Analysis of Alterations in the Human Cancer Genome

by

Scott L. Carter

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN
BIOINFORMATICS AND INTEGRATIVE GENOMICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

ARCHIVES

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SEP 2 1 2011
LIBRARIES

© Scott L. Carter, MMXI. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly
paper and electronic copies of this thesis document in whole or in part.

Author .....
Harvard-MIT Division of Health Sciences and Technology
September 6, 2011

Certified by .....
Gad Getz Ph.D
Director of Cancer Genome Analysis
Broad Institute
Thesis Supervisor

Certified by .....
Matthew Meyerson MD, Ph.D
Professor
Harvard Medical School
Thesis Supervisor

Accepted by .....
Ram Sasisekharan PhD
Director, Harvard-MIT Division of Health Sciences and Technology
Chairman, Department Committee on Graduate Theses

Analysis of Alterations in the Human Cancer Genome

by

Scott L. Carter

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on September 6, 2011, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY IN
BIOINFORMATICS AND INTEGRATIVE GENOMICS

Abstract

Aneuploidy, an abnormal complement of chromosomes, is present in approximately 90% of human malignancies. Despite over 100 years of research, many questions remain regarding the contribution of aneuploidy to the cancer phenotype. In this thesis, we develop computational methods to infer the presence and specific patterns of aneuploidy across thousands of primary cancer tissue specimens. We then combine these inferences with clinical and genomic features of the cancer samples to refine our understanding of both the clinical implications of aneuploidy, and how it evolves in various human cancers. We identified a signature of chromosomal instability from specific genes whose expression was consistently correlated with aneuploidy in several cancer types, and which was predictive of poor clinical outcome multiple cancer types. Current genomic characterization techniques measure somatic alterations in a cancer sample in units of genomes (DNA mass). The meaning of such measurements is highly dependent on the tumors purity and its overall ploidy; they are hence complicated to interpret and compare across samples. Ideally, copy-number should be measured in copies-per-cancer-cell. Such measurements are straightforward to interpret and, for alterations that are fixed in the cancer cell population, are simple integer values. We develop two computational methods to infer tumor purity and malignant cell ploidy directly from allelic analysis of DNA. First we describe HAPSEG, a probabilistic method to interpret bi-allelic marker data in cancer samples in order to produce genome-wide estimates of homologue specific copy-ratios. Second, we describe ABSOLUTE, a method that infers purity, ploidy, and absolute copy-numbers from the estimates produced by HAPSEG. In addition, ABSOLUTE can analyze point mutations to detect subclonal heterogeneity and somatic homozygosity. We used ABSOLUTE to analyze ovarian cancer data and discovered that 54% of somatic point mutations were, in fact, subclonal. In contrast, mutations occurring in key tumor suppressor genes, TP53 and NF1 were predominantly clonal and homozygous. Analysis of absolute allelic copy-number profiles from 3,155 cancer specimens revealed that genome-doubling events are common in human cancer, and likely occur in already aneuploid cells in many cancer types. By correlating genome-doubling status with mutation data, we found that homozygous mutations in NF1 occurred predominantly in non-doubled samples. This finding suggests that genome doubling influences the pathways of tumor progression, with recessive inactivation being less common after genome doubling.

Thesis Supervisor: Gad Getz Ph.D
Title: Director of Cancer Genome Analysis
Broad Institute

Thesis Supervisor: Matthew Meyerson MD, Ph.D
Title: Professor
Harvard Medical School

Acknowledgments

I am deeply indebted to my research supervisor Dr. Gad Getz, who has been an invaluable source of enthusiasm and statistical expertise during my graduate education. Special thanks also go to Dr. Matthew Meyerson, who graciously accepted me into his lab despite my insistence on pursuing my own research agenda in graduate school; this environment provided an important part of my intellectual home during graduate school. In addition, I would like to thank my thesis committee members Drs. Shamil Sunyaev and David Kwiatkowski, as well as Drs. David Pellman, Angelika Amon, Tyler Jacks, and Robert Weinberg for their scientific guidance. I am grateful to Meyerson lab members and alumni; especially Craig H. Mermel, Alex Ramos, Marcin Imielinski, and Elena Helman for many interesting conversations and collaborations over the years.

The majority of the research described in this thesis (Chapters 3 and 4) was conducted at the Broad Institute, and I am indebted to many of its affiliates: in particular Rameen Beroukhim, Joshua Korn, Aron McKenna, Kristian Cibulskis, Nick Patterson, and finally, Dr. Eric Lander, for his expert guidance in scientific writing, and for helping make the Broad Institute such a fantastic setting in which to receive an education. This thesis would not have been possible without the efforts of the data production team at the Broad. In particular Robert Onofrio, Carrie Sougnez, and Wendy Winckler provided key expertise in the generation of the datasets used in this thesis. In addition, I would like to thank my mentors from previous stages of my career, in particular Michael Reich, Drs. Isaac Kohane, and Zoltan Szallasi,

Finally, I would like to thank all of my close friends and family who have supported me over the years, especially Justin and Adam Pieroni and the whole pulse graphics crew, as well as my parents and my brother Keith. Most importantly, I am thankful to my wife Sarah, for her patience, encouragement, and companionship over the years. In addition, our son Marco has helped make the past 6 months a joy with his consistent good nature and amenable sleeping habits. In addition, I have enjoyed many interesting conversations about human genetics with my parents-in-law, the Drs. Joseph and Rita Calvo.

Contents

Chapter 1 - Introduction 9
Chapter 2 - A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers 40
Chapter 3 - Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping 59
Chapter 4 - Absolute quantification of somatic DNA alterations in cancer reveals frequent genome doublings in human cancers 88
Chapter 5 - Discussion and Future Directions 145
Appendix A - Supplementary Material for Chapter 2 165
Appendix B - Markov-chain Monte Carlo simulations for tumor purity/ploidy inference . 189

Chapter 1

—

Introduction

Scott L. Carter

Effective treatment of cancer requires a margin of therapeutic specificity, whereby cancer cells are selectively killed (or inactivated) with sufficient specificity so as to be physiologically tolerated by the patient. Following the early success of Farber developing chemotherapy for childhood leukemia, the history of cancer treatment is strewn with gruesome examples of pushing such tolerance to its limits, and beyond[1].

Targeted approaches to treating cancer

Decades of research have culminated in the modern understanding of cancer [2]. By identifying the genetic alterations unique to cancer cells, which drive the cancer phenotype, modern science has produced cancer treatments precisely targeting cancer cells, offering the promise of cancer eradication with minimal toxicity to the patient.

The first example of such treatment is the use of Imatinib to treat chronic myelogenous leukemia (CML). Nearly all such neoplasms are driven by the BCR-Abl fusion protein, formed by a recurrent chromosomal translocation (the "Philadelphia chromosome") [3], [4]. This success was followed in 2004 by reports describing gain-of-function mutations in exons encoding the kinase domain of the epidermal growth factor receptor (EGFR) which conferred sensitivity to pharmacologic EGFR inhibition [5], [6], [7]. Notably, lung adenocarcinomas harboring *KRAS* mutations exhibit primary resistance to these therapies [8].

Collectively, the dependence of cancer cells on specific mutations which drive the cancer phenotype has been termed "oncogene addiction" [9]. An additional example is the BRAF gene, which is mutated in 50-70% of malignant melanomas[10], predicting enhanced sensitivity to pharmacologic MAP kinase inhibition (e.g., RAF or MEK inhibitors). Pharmacological inhibition of mutant, activated, BRAF [11] was demonstrated to be effective in treatment in metastatic melanoma[12].

Resistance mutations

Unfortunately, all efforts to directly target oncogene addiction in human cancer patients have resulted in the eventual outgrowth of resistant cancer clones, many of which are

refractory to the available targeted and cytotoxic treatment. For example, in CML treatment with BCR-Abl many patients eventually relapse. In 2001, Charles Sawyers and colleagues showed that many such relapses were associated with a single amino acid substitution in the Abl kinase domain, known to form a critical hydrogen bond with the drug [13].

Similar resistance mutations have been reported in the treatment of EGFR-driven cancers with EGFR inhibitors. For example, the mutation EGFR (T790M), which occurs in the kinase domain, is associated with resistance to Gefitinib [14], [8], [15]. Alternately, resistance in these tumors may also be driven by amplification of the MET oncogene [16], which may co-occur with the T790M mutation[17].

Finally resistance to the PLX4720 pharmacologic inhibitor of MEK/BRAF is associated with both MEK1(P124L) and MEK1(Q56P) mutations [18]. In 2010, *MAP3K8* (the gene encoding COT/Tpl2) was identified as a MAPK pathway agonist, the amplification of which drives resistance to RAF inhibition in B-RAF(V600E) cell lines[19]. Also in 2010, another group identified induction of PDGFRb expression and tyrosine phosphorylation as a dominant feature of acquired resistance in human melanoma-derived cells [20].

Thus, in all cases, a single mutational event allows the cancer cells to escape targeted inhibition of driver oncogenes.

Non-oncogene addiction

Another approach to achieve selective killing of cancer cells is to target normal cellular processes that are of greater importance to the survival of cancer cell, relative to normal cells. This concept has been labeled "non-oncogene addiction" [21]. Notably, because such pathways may be targetable by general means, acquired mutations affecting a single gene may be less likely to confer resistance to the cancer cell. Such intervention might therefore be less predisposed to the development of resistance mutations than targeting oncogene addiction directly.

A paradigmatic example of non-oncogene addiction was the demonstration by Susan Lindquist and colleagues in 2007 that the heat shock factor 1 (HSF1) gene is a potent modifier of carcinogenesis[22]. *HSF1* encodes the master regulator of eukaryotic heat shock response [22], and controls the expression of heat-shock proteins that promote protein refolding, prevent protein aggregation, and target mis-folded proteins for destruction. The authors demonstrated that HSF1 deficiency in mice protects against tumorigenesis driven either by mutations of the RAS oncogene or a recurrent mutation in the tumor suppressor p53 (p53R172H) [22],[21].

In contrast to expression of mutant Ras alleles, over-expression of HSF1 was unable to transform immortalized mouse embryonic fibroblasts [22]. In addition, the viability of multiple human cancer cell lines was decreased following knockdown of HSF1 by short hairpin RNAs (shRNAs). However, HSF1 knockdown had only minimal impact on the viability of normal cells such as primary mammary epithelial cells and WI-38 human lung fibroblasts. The heat-shock response thus fulfills the criteria for a targetable non-oncogene addiction in human cancer [21].

The DNA damage repair pathway represents another aspect of normal cellular function which becomes essential to the viability of cancer cells. This has been exploited in BRCA deficient cells by inhibition of the poly(ADP-ribose) polymerase (PARP) enzyme involved in base excision repair, a key pathway in the repair of DNA single-strand breaks. Genetic and pharmacologic inhibition of PARP results in chromosomal instability, cell cycle arrest and subsequent apoptosis in human embryonic stem-cells deficient for BRCA1 or BRCA2, but not in their wild-type counterparts[23]. Both MCF7 (wild-type p53) and MDA-MB-231 (mutated p53) human breast cancer cell lines display a similar sensitivity to PARP inhibition upon depletion of *BRCA2* with short interfering RNAs [24]. Furthermore, pharmacologic PARP inhibition is effective in mouse xenografts seeded with the BRCA2-deficient V-C8 cell-line, but not with the BRCA2 reconstituted V-C8-B2 cell-line [24]. Because patients carrying germline cancer risk

alleles in *BRCA2* ($BRCA2^{+/-}$) are likely to become $BRCA2^{-/-}$ by somatic deletion of the wild-type allele, PARP inhibition offers a selective window to treat such tumors [24], [25]. Indeed, several follow-up studies have shown clinical efficacy of such inhibitors [26], [27], [28].

Thus PARP inhibition in BRCA incompetent cancers represents a distinct aspect of non-oncogene addiction, whereby a specific pathway (BRCA-mediated homologous repair) facilitating cancer development causes increased reliance of the cancer cells on the normal function of a protein (PARP). The demonstration by systematic RNA interference that oncogenic KRAS-driven cancers require TBK1 [29] provides another example of this type of synthetic interaction with a cancer pathway.

The role of aneuploidy in cancer development

In the early 20th century, Theodore Boveri observed that *aneuploidy*, a chromosome complement that is not a multiple of haploid, is a frequent feature of cancer cells and suggested that it might represent a causal mechanism [30]. Recent studies have confirmed that aneuploidy is one of the most ubiquitous features of human cancer, present in greater than 90% of cases [31], [32].

What roles might aneuploidy have in the evolution of human cancers? It has previously been proposed that the distribution of age at cancer diagnosis is consistent with a requirement of genetic instability to generate the number of serial transformations required to evolve the mature cancer clone [33]. A classic model holds that, after an initial period of genetic instability, a stable aneuploid karyotype emerges[34], which may then evolve further by punctuated evolution[35]. Despite its ubiquity in cancer, aneuploidy is not tolerated by normal human cells [36], suggesting that alterations specific to the cancer phenotype may confer viability to aneuploid cells.

How does this frequent aneuploidy arise and what evidence supports a causal role for aneuploidy in cancer? Several mechanisms have been demonstrated by which chromosomal instability and resulting aneuploidy may be generated.

Experimental data have shown that chromosomal instability is functionally associated with defects in mitotic spindle checkpoints [37],[38]. Such defects allow mitoses to proceed in settings of improper chromosomal alignment, kinetochore attachment, or kinetochore tension, [39], [40] thus leading to variable distribution of chromosomes among daughter cells. Despite earlier reports of somatic mutations in spindle-checkpoint genes resulting in chromosomal instability and aneuploidy[38], [41], [42], efforts to attribute chromosomal instability in many human cancers to defects in these genes have been lacking. For example, the *BubR1* mutations identified in a small number of colorectal cancers [38] have not been demonstrated in subsequent genome sequencing studies of colorectal or other human cancers[43]. Thus it remains doubtful whether chromosomal instability is caused by specific genetic alterations in a large fraction of human cancers, as has been previously suggested [41].

Another potential source of genetic instability during tumor evolution is *telomere crisis*, whereby the normal telomeric 'caps' at the ends of chromosomes erode to critically short length, at which point they trigger DNA double-stranded break (DSB) response, which can result in the free DNA ends being randomly joined together. This can lead to the creation of chromatids bearing multiple centromeres, which can then attached to opposite spindle poles during the next cell division. This can result in the di-centric chromosome being broken again, leading to additional DSB repair, exacerbating the process. Importantly, this process may be terminated by expression of the telomerase gene, which can stabilize free DNA ends and halt this cycle [34]. Telomere crisis has been shown to induce genetic instability and tumorigenesis in mouse models [44].

An additional potential source of genetic instability in human cancer could be due to the excess centrosome number sometimes observed in cancer cells. Normal human cells contain 2 centrosomes, which function during cell division to nucleate the assembly of microtubules forming the mitotic spindle. Normally, the replication of centrosomes is controlled by licensing factors permitting duplication only once per cell cycle. Excess centrosome content has been shown to lead to multipolar spindles, typically with disastrous consequences for the resulting cell. One obvious way for a cell to obtain an

excess centrosome complement would be to undergo an abortive mitosis, where both the DNA and centrosomes are duplicated, but the cell fails to complete cytokinesis. Indeed, the frequent observation of chromosome complements far in excess of the diploid number in human cancers has lead to much speculation that tetraploidization is a common event in human cancer[36], [45], [46].

Tetraploidization, because it produces rampant aneuploidy, was first proposed as a carcinogenic mechanism by Theodor Boveri in 1914 [30]. Multiple lines of evidence are consistent with the view that tetraploidization has a role in the development of at least some human cancers: (i) Many mature cancers display chromosome complements far in excess of diploid [46], [47]; (ii) 4N populations of cells have been demonstrated in precancerous lesions, and correlated with p53 loss and increased genetic instability [48], [49]; (iii) A number of cancer causing genetic events – for example Rb loss and APC loss — are known to cause cytokinesis failure and thus genome doublings [36], [50], and additional mechanisms have been demonstrated *in vitro* [51], [52]; (iv) Finally, the principal that cytokinesis failure can promote tumorigenesis has been directly demonstrated using animal models [53], [54].

Recent results have shown that excess centrosomes can occasionally cluster into two poles at mitosis to form a reasonably normal bi-polar spindle [54]; however, excess rates of chromosome mis-segragation were observed in such mitoses. These results suggest a means by which excess centrosomes could contribute to the genetic instability during cancer development.

Despite these insights, several fundamental questions remain regarding the evolution of aneuploidy in human cancer: (i) Is aneuploidy a cause or consequence of the cancer phenotype? (ii) At what stage of cancer development might aneuploidy emerge? (iii) Are specific mutations required to trigger chromosomal instability? (iv) Do specific genes or pathways become essential to viability in the aneuploid cellular state? (v) How does a genetically unstable cell stabilize its karyotype, generating the predominantly

monogenomic cancer clones observed in most mature cancers? (vi) How might aneuploidy be related to polyploidization or genome-doubling?

The advent of high throughput technologies with which to characterize thousands of tumor samples has shed light on some of these questions. For example, previous analysis of cancer karyotypes relied on spectral karyotyping (SKY), a laborious technique to visualize stained metaphase chromosomes. The low resolution and difficulty of analyzing these data have led to some misconceptions about somatic copy alterations (SCNAs) in cancer. For example, it has previously been stated that:

> One barrier to defining the role of aneuploidy in tumour biology is the genetic complexity of aneuploid cells. Like Tolstoy's unhappy families, each aneuploid tumour is aneuploid in its own way[36].

In fact pan-cancer analysis of several thousand tumor copy-profiles using high-density DNA SNP microarrays has revealed that many chromosome-arm level SCNAs are remarkably recurrent within cancer types and lineages [55]. This suggests that selection for gain and loss of specific chromosome arms plays a prominent role in shaping the stable cancer karyotype emerging from chromosomal instability. On the other hand, this also complicates the analysis of aneuploidy observed in human cancers, since it may be difficult or impossible to separate the phenotypic contributions of aneuploidy *per se* from its downstream consequences on proliferation, etc.

Model systems to study the effects of aneuploidy

Over the past several years, Angelika Amon and colleagues have created model systems in which cellular phenotypes associated with specific aneuploidies can be dissected. By engineering yeast strains lacking the karyogamy gene *KAR1* and mating them with wild-type, nuclear fusion is prevented, resulted in abortive matings, in which individual chromosomes are occasionally transferred from one cell to another. By placing different selectable markers in each mating partner at the same genomic location, the authors were able to selectively isolate these rare events and generate strains ("disomics") carrying 13 of the 16 possible single chromosome gains [56]. In order to further discern whether the

phenotypes associated with aneuploidy are dependent on gene dosage effects, these experiments employ yeast artificial chromosomes (YACs), which are constructed from human DNA and do not contain transcribed genes [56].

Characterization of these strains revealed that disomics share a number of phenotypes, including defects in cell cycle progression, increased glucose uptake, and increased sensitivity to conditions interfering with protein synthesis and protein folding (e.g. high temperature)[56]. Importantly, these defects were not observed in the strains bearing one of more YACs, indicating that they result from alterations in gene-dosage, and not due to physiological consequences of abnormal DNA content, such as altered spindle geometry)[56]. Interestingly, the YAC-bearing strains shared a common transcriptional response with several of the disomic lines, suggesting the existence of transcriptional programs induced solely by excess DNA content[56]. Using analogous experimental methods, similar results have been obtained in mouse cells [57].

Recently, it was demonstrated that these same disomic yeast strains display many features of genetic instability, including increased rates of point-mutation and chromosomal mis-segregation[58]. These results suggest that there may not be a hard distinction between aneuploidy and chromosomal instability (i.e. a 'state' vs. a 'rate'), as had previously been proposed [59].

Aneuploidy as a potential non-oncogene addiction

Because of its ubiquity in human cancer, aneuploidy would make a highly attractive type of non-oncogene addiction to target. It is possible that genetic instability is required for generation of the mutational diversity needed for the evolution of the cancer phenotype [33]. While aneuploidy may be an important prerequisite to the development of many cancers, it may also confer specific weaknesses, which may be exploited for cancer treatment. This leads to the question of which normal cellular functions unique to the aneuploid cell-state might be exploited to obtain a therapeutic window. Because it creates stoichiometric imbalances among the components of many protein-complexes, aneuploidy may produce an excess burden for protein folding chaperones and

proteosomal degradation mechanisms; these conditions can induce a partial heat-shock response [60]. As reported by Angelika Amon and colleagues in 2007, aneuploidy of each individual yeast chromosome is sufficient to induce the heat-shock response and proteotoxic stress [56]. Furthermore, these cells are especially sensitive to proteosomal inhibition or high-temperature stresses [56], [32]. These results are consistent with the demonstration by Susan Lindquist and colleagues in 2007 that HSF1 deficiency in mice protects against tumorigenesis driven either by mutations of the RAS oncogene or a recurrent mutation in the tumor suppressor p53 [22],[21].

In 2010, Angelika Amon and colleagues reported the evolution of specific genetic adaptations to tolerate aneuploidy in yeast strains[61]. In particular, a loss-of-function mutation in the gene encoding the deubiquitinating enzyme Ubp6 resulted in increased proliferation of several aneuploid yeast strains [62]. In 2011, the same group reported the identification of chemical compounds with aneuploidy-specific lethality [63].

Targeting of the proteosome has been a successful means to treat malignancies in the past, with demonstrated induction of apoptosis specifically in cancer cells [64]. Clinically therapeutic responses have been obtained in the treatment of multiple myeloma with Bortezomib (Velcade), which targets the 26S proteasome [65]. In addition, the anti-cancer effects of the antibiotic geldanamycin are mediated by its ability to specifically bind an evolutionarily conserved 'pocket' domain of the protein-folding chaperone HSP90 [66]; this pocket likely plays an important role in the binding of HSP90 to its client substrates.

A more thorough of the clinical and genomic correlates of aneuploidy in human cancer might allow for additional insight into mechanisms by which this cancer hallmark might be therapeutically exploited.

Thesis summary

Chapter 2: Gene expression correlates of aneuploidy in human cancer samples

Chapter 2 describes work done in 2005-2006 in collaboration with Dr. Zoltan Szallasi at the Children's Hospital Informatics Program (CHIP). Although increased expression in a panel consisting of 8 major mitotic checkpoint genes in human breast cancer cells with chromosomal instability was reported in early 2006 [67], no systematic evaluation of gene expression associated with aneuploidy had been performed. Because many gene expression datasets were becoming publically available, we used gene expression measurements to infer segmental DNA copy-number alterations. These imbalances could then be directly correlated with the expression of individual genes. This approach was attractive because very few datasets existed with DNA copy-number and mRNA profiling of the same samples.

Inferences of segmental copy-change were performed at fixed resolution, set arbitrarily to cytological sub-bands. The resulting measurements of genome-imbalance were summarized in each sample as a scalar estimate, termed 'functional aneuploidy'[1]. This score was used as a bait to indentify genes co-expressed with aneuploidy over diverse cancer expression profiles. The set of genes with the highest CIN-score over all samples comprised a coherent module of genes, including many known regulators of cytokinesis, spindle checkpoint, DNA damage response[2].

Characterization of aneuploidy is not part of standard clinical practice for most human cancers[3]. Due to the interest in developing clinically prognostic gene expression signatures around 1998 – 2006, many of these samples were annotated with clinical

---

[1] A term coined by Dr. Szallasi, which was inspired by the term 'functional genomics', used to describe analysis of mRNA profiles at the time).

[2] It is the author's opinion that the coherence of this gene-expression module made the limitations of the SCNA detection irrelevant.

[3] An important exception is neuroblastoma, in which hyperdiploidy is closely associated with improved relapse-free survival in some patient groups [42]. Ploidy is thus often evaluated in such cases to inform treatment.

outcome data. We noted that the mean expression of the top CIN-score genes was broadly prognostic of cancer recurrence, consistent with the association between clinical grade and aneuploidy. We demonstrated that the CIN-score was a superior predictor of breast-cancer prognosis than tumor grade however, showing that 'intermediate' grade 2 patients split into two groups by CIN signature, with outcomes equivalent to grade 1 and grade 3. These results were published in 2006, and are included in this thesis as Chapter 2. In the discussion section (Chapter 5), further results based on the CIN signature paper published over the last 5 years are summarized.

Chapter 3: Statistical analysis of DNA SNP-microarray data

In order to better characterize genomic copy-number alterations in human cancer samples, we turn to direct measurement of somatic DNA copy-number alterations (SCNAs) using DNA microarrays. Specifically, we developed a technique to measure the copy-ratio of both homologous chromosomes. This is accomplished by computational identification and quantization of heterozygous loci in the cancer-derived DNA.

Human genomes are normally diploid, with one copy of each homologous chromosome derived from either parent in each cell[4]. Due to pervasive aneuploidy in human cancers, many genomic regions have homologous copy-imbalances. This implies that germline heterozygous loci in the cancer sample may be present at two distinct concentration bands in the measured allelic copy-ratios (for a segment with constant total copy-number).

