Ffiles%2Fgrch38%2Fcosmic%2Fv84%2Fascat_acf_ploidy.tsv, SNP-chip-microarray-based) (22).

In the analysis, WG1 (approximately one copy of a whole genome) is defined as ploidy < 2.5, WG2 as 2.5 < ploidy < 4.5, and WG3 as 4.5 < ploidy. These ploidy-based categories do not definitively reflect whether a WGD event occurred or not but based on published inferences are generally reflective of WGD (21). Notably, tumors with ploidy values close to 2.5 are rare, with this value matching a distinct valley in the ploidy histogram (Fig. 1*A*). In the surprisal analysis, each tumor type was analyzed individually with all tumors included. In some graphs, only the WG1 and WG2 tumors are plotted. In these cases, WG3 tumors are not plotted due to there being too few WG3 samples for trends to be conclusive.

Tumor purity estimates were inferred from the DNA CNA data, based on publicly available calculations using the Absolute algorithm (https://www.synapse.org/#!Synapse:syn1710466, using the file named pancan12.sample_info.txt, whole-exome-sequencing-based) (21, 37). Ploidy unnormalization and purity correction are discussed in detail in *SI Appendix*, section S6.

ACKNOWLEDGMENTS. The work of T.G.G. is supported by the following funding: NIH/National Cancer Institute R01 Grant CA222877 (T.G.G.); UCLA

Specialized Program of Research Excellence (SPORE) in Prostate Cancer NIH P50 CA092131 (T.G.G.); The Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research (T.G.G.), and the Hal Gaba Fund for Prostate Cancer Research (T.G.G.). The work of F.R. is supported by Fonds National de la Recherche Scientifique (F.R.S.-FNRS), Belgium.

1. N. A. Graham *et al*., Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures. *Mol. Syst. Biol.* **13**, 914 (2017).
2. T. Davoli *et al*., Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
3. J. K. Lee, Y. L. Choi, M. Kwon, P. J. Park, Mechanisms and consequences of cancer genome instability: Lessons from genome sequencing studies. *Annu. Rev. Pathol.* **11**, 283–312 (2016).
4. L. Sansregret, B. Vanhaesebroeck, C. Swanton, Determinants and clinical implications of chromosomal instability in cancer. *Nat. Rev. Clin. Oncol.* **15**, 139–150 (2018).
5. S. Mishra *et al*., Cross-talk between lysine-modifying enzymes controls site-specific DNA amplifications. *Cell* **174**, 803–817.e16 (2018).
6. S. F. Bakhoum *et al*., Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* **553**, 467–472 (2018).
7. S. M. Dewhurst *et al*., Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
8. S. Ohshima, A. Seyama, Establishment of proliferative tetraploid cells from telomerase-immortalized normal human fibroblasts. *Genes Chromosomes Cancer* **55**, 522–530 (2016).
9. C. M. Bielski *et al*., Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
10. T. I. Zack *et al*., Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
11. B. S. Taylor *et al*., Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
12. N. Andor *et al*., Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
13. J. Hüfner, D. Mukhopadhyay, Fragmentation of nuclei, stones and asteroids. *Phys. Lett. B* **173**, 373–376 (1986).
14. J. Silberstein, R. D. Levine, Statistical fragmentation patterns in multiphoton ionization: A comparison with experiment. *J. Chem. Phys.* **75**, 5735–5743 (1981).
15. E. Hendell, U. Even, T. Raz, R. D. Levine, Shattering of clusters upon surface impact: An experimental and theoretical study. *Phys. Rev. Lett.* **75**, 2670–2673 (1995).
16. E. E. B. Campbell, T. Raz, R. D. Levine, Internal energy dependence of the fragmentation patterns of C60 and $C60^+$. *Chem. Phys. Lett.* **253**, 261–267 (1996).
17. R. Englman, N. Rivier, Z. Jaeger, Fragment-size distribution in disintegration by maximum-entropy formalism. *Philos. Mag. B Phys. Condens. Matter Stat. Mech. Electron. Opt. Magn. Prop.* **56**, 751–769 (1987).
18. D. Pinkel *et al*., High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**, 207–211 (1998).
19. D. Y. Chiang *et al*., High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).
20. R. Beroukhim *et al*., The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
21. S. L. Carter *et al*., Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
22. P. Van Loo *et al*., Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16910–16915 (2010).
23. F. Favero *et al*., Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
24. B. Zhang, P. G. Wolynes, Genomic energy landscapes. *Biophys. J.* **112**, 427–433 (2017).
25. F. Morcos, J. N. Onuchic, The role of coevolutionary signatures in protein interaction dynamics, complex inference, molecular recognition, and mutational landscapes. *Curr. Opin. Struct. Biol.* **56**, 179–186 (2019).
26. M. Di Pierro, R. R. Cheng, E. Lieberman Aiden, P. G. Wolynes, J. N. Onuchic, De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12126–12131 (2017).
27. F. Remacle, N. Kravchenko-Balasha, A. Levitzki, R. D. Levine, Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10324–10329 (2010).
28. S. Zadran, F. Remacle, R. D. Levine, miRNA and mRNA cancer signatures determined by analysis of expression levels in large cohorts of patients. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19160–19165 (2013).
29. I. Procaccia, R. D. Levine, Potential work: A statistical-mechanical approach for systems in disequilibrium. *J. Chem. Phys.* **65**, 3357–3364 (1976).
30. F. Remacle, R. D. Levine, Statistical thermodynamics of transcription profiles in normal development and tumorigeneses in cohorts of patients. *Eur. Biophys. J.* **44**, 709–726 (2015).
31. N. Kravchenko-Balasha, Y. S. Shin, A. Sutherland, R. D. Levine, J. R. Heath, Intercellular signaling through secreted proteins induces free-energy gradient-directed cell movement. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5520–5525 (2016).
32. S. Zadran, R. Arumugam, H. Herschman, M. E. Phelps, R. D. Levine, Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 13235–13240 (2014).
33. Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
34. A. Gross, R. D. Levine, Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype. *PLoS One* **8**, e61554 (2013).
35. N. Kravchenko-Balasha *et al*., On a fundamental structure of gene networks in living cells. *Proc. Natl. Acad. Sci. USA* **109**, 4702–4707 (2012).
36. G. H. Golub, C. F. V. Loan, *Matrix Computations*, (John Hopkins University Press, Baltimore, ed. 3, 1996).
37. K. A. Hoadley *et al*.; Cancer Genome Atlas Research Network, Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).