Chapter 3 describes work done in 2008-2010 in collaboration with Dr. Gad Getz. We developed rigorous statistical methodology (HAPSEG) for analysis of allelic copy data at polymorphic loci in cancer-derived DNA. HAPSEG was originally developed in order to

[4] Although it is becoming increasingly apparent that polyploid cells also play a role in normal physiological processes [33], notably, as a response to genotoxic stress in mature hepatocytes [43].

produce highly accurate estimates of homologue-specific copy ratios (HSCRs) in cancer samples. As we point out, gross genomic aberrations common in cancer samples require specialized inference methods. Specifically, we estimate the germline genotypes and genome-wide copy-ratios of both parental homologues. This is accomplished by statistical analysis of bi-allelic DNA concentration measurements at many polymorphic loci using either SNP microarray or massively parallel sequencing technology, in order to infer germline genotypes, phase of imbalanced heterozygous markers, and copy ratio of both homologues in the cancer clone. We develop specialized error-models for use with DNA microarrays.

Accurate inference of homologous copy-ratios is crucial for interpretation of absolute copy-number states in cancer-cell populations, which may be obscured by unknown tumor purity and ploidy. Indeed, HSCR estimates from HAPSEG formed the input data for subsequent joint inference of tumor purity, ploidy, and absolute allelic copy-numbers (described in Chapter 4).

The use of HSCRs reduces the ambiguity of observed copy- profiles with regard to their interpretation as fixed SCNAs having simple integer copy-numbers. For example, the total copy ratio profile of a sample without SCNAs would be equivalent for ploidy values of 1,2,3, etc., however the HSCR profile would rule out odd ploidy values, since these would not be consistent with equal allelic copy-numbers. In addition, since subclonal alterations will affect only a single HSCR value in a given genomic segment, the ratio of clonal to subclonal SCNAs genome-wide is generally higher than when only total copy-numbers are considered.

Because recombination at meiosis is highly non-random[68], and tends to occur at 'hotspots' along the genome where recombination is much more likely, the genotypes of adjacent SNPs are not statistically independent of one another. Rather, they show patterns of *linkage disequilibrium*, whereby nearby genotypes are correlated.

When the phase of adjacent heterozygous markers is known, the resulting haploid genotypes (or, *haplotypes*), display an even stronger statistical dependency, the structure of which is intimately connected to the recombination rate and the evolutionary history of the patient population. Specifically, it is possible to statistically describe phased haplotypes involving common SNPs as a 'mosaic model' [69], which summarizes each sequence of phased markers as a 'mosaic' of reference haplotypes, with occasional jumps from one chromosome to another; the probability of such jumps is governed by the local recombination rate, which is often estimated using family-trio genotyping[70].

Such a panel of reference genotyped chromosomes were produced by the international HAPMAP project[70]. Because only a few hundred such chromosomes are needed to describe the genetic variation in a given population (down to 5% MAF), this provides a more powerful way to summarize human population-genetic variation than can be obtained using minor allele frequencies, or the commonly used r^2 metric of correlation between nearby SNP genotypes.

These dependencies between phased adjacent genotypes are exploited by HAPSEG to obtain optimized estimates of HSCR and germline genotypes. This is useful to accurately recover the germline genotypes when they have become obscured in the cancer sample, as can happen in regions of somatic LOH when the cancer sample is very pure[5]. Because estimation of HSCRs requires jointly estimating marker genotypes, exploiting genotype dependencies also improves the HSCR estimates; specifically the additional genotype information is invaluable in preventing spurious over-fitting of the HAPSEG statistical model at regions of allelic-balance, which can occur in some samples. Because the haploid genotypes implied by such over-fitting are not at all consistent with a plausible mosaic of phased chromosomes, HAPSEG will reject the spurious phasing and converge at allelic balance for such regions.

[5] In this scenario, the concentration of (germline) heterozygous alleles is nearly identical to that of germline homozygous alleles.

The separation between the allelic concentrations of each homologous chromosome provides a means of direct haplotyping within such regions, which we demonstrate anecdotally. Implications of this capability for medical and population genetic analysis are discussed further in Chapter 3 and in the discussion section (Chapter 5).

Chapter 4: Resolution of integral copy numbers for fixed SCNAs

Chapter 4 describes work done in 2007-2011 in collaboration with Drs. Gad Getz and Matthew Meyerson. We developed a novel method, called ABSOLUTE, which can accurately infer the number of copies per-cancer-cell of somatic DNA alterations detected in (commonly impure) tumor tissue samples. In order to accomplish this, ABSOLUTE performs joint inference of tumor purity and ploidy, based on the HSCR profiles produced by HAPSEG (Chapter 3).

The shift from the currently used relative measurements to absolute per-cancer-cell measurements greatly enhances our ability to analyze cancer samples because absolute units reflect the true underlying discrete nature of the DNA alterations that occurred during cancer evolution (e.g. amplifications, deletions, point mutations), and enables us to distinguish clonal from subclonal events.

Although we and others have been attempting to solve this problem for several years, no currently described methods are capable of large-scale deployment to the analysis of thousands of tumor samples. The key improvements that contributed to increasing the accuracy were: (i) Enhancing our model to allow for a small fraction of copy-number events to be subclonal; and, most importantly, (ii) Disambiguation of several equally-likely candidate solutions by comparison to a large set of tumor absolute allelic copy-profiles (these were constructed directly from the input dataset of 3,155 cancer samples, by using the least ambiguous profiles in a 'boot-strapping' fashion.)

We present the results of the systematic application of ABSOLUTE to 3,155 cancer copy-profiles and 214 paired whole-exome sequencing samples, revealing several novel insights regarding human cancer evolution. Analysis of absolute allelic copy-profiles

from a diverse panel of human cancers revealed that whole genome doubling events occur frequently during somatic evolution, ultimately resulting in mature cancers descended from doubled cells, bearing complex karyotypes.

In addition to the discussion of our use of ABSOLUTE to characterize human cancer evolution with respect to aneuploidy and genome doubling in Chapter 4, additional scientific context and future directions are presented in Chapter 5.

Statistical inference techniques

Formal coursework in graduate school statistics consisted of Machine learning at MIT with Professor Leslie P. Kaelbling, Introduction to Computational Biology at MIT with Professors Tommi Jaakkola, Richard Young, and David Gifford, and attendance of Bayesian statistics lectures by Professor Jun Liu at Harvard. Starting in the fall of 2007, I began working with Dr. Gad Getz who served as an invaluable mentor in translating statistical ideas covered in this coursework into practically applicable research tools. Dr Getz's background in the physical sciences naturally inclined him towards Bayesian methods, an inclination shared by his new student.

Classical statistical analysis allows for inference of unknown parameters underlying the generation of observed data by a modeled stochastic process. These inferences may be expressed as point-estimates (e.g. maximum likelihood estimates), or as confidence intervals, the interpretation of which is restricted to the *frequentist* definition of probability as the limiting distribution of a sampling process, (e.g., as the frequency of a 'heads' in an infinite number of coin-tosses). The classic application of frequentist statistics is to answer questions of the form: *what is the probability of a obtaining a test-statistic of this magnitude or greater, given a specific parameter setting?* Where the test-statistic is computed from a probability distribution over the data with parameter set to model the 'null effect' hypothesis, to be rejected if this probability (or *P-value*) falls below 0.05.

While this definition is often useful for simple statistical tests,[6] it does not easily support the integration of uncertainty from multiple sources, which can be invaluable in leveraging all available data to obtain optimally sensitive inferences. For a demonstration of inconsistencies implied by the frequentist interpretation of probability, as well as a historical perspective explaining its dominance in 20th century statistics, see the 1986 paper by B. Efron, "Why isn't everyone a Bayesian?" [71].

If data X represent random variables generated from a probability distribution parameterized by θ, then the likelihood of the data $P(X \mid \theta)$ can be directly calculated. By inversion of this conditional probability using Bayes' rule we obtain:

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{\int P(X \mid \theta)P(\theta)d\theta}$$

Where $P(\theta \mid X)$ gives the *posterior* probability distribution on θ, summarizing the state of uncertainty on the value of θ given the prior specified as $P(\theta)$ and observed data X[7]. The key innovation of Bayesian statistical analysis is the modeling of parameters specifying random processes as themselves generated from additional probability distributions of specified form. In this formulation, the interpretation of probability is ultimately subjective, rather than as the limiting sampling distribution of the frequentist[8]. This difference in interpretation has no bearing on the mathematical properties of probability, however. At the end of the day, it is up to the practitioner to decide what the resulting probabilities mean, although this is certainly not unique to Bayesian inference.

---

[6] in which cases it is generally equivalent to a Bayesian procedure, e.g. the Student's *t*-test.

[7] Although it is often assumed that the use of Bayes' rule is equivalent to Bayesian statistics, this is incorrect, as the use of Bayesian statistics in the modern technical sense precedes Bayes himself, with the marquis de Laplace generally acknowledged as the first Bayesian.

[8] Arguments related to this particular difference in interpretation have reportedly resulted in several brawls among professional statisticians.

Although it may seem like a this difference is therefore a subtle semantic point, this shift in viewpoint allows for the development of complex statistical models, in which probabilistic dependencies among multiple variables are linked together to optimally exploit all available data. This process is formalized in the modern technique of probabilistic modeling with graphical models [72].

In particular, Hierarchical Bayesian models (HBMs), as described in [73], allow pooling of information from many instances of a phenomenon in order to estimate parameters which are modeled as shared among them. This strategy is employed directly to estimate parameters quantifying the extent of cross-hybridization between probes to either allele in SNP microarray data (Chapter 3) and in various aspects of the modeling described in Chapter 4.

In addition, we favor the 'Occam Factor' interpretation of Bayesian evidence [74], which naturally embodies the scientific precept that simpler models should be favored over more complex models with equivalent explanatory power. This principle can be used to quantitatively evaluate the evidence supporting models of differing complexity, given the available data (another major advantage of the Bayesian framework). This technique was used in the HAPSEG method (Chapter 3) in order to choose whether adjacent genomic segments truly had differing copy-number. For further ease of computation, the Laplace approximation[74] was used to approximate the posterior volume of the parameter space of the locations in each model. In addition, the formulation proposed by Spiegelhalter[75] is attractive, although we have not yet implemented such techniques.

An attractive feature of explicit modeling using HBMs is quantification of the requisite amount of prior belief required for acceptance of a given inference from the model. This allows for a rigorous examination of the importance of one's prior assumptions on the conclusions drawn from a given dataset. That being said, in the case of the work with ABSOLUTE presented in Chapter 4, there is very little doubt as to the robustness of the main biological conclusions, which are qualitatively consistent with many aspects of cellular biology and prior cytological analysis of cancer samples.

An important aspect of using complex probabilistic models is that they be *calibrated*, that is, the posterior certainty obtained by fitting to the data truly reflects the probability of the modeled event (e.g., a parameter value lying within a 95% posterior-fit confidence interval should contain the true value 95% of the time). We note that this is only generally possible *when the model describes the exact stochastic process by which the data were generated.* We note that this condition describes the ideal modeling situation, whereby all sources of uncertainty can be normalized to universally interpretable probability distributions, which can then be combined for optimal inference. While this condition is never truly satisfied in practice, it is often the case that models accurate enough to be useful can be readily constructed and fit to datasets using modern Bayesian methods, which involve various types of approximation to aid computation.

Importantly, if the a given model under consideration is inadequate to produce calibrated estimates of the uncertainty in fit parameters, then the construction of an elaborate computational scheme to estimate posterior distributions of these parameters will not result in useful inferences. Rather, in such cases effort should be directed at refining the model specification to capture more of the relevant variation in the data, leveraging as much prior information regarding the processes generating the data as is possible.

The use of hierarchically constructed karyotype models to regularize the inference of tumor purity/ploidy (Chapter 4) is an example of this philosophy. The mode-finding algorithm used to find all possible interpretations of sample allelic copy-profiles[9] finds maximum likelihood solutions; estimation of the uncertainty around these modes is irrelevant to the inference of main interest: which of the modes represents the unique correct solution. This is because the likelihood used in the mode-finding is based on a Gaussian mixture-model describing the generation of observed HSCRs from fixed SCNAs, which are simple integer values, with some allowance for subclonal alterations (modeled with uniform density over a moderate copy-ratio range). Such models are

[9] Formed from the homologue-specific copy-ratios (HSCRs) estimated using HAPSEG (Chapter 3).

vastly simpler than the biological processes which generated the data, and likely leave much of the relevant variation in the data un-modeled.

Early versions of purity/ploidy inference used Markov chain Monte Carlo (MCMC) sampling to fit an HBM with many parameters (including shared parameters among samples) to large tumor datasets (details of the implementation are presented as Appendix B of this thesis). Although this approach was successful, it was computationally cumbersome to use, and the stochastic nature of the MCMC sampling process made some results difficult to interpret. By further consideration of the problem, it was discovered that many of the parameters represented in this model could be fixed for the purpose of mode-finding, resulting in a simple two-dimensional optimization problem which was solved using numerical coordinate ascent of the model likelihood, initiated at points in regular lattice over the purity/ploidy domain (Chapter 4). Importantly, this process can be performed for each sample independently, dramatically simplifying the fitting procedure.

Instead of explicit hierarchical modeling, we incorporated pooling of sample-level parameters by clustering cancer karyotypes to construct a database of recurrent motifs. This was accomplished using the Expectation Maximization method on multinomial mixture models (Chapter 4, Online methods). This database was constructed directly from the input data itself in a boot-strapping manner, whereby less ambiguous copy-profiles were used to seed models producing provisional priors, which further reduced the ambiguity of remaining profiles, and so on. This process was analogous the use of an HBM, even though it was not necessary to explicitly construct such a model to perform the relevant inferences. In some sense, the fitting of hierarchical parameters took place in the minds of the researchers [76], [77].

Statistical excursions

In addition to the above mentioned MCMC-sampling of the multi-sample HBM for purity / ploidy inference mentioned above (and described in Appendix B), collaboration with Dr. Gad Getz resulted in several other statistical excursions over the course of my Ph.D.

In 2007-2008, we worked on a mutually devised project to predict SCNA status in genomic regions pre-defined as being significantly recurrent in human cancer using GISTIC[78] using gene expression data. We adopted a Bayesian factor analysis model, with missing indicators for each gene modeling its status as copy-dependant or not[10]. This model was fit using Gibbs sampling, which resulted in predictive models with average sensitivity and specificity of ~85% for modeled GISTIC regions (not shown). Although these results were encouraging, the production of large-scale datasets by The Cancer Genome Atlas consortium made the need to infer DNA SCNAs from gene expression data less compelling. Furthermore, although the dissection of gene-expression correlation with copy-number has since been exploited to identify putative targets of multi-genic SCNAs [79], we felt it was more pressing at the time to develop methods for absolute copy-number inference (Chapter 4).

In 2008-2009, we investigated the Dirichlet Process (DP) model [80], [81], [82], [83], and its use to integrate over all clusterings of genomic segments by copy-ratio. This led to what we believe is a novel extension of the DP to the case whereby data atoms have variable 'weights' in the clustering process. These weights corresponded to genomic segment sizes, and were allowed reweighting of the DP clustering process to favor large segments (details are given in Appendix B). These techniques may become useful for the future analysis of tumor copy-profiles, particularly in the case where several clones are superimposed in the cancer cell population.

During the same time period, we investigated extended ensemble MCMC sampling [84] for use in fitting HBMs with strongly interacting degrees of freedom. In particular, multicanonical MCMC sampling, as previously described [85], [86],[87] was implemented on several real and 'toy' problems. Although this technique is extremely powerful, and has been used to estimate key parameters of model systems in statistical physics[87], the problems addressed in this thesis did not require this level of

[10] For example, many genes expressed at or below the detection limit of the microarrays used would be better modeled as uninformative of copy-number.

sophistication. These techniques may be useful for future computational problems in cancer genomics, however.

References

1. Mukherjee S (2010) The emperor of all maladies: a biography of cancer.

2. Hanahan D, Weinberg RA (2011) Hallmarks of Cancer: The Next Generation. Cell 144: 646–674. doi:10.1016/j.cell.2011.02.013

3. Druker B, Tamura S, Buchdunger E, Ohno S (1996) Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. Nature Medicine.

4. Sawyers CL (2002) Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. Blood 99: 3530–3539. doi:10.1182/blood.V99.10.3530

5. Lynch T, Bell D, Sordella R (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. New England Journal of Medicine

6. Pao W, Miller V, Zakowski M (2004) EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proceedings of the national academy of science.

7. Paez JG (2004) EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. Science 304: 1497–1500. doi:10.1126/science.1099314

8. Pao W, Wang TY, Riely GJ, Miller VA, Pan Q, et al. (2005) KRAS Mutations and Primary Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib. Plos Med 2: e17. doi:10.1371/journal.pmed.0020017.sd003

9. Weinstein I (2002) Addiction to oncogenes--the Achilles heal of cancer. Science.

10. Davies H, Bignell G, Cox C, Stephens P, Edkins S (2002) Mutations of the BRAF gene in human cancer. Nature.

11. Marimuthu A, Nguyen H, Lam B (2008) Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. Proceedings of the national academy of science.

12. Flaherty K, Puzanov I, Kim K, Ribas A (2010) Inhibition of mutated, activated BRAF in metastatic melanoma. New England Journal of Medicine

13. Gorre M, Mohammed M, Ellwood K, Hsu N (2001) Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. Science.

14. Carter T, Wodicka L, Shah N (2005) Inhibition of drug-resistant mutants of ABL, KIT, and EGF receptor kinases. Proceedings of the national academy of science.

15. Pao W, Miller VA, Politi KA, Riely GJ, Somwar R, et al. (2005) Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. Plos Med 2: e73. doi:10.1371/journal.pmed.0020073.sg003

16. Engelman JA, Zejnullahu K, Mitsudomi T, Song Y, Hyland C, et al. (2007) MET Amplification Leads to Gefitinib Resistance in Lung Cancer by Activating ERBB3 Signaling. Science 316: 1039–1043. doi:10.1126/science.1141478

17. Bean J, Brennan C, Shih J (2007) MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib. Proceedings of the national academy of science.

18. Emery C, Vijayendran K (2009) MEK1 mutations confer resistance to MEK and B-RAF inhibition. Proceedings of the national academy of science.

19. Johannessen CM, Boehm JS, Kim SY, Thomas SR, Wardwell L, et al. (2010) COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. Nature 468: 968–972. doi:10.1038/nature09627

20. Nazarian R, Shi H, Wang Q, Kong X, Koya RC, et al. (2010) Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. Nature 468: 973–977. doi:10.1038/nature09626

21. Solimini NL, Luo J, Elledge SJ (2007) Non-Oncogene Addiction and the Stress Phenotype of Cancer Cells. Cell 130: 986–988. doi:10.1016/j.cell.2007.09.007

22. Dai C, Whitesell L, Rogers AB, Lindquist S (2007) Heat Shock Factor 1 Is a Powerful Multifaceted Modifier of Carcinogenesis. Cell 130: 1005–1018. doi:10.1016/j.cell.2007.07.020

23. Farmer H, McCabe N, Lord C, Tutt A (2005) Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature.

24. Bryant H, Schultz N, Thomas H, Parker K (2005) Specific killing of BRCA2-deficient tumours with inhibitors of poly (ADP-ribose) polymerase. Nature.

25. Helleday T, Petermann E, Lundin C, Hodgson B, Sharma RA (2008) DNA repair pathways as targets for cancer therapy. Nat Rev Cancer 8: 193–204. doi:10.1038/nrc2342

26. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, et al. (2009) Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCAMutation Carriers. N Engl J Med 361: 123–134. doi:10.1056/NEJMoa0900212

27. Tutt A, Robson M, Garber JE, Domchek SM, Audeh MW, et al. (2010) Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. The Lancet 376: 235–244. doi:10.1016/S0140-6736(10)60892-6

28. Patel A, Sarkaria J (2011) Nonhomologous end joining drives poly (ADP-ribose) polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells. Proceedings of the national academy of science.

29. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, et al. (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 461: 108–112. doi:10.1038/nature08460

30. Boveri T (1914) Boveri: Zur frage der entstehung maligner tumoren - Google Scholar.

31. Weaver BA, Cleveland DW (2006) Does aneuploidy cause cancer? Current Opinion in Cell Biology 18: 658–667. doi:10.1016/j.ceb.2006.10.002

32. Torres EM, Williams BR, Amon A (2008) Aneuploidy: Cells Losing Their Balance. Genetics 179: 737–746. doi:10.1534/genetics.108.090878

33. Loeb L (1991) Mutator phenotype may be required for multistage carcinogenesis. Cancer Research.

34. Albertson D, Collins C, McCormick F, Gray J (2003) Chromosome aberrations in solid tumors. Nat Genet 34: 369–376.

35. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90–94. doi:10.1038/nature09807

36. Storchova Z, Pellman D (2004) From polyploidy to aneuploidy, genome instability and cancer. Nature Reviews Molecular Cell Biology 5: 45–54.

37. Bunz F (1998) Requirement for p53 and p21 to Sustain G2 Arrest After

DNA Damage. Science 282: 1497–1501. doi:10.1126/science.282.5393.1497

38. Cahill D, Lengauer C, Yu J, Riggins G, Willson J (1998) Mutations of mitotic checkpoint genes in human cancers. Nature.

39. Amon A (1999) The spindle checkpoint. Current opinion in genetics & development.

40. Nasmyth K (2002) Segregating Sister Genomes: The Molecular Biology of Chromosome Separation. Science 297: 559–565. doi:10.1126/science.1074757

41. Rajagopalan H, Jallepalli P, Rago C (2004) Inactivation of hCDC4 can cause chromosomal instability. Nature.

42. Wang Z (2004) Three Classes of Genes Mutated In Colorectal Cancers with Chromosomal Instability. Cancer Research 64: 2998–3001. doi:10.1158/0008-5472.CAN-04-0587

43. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, et al. (2006) The Consensus Coding Sequences of Human Breast and Colorectal Cancers. Science 314: 268–274. doi:10.1126/science.1133427

44. Artandi SE, Chang S, Lee SL, Alson S, Gottlieb GJ, et al. (2000) Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. Nature 406: 641–645. doi:10.1038/35020592

45. Ganem NJ, Pellman D (2007) Limiting the Proliferation of Polyploid Cells. Cell 131: 437–440. doi:10.1016/j.cell.2007.10.024

46. Storchova Z, Kuffer C (2008) The consequences of tetraploidy and aneuploidy. Journal of Cell Science 121: 3859–3866. doi:10.1242/jcs.039537

47. Holland A, Cleveland D (2009) Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nature Reviews Molecular Cell Biology 10: 478–487.

48. Galipeau PC, Cowan DS, Sanchez CA, Barrett MT, Emond MJ, et al. (1996) 17p (p53) allelic losses, 4N (G2/tetraploid) populations, and progression to aneuploidy in Barrett's esophagus. Proceedings of the National Academy of Sciences of the United States of America 93: 7081–7084.

49. Barrett MT, Sanchez CA, Prevo LJ, Wong DJ, Galipeau PC, et al. (1999)

Evolution of neoplastic cell lineages in Barrett oesophagus. Nat Genet 22: 106–109. doi:10.1038/8816

50. Chandhok N, Pellman D (2009) A little CIN may cost a lot: revisiting aneuploidy and cancer. Current opinion in genetics & development.

51. Shi Q, King R (2005) Chromosome nondisjunction yields tetraploid rather than aneuploid cells in human cell lines. Nature.

52. Davoli T, Denchi EL, Lange T de (2010) Persistent Telomere Damage Induces Bypass of Mitosis and Tetraploidy. Cell 141: 81–93. doi:10.1016/j.cell.2010.01.031

53. Fujiwara T, Bandi M, Nitta M, Ivanova E, Bronson R, et al. (2005) Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. Nature 437: 1043–1047.

54. Ganem NJ, Godinho SA, Pellman D (2009) A mechanism linking extra centrosomes to chromosomal instability. Nature 460: 278–282. doi:10.1038/nature08136

55. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905. doi:10.1038/nature08822

56. Torres EM, Sokolsky T, Tucker CM, Chan LY, Boselli M, et al. (2007) Effects of Aneuploidy on Cellular Physiology and Cell Division in Haploid Yeast. Science 317: 916–924. doi:10.1126/science.1142210

57. Williams BR, Prabhu VR, Hunter KE, Glazier CM, Whittaker CA, et al. (2008) Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. Science 322: 703–709. doi:10.1126/science.1160058

58. Sheltzer JM, Blank HM, Pfau SJ, Tange Y, George BM, et al. (2011) Aneuploidy Drives Genomic Instability in Yeast. Science 333: 1026–1030. doi:10.1126/science.1206412

59. Lengauer C, Kinzler K, Vogelstein B (1998) Genetic instabilities in human cancers. Nature.

60. Whitesell L, Lindquist SL (2005) HSP90 and the chaperoning of cancer. Nat Rev Cancer 5: 761–772. doi:10.1038/nrc1716

61. Torres EM, Dephoure N, Panneerselvam A, Tucker CM, Whittaker CA, et al. (2010) Identification of Aneuploidy-Tolerating Mutations. Cell 143: 71–83. doi:10.1016/j.cell.2010.08.038

62. Torres EM, Dephoure N, Panneerselvam A, Tucker CM, Whittaker CA, et al. (2010) Identification of Aneuploidy-Tolerating Mutations. Cell: 1–13. doi:10.1016/j.cell.2010.08.038

63. Tang Y-C, Williams BR, Siegel JJ, Amon A (2011) Identification of Aneuploidy-Selective Antiproliferation Compounds. Cell: 1–14. doi:10.1016/j.cell.2011.01.017

64. Almond J (2002) The proteasome: a novel target for cancer chemotherapy. Official journal of the Leukemia Society

65. Rajkumar SV (2004) Proteasome Inhibition As a Novel Therapeutic Target in Human Cancer. Journal of Clinical Oncology 23: 630–639. doi:10.1200/JCO.2005.11.030

66. Stebbins C, Russo A, Schneider C, Rosen N (1997) Crystal structure of an Hsp90-geldanamycin complex: targeting of a protein chaperone by an antitumor agent. Cell.

67. Yuan B (2006) Increased Expression of Mitotic Checkpoint Genes in Breast Cancer Cells with Chromosomal Instability. Clinical Cancer Research 12: 405–410. doi:10.1158/1078-0432.CCR-05-0903

68. Winckler W, Myers S, Richter D, Onofrio R, McDonald G, et al. (2005) Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees. Science 308: 107–111. doi:10.1126/science.1105322

69. Browning BL, Yu Z (2009) Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. The American Journal of Human Genetics 85: 847–861. doi:10.1016/j.ajhg.2009.11.004

70. Consortium TIH3, investigators P, leaders PC, group MW, QC GA, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58. doi:10.1038/nature09298

71. Efron B (1986) Why isn't Everyone a Bayesian? American Statistician.

72. Jordan M (1998) Learning in graphical models - Google Books.

73. Gelman A (2004) Bayesian data analysis. CRC Press. 668 p.

74. MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge Univ Pr. 628 p.

75. Spiegelhalter D, Best N, Carlin B, van der Linde A (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical

Society: Series B (Statistical Methodology) 64: 583–639.

76. Kemp C (2008) The discovery of structural form. Proceedings of the National Academy of Science.

77. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to Grow a Mind: Statistics, Structure, and Abstraction. Science 331: 1279–1285. doi:10.1126/science.1192788

78. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences 104: 20007.

79. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, et al. (2010) An Integrated Approach to Uncover Drivers of Cancer. Cell 143: 1005–1017. doi:10.1016/j.cell.2010.11.013

80. Ferguson T (1973) A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1: 209–230.

81. Blackwell D (1973) Discreteness of Ferguson selections. The Annals of Statistics 1: 356–358.

82. Antoniak C (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2: 1152–1174.

83. Escobar M, West M (1995) Bayesian Density Estimation and Inference Using Mixtures. Journal of the american statistical association 90.

84. Iba Y (2000) Extended Ensemble Monte Carlo. arXiv cond-mat.dis-nn. doi:10.1142/S0129183101001912

85. Getz G, Shental N, Domany E (2006) Semi-Supervised Learning--A Statistical Physics Approach. Arxiv preprint cs/0604011.

86. Wang F, Landau D (2001) Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. Physical Review E 64: 56101.

87. Wang F, Landau D (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. Physical Review Letters 86: 2050–2053.

Chapter 2

—

A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers

Scott L. Carter, Aron C Eklund, Isaac S Kohane, Lyndsay N Harris, and Zoltan Szallasi

This chapter is presented in the context of its contemporary science and originally appeared in *Nature Genetics* 38, 1043 - 1048 (2006).

Corresponding Supplementary Material can be found in Appendix A.

A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers

Scott L Carter, Aron C Eklund, Isaac S Kohane, Lyndsay N Harris & Zoltan Szallasi

We developed a computational method to characterize aneuploidy in tumor samples based on coordinated aberrations in expression of genes localized to each chromosomal region. We summarized the total level of chromosomal aberration in a given tumor in a univariate measure termed total functional aneuploidy. We identified a signature of chromosomal instability from specific genes whose expression was consistently correlated with total functional aneuploidy in several cancer types. Net overexpression of this signature was predictive of poor clinical outcome in 12 cancer data sets[1–12] representing six cancer types. Also, the signature of chromosomal instability was higher in metastasis samples than in primary tumors and was able to stratify grade 1 and grade 2 breast tumors according to clinical outcome. These results provide a means to assess the potential role of chromosomal instability in determining malignant potential over a broad range of tumors.

One of the most consistent characteristics of human solid tumors is chromosomal instability (CIN)[13]. Despite its fundamental importance in cancer biology, the molecular mechanisms underlying CIN are poorly understood[14,15], and it is rarely measured in the clinic owing to the technical difficulty of ascertainment and a lack of clear therapeutic application[16].

Because aneuploidy is a consequence of CIN, genes with expression levels consistently associated with aneuploidy might provide insight into molecular mechanisms underlying CIN. Currently, few clinical tumor samples exist for which the extent of aneuploidy has been directly observed and for which high-resolution gene expression data are available. However, significant correspondence has been observed between chromosome copy number alterations and gene expression changes in the affected regions[17,18]. Therefore, we used gene expression data to infer chromosomal imbalance at sub-band resolution (**Fig. 1**). This 'functional aneuploidy' profile describes the net deviation in expression of genes contained in each chromosomal region relative to the remainder of the sampled transcriptome (see Methods).

As a measure of overall chromosomal imbalance, we defined the 'total functional aneuploidy' (tFA) of a sample as the sum of the magnitudes of its functional aneuploidy features.

We assessed tFA across the NCI60 cell line panel, for which gene expression profiles, DNA copy number profiles (SNP-chip)[19] and spectral karyotyping (SKY) data[20] are available, providing independent DNA-based characterization of chromosomal aberrations. The tFA measure correlated with the total number of structural chromosomal aberrations as assessed by SKY ($P = 0.0009$) and with total aneuploidy measures derived from DNA copy number measurements ($P = 0.0004$; **Supplementary Fig. 1** online). Thus, tFA is a proxy for the overall level of aneuploidy assessed at the transcriptional level (**Fig. 1b,c**).

We analyzed 18 gene expression data sets[1–12] (**Supplementary Fig. 2** online) of various cancer types and found that high tFA was a significant predictor of poor clinical outcome in four data sets profiling breast, lung and brain tumors[2,4,6,7] (**Supplementary Fig. 2**). Given this association with poor prognosis, we reasoned that genes associated with tFA might uncover specific molecular mechanisms connecting high levels of CIN with increased tumor malignancy. Within each data set, we calculated the correlation between tFA and expression level for each gene. The correlation of expression of each gene to tFA was similar across data sets (**Supplementary Fig. 3** online), suggesting the existence of a common mechanism involved in the generation or maintenance of aneuploidy.

We derived a 'CIN score' for 10,151 genes present on multiple platforms by combining the ranked level of correlation of each gene to tFA in multiple data sets from diverse tumor types (see Methods). The gene with highest CIN score, *TPX2*, has a key role in chromosomal segregation and causes polyploidization when overexpressed[21]. The second–highest ranked, *PRC1*, encodes a regulator of cytokinesis[22] and a putative substrate of the product of another top-ranking gene, *CDC2* (ref. 23). *FOXM1* encodes a transcriptional regulator of genes involved in chromosomal segregation and cytokinesis[24], such as *KIF20A* (ref. 25), which also has one of the highest CIN scores. In fact, the list of 70 top-ranking genes contain an array of key regulators that maintain faithful replication and segregation of chromosomes, including *AURKB*, *NEK2*, *H2AFX*, *CDC20*, *ZWINT*, *CCNB1*, *CCNB2* and *AURKA* (**Supplementary Table 1** online). Therefore, these genes may represent members of one or more functional modules, the overexpression of which causes or is associated with CIN.

We examined the ability of the top-ranked genes to predict clinical outcome in the 18 aforementioned cancer data sets. We stratified patients into two groups, based on whether the top 25 or 70 genes with the highest CIN score (designated the CIN25 and CIN70 signatures,

respectively) in each individual were above or below mean expression of the CIN25 or CIN70 signature in the cohort (**Fig. 1d,e**). This simple classifier (CIN25) was a significant predictor of clinical outcome in 12 of the 18 data sets representing breast cancer[1–5], lung cancer[6], medulloblastoma[10], glioma[7,9,12], mesothelioma[8] and lymphoma[11] (**Fig. 2**). (Nonsignificant predictions are shown in **Supplementary Fig. 4** online.) We obtained similar results using the CIN70 signature (13 of 18 data sets were stratified according to clinical outcome; data not shown).

We next evaluated the CIN25 and CIN70 signatures in data from a study that examined primary tumors and metastatic foci from several solid tumor types[26]. The metastasis samples expressed a higher level of the CIN25 signature than did the primary tumors ($P = 0.0067$, **Fig. 3** and **Supplementary Methods** online). The same analysis using the CIN70 signature gave a similar result ($P = 0.0035$; data not shown). Furthermore, we observed a subset of primary tumors with higher relative expression of the signatures, consistent with the hypothesis that the metastatic expression profile originates in primary tumors and is maintained in the distant metastases.

Clinically, the degree of malignancy is thought to be reflected by tumor grade, which is a subjective histological assessment of cellular differentiation and proliferation. We examined the relationship between the CIN25 signature and tumor grade in three breast cancer data sets[1–3] for which tumor grade data were available. We stratified grade 1 and grade 2 tumors according to their CIN25 signature expression and found that high-CIN25 tumors within each of these grade categories were associated with a worse clinical outcome (**Fig. 4**). However, grade 3 tumors were uniformly associated with high CIN25 signature expression and could not be separated using this measure. This result supports the recent suggestion that histopathological grading might be refined or replaced by gene expression signatures, which more accurately reflect patient outcome[1].

Because many of the genes we identified as associated with CIN are periodically expressed during the mitotic cell cycle, it was formally possible that the prognostic ability of the CIN signature was due to its detection of proliferation rate rather than to its detection of CIN itself. Previous work has examined genome-wide patterns of gene expression over time in synchronized HeLa cells[27]. Each gene was assigned a score proportional to the periodicity of its expression over the time points after synchronization[27]. We considered 888 unique genes

(represented by 1,133 clones on the cDNA microarrays used) to be cell cycle regulated; these had cell cycle scores exceeding an empirically determined false discovery rate (FDR) threshold[27]. The total expression level of these genes in a tissue sample is thought to be a surrogate measure of cell proliferation rate, which is associated with poor prognosis in human cancer[27].

The CIN score and cell cycle scores were modestly correlated over 10,795 matched measurements representing 5,484 unique genes ($r = 0.186$), although the genes with the highest CIN scores were more likely to have high cell cycle scores (**Fig. 5a**). To compare the utility of the CIN score and the cell cycle score for choosing predictive genes, we systematically dissected the two signatures.

First, we divided the 70 genes with the highest CIN scores (CIN70) into a set that was considered cell cycle regulated (CIN70_CC), and one that was not (CIN70_noCC) (**Fig. 5a**). We found that the CIN70_noCC group generated significant outcome predictions in 10 of the 18 tested cohorts, whereas the CIN70_CC group did so in 11. Therefore, the predictive ability of the CIN70 signature was not solely due to the inclusion of cell cycle–regulated genes. We then split the set of cell cycle–regulated genes into two groups based on their mean CIN score (**Fig. 5a**). Although the high-CIN cell cycle group generated significant predictions in 11 of 18 cohorts tested, the low CIN-score group worked in only five (**Fig. 5b** and **Supplementary Fig. 5** online).

We evaluated functional categories representing specific aspects of the cell cycle implicated in CIN[14,15] for CIN score enrichment relative to the set of cell cycle genes. Gene sets representing the mitotic spindle assembly ($P = 0.003$), spindle checkpoint ($P = 0.00078$) and DNA damage checkpoint ($P = 0.048$) had significantly greater CIN scores than the cell cycle genes (**Fig. 5c**), consistent with the hypothesis that overexpression of these specific cell cycle control genes is associated with CIN and is not simply due to an increased proportion of mitotic cells. Gene sets representing DNA repair, centrosome duplication and chromosome segregation were not significantly enriched in CIN score relative to cell cycle–regulated genes ($P = 0.52$, 0.64, 0.99, respectively).

We also evaluated the gene sets with CIN score enrichment for enrichment in cell cycle score[27] (**Fig. 5d**). While the mitotic spindle assembly and spindle checkpoint gene sets showed strong enrichment for cell cycle score, the DNA damage checkpoint gene set showed only modest enrichment. The fact that DNA damage checkpoint genes are not cell cycle regulated but

tend to be overexpressed in aneuploid tumors suggests that transcriptional induction of the DNA damage checkpoint is specifically associated with CIN. This result was in agreement with a previous finding that activation of the DNA damage checkpoint in preneoplastic and neoplastic lesions is not directly related to the proliferative status of tumor cells[28].

There was also an overlap of four genes between our CIN25 signature and a previously described 44-gene 'proliferation cluster'[29] that is overexpressed in high-grade tumors[29]. Multivariate analysis confirmed that the CIN signature had greater prognostic relevance (**Supplementary Fig. 6** online), and systematic dissection of the proliferation signature demonstrated that its prognostic utility was due to its inclusion of CIN-associated genes (**Supplementary Fig. 7** online).

The overexpression of the CIN signature in cells with high levels of aneuploidy may be due to at least three distinct mechanisms. It is possible that cells with aberrant DNA content produce more of the machinery required for chromosomal duplication and segregation. Alternatively, the expression level of the CIN signature may reflect a compensatory mechanism for impaired functioning of the machinery responsible for maintaining the integrity of genetic information at the chromosomal level. This possibility would provide potential targeting strategies for cancer treatment. A third possibility is that the overexpression of the CIN genes allows cells to complete mitosis unimpeded by the usual checkpoints, thus conferring both direct proliferative advantage and increased cellular heterogeneity.

Our results focus attention on a specific set of genes involved in cellular processes critical for maintenance of genome integrity that are relevant both biologically and clinically. The CIN signature stratifies multiple tumor types according to clinical outcome across tumor grade and shows elevated expression in metastatic foci, suggesting the importance of these genes in defining a more aggressive cancer phenotype. Dissecting the mechanisms by which CIN arises and contributes to the malignant phenotype remains a major challenge in cancer research. By identifying specific genes associated with CIN, we hope to elucidate a key process underlying malignant transformation. In addition, these results support the more widespread application of chromosomal instability measures in clinical diagnostics, which have been rarely used owing to technical difficulty and unclear significance[16]. Finally, this work identifies important cellular targets that may yield broadly applicable therapies.

METHODS

Integration of gene expression data sets. The methods used to integrate and normalize the microarray data sets are given in **Supplementary Methods**.

Functional aneuploidy measure. For a given data set, all of the normalized gene expression measurements present on the microarray and mapping to a given chromosomal cytoband region were grouped into a set designated 'B' (short for band). The rest of the genes, localized elsewhere on the genome, were grouped into a set 'G' (short for genome). The functional aneuploidy measure for the given cytoband is the value of Student's *t* statistic comparing sets B and G:

$$t = \frac{\mu_B - \mu_G}{\sqrt{\left(\sigma_B^2 / N_B\right) + \left(\sigma_G^2 / N_G\right)}}$$

where m is the mean, s^2 is the variance and N is the number of genes within a set. This process was iterated for each gene expression profile in a given data set such that we obtained a matrix of *t* statistics for each of 300 cytobands per array. If fewer than ten genes were present in a given cytoband, then we considered the statistical measure unreliable, and that cytoband was eliminated from further analysis. This left 200–250 features per sample in a given data set, depending on the coverage of the microarray platform used.

'Total functional aneuploidy' (tFA) is the sum of all functional aneuploidy magnitudes (the absolute *t* statistics) in a given tumor sample, giving a single measure of spatially coordinated transcriptional activity.

Derivation of the CIN25 signature. First, we calculated, for each data set, the tFA for each sample. We then computed for all genes the correlation coefficient across all samples between each gene's expression vector and the tFA vector. We considered only the 10,151 genes measured on each of the Affymetrix U133A, U133+2 and Rosetta 25k microarray platforms. We then ranked the genes in each data set by their correlation with tFA. Ranks were normalized within each data set such that the gene with the highest (positive) correlation had rank 1.0 and the gene with lowest (negative) correlation had rank 0. We then selected three of the data sets representing diverse tumor types (breast[10], ovarian[13] and small-cell lung carcinoma[13]), with the highest coverage of the transcriptome (for which we had greater confidence in our computations of functional aneuploidy and tFA), and we summed the ranks of each gene to derive the final integrated ranking (**Supplementary Fig. 3**). The top 25 genes from this ranking formed the

CIN25 signature. The CIN70 signature was derived in a similar manner using the 70 top-ranked genes.

ACKNOWLEDGMENTS

We thank A. Amon, D. Botstein, M. Meyerson and T. Ried for helpful suggestions, and Novartis for the NCI60 expression data. This work was supported in part by the US National Institutes of Health through grant 1PO1CA-092644-01 and the Department of Defense through grant W81XWH-04-1-0549. I.S.K. was supported in part by National Institutes of Health National Center for Biomedical Computing grant 5U54LM008748-02.

AUTHOR CONTRIBUTIONS

S.L.C. and Z.S. conceived of and designed the study. S.L.C. carried out all the analysis. S.L.C. and Z.S. wrote the manuscript. A.C.E., I.S.K. and L.N.H. provided guidance and participated in the preparation of the manuscript.

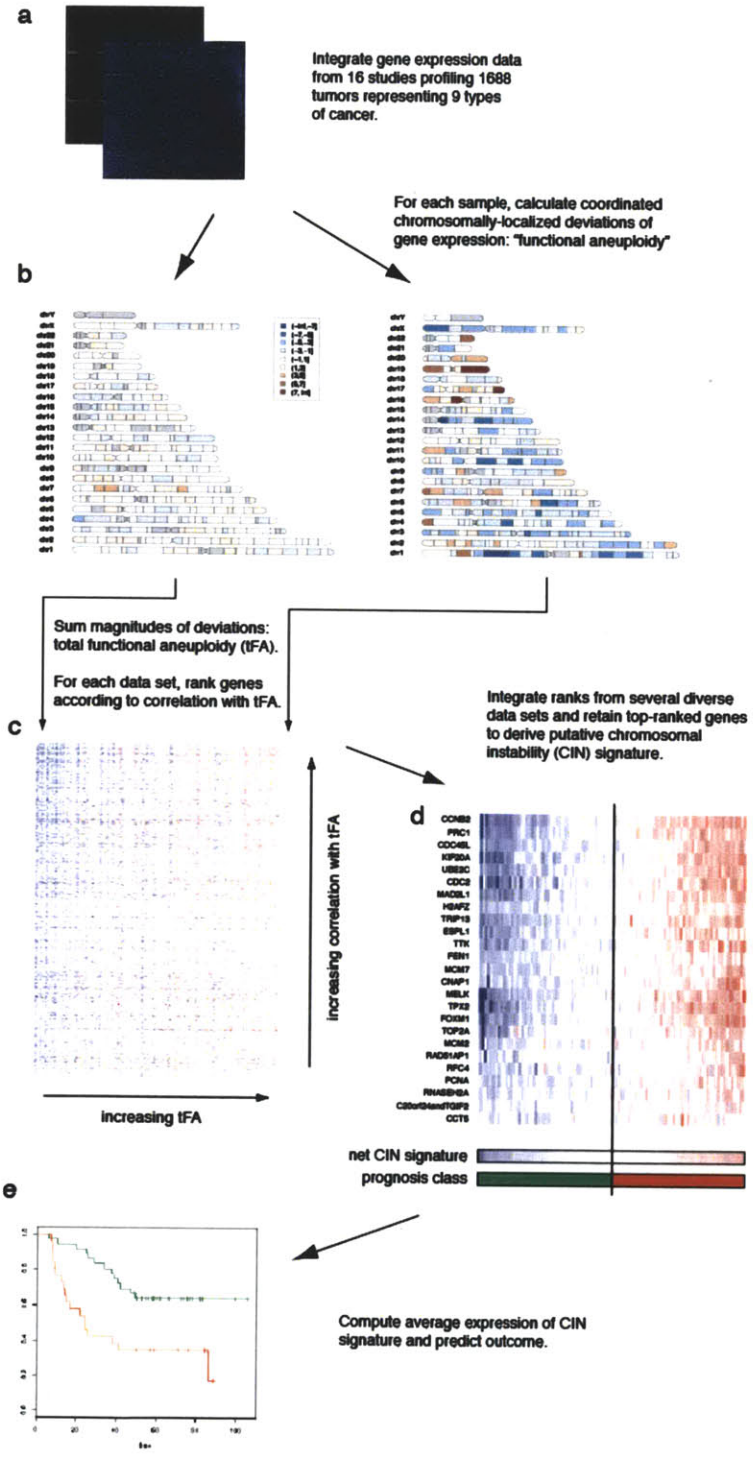


Figure 1 Schematic overview of the process used to derive and apply a gene expression signature of chromosomal instability. (**a**) Published microarray data sets from various microarray

platforms were integrated for the analysis. (**b**) Profiles of functional aneuploidy were computed for each tumor as described in the text. Shown are two examples with low (left) and high (right) levels of total functional aneuploidy (tFA). The color and intensity of the individual cytobands indicate the direction and magnitude of the differential expression of the genes contained in the given chromosomal region. (**c**) Heat map showing expression of the 500 genes most correlated with tFA, the top 25 of which define the CIN25 signature. (**d**) Heat map showing normalized expression levels of the CIN25 signature for a cancer-associated data set. Tumors with an above-average level of CIN25 signature net expression are predicted to have a poor clinical outcome (bar at the bottom indicates clinical outcome). (**e**) Censored survival (Kaplan-Meier) curves for each predicted category.

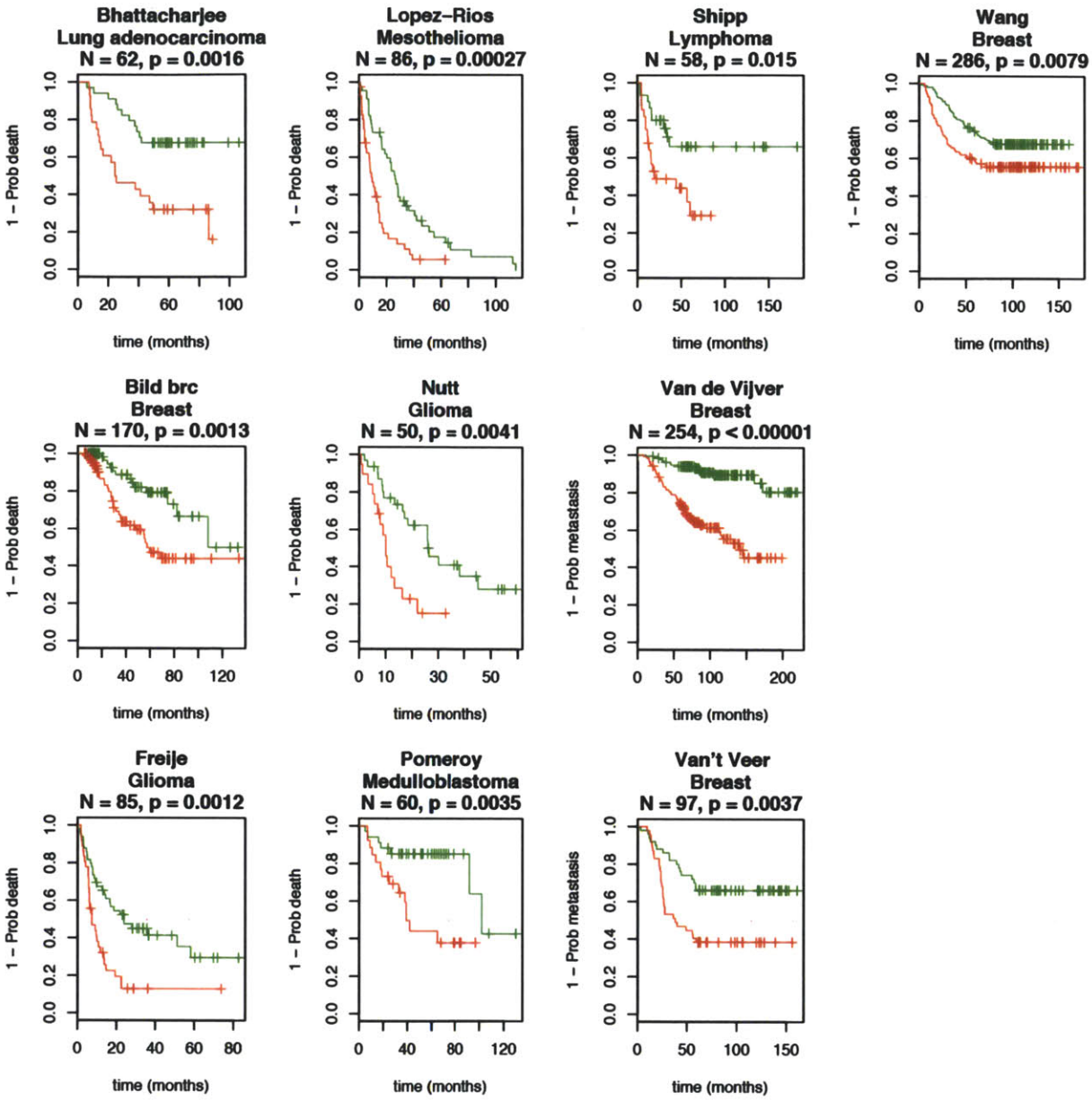


Figure 2 The CIN25 signature predicts survival in 12 independent cohorts representing six cancer types. Data sets are identified by the first author of the published manuscript and the type of cancer profiled. *P* values correspond to the log-rank test comparing the survival curves generated from the CIN25 signature as a categorical predictor.

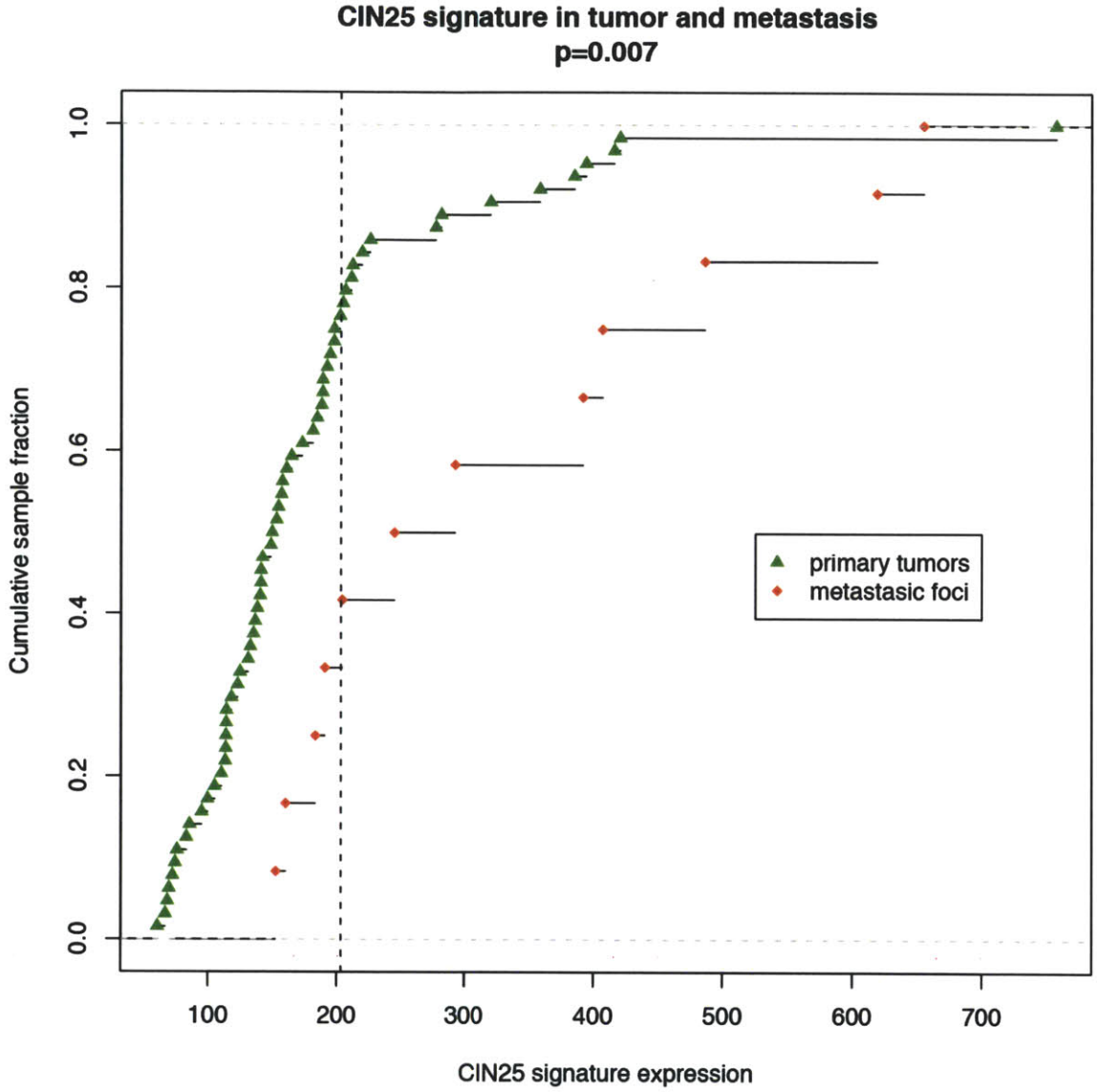


Figure 3 Metastatic foci expressed higher levels of the net CIN25 signature than primary tumors of diverse origin. Shown are cumulative density plots of CIN25 expression for the primary tumors (green) and metastases (red). The vertical dashed line indicates the average level of the CIN25 signature's net expression for all samples. *P* values assessing the significance of increased CIN25 signature expression in metastases were calculated from one-sided *t* tests, as decreased expression would contradict our previous results that the CIN signature is associated with a more advanced tumor phenotype. A subset of the primary tumors showed elevated levels of CIN25 signature expression, possibly indicating higher malignant potential in these tumors.

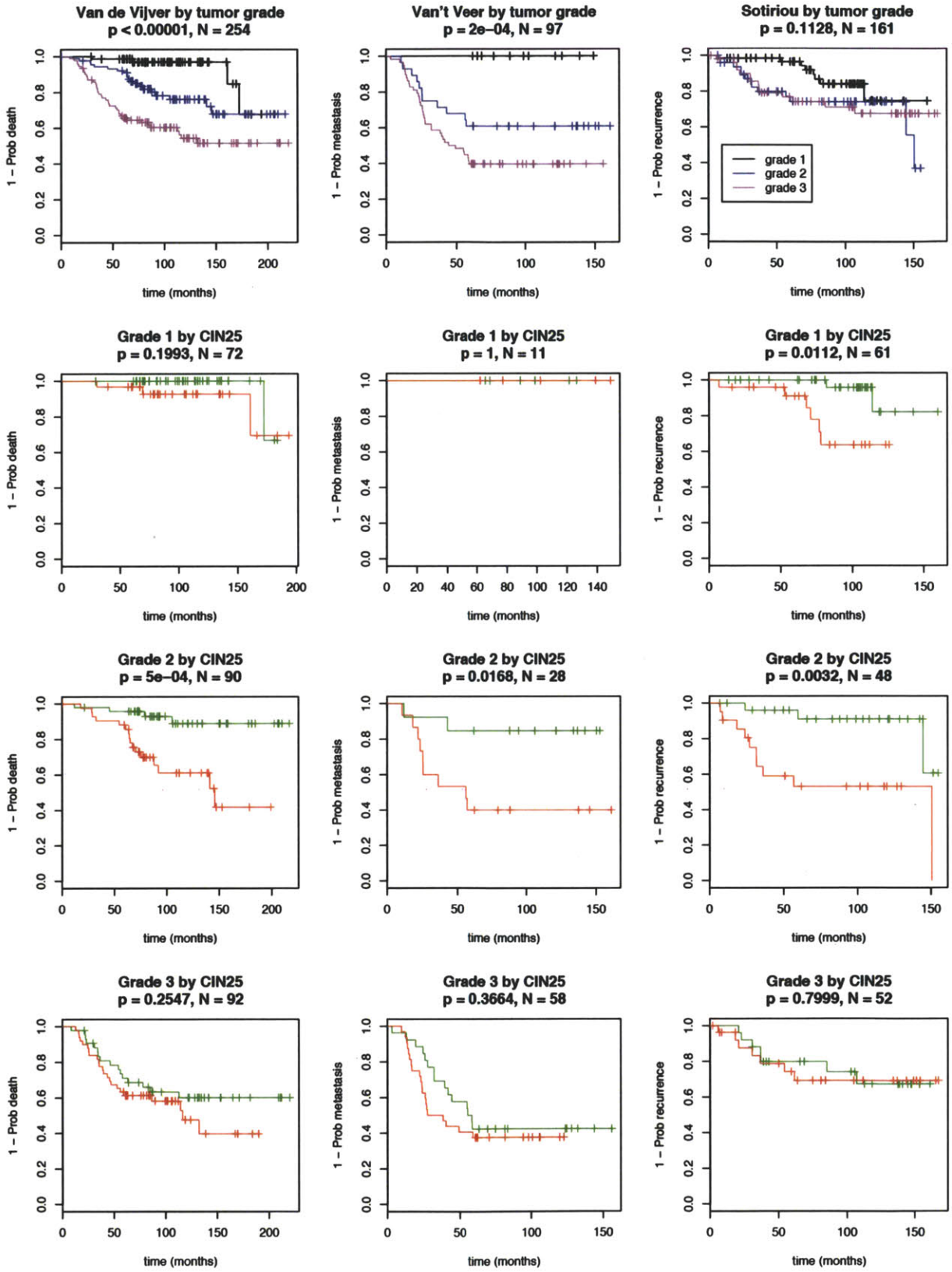

Van de Vijver by tumor grade
p < 0.00001, N = 254
Van't Veer by tumor grade
p = 2e−04, N = 97
Sotiriou by tumor grade
p = 0.1128, N = 161
grade 1
grade 2
grade 3
Grade 1 by CIN25
p = 0.1993, N = 72
Grade 1 by CIN25
p = 1, N = 11
Grade 1 by CIN25
p = 0.0112, N = 61
Grade 2 by CIN25
p = 5e−04, N = 90
Grade 2 by CIN25
p = 0.0168, N = 28
Grade 2 by CIN25
p = 0.0032, N = 48
Grade 3 by CIN25
p = 0.2547, N = 92
Grade 3 by CIN25
p = 0.3664, N = 58
Grade 3 by CIN25
p = 0.7999, N = 52
1 − Prob death
1 − Prob metastasis
1 − Prob recurrence
time (months)

Figure 4 The CIN25 signature stratifies three independent cohorts of grade 2 breast tumors according to clinical outcome, and one cohort of grade 1 breast tumors. The CIN25 signature did not stratify any of the grade 3 cohorts according to clinical outcome, despite the presence of heterogeneity in clinical outcome among these tumors. *P* values correspond to the log-rank test comparing each pair of survival curves.

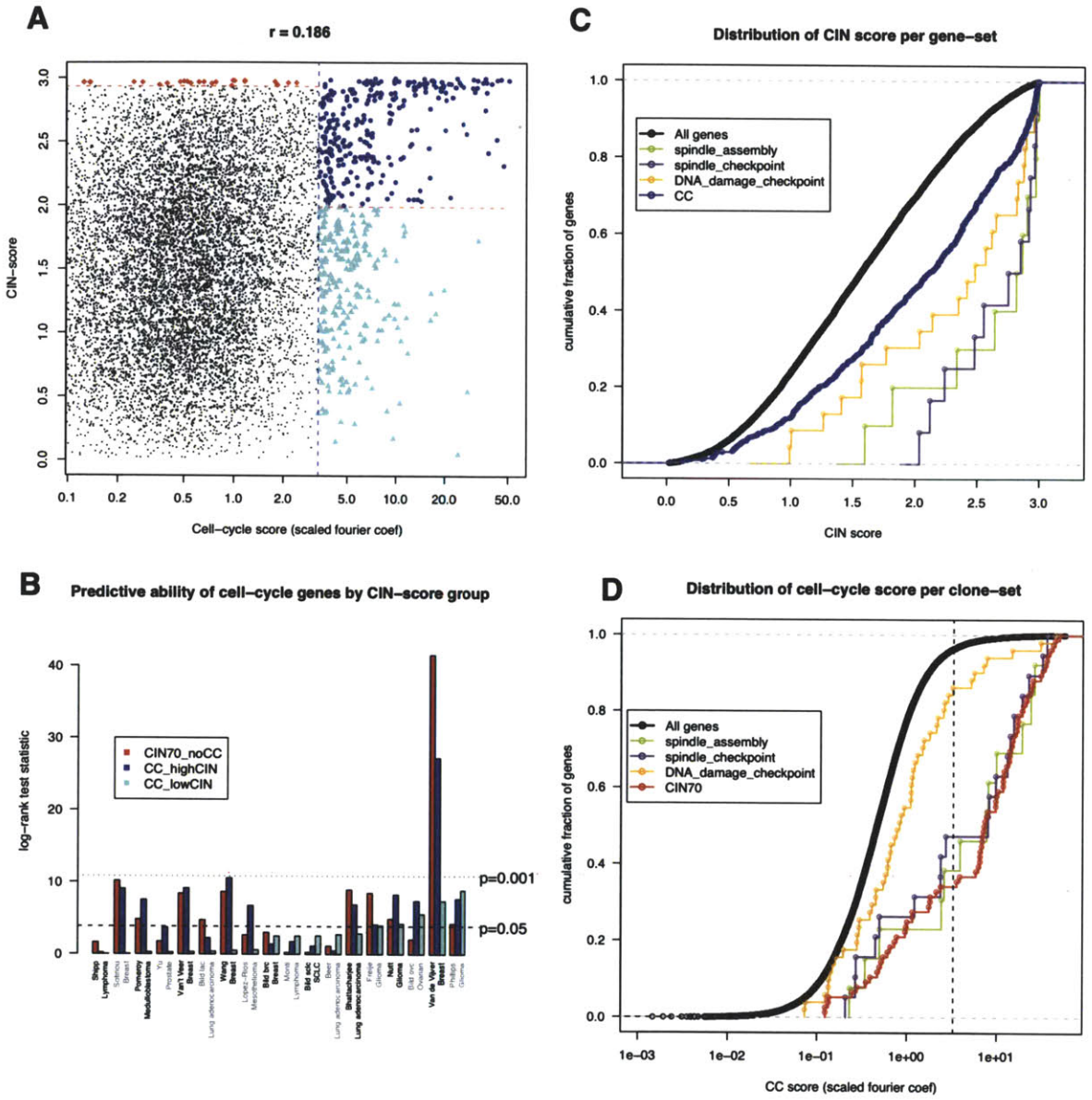


Figure 5 The ability of CIN genes to predict clinical outcome is independent of their cell cycle score. (**a**) A plot of cell cycle score versus CIN score shows modest overall correlation. Gene sets were formed by imposing combined thresholds on CIN score and cell cycle score: CIN70_noCC (red, the CIN70 genes with the 41 cell cycle–regulated genes removed). The 764 unique cell cycle–regulated genes were divided into two groups, based on whether their CIN score was above or below average relative to the cell cycle group (blue and turquoise, termed CC_highCIN and CC_low CIN, respectively). (**b**) Comparison of prediction performance of

various gene sets. Each bar indicates the log-rank test statistic comparing the survival curves corresponding to the two prognosis classes generated using the indicated gene set for outcome prediction. The CIN70-CC signature generated significant predictions for 11 data sets, and the CIN70-noCC was effective in 10, confirming that the general prognostic power of the CIN genes is not due to a direct association with proliferation. The CC_lowCIN signature classified only five data sets significantly. (**c**) Gene categories implicated in the faithful propagation of genomic information, such as mitotic spindle assembly, spindle checkpoint and DNA damage checkpoint, are enriched for high CIN scores, relative to all cell cycle–regulated genes (one-sided Wilcoxon test, $P = 0.003$, $P = 0.00078$, $P = 0.048$, respectively). (Gene category membership is based on NCBI Entrez Gene as of 25 April 2006.) The cumulative density of CIN scores is plotted for the genes in the indicated categories. (**d**) Although the mitotic spindle assembly and spindle checkpoint gene sets show high enrichment for cell cycle score, the DNA damage checkpoint gene set showed only modest enrichment. Shown are cumulative density plots of cell cycle score for the clones mapping to genes in the indicated categories.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics*) website for details.

Published online at http://www.nature.com/naturegenetics

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262–272 (2006).

2. van de Vijver, M.J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).

3. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).

4. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679 (2005).

5. Bild, A.H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357 (2006).

6. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790–13795 (2001).

7. Phillips, H.S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157–173 (2006).

8. Lopez-Rios, F. *et al.* Global gene expression profiling of pleural mesotheliomas: overexpression of aurora kinases and P16/CDKN2A deletion as prognostic factors and critical evaluation of microarray-based prognostic prediction. *Cancer Res.* **66**, 2970–2979 (2006).

9. Freije, W.A. *et al.* Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.* **64**, 6503–6510 (2004).

10. Pomeroy, S.L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).

11. Shipp, M.A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68–74 (2002).

12. Nutt, C.L. *et al.* Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* **63**, 1602–1607 (2003).

13. Lengauer, C., Kinzler, K.W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).

14. Gollin, S.M. Mechanisms leading to chromosomal instability. *Semin. Cancer Biol.* **15**, 33–42 (2005).

15. Draviam, V.M., Xie, S. & Sorger, P.K. Chromosome segregation and genomic stability. *Curr. Opin. Genet. Dev.* **14**, 120–125 (2004).

16. Diaz, L.A., Jr. The current clinical value of genomic instability. *Semin. Cancer Biol.* **15**, 67–71 (2005).

17. Pollack, J.R. *et al.* Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA* **99**, 12963–12968 (2002).

18. Tonon, G. *et al.* High-resolution genomic profiles of human lung cancer. *Proc. Natl. Acad. Sci. USA* **102**, 9625–9630 (2005).

19. Garraway, L.A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).

20. Roschke, A.V. *et al.* Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer Res.* **63**, 8634–8647 (2003).

21. Heidebrecht, H.J. *et al.* Repp86: a human protein associated in the progression of mitosis. *Mol. Cancer Res.* **1**, 271–279 (2003).

22. Kurasawa, Y., Earnshaw, W.C., Mochizuki, Y., Dohmae, N. & Todokoro, K. Essential roles of KIF4 and its binding partner PRC1 in organized central spindle midzone formation. *EMBO J.* **23**, 3237–3248 (2004).

23. Mollinari, C. *et al.* PRC1 is a microtubule binding and bundling protein essential to maintain the mitotic spindle midzone. *J. Cell Biol.* **157**, 1175–1186 (2002).

24. Costa, R.H. FoxM1 dances with mitosis. *Nat. Cell Biol.* **7**, 108–110 (2005).

25. Wonsey, D.R. & Follettie, M.T. Loss of the forkhead transcription factor FoxM1 causes centrosome amplification and mitotic catastrophe. *Cancer Res.* **65**, 5181–5189 (2005).

26. Ramaswamy, S., Ross, K.N., Lander, E.S. & Golub, T.R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).

27. Whitfield, M.L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).

28. Gorgoulis, V.G. *et al.* Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature* **434**, 907–913 (2005).

29. Whitfield, M.L., George, L.K., Grant, G.D. & Perou, C.M. Common markers of proliferation. *Nat. Rev. Cancer* **6**, 99–106 (2006).

Chapter 3

—

Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping

Scott L. Carter, Matthew L. Meyerson, Gad Getz

This chapter was submitted in its present form for publication on August 10th, 2011.

Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping

Scott L. Carter[1,3,*], Matthew Meyerson[2,3], Gad Getz[3,*]
1 Division of Health Science and Technology, MIT, Cambridge, MA, USA
2 Dana-Farber Cancer Institute, Boston, MA, USA
3 Broad Institute, Cambridge, MA, USA
*** E-mail: scarter@broadinstitute.org, gadgetz@broadinstitute.org**

Abstract

Interpretation of allelic copy measurements at polymorphic markers in cancer samples presents distinctive challenges and opportunities. Due to frequent gross chromosomal alterations occurring in cancer (aneuploidy), many genomic regions are present at homologus-allele imbalance. Within such regions, the unequal contribution of alleles at heterozygous markers allows for direct phasing of the haplotype derived from each individual parent. In addition, genome-wide estimates of homologue specific copy-ratios (HSCRs) are important for interpretation of the cancer genome in terms of fixed integral copy-numbers. We describe HAPSEG, a probabilistic method to interpret bi-allelic marker data in cancer samples. HAPSEG operates by partitioning the genome into segments of distinct copy number and modeling the four distinct genotypes in each segment. We describe general methods for fitting these models to data which are suitable for both SNP microarrays and massively parallel sequencing data. In addition, we demonstrate a specially tailored error-model for interpretation of systematic variations arising in microarray platforms. The ability to directly determine haplotypes from cancer samples represents an opportunity to expand reference panels of phased chromosomes, which may have general interest in various population genetic applications. In addition, this property may be exploited to interrogate the relationship between germline risk and cancer phenotype with greater sensitivity than is possible using unphased genotype. Finally, we exploit the statistical dependency of phased genotypes to enable the fitting of more elaborate sample-level error-model parameters, allowing more accurate estimation of HSCRs in cancer samples.

Author Summary

The human genome typically exists in two copies of each chromosome, with one copy, or *homologue* derived from either parent. One of the most fundamental hallmarks of human cancers is their tendency to have *aneuploid* genomes, that is, unbalanced copy-number alterations in the genetic material at various blocks of the normal human genome. This often results in unequal contributions of the homologues derived from either parent to the genomes of cancer cells. Estimation of the precise contribution of each homologue in a DNA sample obtained from cancer tissue is crucial to understand the genetic alterations occurring specifically in the cancer cells. Such estimation requires the identification and interpretation of alterations in the genetic sequence of each homologue, such as they appear in the cancer sample, which contains a mixture of DNA derived from both

cancerous and normal cells. We have developed a statistical method to accurately model the resulting data, producing summarizations of the information from hundreds of thousands of data points into hundreds of segments, describing the concentration of both homologues for large genomic regions of equal copy-number. In addition, we show that the tendency of cancer samples to be aneuploid may be exploited to attribute contiguous chromosomal blocks of genetic variation to one or the other homologue, representing the contribution of a single parent. Because they can help reduce the dimensionality of the data, such summarizations will be useful to further understand the interaction between inherited genetic variation and cancer development.

Introduction

The genomes of human cancer cells frequently harbor copy-number alterations, ranging from focal gain and loss of small regions to widespread chromosomal aneuploidy [1], [2], in many cases exacerbated by DNA ploidy increases followed by predominant attrition of the fixed DNA in the evolving somatic clone [3]. Allelic analysis of cancer genomes offers several advantages in analysis of cancer genomes [4], [5]. Detection of genomic regions with fixed somatic loss of heterozygosity - (LOH) helps identify recessively inactivated tumor suppressors, carrying mutations on the retained allele [6]. In particular, allelic measurements are required for detection of copy-neutral LOH, which may occur either due to compensatory gain of the retained alleles, or by homologous recombination. Finally, genome-wide HSCR estimates provide the foundation for inference of tumor-nuclei percentage, cancer-genome ploidy, and integral allelic copy-numbers [7], [8], [9], [3].

Human genomes are normally diploid, with one haploid genome inherited from each parent (although tetraploid cells are also involved in physiological processes [10].) As a result of widespread chromosomal aneuploidy, many genomic regions are in homologos-allele imbalance, with the two homologues fixed at unequal copy-numbers within the somatic clone [3]. Methods for genotyping diploid samples are unsuitable for analysis of aneuploid cancer samples. For example, discovery of both germline and somatic point-mutations in cancer samples using massively paralell sequencing typically requires analysis of paired cancer-normal DNAs [11], [12]. For SNP microarrays, probes for each SNP allele are calibrated using diploid control samples [13], making genotype calls unreliable in aneuploid samples [14].

Accurate estimation of homologue-specific DNA concentrations in tumor samples is challenging due to substantial marker-level noise occurring on a background of complex biases related to the genotypes, sequence context, allelic copy-numbers at which loci are observed. The particular manifestation of these factors in the resulting data depends on physical properties of the technology used. Development of explicit generative models describing these effects in particular samples can provide greater sensitivity to detect true alterations in the underlying cancer genome. We extend previous physically grounded error models for gene expression microarrays [15], [16] for use with Affymetrix SNP arrays, facilitating our ability to estimate HSCRs in multiple cancer-derived datasets.

Early work with SNP microarrays demonstrated the feasibility of allelic cancer-genome analysis [17, 18, 4], setting the stage for development of methods to detect allele-specific amplification [19], and LOH in unpaired samples [20]. Development of high-resolution Affymetrix SNP6.0 microarrays was enabled by the development of a method to accurately calibrate allelic probes, allowing for highly accurate genotyping without the need for mismatch probes [13]. These calibrations became

the basis for modern cancer total copy-ratio analysis using these arrays [21], [22], [23].

The development of methods for allelic analysis in cancer samples has been further pursued to identify genomic regions of LOH [24], and as part of solutions to the tumor purity/ploidy problem using Illumina bead-arrays [7], [8], and in Affymetrix SNP6.0 microarrays [25], [14], [9]. Our view is that this problem is best treated separately, allowing for detailed treatment of the HSCR problem and facilitating comparison of various methods specifically addressing this problem [26], [27], [28].

We present a novel computational method, HAPSEG, for accurate estimation of haplotype-specific DNA copy-ratios (HSCRs), offering several specific advantages over existing methods: (i) implementation of an advanced error model tailored to the basic physics of Affymetrix SNP microarray measurements; (ii) Internal recomputation of genomic segmentation using error-model fit; (iii) utilization of LD information from phased haplotype panels [29], improving inference of genotypes by exploiting statistical dependencies between the genotypes of adjacent markers, and improving HSCR estimation.

We present a novel demonstration of direct haplotype phasing using allelic imbalances in cancer-tissue samples. These are of interest to further characterize the influence of germline risk in cancer development. In addition, this capability may be used to generate densley typed reference panels of phased chromosomes for use with imputation of rare alleles in whole exome sequencing

Results

Method overview

Consider a defined set of polymorphic bi-allelic markers (SNPs), for which data proportional to the concentration of each allele (channels a and b) has been generated. For microarray data, the set of SNPs is fixed by the array design. We define the *copy-ratio* as the ratio of allelic concentration in a cancer-derived DNA sample to that of the haploid locus concentration in an equivalent DNA aliquot derived from diploid cells. The copy-ratio of a given allele depends on both the (germline) genotype and on the concentration in the cancer-derived DNA of the homologue on which it resides (the location).

Calibration is the process by which SNP measurements are standardized to copy-ratios. The specific manner in which calibration is performed depends upon the measurement technology being considered (e.g. microarray or sequencing). After calibration, the expected a and b values for a SNP in a diploid sample (at a region of 2 copies) are (0,2), (1,1), or (2,0), corresponding to genotypes BB, AB/BA, and AA, respectively. These values are observed with substantial noise, the distribution of which depends upon the measurement platform being used.

Because DNA copy-number is expected to be locally constant along the genome, superior HSCR estimates can be obtained by pooling datapoints within segments of constant total copy-ratio. For the purpose of initializing such a segmentation, calibrated signal from the a and b channels is added together genome-wide to produce copy-ratio estimates independently of marker genotypes. These values are input to the a segmentation algorithm, (e.g. CBS [30]), which fits a piecewise constant regression function to these values with respect to their location along the reference genome.

Figure 1 shows an example of HAPSEG on a cancer sample with pervasive allelic imbalance. Calibrated allelic copy-ratio estimates for each marker (fig. 1a) were modeled using HAPSEG,

which performed haploid genotyping and segmentation (fig. 1b). The resulting HSCR estimates are summarized at segmental level (fig 1c), revealing discrete levels occurring at regular copy-ratio intervals, consistent with fixed SCNAs in the cancer clone (fig 1d). These estimates formed the basis of a large-scale inference of absolute copy-numbers in cancer [3].

We assume that, within such a segment, the copy-ratios of both haplotypes will nearly always remain constant. Violations from this assumption would correspond to coincident breakpoints, with compensatory changes in the homologous copy numbers such that the total is unchanged. We note that such cases are theoretically possible (homologous recombination), and have been previously reported using related methods [28]; we leave the treatment of such alterations to future development.

A maximum of two distinct homologous copy-ratios can exist in a given segment, and the sum of A and B alleles at (germline) heterozygous sites must equal the number of A and B alleles at homozygous AA and BB sites, respectively. The allelic concentration ratios therefore generally correspond to four clusters the possible phased genotype of each SNP. The locations of these components, denoted $\boldsymbol{\mu}_{\mathrm{AA}_i}, \boldsymbol{\mu}_{\mathrm{AB}_i}, \boldsymbol{\mu}_{\mathrm{BA}_i}, \boldsymbol{\mu}_{\mathrm{BB}_i}$, are specified by two free parameters: the total copy-ratio τ_i, and the difference of homologous copy-ratios, denoted δ_i. The locations of the homozygous components are then:

$$\boldsymbol{\mu}_{\mathrm{AA}_i} = (0,\, \tau_i)\,,\; \boldsymbol{\mu}_{\mathrm{BB}_i} = (\tau_i,\, 0)\,. \tag{1}$$

The locations of the heterozygous components are:

$$\boldsymbol{\mu}_{\mathrm{AB}_i} = \left(\frac{\tau_i - \delta_i}{2},\, \frac{\tau_i + \delta_i}{2}\right),\; \boldsymbol{\mu}_{\mathrm{BA}_i} = \left(\frac{\tau_i + \delta_i}{2},\, \frac{\tau_i - \delta_i}{2}\right). \tag{2}$$

As segments approach allelic-balance ($\delta_i \to 0$), the two heterozygous clusters become superimposed, with the natural corollary that no information is provided regarding the phase of SNPs in that segment (as when genotyping diploid samples). To model observed allelic data, we represent the four possible phased genotypes using a mixture model with components for each genotype {AA, AB, BA, BB}. Figure 2 demonstrates HSCR estimation on 3 example segments (from a single sample,) at differing homologous copy-numbers (fig 2a,b). The relationship between homologous imbalance and phased genotyping is demonstrated (fig 2c-h).

The calibrated data for M measured markers, $\mathbf{X}$, are a $2 \times M$ matrix of bi-allelic (haploid) copy-ratio point estimates. Segmentations of $\mathbf{X}$ are denoted by $\mathcal{S}$, which specify partitions of $\mathbf{X}$ into N successive matrices $\mathbf{X}_{1\ldots N}$. For each segment i, $\mathbf{X}_i$ has dimension $2 \times M_i$, where M_i refers to the number of SNPs in the segment. The model parameter $\mathbf{C}$ represents the 4 phased genotypes of SNPs $\mathbf{X}$. Denote by $\mathbf{X}_{ij}$ the a and b channels for SNP j in segment i, and by $\mathbf{C}_{ij}$ the genotype of SNP j. Prior information about the genotype of each SNP may be available from population allele-frequencies, from the analysis of a paired-normal sample, or from phased haplotype panels, and is represented as $\mathcal{G}$.

The conditional distribution of $\mathbf{X}$, representing the mapping from segment locations $\boldsymbol{\mu}_i$ to probability densities over observed $\mathbf{X}_i$, is denoted $\mathcal{P}\left(\mathbf{X}_i | \boldsymbol{\mu}_i, \boldsymbol{\Theta}\right)$, where $\boldsymbol{\Theta}$ represents the set of sample-level parameters representing specific sources of experimental fluctuation. Estimation of $\boldsymbol{\Theta}$ at the sample-level increases our power to fit realistically complex error-models without the risk of over-fitting small segments. The function $\mathcal{P}$ and parameters $\boldsymbol{\Theta}$ are specific to the measurement platform in use. We developed a novel error-model for Affymetrix SNP microarrays (Methods) and applied it throughout.

We specify the probability density of an observed point, conditional on g as:

$$P\left(\mathbf{X}_{ij}|\mathbf{C}_{ij}=g,\delta_i,\tau_i,\boldsymbol{\Theta}\right)\equiv\mathcal{P}\left(\mathbf{X}_{ij}|\boldsymbol{\mu}_{ig},\boldsymbol{\Theta}\right). \tag{3}$$

The complete likelihood of segment i is therefore:

$$\mathcal{L}_i\left(\mathbf{X}_i|\delta_i,\tau_i,\boldsymbol{\Theta},\mathcal{G}\right)=\prod_{j=1}^{M_i}Z_{ij}, \tag{4}$$

where Z_{ij} denotes the complete likelihood of SNP j:

$$Z_{ij}=\sum_{g=1}^{4}P\left(\mathbf{X}_{ij}|\mathbf{C}_{ij}=g,\delta_i,\tau_i,\boldsymbol{\Theta}\right)P\left(\mathbf{C}_{ij}=g|\mathcal{G}\right). \tag{5}$$

The full sample-likelihood is:

$$\mathcal{L}_f\left(\mathbf{X}|\boldsymbol{\delta},\boldsymbol{\tau},\boldsymbol{\Theta},\mathcal{S},\mathcal{G}\right)=\prod_{i\in\mathcal{S}}\mathcal{L}_i\left(\mathbf{X}_i|\delta_i,\tau_i,\boldsymbol{\Theta},\mathcal{S},\mathcal{G}\right). \tag{6}$$

We implemented model-fitting using general numerical optimizations which are independent of the specific likelihood; support for alternate measurement platforms may be added by implementing an appropriate density for $\mathcal{P}\left(\mathbf{X}_i|\boldsymbol{\mu}_i,\boldsymbol{\Theta}\right)$. We describe an general algorithm for joint estimation of sample-level error-model parameters Θ and segmental allelic copy-ratios δ,τ, denoted $\mathbf{C^*}$, allowing updates of these parameters based on $\{\mathbf{X},\hat{\mathcal{S}},\mathcal{G}\}$: $\{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\tau}}\}=\mathbf{C^*}\left(\hat{\boldsymbol{\Theta}}^{(n)}|\mathbf{X},\hat{\mathcal{S}},\mathcal{G}\right)$.

We use the $\mathbf{C^*}$algorithm to iteratively update the genomic segmentation $\mathcal{S}$, genotype probabilities $\mathcal{G}$, sample error-model parameters $\boldsymbol{\Theta}$, and HSCR locations $\boldsymbol{\delta},\boldsymbol{\tau}$.

Method HAPSEG:

1. **Initialize:** Compute $\hat{\mathcal{S}}^{(1)}$ from $\mathbf{X}$. Set $\boldsymbol{\Theta}^{(0)},\boldsymbol{\delta}^{(0)},\boldsymbol{\tau}^{(0)},\mathcal{G}^{(0)}$.
2. **Fit error-model / segment-locations:** $\{\hat{\boldsymbol{\Theta}}^{(1)},\hat{\boldsymbol{\delta}}^{(1)},\hat{\boldsymbol{\tau}}^{(1)}\}=\mathbf{C^*}\left(\boldsymbol{\Theta}^{(0)}|\mathbf{X},\hat{\mathcal{S}}^{(1)},\mathcal{G}^{(0)}\right)$.
3. **Refine segmentation:** Compute $\hat{\mathcal{S}}^{(2)}$ from $\{\hat{\boldsymbol{\Theta}}^{(1)},\hat{\boldsymbol{\delta}}^{(1)},\hat{\boldsymbol{\tau}}^{(1)}\}$ Estimation of $\hat{\boldsymbol{\Theta}}^{(1)}$ allows for more sensitive and specific evaluation of proposed segmental breakpoints. Consider a pair of adjacent segments in the reference genome. Because the initial segmentation $\hat{\mathcal{S}}^{(1)}$ may have introduced a spurious breakpoint, we develop a probabilistic criterion for their merger using Bayesian model comparison (Methods).
4. **Update:** $\{\hat{\boldsymbol{\Theta}}^{(2)},\hat{\boldsymbol{\delta}}^{(2)},\hat{\boldsymbol{\tau}}^{(2)}\}=\mathbf{C^*}\left(\hat{\boldsymbol{\Theta}}^{(1)}|\mathbf{X},\hat{\mathcal{S}}^{(2)},\mathcal{G}^{(0)}\right)$.
5. **Haploid genotype estimation:** compute $\hat{\mathbf{C}}^{(1)}$ from $P(\mathbf{C}|\mathbf{X},\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\tau}},\mathcal{G}^{(0)})$. For a segment i, posterior distributions $\hat{\mathbf{C}}_i$ are computed for each SNP j as follows:

$$\hat{\mathbf{C}}_{ij}=\frac{1}{Z_{ij}}P\left(\mathbf{X}_{ij}|\mathbf{C}_{ij}=g,\hat{\delta}_i,\hat{\tau}_i,\hat{\boldsymbol{\Theta}}\right)P\left(g|\mathcal{G}^{(0)}\right), \tag{7}$$

where Z_{ij} is computed as in (5). $\hat{\mathbf{C}}^{(1)}$ contains the probability distribution of the haploid genotypes, with the phase information provided by homologous copy-imbalance in the somatic clone.

6. **Reconciliation of phased genotype estimates with reference panels:** compute $\hat{\mathcal{G}}^{(1)}$ from $\hat{\mathbf{C}^{(1)}}$, $\mathbf{D}$, $\mathbf{m}$.

   This procedure examines the evidence for the phased genotypes in a given segment in a panel of phased reference chromosomes characterizing population diversity of haplotypes Consortium:2010en. This is accomplished using the statistical program BEAGLE [31] to phase diploid genotypes estimates computed from $\hat{\mathbf{C}^{(1)}}$. This procedure utilizes phased panels $\mathbf{D}$, and an estimate of the genetic recombination rate $\mathbf{m}$ to compute a maximum-likelihood estimate of phase. HAPSEG then identifies and corrects 'switch-errors' in this phasing by reconciling the estimates with those based on homologous imbalance ($\hat{\mathbf{C}^{(1)}}$). The phasing ability of HAPSEG was validated by examining the local concordance with maximum likelihood phase estimates produced via BEAGLE [31], a statistical method based on a population reference-panel of phased chromosomes (fig 3). We demonstrate that phase estimates produced by BEAGLE are locally concordant with those implied by HAPSEG, but that switch-errors tend to occur at regions of high recombination rate (fig 3a,b).

7. **Final estimates:**

$$\{\hat{\mathbf{\Theta}}, \hat{\delta}, \hat{\tau}\} = \mathbf{C^*}\left(\hat{\mathbf{\Theta}}^{(2)} | \mathbf{X}, \hat{\mathcal{S}}^{(2)}, \hat{\mathcal{G}}^{(1)}\right) \tag{8}$$

$$\hat{\mathbf{C}}^{(2)} = P\left(C_{ijk} | \mathbf{X}_{ij}, \hat{\delta}_i, \hat{\tau}_i, \hat{\mathbf{\Theta}}, \mathcal{G}^{(1)}\right) \tag{9}$$

8. **Standard error of location estimates.** With the Hessian $\mathbf{A}_i$ as in eq.(14), standard errors on the segment locations are:

$$\sigma_{\delta_i} = \sqrt{|\mathbf{A}_{i_{11}}^{-1}|},\ \sigma_{\tau_i} = \sqrt{|\mathbf{A}_{i_{22}}^{-1}|}.$$

   This approximation is valid given that the posterior distribution of δ and τ is multivariate normal.

Affymetrix error model. Following a classic microarray error-model [15] we consider the observed calibrated signal X to be distributed according to:

$$X = \mu e^{\eta} + \epsilon, \tag{10}$$

$$\eta \sim \mathcal{N}(0, \sigma_\eta),\ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon),$$

where μ represents the true copy-ratio, η and ϵ represent multiplicative and additive errors, respectively. Importantly, this allows for specifying cluster variance as a function of the mean, without the need to fit additional segment-level variance parameters, as in [28].

We generalized this error model to the two-dimensional case appropriate for fitting two-channel microarray data (Methods, eq. (17)). Because an explicit formula for likelihood calculations could not be obtained, we used an approximation whereby a transformation is applied to the data, after which its distribution becomes approximately bi-variate Gaussian. Specifically, we extended the variance stabilization technique described for the one-dimensional case [32] to the 2D error model, accounting for additional observed correlations in heterozygous SNP clusters, which were presumed to arise from the PCR amplification. Additional parameters are included in $\mathbf{\Theta}$ taking into account attenuation and background fluctuation (Methods). The quality of the model-fit to the marker-

level data is shown in figure 4, demonstrating that much of the variation in the data is accurately captured.

Discussion

We have demonstrated the ability of HAPSEG to accurately estimate haplotype-specific DNA concentrations in tumor samples, including those for which no patient-matched normal sample exists. We demonstrated the utility of HSCR estimates from HAPSEG to estimate tumor purity, ploidy, and absolute copy-numbers using ABSOLUTE [3]. By factorizing general inference methods from platform-specific error models, greater generality was achieved, allowing us to easily adapt HAPSEG to the analysis of massively parallel sequencing data. Although promising results have been obtained from whole-exome hybrid capture sequencing (WES)[33] data (not shown), further development is needed to derive calibrated copy-estimates from multiplexed WES using bar-coded reads.

In addition to providing accurate estimates of haplotype-specific copy-ratios, HAPSEG can produce partially phased haplotypes by exploiting the widespread allelic-imbalance common in aneuploid cancer genomes, without the need for methods based on phased reference panels. This will have several applications in cancer genetics, such as establishing compound heterozygosity of germline alleles. In addition, when used with sequencing data, long-range haplotyping using HAPSEG will aid reconstruction of complete somatic karyotype and history of aberrations transforming from the diploid state[34]. Furthermore, these methods may aid detection of specific SCNAs subject to allelic bias in cancer [35, 36, 8].

The availability of partially phased genotypes will allow improvements in panel-based imputation methods, which attempt to model observed diploid genotypes as a mosaic of phased reference haplotypes [37], [38], [39], [40], [31], (reviewed[41]). Optimal exploitation of will require extension of these methods to handle partially phased genotype data containing variable gaps (due to somatic allele-balance.)

Similar considerations apply to admixture mapping, whereby long-range phasing may be inferred by differences in parental allele-frequency along the genome, due to recent mixing of outbred populations [42], [43]. The phased homologues provided by HAPSEG, each of which is derived from a single parent, will allow for more sensitive detection of recent admixture, possibly supporting mapping with less divergent populations than permitted in the purely diploid case. Since admixture inferences allow for inter-chromosomal phasing of homologues, they may be of interest in the analysis of germline risk loci.

Long-range haplotyping has utility to uncover interactions between germline and somatic genetics in cancer. Examples include variants at the 8q24 locus, affecting inherited risk of development of multiple cancer types [42], and later shown to interact with MYC [44]. Furthermore, these methods may facilitate identification of additional germline alleles mechanistically predisposed to somatic alteration in cancer, such as JAK2 in myeloproliferative neoplasms [45]. In contrast, to the above examples, the genetic basis by which EGFR mutations in lung adenocarcinoma occur predominately in female non-smoking patients of East Asian ethnicity [46] remains unclear.

As whole-genome sequencing of aneuploid cancer samples becomes increasingly routine, directed long-range haplotyping using HAPSEG will increase the available panels of densely genotyped chromosomes, without the need for genotyping parent-child trios. This resembles the precedent of geno-

typing uniparental-disomic samples derived from complete hydatidiform moles [47], which formed an important component of the phased Japanese HAPMAP panel [48]. HAPSEG generalizes this requirement to allow direct phasing from homologous imbalance, extending the samples on which it may be applied to tens of thousands of cancer samples [49].

Long-range haplotyping is an important technique for human population genetics, with wide ranging applications including understanding haplotype diversity and genetic recombination rates [29], illuminating recent positive selection in the human lineage [50] and human history [51].

Partially phased genotyping by analysis of cancer-derived DNA using HAPSEG represents a novel technique to obtain such data. This may be particularly useful in the study of poorly characterized populations, for which phased reference panels do not exist, and for which parent-child trio data may be impossible to obtain. In such cases, this strategy may present a cost-effective alternative to recently described molecular methods for long-range haplotyping [52], [53].

Methods

Algorithm C*

C* uses the data, segmentation $\mathcal{S}$, and genotype priors $\mathcal{G}$ to update estimates of the sample error-model $\boldsymbol{\Theta}$ and segment locations $\{\boldsymbol{\delta}, \boldsymbol{\tau}\}$. This is accomplished by iterating two conditional updates until convergence of $\mathcal{L}_f$:

1. **Conditional update of locations $\{\boldsymbol{\delta}, \boldsymbol{\tau}\}$:**

$$\forall_i,\ \left\{\delta_i^{(t+1)}, \tau_i^{(t+1)}\right\} = \underset{\{\delta_i, \tau_i\}}{\operatorname{argmax}} \log \mathcal{L}_i \left(\mathbf{X}_i | \delta_i, \tau_i, \boldsymbol{\Theta}^{(t)}, \hat{\mathcal{S}}, \mathcal{G}\right) \tag{11}$$

2. **Conditional update of sample error-model $\boldsymbol{\Theta}$.** $\boldsymbol{\Theta}$ is a vector of platform-specific parameters upon which conditional updates are performed serially. We found that the results are insensitive to ordering, and suppress indexing:

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \log \mathcal{L}_f \left(\mathbf{X} | \boldsymbol{\delta}^{(t)}, \boldsymbol{\tau}^{(t)}, \boldsymbol{\Theta}, \hat{\mathcal{S}}, \mathcal{G}\right) P(\boldsymbol{\Theta}). \tag{12}$$

   Note the superscripts (t) above denote the internal iterations of the **C***algorithm.

3. **Stop if** the full likelihood $\mathcal{L}_f$ does not change significantly.

Probabilistic segment merging

Let $\mathcal{H}_0$ and $\mathcal{H}_1$ denote the seperate and merged segment hypotheses. In order to compare the evidence supporting each model, we compute an approximation to the Bayesian evidence (Laplace) by approximating the normalizing constant of the posterior distribution $P(\delta_i, \tau_i | \mathbf{X}_i, \hat{\boldsymbol{\Theta}})$. If the posterior is multivariate Gaussian, the normalization constant may be calculated from the likelihood

and curvature at the mode:

$$\mathrm{Ev}(\mathbf{X}_i) = \int_{\delta_i}\int_{\tau_i} \mathcal{L}_i(\mathbf{X}_i|\delta_i,\tau_i,\hat{\mathbf{\Theta}},\mathcal{G}^{(0)}) \tag{13}$$

$$\simeq \hat{\mathcal{L}}_i \times \left|\frac{\det \mathbf{A}_i}{2\pi}\right|^{-1/2} \times (5/2)^{-2}, \tag{14}$$

where $\hat{\mathcal{L}}_i$ denotes the maximum likelihood (at $\hat{\delta}_i, \hat{\tau}_i$), and $\mathbf{A}_i = -\nabla\nabla \log \mathcal{L}_i(\mathbf{X}_i|\delta_i,\tau_i,\hat{\mathbf{\Theta}})$ denote the Hessian matrix around the mode of a segment location: $\{\hat{\delta}_i, \hat{\tau}_i\}$. We verified the accuracy of the approximation using quadrature on a subset of segments (not shown).

The constant third term gives the volume of the domain: $\delta_i \times \tau_i$. Note that the first and second-two terms correspond to the 'Best fit likelihood' × 'Occam factor' formulation of Evidence, as described in [54].

Following Bayes' rule: $P(\mathcal{H}|\mathbf{X}_i) \propto P(\mathbf{X}_i|\mathcal{H})P(\mathcal{H})$, we compute the Bayes factor $\mathrm{BF}(\mathcal{H}_0)$:

$$\frac{P(\mathcal{H}_0|\mathbf{X}_{1,2})}{P(\mathcal{H}_1|\mathbf{X}_{12})} = \frac{\mathrm{Ev}(\mathbf{X}_1)\mathrm{Ev}(\mathbf{X}_2)}{\mathrm{Ev}(\mathbf{X}_{12})}. \tag{15}$$

Our experience suggests that the cancer genome is often heavily over-segmented using the initial segmentation $\hat{\mathcal{S}}^{(1)}$ described above. Adjacent segments are therefore merged if $\mathrm{BF}(\mathcal{H}_0) < 1 \times 10^{-10}$. $\mathrm{BF}(\mathcal{H}_0)$ is computed for all breakpoints in $\hat{\mathcal{S}}^{(1)}$ and segments are merged greedily by joining the adjactent pair with the lowest $\mathrm{BF}(\mathcal{H}_0)$ value. The merge-probability for any breakpoints adjacent to the resulting combined segment are computed at each step. The procedure is finished when no pairs exist with $\mathrm{BF}(\mathcal{H}_0)$ below the threshold, resulting in a refined segmentation $\hat{\mathcal{S}}^{(2)}$. We note that the Bayes factor defined above could equivalently be applied to accept or reject novel proposed breakpoints in $\mathcal{S}$. We leave the development of an efficient method to identify valid breakpoints to the future.

Reconciliation of phased genotype estimates with reference panels

1. Compute diploid genotype estimates $\hat{d}$ from haploid estimates $\hat{\mathbf{C}}^{(1)}$ by collapsing (marginalizing) the two heterozygous clusters.

2. Compute the maximum likelihood statsitical phasing of d using BEAGLE [31]:

   $$\hat{\mathbf{C}}^{(\mathrm{B})} = \mathrm{BEAGLE}(\hat{d}, \mathbf{D}, \mathbf{m})$$

   If evidence exists for the haploytype configurations $\hat{\mathbf{C}}^{(1)}$ as a mosaic of phased chromosomes in $\mathbf{D}$ with recombination rate $\mathbf{m}$, then these configurations will be present in $\hat{\mathbf{C}}^{(\mathrm{B})}$, up to occasional spurious phase reversals (switch-errors), which are expected to occur at frequency related to $\mathbf{m}$.

3. Reconcile haploid estimates from allelic-imbalance $\hat{\mathbf{C}}^{(1)}$ with $\hat{\mathbf{C}}^{(\mathrm{B})}$ by allowing for occasional switch-errors:

   $$\hat{\mathcal{G}}^{(1)} = \mathrm{Viterbi}(\hat{\mathbf{C}}^{(1)}, \hat{\mathbf{C}}^{(\mathrm{B})}, \mathbf{\Phi}_{\mathbf{m}})$$

   This is accomplished using the Viterbi algorithm, which finds the best path through adjacent-marker phased-genotype probabilities $\hat{\mathbf{C}}^{(1)}$ such that they are concordant with $\hat{\mathbf{C}}^{(\mathrm{B})}$, with

occasional switches to the phase-reversed $\hat{\mathbf{C}}^{(-\mathrm{B})}$ (obtained by swapping of the AB and BA genotype probabilities). Switching dynamics are governed by the transition-probability matrix $\mathbf{\Phi_m}$:

$$\mathbf{\Phi_m} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

Thus, switches between $\hat{\mathbf{C}}^{(\mathrm{B})}$ and $\hat{\mathbf{C}}^{(-\mathrm{B})}$ are expected to occur with marginal probability 0.1. We note the adjusting the switch-error probability according to the recombination rate $\mathbf{m}$ would likely be more sensitive, however these values appear to work well on our data (fig. 3). Upon termination, $\hat{\mathcal{G}}^{(1)}$ contains updated knowledge about haploid genotypes $\mathbf{C}$. If no phase information is implied in $\hat{\mathbf{C}}^{(1)}$ (due to allelic balance), $\hat{\mathcal{G}}^{(1)} = \hat{\mathbf{C}}^{(\mathrm{B})}$, since the phase of either configuration in $\hat{\mathbf{C}}^{(1)}$ will be equally probable. Likewise, if $\hat{\mathbf{C}}^{(1)}$ reflects a spurious phasing from over-fitting the locations of a given segment, then the switch probability 0.1 is not adequate to form a path with expected switches at every other marker. In such cases, the locations will be corrected by recomputing the segment HSCRs using $\hat{\mathcal{G}}^{(1)} = \hat{\mathbf{C}}^{(\mathrm{B})}$.

Calibration

(Note: The remaining sections describe methods specific to SNP microarrays)

Calibration of Affymetrix SNP-array measurements was accomplished using the Birdseed algorithm [13], which combines population allele-frequency estimates with location estimates on training samples with known genotypes. These prior estimates are adjusted by Birdseed to fit data from batches of samples. These adjusted location estimates are then used to define the three diploid genotype hybridization-intensity clusters, with locations denoted I_{AA}, I_{AB}, I_{BB}. A large amount of measurement error is thought to arise from variability in the conditions under which the PCR reaction is run. We therefore define batches of PCR (eg., 96-well plates for which the reactions were run in parallel.) Calibrations are performed independently for each of such batches. Genotype locations are estimated using only the normal (non-cancer) samples in the batch, which need not be paired to the cancer samples. Because Birdseed jointly determines the genotypes of the normal samples along with the locations of the genotype clusters for the batch, it is not necessary to use controls with known genotypes. Because normal samples are essential for calibration, we generally recommend that at least 10 such samples be included in each batch.

With the intensity of the 3 genotype-clusters determined by Birdseed, background and scale parameters for the a and b channels of each probe-set can be obtained. The background for each channel is estimated as the observed intensity corresponding to zero copies of that allele:

$$I_{0_a} = I_{BB_a},\ I_{0_b} = I_{AA_b}.$$

The scale for each channel is estimated as the difference in observed intensities corresponding to one and zero copies of that allele:

$$d_{1_a} = I_{AB_a} - I_{BB_a},\ d_{1_b} = I_{AB_b} - I_{AA_b}.$$

Calibrations of observed intensities in cancer samples are then computed as:

$$X_a = \frac{I_a - I_{0_a}}{d_{1_a}},\ X_b = \frac{I_b - I_{0_b}}{d_{1_b}}. \tag{16}$$

The calibrated signals may therefore be interpreted as ratios of the locus-concentration in the tumor-sample to the concentration corresponding to 1 copy in a sample derived from diploid cells.

The above calibration procedure is strictly valid given linear responses throughout the domain of locus-concentration. In fact, attenuation effects are observed for many probe-responses at locus-concentrations corresponding to 2-copies / diploid cell, and become more pronounced with increasing concentration. By performing calibration using only locus concentrations corresponding to 0 and 1 copy, we have sought to base our estimates on values from the linear response range in karyotypically normal samples. Attenuation effects occurring in tumor samples are dealt with in a subsequent section.

Following the procedure described above, we perform an additional probe-set level calibration step designed to remove spurious correlation between the a and b channels arising due to cross-hybridization. This procedure is described in a subsequent section.

Error model

Modern SNP microarrays measure on the order of 100-1000K polymorphic sites present in an input DNA aliquot. Specifically, these arrays allow for the measurement of hybridization occurring between input DNA and substrate-bound oligonucleotides corresponding to each of two forms for each polymorphic locus interrogated (an a and a b channel for each SNP). In ideal cases, the experimental conditions under which the array is run result in hybridization proportional to the concentration of the specific locus being interrogated, plus some unknown background (presumed to arise from non-specific hybridization.)

The motivation for the model in eq. (10) stems from the consideration of two major noise sources inherent in the measurement procedure. A molecule with concentration proportional to μ is subjected to n rounds of PCR amplification, to yield 2^n molecules. Assume $n \sim \mathcal{N}(m, \sigma_m)$, then $\log(\mu) \sim \mathcal{N}$. Hybridization of species to the microarray then results in measured signal $X \sim \mathcal{N}(\mu e^{\eta}, \sigma_\epsilon)$. The distribution of X is therefore the convolution of a normal and a log-normal distribution, representing the measurement of a process subject to multiplicative noise using a process (hybridization) subject to additive noise.

Although the error-model of eq. 10 fits the observed data reasonably well, consistent positive correlation is observed between the a and b channels of the heterozygous clusters AB and BA. Furthermore, the magnitude of this correlation is directly proportional to $\hat{\sigma}_\eta$ (not shown). Noting that the two alleles of any given SNP lie on restriction fragments that typically differ by a single base, we surmise that any variations in PCR efficiency arising due to sequence composition are likely to be shared. This phenomenon affects only the heterozygous clusters because, by definition, one allele is absent for any homozygous SNP, and is not subject to PCR amplification. We therefore generalize the error-model of equation 10 to the 2-dimensional case as follows:

$$\begin{bmatrix} X_a \\ X_b \end{bmatrix} = \alpha + \begin{bmatrix} \mu_a e^{\eta_a} \\ \mu_b e^{\eta_b} \end{bmatrix} + \begin{bmatrix} \epsilon_a \\ \epsilon_b \end{bmatrix}$$

$$\begin{bmatrix} \eta_a \\ \eta_b \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Sigma}_\eta); \mathbf{\Sigma}_\eta = \begin{bmatrix} \sigma_\eta & \rho\sigma_\eta \\ \rho\sigma_\eta & \sigma_\eta \end{bmatrix} \tag{17}$$

$$\begin{bmatrix} \epsilon_a \\ \epsilon_b \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Sigma}_\epsilon); \mathbf{\Sigma}_\epsilon = \begin{bmatrix} \sigma_\epsilon & 0 \\ 0 & \sigma_\epsilon \end{bmatrix}$$

Note that the assumption of diagonal additive covariance (Σ_ϵ) in the above equation is reasonable only because of the calibration procedure which removes cross-hybridization signal prior to fitting the model. This procedure is described in a subsequent section.

Attenuation

The heterozygous segment locations are overdetermined with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$ (2). For each segment, 8 genotype-cluster coordinates are determined by two free parameters. Observed discrepancies from these dependencies can be exploited for improved estimates of microarray attenuation.

As in [55], we adopt the Langmuir isothermal adsorption model [56] to deal with attenuation effects observed as the concentration of a given hybridized species increases:

$$\theta = \frac{\phi C}{1 + \phi C},$$

where θ indicates the proportion of bound species (presumed to be equivalent to the observed hybridization signal), C indicates the concentration of a particular species, and ϕ is a constant related to the binding affinity of the species to the oligonucleotide probes targeting it. This model has been previously shown to accurately model the attenuation characteristics of microarray probes [16].

Due to the calibration procedure described in eq (16), the observed copy-ratio X is actually the ratio of two isotherms:

$$X = \frac{\phi C_T}{1 + \phi C_T} / \frac{\phi C_N}{1 + \phi C_N},$$

where C_T and C_N are proportional to concentration in the tumor and normal sample, respectively. Let $C_R = C_T / C_N$. Because C_N is defined to be the hybridization intensity corresponding to 1 copy for a given probe-set, we can rewrite the above equation as:

$$X = \frac{C_R(1 + \phi)}{(1 + C_R\phi)}.$$

We can then define a transformation of genotype-cluster coordinates:

$$g(\mu) = \frac{\mu(1 + \phi)}{(1 + \mu\phi)} \tag{18}$$

This transformation is applied to the locations of the 4 genotype-clusters defined in equations 1, 2. This ensures that the physical constraints defined by the model are compatible with the observed data.

Noting that

$$\lim_{C_R \to \infty} \frac{C_R(1 + \phi)}{(1 + C_R\phi)} = \frac{1 + \phi}{\phi}$$

corresponds to the asymptotic saturation copy-ratio, m, gives a direct physical interpretation for the value of ϕ; $\phi = \frac{1}{m-1}$.

Variance-stabilizing transformation

Following [32], we make use of a variance-stabilizing transformation h, such that

$$h(x) \simeq h(\mu) + \epsilon_h;\ \epsilon_h \sim \mathcal{N}(0, \sigma_h), \tag{19}$$

where σ_h is constant over the domain of X. Writing the variance of X as a function of the mean μ : $\mathrm{var}(X) = v(\mu)$, a general form for h can be expressed as [57]:

$$h(x) = \int^x \frac{1}{\sqrt{v(u)}} du.$$

The specific form of $h(x)$ for use with the error-model defined by eq. (10) is derived in [32]:

$$\begin{aligned} h(x) &= \log\left(bx + \sqrt{1 + bx^2}\right) \\ &= \sinh^{-1}(bx). \end{aligned} \tag{20}$$

The transformation has a parameter b that is calculated from σ_η and σ_ϵ as follows:

$$b = \frac{\sqrt{e^{\sigma_\eta^2} - 1}}{\sigma_\epsilon}. \tag{21}$$

The transformed data, denoted $\mathbf{X}'$, will have variance, denoted σ_h^2, approximately independent of $\mathbf{X}$, with

$$\begin{aligned} \sigma_h &= \sigma_\epsilon \frac{d}{dx} h(x)|_{x=0} \\ &= \sigma_\epsilon b. \end{aligned} \tag{22}$$

Using simulated data, we verified that the approximation in eq. 19 is accurate for values of σ_η and σ_ϵ within the range of typical estimates (data not shown).

Likelihood calculation

We construct an approximation to the model defined in eq. (17) based on application of the variance-stabilizing transformation (eq. 20) applied to both data channels. We use the scaled bivariate t-distribution on the transformed data $\mathbf{X}'$, which has an additional parameter $\nu \in \boldsymbol{\Theta}$ allowing excess density in the tails.

$$\begin{aligned} \mathcal{P}\left(\mathbf{X}_{ij} | \boldsymbol{\mu}_{ik}, \boldsymbol{\Theta}\right) &\equiv P(\mathbf{X}_{ij} | \boldsymbol{\mu}_{ik}, \nu, \boldsymbol{\Sigma}_i) \\ &= \frac{1}{2\pi |\boldsymbol{\Sigma}_i|^{1/2}} \left(1 + \nu \left(\mathbf{X}'_{ij} - \boldsymbol{\mu}'_{ik}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{X}'_{ij} - \boldsymbol{\mu}'_{ik})\right)^{-(\nu/2+1)} |\mathbf{J}_{ij}|, \end{aligned} \tag{23}$$

where

$$\mathbf{X}'_{ij} = h(\mathbf{X}_{ij}),$$

and

$$\boldsymbol{\mu}'_{ik} = h(g(\boldsymbol{\mu}_{ik})),$$

with g and h given by eqns. 18 and 20, respectively. Define

$$\mathbf{X}'_{ij} = \begin{bmatrix} X'_a \\ X'_b \end{bmatrix},$$

The Jacobian of the transformation h is:

$$\begin{aligned} |\mathbf{J}_{ij}| &= \begin{vmatrix} \frac{\partial h(X'_a)}{\partial X'_a} & \frac{\partial h(X'_a)}{\partial X'_b} \\ \frac{\partial h(X'_b)}{\partial X'_a} & \frac{\partial h(X'_b)}{\partial X'_b} \end{vmatrix} \\ &= \frac{\partial h(X'_a)}{\partial X'_a} \frac{\partial h(X'_b)}{\partial X'_b} \\ &= \frac{b}{\sqrt{1+(bX'_a)^2}} \frac{b}{\sqrt{1+(bX'_b)^2}}, \end{aligned}$$

where b and σ_h may be calculated from σ_η and σ_ϵ via equations (21) and (22), respectively.

Introducing a sample-level parameter $0 < \rho_h < \frac{3}{4}\sigma_h^2 \in \boldsymbol{\Theta}$, we approximate the covariance-matrix of the transformed data in a given segment i as:

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_h^2 & s\rho_h \\ s\rho_h & \sigma_h^2 \end{bmatrix}, \text{ with } s = \begin{cases} \mu_{ik_a}\mu_{ik_b} & \text{if } \mu_{ik_a}\mu_{ik_b} < 1 \\ 1 & \text{otherwise.} \end{cases}$$

Thus, the covariance is spherical for the homozygous clusters. The scaling of ρ_h by s provides a consistent approximation to the covariance implied by eq. (17) for heterozygous clusters.

Substitution of eq. (23) into the generic form for $\mathcal{P}(\mathbf{X}_{ij}|\boldsymbol{\mu}_{ik}, \boldsymbol{\Theta})$ in eq. (3) completes the specification of our HSCR estimation based on SNP microarray data. The sample-level error model parameters which must be estimated from the data are: $\boldsymbol{\Theta} = \{\sigma_\epsilon, \sigma_\eta, \rho_h, \nu, \phi\}$.

Calibration of cross-hybridization effects

We develop a calibration step to remove allelic-crosstalk from copy-ratio data $\mathbf{X}$ in a SNP-specific manner as follows. Starting with large collection of normal samples, genotyped using Birdseed [13], we compute the cluster-centers for each genotype for each SNP. We attempt to estimate the covariance matrix for each SNP, denoted $\boldsymbol{\Sigma}_i$. We make use of the factorization $\boldsymbol{\Sigma} = \mathbf{SRS}$, where

$$\mathbf{R} = \begin{pmatrix} 1 & r_{AB} \\ r_{AB} & 1 \end{pmatrix}, \mathbf{S} = \begin{pmatrix} \sigma_2 & 0 \\ 0 & \sigma_0 \end{pmatrix},$$

with r_{AB} denoting the correlation between the A and B allele. σ_0 and σ_2 denote the standard-deviation of genotype-cluster axes corresponding to 0 and 2 copies, respectively.

Because we are attempting to estimate a large number of parameters from data which may be limited, we make use of a Bayesian regularization technique. Population priors describing the distribution of scales and correlations are estimated from all SNPs, and these are then used to compute the most probable (MP) estimates for each SNP. This follows standard use of a hierarchical

model, as in [58].

We use scaled-inverse-χ^2 prior distributions for σ_0 and σ_2, defined as follows:

$$\text{Let } X \sim \chi^2(\nu) \text{ and } Y = \frac{\sigma^2 \nu}{X}, \text{ then } Y \sim \text{Scale-inv-}\chi^2(\nu, \sigma^2).$$

We use a normal distribution for the Fisher-transform of r_{AB}, $z_{AB} = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$. We denote the standard error of z as $z_\sigma = \frac{1}{\sqrt{N-3}}$. We define population-level distributions on the SNP covariance parameters as:

$$z_{AB} \sim \mathcal{N}(\pi_{\mu_z}, \pi_{\sigma_z})$$

$$\sigma_0^2 \sim \text{Scaled-Inverse-}\chi^2(\pi_{s_0}, \pi_{\nu_0})$$

$$\sigma_2^2 \sim \text{Scaled-Inverse-}\chi^2(\pi_{s_2}, \pi_{\nu_2})$$

π_{μ_z} and π_{σ_z} are estimated via maximum likelihood from the calculated SNP correlation-coefficients.

The MP estimate of z is the mode of the posterior (Gaussian) distribution:

$$\hat{z}_{\text{MP}} = \frac{\frac{\pi_{\mu_z}}{\pi_{z_\sigma}^2} + \frac{\hat{z}_{\text{ML}}}{z_\sigma^2}}{\frac{1}{\pi_{z_\sigma}^2} + \frac{1}{z_\sigma^2}}$$

The MP estimate of r is the inverse of the Fisher transform on $\hat{z}_{\text{MP}}$:

$$\hat{r}_{\text{MP}} = \frac{e^{2\hat{z}_{\text{MP}}} - 1}{e^{2\hat{z}_{\text{MP}}} + 1}$$

The posterior distribution for the scales σ_0 and σ_2 is:

$$\sigma_0^2 | X \sim \text{Scaled-Inverse-}\chi^2\left(\theta_\sigma, \theta_\nu\right),$$

with

$$\theta_\sigma^2 = \frac{\pi_{\nu_0}\pi_{s_0}^2 + N\hat{\sigma}_{0\text{ML}}^2}{\pi_{\nu_0} + N}, \; \theta_\nu = \pi_{\nu_0} + N$$

The posterior mode is therefore

$$\hat{\sigma}_{0\text{MP}} = \sqrt{\frac{\theta_\nu \theta_\sigma^2}{\theta_\nu + 2}}.$$

$\hat{\sigma}_{2\text{MP}}$ is computed in a similar fashion.

We then construct the MP covariance matrix for each SNP:

$$\hat{\mathbf{\Sigma}}_{\text{MP}} = \begin{pmatrix} \hat{\sigma}_{2\text{MP}} & 0 \\ 0 & \hat{\sigma}_{0\text{MP}} \end{pmatrix} \begin{pmatrix} 1 & \hat{r}_{\text{MP}} \\ \hat{r}_{\text{MP}} & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_{2\text{MP}} & 0 \\ 0 & \hat{\sigma}_{0\text{MP}} \end{pmatrix}.$$

Denote the slope of the 1st eigen-vector of $\hat{\mathbf{\Sigma}}_{\text{MP}}$ as $\hat{M}_{BB}$. We construct an affine transformation as follows:

$$\mathbf{T} = \begin{pmatrix} 1 & -\hat{M}_{BB} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\hat{M}_{BB} & 1 \end{pmatrix}.$$

The corrected data is then:

$$\mathbf{X'} = \mathbf{TX}.$$

Acknowledgments

We are indebted to Joshua Korn for several helpful conversations.

References

1. Albertson D, Collins C, McCormick F, Gray J (2003) Chromosome aberrations in solid tumors. Nature Genetics 34: 369–376.

2. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.

3. Carter S (2011) Ph.D Thesis, Chapter 4: Absolute quantification of somatic DNA alterations in cancer reveals frequent genome doublings in human cancers .

4. Zhao X, Li C, Paez J, Chin K, Jänne P, et al. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Research 64: 3060.

5. Weir B, Zhao X, Meyerson M (2004) Somatic alterations in the human cancer genome. Cancer Cell 6: 433–438.

6. Knudson A (1971) Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences 68: 820.

7. Popova T, Manié E, Stoppa-Lyonnet D, Rigaill G, Barillot E, et al. (2009) Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biology 10: R128.

8. Van Loo P, Nordgard S, Lingjærde O, Russnes H, Rye I, et al. (2010) Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences 107: 16910.

9. Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, et al. (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biology 11: R92.

10. Storchova Z, Pellman D (2004) From polyploidy to aneuploidy, genome instability and cancer. Nature Reviews Molecular Cell Biology 5: 45–54.

11. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics .

12. Cibulskis K (2011) muTect: Reliable and Accurate Somatic Mutation Detection in Next Generation Cancer Genome Sequencing, manuscript in preparation .

13. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics 40: 1253–1260.

14. Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, et al. (2010) PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics 11: 164–175.

15. Rocke DM, Durbin B (2001) A model for measurement error for gene expression arrays. Journal of computational biology : a journal of computational molecular cell biology 8: 557–569.

16. Hekstra D (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. Nucleic acids research 31: 1962–1968.

17. Mei R (2000) Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays. Genome research 10: 1126–1137.

18. Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, et al. (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nature Biotechnology 18: 1001–1005.

19. LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, et al. (2005) Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis. PLoS Computational Biology 1: e65.

20. Beroukhim R, Lin M, Park Y, Hao K, Zhao X, et al. (2006) Inferring Loss-of-Heterozygosity from Unpaired Tumors Using High-Density Oligonucleotide SNP Arrays. PLoS Computational Biology 2: e41.

21. McLendon R, Friedman A, Bigner D, van Meir E (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature .

22. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, et al. (2010) Signatures of mutation and selection in the cancer genome. Nature 463: 893–898.

23. Network TCGAR, Network TCGAR, sites Dwg, tissue source, Medicine GscBCo, et al. (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.

24. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Goransson H, et al. (2008) Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biology 9: R136.

25. Attiyeh EF, Diskin SJ, Attiyeh MA, Mosse YP, Hou C, et al. (2008) Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome research 19: 276–283.

26. Bengtsson H, Neuvial P, Speed TP (2010) TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. BMC bioinformatics 11: 245.

27. Olshen AB, Bengtsson H, Neuvial P, Spellman P, Olshen RA, et al. (2011) Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. Bioinformatics .

28. Chen H, Xing H, Zhang NR (2011) Estimation of Parent Specific DNA Copy Number in Tumors using High-Density Genotyping Arrays. PLoS Computational Biology 7: e1001060.

29. Consortium TIH, investigators P, leaders Pc, group Mw, QC G, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.

30. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657–663.

31. Browning BL, Yu Z (2009) Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. The American Journal of Human Genetics 85: 847–861.

32. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1: S96–104.

33. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27: 182–189.

34. Ozery-Flato M, Shamir R (2009) Sorting Cancer Karyotypes by Elementary Operations. Journal of Computational Biology 16: 1445–1460.

35. LaFramboise T, Dewal N, Wilkins K, Pe'er I, Freedman ML (2010) Allelic Selection of Amplicons in Glioblastoma Revealed by Combining Somatic and Germline Analysis. PLoS Genetics 6: e1001086.

36. Dewal N, Freedman ML, LaFramboise T, Pe'er I (2010) Power to detect selective allelic amplification in genome-wide scans of tumor data. Bioinformatics 26: 518–528.

37. Li Y (2006) Li: Mach 1.0: rapid haplotype reconstruction and... - Google Scholar. Am J Hum Genet.

38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81: 559–575.

39. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39: 906–913.

40. Howie B, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.

41. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nature Publishing Group 11: 499–511.

42. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proceedings of the National Academy of Sciences of the United States of America 103: 14068–14073.

43. Price A, Tandon A, Patterson N, Barnes K, Rafaels N, et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics 5.

44. Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, et al. (2010) 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proceedings of the National Academy of Sciences 107: 9742–9746.

45. Kilpivaara O, Mukherjee S, Schram AM, Wadleigh M, Mullally A, et al. (2009) A germline JAK2 SNP is associated with predisposition to the development of JAK2V617F-positive myeloproliferative neoplasms. Nature Genetics 41: 455–459.

46. Yatabe Y, Mitsudomi T (2007) Epidermal growth factor receptor mutations in lung cancers. Pathology International 57: 233–244.

47. Fan JB, Surti U, Taillon-Miller P, Hsie L, Kennedy GC, et al. (2002) Paternal Origins of Complete Hydatidiform Moles Proven by Whole Genome Single-Nucleotide Polymorphism Haplotyping. Genomics 79: 58–62.

48. Kukita Y (2005) Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles. Genome research 15: 1511–1518.

49. Hudson Chairperson TJ, Anderson W, Aretz A, Barker AD, Bell C, et al. (2010) International network of cancer genome projects. Nature 464: 993–998.

50. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science 327: 883–886.

51. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature : 1–5.

52. Fan HC, Wang J, Potanina A, Quake SR (2010) Whole-genome molecular haplotyping of single cells. Nature Biotechnology 29: 51–57.

53. Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan RP, et al. (2010) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nature Biotechnology 29: 59–63.

54. MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge Univ Pr.

55. Chiang D, Getz G, Jaffe D, O'Kelly M, Zhao X, et al. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature methods 6: 99.

56. Langmuir I (1916) . Journal of the American Chemical Society 38: 2221–2295.

57. Tibshirani R (1988) Estimating transformations for regression via additivity and variance stabilization. Journal of the american statistical association .

58. Gelman A (2004) Bayesian data analysis. CRC Press.

Figures

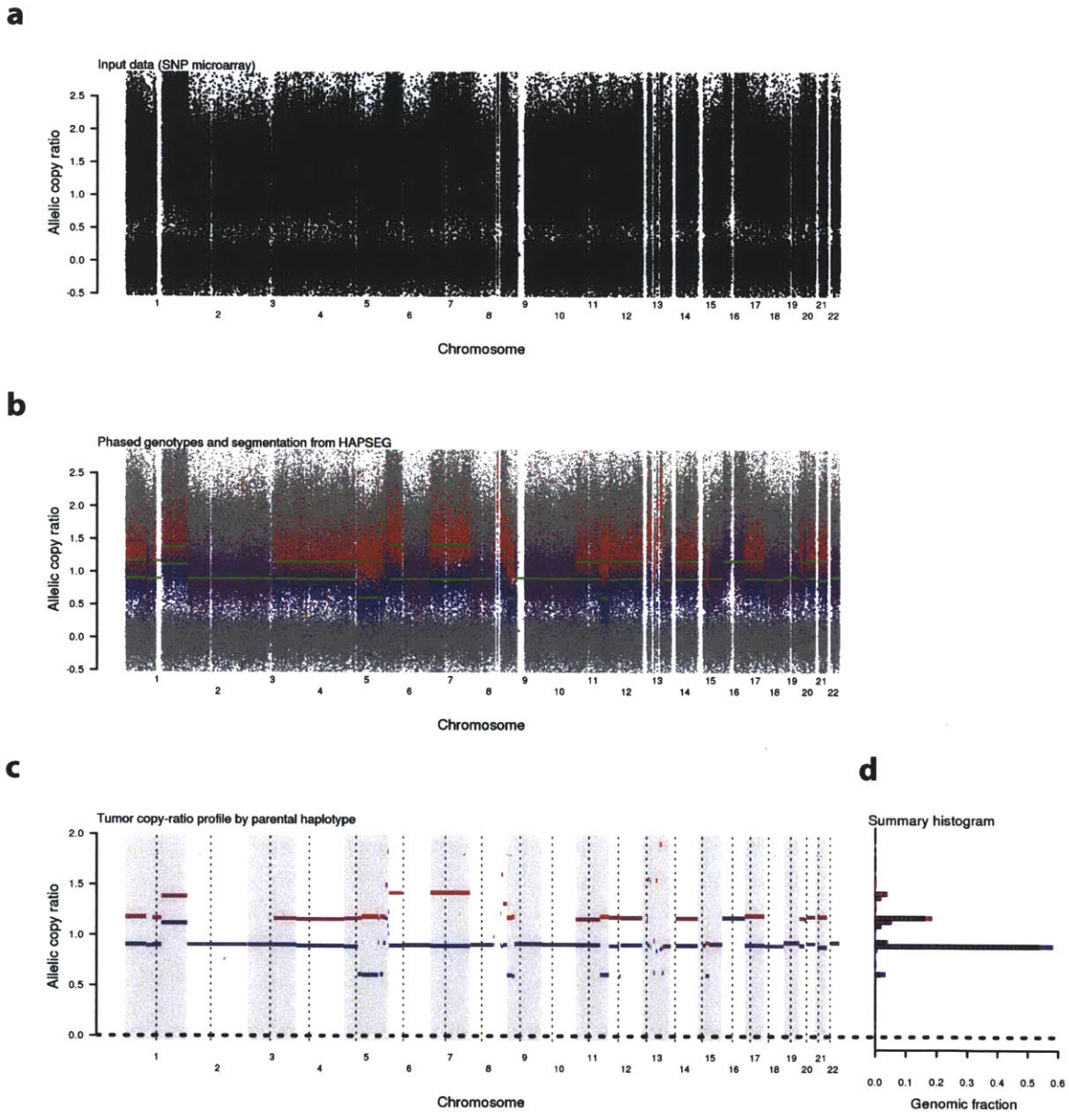


Figure 1

Figure 1. Overview of analysis with HAPSEG a., Input data from an Affymetrix SNP microarray hybridization. The calibrated allelic copy-ratios of A and B channels is shown for each SNP vs. its genomic location. **b.**, The data in (a), after processing with HAPSEG. SNPs are colored according to their phased genotype: grey homozygous, red/blue heterozygous (phased), purple heterozygous (unphased). The HSCR segmentation is indicated by green horizontal lines. **c.**,

Summary of the data in (b), where the individual SNPs have been removed and each HSCR segment is colored by the average phasability of the heterozygous SNPs from which it was estimated. **d.**, A histogram summarizing the data in (c), by marginalizing over the genome. Although the HSCR locations are independent between segments, only four discrete levels are apparent in the histogram. These corresponding to fixed SCNAs fixed in the tumor sample, and are the basis for analysis with ABSOLUTE [3].

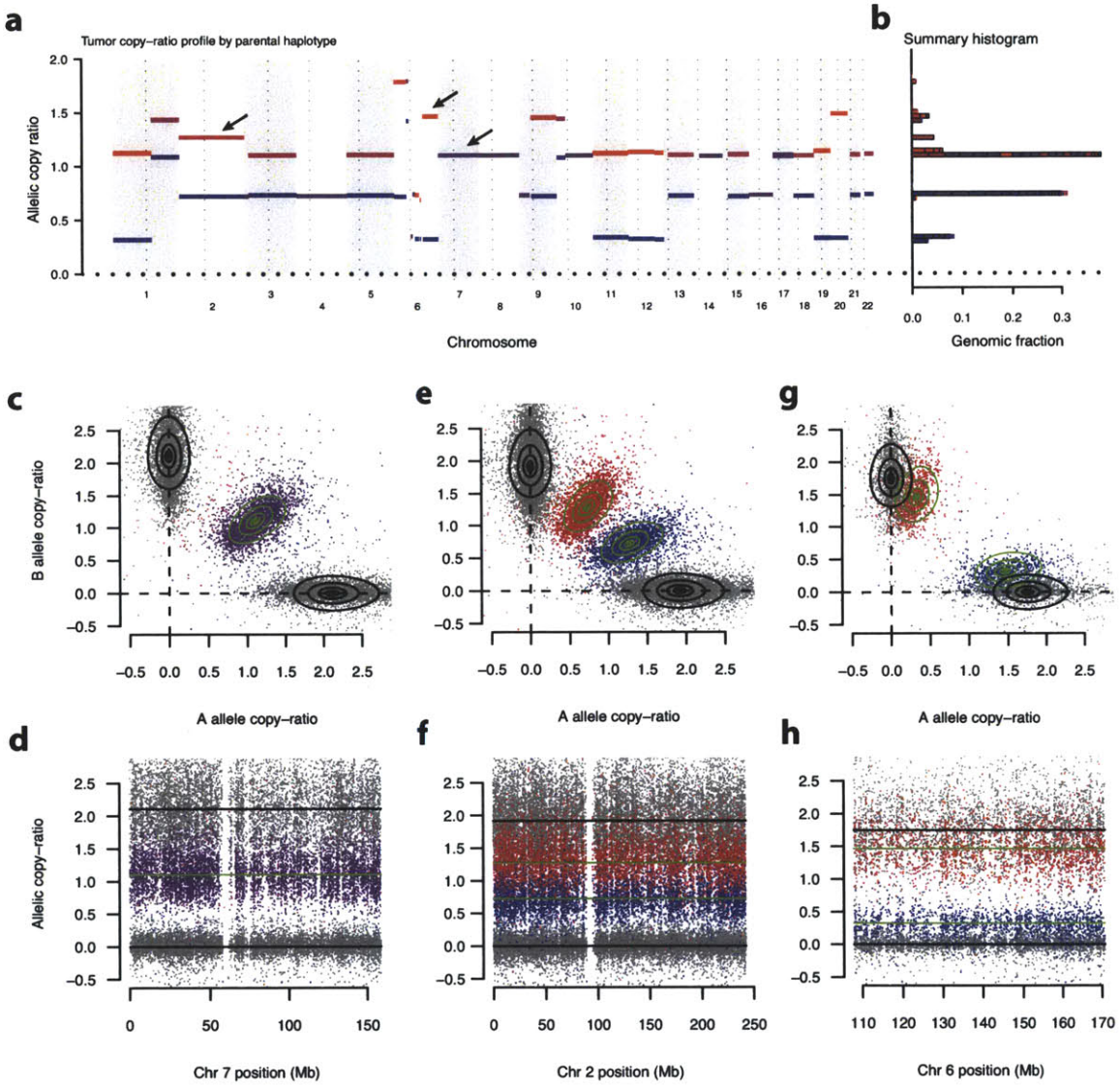


Figure 2

Figure 2. Examples of segmental HSCR inference by HAPSEG **a.**, A tumor sample with genome-wide HSCRs as inferred by HAPSEG, as in fig. 1c. **b.**, Summary histogram of estimated HSCR values, as in fig. 1d. **c.**, **e.**, **g.**, Plots of calibrated A vs. B-allele copy-ratios for SNPs in three genomic locations indicated in (a) (arrows). SNPs are colored by inferred (haploid) genotype, as in fig. 1b. Contours denote the error-model fit for each genotype cluster: black - homozygous, green - heterozygous. **d.**,**f.**,**h.**, Plots of A and B-allele copy-ratios vs. genomic position for the segments in (c,e,g). SNPs are colored as in (c,e,g). Horizontal lines denote genotype cluster locations, colored as in (c,e,g) The heterozygous locations (HSCRs) correspond to the allelic copy-ratios in (a). **c.**,**d.**, A segment (chr 7) at allelic balance, with equal copy-numbers of both homologues. No phase information is available for heterozygous SNPs in this segment. **e.**,**f.**, A segment (chr 2) at allelic imbalance, with unequal homologous copy-numbers. Note that the SCNA affecting this segment

was predicted to be subclonal by ABSOLUTE [3]. **g.**,**h.**, A segment (chr 6) at more extreme allelic imbalance. Note that the lower HSCR here corresponds to LOH in this tumor sample [3]; the DNA contributing to the heterozygous alleles is derived from normal contaminating cells.

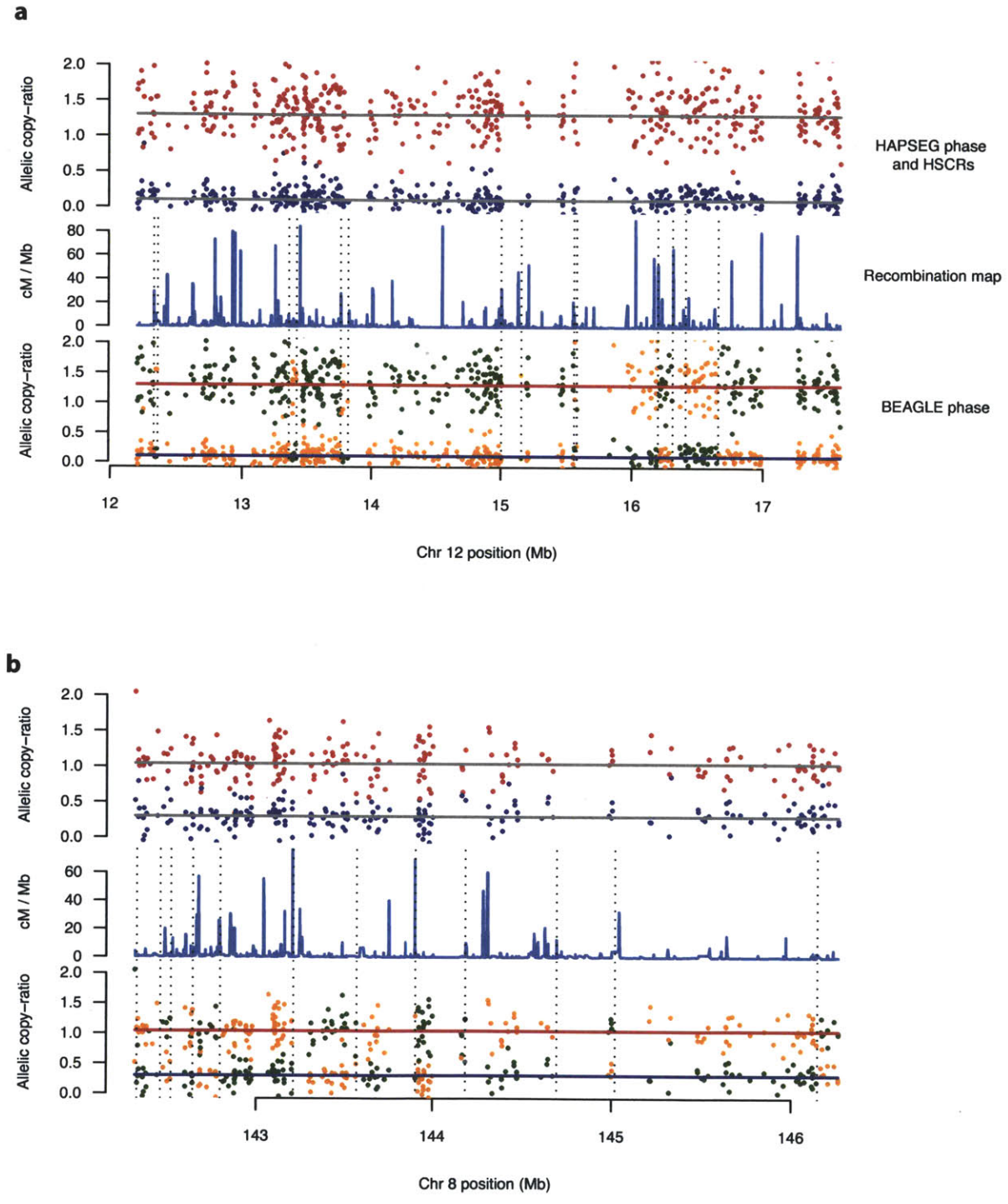

a
Allelic copy−ratio
2.0
1.5
1.0
0.5
0.0
HAPSEG phase
and HSCRs
cM / Mb
80
60
40
20
0
Recombination map
Allelic copy−ratio
BEAGLE phase
12
13
14
15
16
17
Chr 12 position (Mb)
b
Allelic copy−ratio
cM / Mb
60
40
20
0
143
144
145
146
Chr 8 position (Mb)

Figure 3. Demonstration of direct haplotyping by homologous imbalance a.,b., Comparison of HAPSEG and BEAGLE phase in example segments. Top - Allelic copy-ratios of heterozygous SNPs are shown at their genomic coordinates. Color indicates phase as estimated by HAPSEG ($\hat{\mathbf{C}}^{(2)}$), eq. (9). Grey horizontal lines indicate the HSCR estimates for the segment, estimated by HAPSEG, eq. (8). Middle - The genetic recombination rate is plotted vs. the genome. Bottom - As in (**a**), but replacing the HAPSEG phase estimates by those obtained using the statistical phasing program BEAGLE [31]. Switch errors in the BEAGLE phasing, detected by HAPSEG, are indicated by dotted vertical lines.

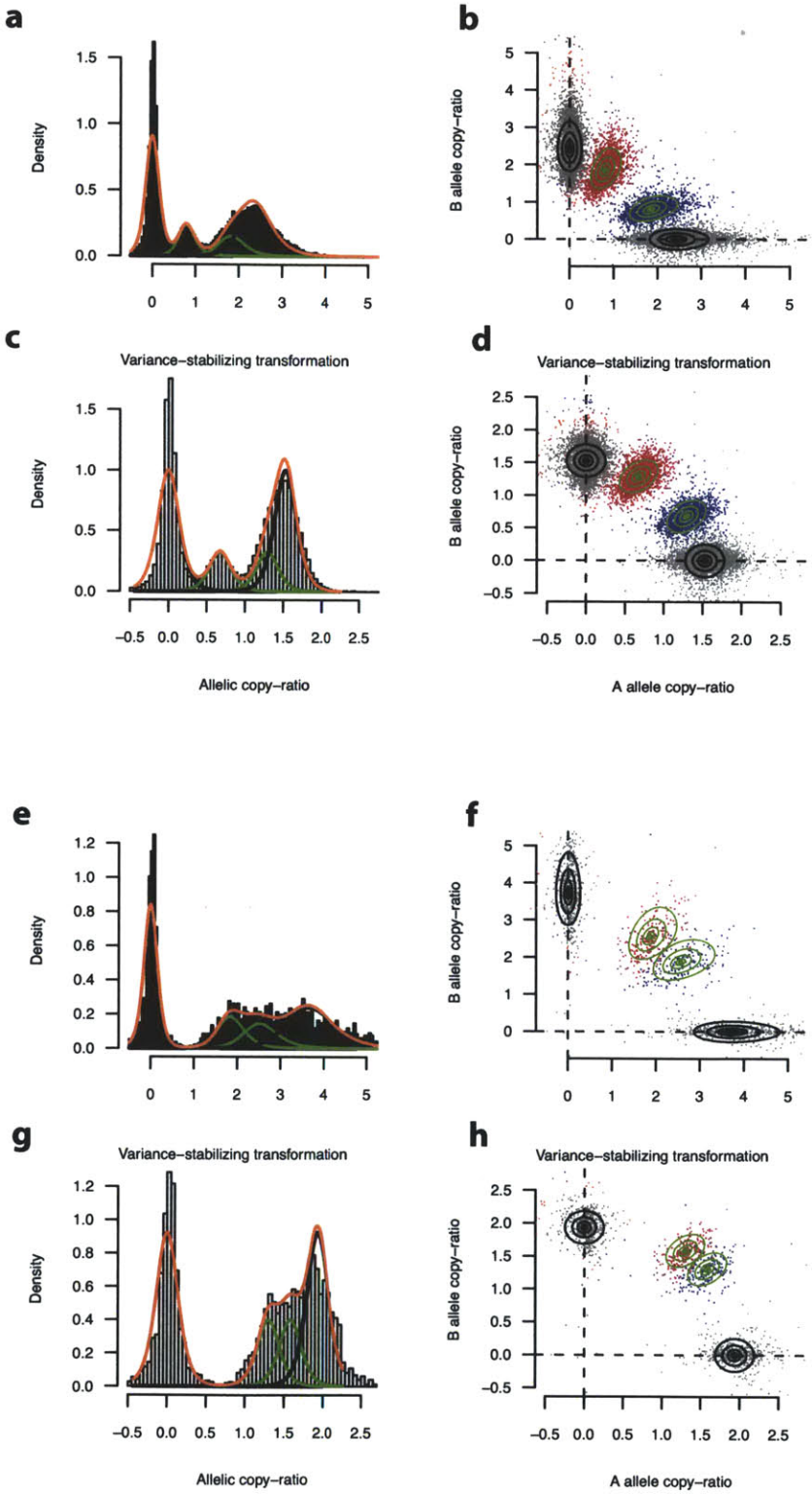

a
b
c
d
e
f
g
h
Density
B allele copy−ratio
Variance−stabilizing transformation
Allelic copy−ratio
A allele copy−ratio

Figure 4. Demonstration of error-model for Affymetrix SNP microarrays. a-d., The fit of the error-model to marker-level data is shown in a single example segment. **a.**, The histogram summarizes the A and B channels for assayed markers within the genomic segment. The marginal fit of the A/B marker copy-ratios is denoted by the green (het) and coral (het+hom) curves. **b.**, Marker copy-ratios in (**a**) are shown separately for the A vs. B channel. Contours denote fit of the error model to the 4 modeled genotype clusters, eq. (1), (2) (green - het, black - hom). **c,d.**, as in **a,b.**, but using the variance stabilizing transformation, eq. (20). Likelyhood calculations are performed in this space using a bi-variate t distribution, eq. (23). Positive covariance of the A/B channels is modeled in the heterozygous clusters (green), this aspect of the fit cannot be dispalyed in one dimension, as in (**a,c**). Contours show the fit of this density to the data, as in (**b**), which were derived by inverting the variance stabalizing transformation. We note that the covariance matrix $\mathbf{\Sigma}_i$ in the density used here is fully determined by the segment HSCR locations, conditional on error model parameters $\mathbf{\Theta}$, which are fit at the sample level, eq. (12). Only two degrees of freedom are fit to the data specifically shown here, eq. (2). **e-h.**, an additional example segment is shown, as in **a-d**.

Chapter 4

—

Absolute quantification of somatic DNA alterations in cancer reveals frequent genome doublings in human cancers

Scott L. Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W. Laird, Robert C. Onofrio, Wendy Winckler, Barbara Weir, Rameen Beroukhim, David Pellman, Douglas A. Levine, Eric S. Lander, Matthew Meyerson, Gad Getz

This chapter was submitted in its present form for publication on August 9th, 2011.

Absolute quantification of somatic DNA alterations in cancer reveals frequent genome doublings in human cancers

Scott L. Carter[1,2*], Kristian Cibulskis[1#], Elena Helman[1,2#], Aaron McKenna[1#], Hui Shen[3#], Travis Zack[4,5#], Peter W. Laird[3], Robert C. Onofrio[1], Wendy Winckler[1], Barbara A. Weir[1], Rameen Beroukhim[1,5,6], David Pellman[7], Douglas A. Levine[8], Eric S. Lander[1,10,11], Matthew Meyerson[1,5], Gad Getz[1*]

[1]*The Broad Institute of Harvard and MIT, Cambridge, MA 02142*
[2]*Division of Health Sciences and Technology, MIT, Cambridge, MA 02139*
[3]*USC Epigenome Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033*
[4]*Biophysics Program, Harvard University, Cambridge, 12 Oxford St., MA 02138*
[5]*Divisions of Medical Oncology and Cancer Biology and Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts, 02115*
[6]*Department of Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA, 02115*
[7]*Howard Hughes Medical Institute, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Children's Hospital, Department of Cell Biology, Harvard Medical School, Boston, MA 02115*
[8]*Gynecology Service, Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, NY 10065.*
[9]*Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142*
[10]*Department of Systems Biology, Harvard Medical School, 25 Shattuck Street, Boston, MA, 02115*
[11]*MIT Department of Biology, 31 Ames Street, Cambridge MA 02139*

[#]These authors contributed equally to this work
***To whom correspondence should be addressed**

ABSTRACT

Defining chromosome copy number and allele ratios is fundamental to understanding the biological consequences of the mutations and the evolution of cancer. Current genomic characterization techniques measure somatic alterations in a cancer sample in units of genomes (DNA mass). The meaning of such measurements is highly dependent on the tumor's purity and its overall ploidy; they are hence complicated to interpret and compare across samples. Ideally, copy-

number should be measured in copies-per-cancer-cell. Such measurements are straightforward to interpret and, for alterations that are fixed in the cancer cell population, are simple integer values. Here, we describe a novel computational method (ABSOLUTE) that infers tumor purity and malignant cell ploidy directly from allelic analysis of DNA. Using these values, ABSOLUTE can detect subclonal heterogeneity, somatic homozygosity, and calculate statistical sensitivity to detect specific aberrations. We used ABSOLUTE to analyze ovarian cancer data and discovered that 54% of somatic point mutations were, in fact, subclonal. In contrast, mutations occurring in key tumor suppressor genes, *TP53* and *NF1* were predominantly clonal and homozygous, as were mutations in a candidate tumor suppressor gene, *CDK12*. Analysis of absolute allelic copy-number profiles from 3,155 cancer specimens revealed that genome-doubling events are common in human cancer, and likely occur in already aneuploid cells in many cancer types. By correlating genome-doubling status with mutation data, we found that homozygous mutations in *NF1* occurred predominantly in non-doubled samples. This finding suggests that genome doubling influences the pathways of tumor progression, with recessive inactivation being less common after genome doubling.

Introduction

Interpreting the mutational drivers of cancer requires the ability to accurately infer the absolute copy number of somatic copy-number alterations (SCNAs) across the human genome, corresponding to integral copy-numbers per cell in a cancer clone. This is considerably more challenging than measure relative copy-number in units of diploid DNA mass in a tumor-derived sample.

Measuring copy number alterations on a relative basis is straightforward using microarrays [1, 2, 3, 4, 5] or massively parallel sequencing technology [6, 7]; it has been the standard approach for copy-number analysis since the development of comparative genomic hybridization (CGH) [8].

Inferring absolute copy-number is more difficult because: (i) cancer cells are nearly always intermixed with an unknown fraction of normal cells (tumor purity); (ii) the actual DNA content of the cancer cells (ploidy), resulting from gross numerical and structural chromosomal abnormalities, is unknown [9], [10], [11], [12], [13] and (iii) the cancer cell population may be heterogeneous, perhaps due to ongoing subclonal evolution [14], [15]. In principle, one could infer absolute copy-numbers by rescaling relative data based on cytological measurements of DNA mass per cancer cell [16], [17], [18], or by single-cell sequencing approaches [15]. However, such approaches do not scale well for high-throughput characterization of cancer genomes [19].

We began focusing on this issue several years ago, initially developing *ad hoc* techniques [20], [21]. We subsequently developed a fully quantitative method (ABSOLUTE) and applied it to several cancer genome analysis projects, including The Cancer Genome Atlas (TCGA) consortium. We used these methods to correlate purity and ploidy estimates with expression subtypes {Verhaak, et al., submitted} and to develop statistical power calculations and use them to select well-powered samples for whole-genome sequencing in several published [22], [23], {Bass, A., et al. (accepted)}, and numerous ongoing projects, including breast, prostate, and skin cancer genome characterization. Recently, we extended ABSOLUTE to infer the multiplicity of somatic point-mutations in integral allelic units per cancer-cell.

Our purpose here is to: (i) to present the mathematical inference framework of the ABSOLUTE method, as well as experimental validation of its predictions; (ii) to apply it to analyze a large cancer dataset, enabling novel characterization of the incidence and timing of whole-genome doublings during tumor evolution; and (iii) to describe a novel integrated analysis of point-mutation and copy-number estimates and its application to ovarian carcinoma.

We describe two key mathematical features of ABSOLUTE. First, it jointly estimates tumor purity and ploidy directly from observed relative copy profiles. Second, because

joint estimation may not be fully determined on a single sample, it uses a large and diverse sample collection to help resolve ambiguous cases.

We then report the first large-scale 'pan-cancer' analysis of copy-number alterations on an absolute basis, across 3,155 cancer samples, representing 25 diseases with at least 20 samples. The analysis reveals that whole-genome doubling events occur frequently during tumorigenesis, ultimately resulting in mature cancers descended from doubled cells, bearing complex karyotypes. Despite evidence that genome doublings can result in genetic instability and accelerate oncogenesis [24],[13],[25] the incidence and timing of such events had not been characterized in human cancer.

We then describe how estimates of tumor purity and absolute copy-number allow us to analyze sequencing data to distinguish clonal and subclonal point-mutations. Clonal events may be classified as homozygous or heterozygous in the cancer cells, guiding interpretation of their function. In addition, the ability to quantify integral multiplicity of point mutations distinguishes events occurring prior to DNA amplification involving the mutated locus from those occurring later.

Finally, our data allows characterization of somatic cancer evolution with respect to genome doubling, which we demonstrate in ovarian carcinoma and associate with clinicopathological parameters.

ABSOLUTE provides a foundation for integrative genomic analysis of cancer genome alterations on an absolute (cellular) basis.

Results

Inference of sample purity and ploidy in cancer copy-profiles

We begin by describing the inference framework used in the ABSOLUTE method. Suppose a cancer-tissue sample consists of a mixture of a proportion α of cancer cells (assumed to be monogenomic – that is, with homogenous SCNAs in the cancer cells) and

a proportion (1-α) of contaminating normal (diploid) cells. For each locus x in the genome, let q(x) denote the integral copy-number of the locus in the cancer cells. Let τ denote the mean ploidy of the cancer-cell fraction, defined as the average value of q(x) across the genome. In the mixed cancer sample, the average absolute copy number of locus x is α q(x) + 2(1 - α) and the average ploidy is D = $\alpha\tau$ + 2(1 - α), measured in units of haploid genomes.

The relative copy number of locus x is therefore:

$$R(x) = [\alpha\, q(x) + 2(1 - \alpha)]/D = [\alpha/D]\, q(x) + [2(1 - \alpha)/D] \quad (1)$$

Because q(x) takes integral values, R(x) takes discrete values. The smallest possible value is [2(1 - α)/D], which occurs at homozygously deleted loci and corresponds to the fraction of DNA from normal cells. The spacing between values is [α/D] corresponds to the concentration ratio of alleles present at one copy per cancer cell and 0 copies per normal cell (e.g. somatic point-mutations). Importantly, if a cancer cell is not monogenomic, copy-number alterations occurring in substantial subclonal fractions will appear as outliers from this pattern (Fig. 1a-c, arrows).

The ABSOLUTE algorithm examines possible mappings from relative to integral copy numbers by jointly optimizing the two parameters α and τ (Fig. 1c-d,h-i; Methods eq. 4). In many cases, several such mappings are possible, corresponding to multiple optima. In such cases, we use recurrent cancer karyotype models based on large data sets of samples (Supplementary figure 1; Methods eq. 6) to identify the simplest (that is, most common) karyotype that can adequately explain the data. The method selects the simplest solution, while preserving the flexibility to identify unexpected karyotypes given sufficient evidence from fitting the copy-profile (Fig. 1e,j). Indeed, several unusual karyotypes, including near-haploid (<1.2n) and hyper-aneuploid (>6n) genomes, were identified using ABSOLUTE (Supplemental Figure 2).

Validation

We validated the purity and ploidy predictions made by ABSOLUTE by several approaches: (i) mixing experiments, in which cancer cell lines were mixed with paired normal B-lymphocyte-derived DNAs in varying mass proportions (mean absolute differences of 6%, Fig. 2c, Methods); (ii) ploidy analysis of 37 TCGA ovarian carcinoma samples by fluorescence-activated cell sorting [26] (Methods) (mean absolute difference = 0.32, and 87% of samples concordant within one unit, Fig. 2a); and (iii) ploidy analysis of the NCI60 cell lines by spectral karyotyping (SKY) [27] (mean absolute difference = 0.36, and 91% of samples concordant within one unit, Fig. 2b);

Notably, the purity estimates produced by ABSOLUTE appeared to be more accurate for the bulk tumor than those derived from histological examination of frozen tumor sections (Methods). Estimates of the proportion of contaminating normal cells for 458 ovarian carcinoma samples[26] produced by ABSOLUTE were strongly correlated with a molecular signature of genomic methylation (Methods) seen in leukocytes ($r^2 = 0.59$, $P < 2.2\text{x}10^{-16}$, Fig. 2d), but only weakly correlated with estimates of contamination from histological examination ($r^2 = 0.1$, $P = 2.4\text{x}10^{12}$; Methods; Fig. 2d x-axis scale, Supplemental figure 3). These results suggest that leukocyte infiltration may not be reliably detected by histological examination.

Estimation of tumor purity and ploidy across cancer types

We used ABSOLUTE to analyze copy-ratio profiles derived from SNP arrays from 3,155 cancer samples, comprising 2,791 tissue specimens and 364 cancer cell lines. The samples came from two TCGA pilot studies describing glioblastoma (GBM; 192 samples) [21] and ovarian carcinoma (488 samples) [26], as well as 2,445 profiles incorporated from a previous pan-cancer copy-number analysis [28] (Methods). A minority of these samples (519 or 16.4%) could not be analyzed because they lacked clearly identifiable SCNAs, either because they were nearly euploid ("non-aberrant"), or were excessively contaminated with normal cells ("insufficient purity"), (fig 2e).

For the 2,636 samples with detectable SCNAs, ABSOLUTE provided purity and ploidy calls for 92% of cases, and designated the remaining samples as "polygenomic" (genomically heterogeneous) (Fig. 2e), (Methods; Supplemental figure 5). The fraction of called samples varied by disease type, from 34.6% (myeloproliferative disease; mostly non-aberrant genomes) to 96.7% (ovarian carcinoma, 100% aberrant genomes), with a median call-rate of 79.2% (Fig. 2e).

The distributions of estimated purity varied among cancer types, with the tested lung, esophageal, and breast cancer samples being the least pure on average in our dataset. The effect of contamination was readily visible in the copy ratios of impure tumor types (Supplemental figure 6). Distributions of estimated ploidy (Supplemental figure 7a) were qualitatively consistent with those derived from previously obtained cytological data for each tumor type [13].

Coverage requirements for detection of somatic point-mutations by sequencing

We used ABSOLUTE to determine the depth of sequencing necessary to detect point mutations. For example, suppose that a region is amplified to 6 copies with only 1 copy carrying a mutation, in a sample that has 50% contamination with normal cells. In this case, only 1 of 8 alleles at this locus (6 from the cancer cells and 2 from the normal cells) carry the mutation (fig 3a). We therefore expect that the mutation will be observed in only 12.5% of reads. Given this allelic fraction, sequence coverage of 33-fold is required to detect the mutation with 80% sensitivity, assuming a sequencing error rate of 10^{-3} per base and a false positive rate controlled at $<5x10^{-7}$, (Methods, eq. 7, Fig. 3b).

Using ABSOLUTE's estimates of purity and genome-wide integral copy-numbers, we can calculate the required coverage to confidently detect mutations present at specified allelic multiplicity per cancer cell. Similar considerations apply to detecting subclonal mutations, present in a fraction of cancer cells, by using fractional multiplicities (Fig. 3c). We note that consideration of tumor purity in units of cells, rather than DNA mass, is preferred for devising power calculations for sequencing experiments, since many somatic alterations of interest are expected to occur at a single copy *per cancer cell.*

We analyzed the distribution of purity and ploidy values in cancer samples analyzed for allelic copy number [28], [21], [26] to determine an appropriate depth of sequencing coverage needed to detect clonal mutations with power 0.8 in each sample. For this purpose, we calculated the number of reads needed to detect a mutation present in one copy at a locus present at the average ploidy level given the sample's purity. (One could alternatively choose a particular percentile on the ploidy distribution.) For such a locus, we found that 30X coverage sufficed for most samples (Fig. 3d). By contrast, such locus carried in a subclone at 20% frequency would require coverage of ~100-fold to allow detection in about half of the samples (fig 3e).

We then examined whole-exome sequencing (WES) data (~150X coverage) from 214 TCGA ovarian carcinoma samples [26] to determine whether detection power was related to the number of mutations actually observed. For each sample, we calculated the proportion of loci for which the sequence data provided at least 80% power to detect mutations present at single copy in a subclone present at 5%. Those samples with the lowest proportion of such well-powered loci tended to be those in which the least mutations were detected ($r^2 = 0.24$, $P = 2.7\text{x}10^{-13}$; Fig. 3f), suggesting that the failure to find such mutations was due to the lack of power. This result also demonstrates the importance of power calculations for characterization of the subclonal frequency spectrum.

Cellular multiplicity analysis of somatic point mutations from sequencing data

We next used ABSOLUTE to infer the absolute copy-number not simply of loci, but of specific mutations. For this purpose, we examined 29,268 somatic mutations identified in whole-exome hybrid capture Illumina sequencing (WES)[29] data from 214 ovarian carcinoma tumor-normal pairs [26] (Methods, Fig. 4a). Tumor purity, ploidy, and absolute copy-number values were obtained from Affymetrix SNP6.0 hybridization data on the same DNA aliquot that was sequenced, allowing the rescaling of allelic-fractions to units of multiplicity (average cancer-cell units) (Fig. 4a,b; Methods, eq. 10).

This procedure identified pervasive subclonal point-mutations in ovarian carcinoma samples. While many of the mutations were clustered around integer multiplicities, the majority of variants (54%) occurred at multiplicities substantially less than 1 copy per average cancer cell, implying the variants were subclonal (Fig. 4b).
As a control, we applied multiplicity analysis to known germline polymorphisms; as expected, these polymorphisms showed no indication of subclonal multiplicity (not shown). The mutation spectrum seen for clonal and subclonal mutations was similar (RMSE = 0.02, Fig. 4c), consistent with a common mechanism of origin. (The power calculations show that the percentage of subclonal mutations that can be detected with 80% power in these sample ranges from 10% to 53%, with a median of 19%, Fig. 3e.)

We next used ABSOLUTE to analyze the multiplicity of both the reference and alternate alleles, to classify point-mutations as either heterozygous or homozygous in the affected cell fraction. We considered 15 genes with mutations recently indentified in these data [26], including 5 known tumor suppressor genes (TSGs) and 5 oncogenes (Fig. 4g). The frequency of homozygous mutations in known TSGs and oncogenes were significantly different, with a significantly elevated fraction of homozygous mutations in the TSGs (P = 0.006, Fig. 4g) and no homozygous mutations in the oncogenes: (P = 0.012, Fig. 4g). This result provides additional evidence that *CDK12* is a putative TSG in ovarian carcinoma [26], since 7 of 12 carrying mutations were homozygous (P= 6.5×10^{-5}; Fig. 4g).

Overall, *TP53* had among the greatest fraction of clonal, homozygous, and amplified mutations of any gene in the coding exome (Fig. 4h), demonstrating the clear identification of a key initiating event in HGS-OvCa carcinogenesis [30] directly from genomic data and independently of statistical recurrence analysis.

Whole genome doubling occurs frequently in human cancer

For many cancer types, the distribution of total copy number (ploidy) was markedly bi-modal (Supplemental figure 7a), consistent with chromosome-count profiles derived from SKY[10], [13]. Although these results are consistent with whole-genome doubling during

their somatic evolution, it has been difficult to rule out the alternative hypothesis that evolution of high-ploidy karyotypes result from a process of successive partial amplifications [12].

To study genome doublings, we used the fact that ABSOLUTE provides not just *total* copy-number information, but *allelic* copy-number information – that is, the copy numbers, b_i and c_i, of the two homologous chromosome segments at each locus. By looking at the distribution of b_i, c_i across the genome, we can draw inferences regarding genome doubling. Immediately following genome doubling, both, b_i and c_i would be even numbers. Following the loss of single copy, the larger of the two numbers will remain even, but the smaller would be odd. In fact, when we looked at high-ploidy samples, we discerned a clear bias for the higher of the two numbers to be even. This bias was coincident with the transition from low to high-ploidy samples (Supplemental figure 7) being due to genome doubling. Using simulations, we found that the observed profiles were unlikely to arise due to SCNAs occurring in serial fashion at multiple independent chromosomes ($P < 1e\text{-}3$; Methods).

Using such information, we could classify samples into three groups, which we interpreted as corresponding to 0, 1 and > 1 genome doubling events in the clonal evolution of the cancer. These three groups had modal ploidy values of 1.75, 2.75, and 4.0, respectively (Fig. 5a), and also segregated into three clusters by ploidy and mean allelic copy-number imbalance (Fig. 5b). We interpreted this as evidence of SNCAs occurring with net losses, interspersed with the genome doublings. This process resulted in intermediate ploidy values for the doubled clones (2.2 – 3.4N), with pervasive imbalance of homologous chromosomes (Fig. 5b).

The frequency of genome doubling varied across tumor types (Fig. 5c), reflecting differences in disease specific biology and clinical progression status. Hematopoietic neoplasms (MPD, ALL) had nearly no doubling events, whereas GBM, RCC, prostate cancer, various sarcomas, HCC, and medulloblastoma all had ~25% incidence of doubling. Genome doubling was more common in epithelial cancers, with colorectal,

breast, lung, ovarian, and esophageal cancers all having > 50% incidence of doubling (Fig. 5c). Esophageal adenocarcinoma had the greatest incidence of genome doublings, consistent with previous reports of frequent "4N" populations at various stages of Barrett's esophagus progression [31,32].

Specific aneuploidies precede genome doubling

We then used ABSOLUTE to infer the temporal order of genome doubling in tumorigenesis, relative to SNCAs involving specific chromosome arms. In many cancer types, the fixation of arm-level SCNAs was inferred to occur prior to genome doubling, since both doubled and non-doubled samples had similar frequencies of specific arm-level SNCAs (Fig. 5d, Supplementary figure 8).

In GBM samples, LOH involving chromosomes 9 and 10, and gains of chromosome 7 occurred at equivalent frequencies (Fig. 5d), demonstrating that the most common broad SCNAs in GBM occur prior to genome doubling. Gain of chromosomes 19 and 20 was nearly exclusive to non-doubled samples, and several arms had greater LOH frequency in doubled samples (Fig. 5d), suggesting that additional biological differences underlie these samples.

Because ABSOLUTE could not distinguish between ploidy 2N and 4N in cases with no observed SCNAs, we discarded such non-aberrant samples from our analysis (Fig. 2e). For many tumor types, such cases were rare, due to the tendency for chromosomal losses following doubling (Fig. 2e, Fig. 5a,b, Supplemental figure 7). The representation of specific cancer subtypes may be biased by differences in ascertainment, however.

In contrast to broad chromosomal alterations, focal SCNA events occurred at greater frequency in doubled genomes (Fig. 5e). The frequency of focal SCNAs as a function of their length (L) appeared to follow power-law scaling: $P(L) \propto L^{-\alpha}$, for L between 0.5 and 20 Mb (Fig. 5e). The observation of power-law scaling was consistent with previous reports [28, 33]. Although samples with genome doubling showed a higher number of SCNAs, we obtained equivalent estimates of $\alpha \cong 0.71$ for each group (Fig. 5e),

suggesting that the mechanism(s) by which they were generated did not depend on ploidy.

Genome doubling interactions with somatic genetics in ovarian carcinoma

We next sought to correlate whole-genome doubling occurrence in ovarian carcinoma with other genetic and clinical features. Genome-doubled samples showed a higher incidence of heterozygous mutations, but correcting for sample ploidy removed this effect (Fig. 6a), suggesting that the per-base mutation rates are equivalent. Clonal amplified mutations were approximately ten-fold more prevalent in doubled samples; many of these events likely occurred prior to the doubling event. Genome-doubled samples had a lower frequency of homozygous deletions (Fig. 6b) and a two-fold lower rate of clonal homozygous mutations ($P = 1.55 \times 10^{-8}$, Fig. 6c). We expect that many of the observed homozygous alterations in the doubled samples were fixed prior to genome doubling.

The lower incidence of homozygous mutations in genome-doubled samples may reflect the fact that more events are required to render a mutation homozygous in a genome-doubled sample (although the effect may be partially offset, however, by a possible increase in genetic instability following doubling, e.g., by centrosome duplication [34]). These considerations suggest that genome-doubled samples evolve via distinct trajectories, because inactivation of tumor suppressors may occur less frequently following doubling.

We note that 13 of the 15 detected point mutations in the tumor suppressor *NF1* occurred in the 93 samples that had not undergone genome doubling ($P = 0.002$), and these mutations were uniformly homozygous (not shown). This is consistent with selection for recessive inactivation of *NF1*, a typical pattern for a tumor suppressor gene. It also suggests that non-genome-doubled ovarian carcinoma samples evolved via a distinct trajectory, rather than being precursors to doubled samples. If not, many *NF1* mutations would be homozygous and amplified in doubled samples, as is seen for *TP53*.

Although the genome doubled and non-doubled samples showed significant overall differences in the frequency of subclonal mutations, this number did not provide a useful way to characterize individual samples, because it was too strongly dependent on the sequence coverage, and therefore on the sensitivity to detect mutations (Fig. 3c,e,f).

Finally, we noted that genome-doubled samples were associated with a significant increase in the age at pathological diagnosis (Fig. 6d) and with a significantly greater incidence of cancer recurrence (Fig. 6e).

Discussion

The development of a reliable high-throughput method to infer absolute allelic copy-number from DNA derived from tumor samples represents an important advance in cancer genome analysis. ABSOLUTE is currently the most technically complete approach to analyzing cancer genomic data in cellular units, both for point mutations and SCNAs, and has become an established step in ongoing integrative genomic analysis projects. It may be possible to extend ABSOLUTE to other types of genomic alterations, such as structural rearrangements and small insertions/deletions, although this may require longer sequence reads in order to ensure accurate sequence alignment.

We have demonstrated that the classification of point-mutations based on their multiplicities, as estimated by ABSOLUTE, can help distinguish tumor suppressors and oncogenes beyond traditional evidence based on mutation types (nonsense, frame-shift indels and splice site mutations are more common in tumor suppressors) and their distribution along the gene (clustered mutations are more common in oncogenes) (Fig. 4). Further development of techniques utilizing these classifications may contribute clues to the function of novel recurrent somatic alterations in the cancer genome.

ABSOLUTE provides a framework for the design of genomic sequencing efforts, based on calculation of sensitivity to detect mutations as a function of a sample's purity, local copy-number and sequencing depth (Fig. 3). The high accuracy of tumor purity and

ploidy estimates produced by ABSOLUTE based on SNP microarray data (Fig. 2) make it possible to determine the sequencing depth required for a given sample or to select suitable samples given a fixed sequencing depth. Furthermore, these considerations are important for interpretation of apparent differences in mutation rates across samples of differing purity, such as primary tumors and cell lines, metastases, or xenografts [35],[36]. Finally, such considerations are vital to the interpretation of subclonal point-mutations (Fig. 3f, Supplementary figure 9).

The detection of subclonal mutations in primary tumors will further enable the study of tumor heterogeneity, as in [37], [38]. Further analysis of the subclonal multiplicities, provided by ABSOLUTE, may expose discrete structure in the tumor subclonal populations reflecting distinct fractions of clonal sub-populations in polyclonal tumors. Such analysis will require adequate sequence coverage, as suggested by our sensitivity calculations (figure 3, Supplementary figure 9).

ABSOLUTE analysis of SCNAs demonstrated that many of the copy-number profiles analyzed in this study could be explained as derived from approximately monogenomic tumor samples. In many other cases, however, ABSOLUTE predicts that some SCNAs are, in fact, present in different number of copies in minor subclones. These predictions present opportunities to study tumor subclonal evolution.

Analysis of the predicted absolute allelic copy-number profiles across human cancers produced by ABSOLUTE shed new light on our understanding of cancer genome evolution. The profiles (Supplemental figure 7) were consistent with a common trajectory consisting of an early period of chromosomal instability followed by the emergence of a stable aneuploid clone, as previously described [11]. Our data further indicate that genome doublings occur in a subset of cancer cells already harboring arm-level SCNAs characteristic of the corresponding tumor type. The genomes of these cancers were therefore shaped by selection at chromosomal arm-level resolution prior to doubling and further clonal outgrowth (Fig. 5g).

These findings are broadly consistent with an earlier interpretation of primary breast cancer FACS/SKY profiles [39], and has recently been recapitulated in studies of macro-dissected and ploidy-sorted cell populations [14], and single cell sequencing [15] of primary breast tumors. A notable exception occurred in prostate cancer samples, a subset of which were predicted as near-tetraploid, with few broad SCNAs (Supplemental figure 7a, 8); further validation is required to confirm these predictions. We note that this model represents a departure from the idea that tetraploidization is an *initiating* event [40, 13, 25, 41, 42]. In addition, the association of genome doubling with epithelial lineage (Fig. 5c), and with age at diagnosis in ovarian carcinoma (Fig. 6d) is consistent with a recently described mechanism linking telomere crisis, DNA damage response, and genome doubling in cultured mouse embryonic fibroblasts (MEFs) [41].

Further analysis of absolute allelic copy-number profiles and the classification of samples based on the number of genome doubling events they have undergone may help improve methods for implicating specific SCNAs as driver events in cancer [28, 33]. For example, background alteration rates could be adjusted to account for differing alterations frequencies during various phases of cancer genome evolution (eg. pre/post doubling). Because genome doubled cancers had a greater incidence of focal SCNAs (Fig. 5e) and point mutations (Fig. 6a), it may be valuable to use different background mutation rates corresponding to different phases of cancer genome evolution (e.g., pre- vs. post-doubling).

Synthetic lethality screens have identified genetic dependencies with yeast polyploidy [43, 44] and aneuploidy [45]. In addition, chemical compounds have been identified with pro-apoptotic function specifically in primary MEFs bearing specific aneuploidies [46]. Analysis of aneuploidy and genome doubling status in large collections of human cancer cell-lines using ABSOLUTE will inform further chemical and genetic screening of these cells [47].

Finally, knowledge of discrete tumor copy-states and genome doubling status provides an important foundation for reconstruction of the temporal sequence of alterations by which

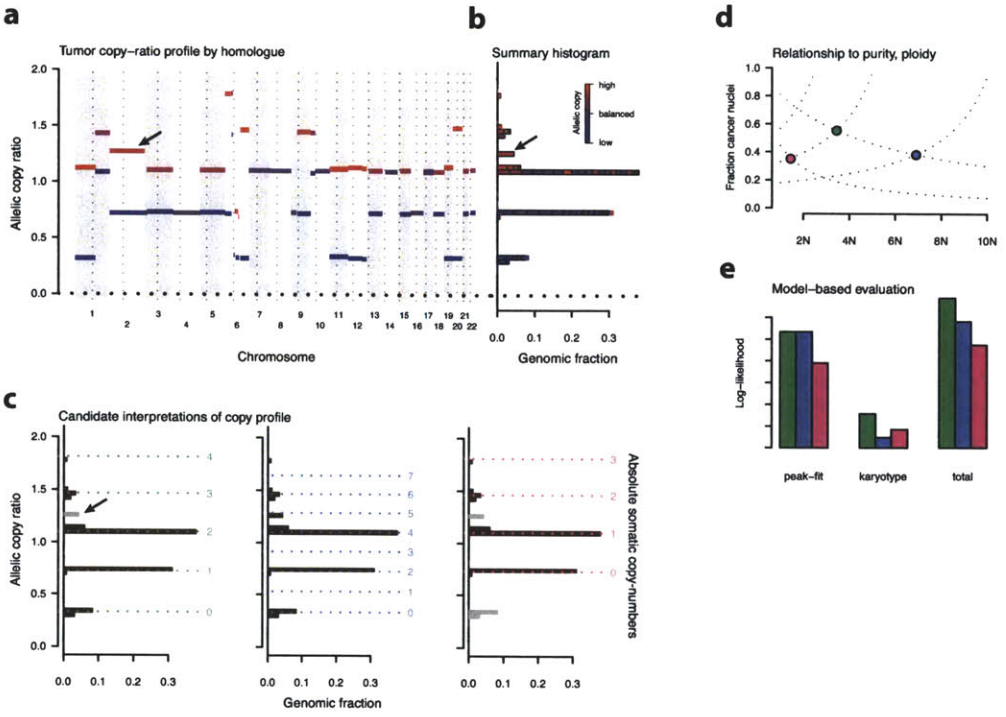
the cancer genome arose. The idea of using the fraction of homozygous mutations in regions of copy-neutral LOH to time duplication events has been previously proposed [48], and applied to ovarian cancer {Durnick 2011, in press}, where it confirmed the known early occurrence of TP53 mutations[30]. The use of ABSOLUTE will extend these techniques by providing information about the timing of events that do not occur in regions of copy-neutral LOH, thereby extending the applicability of these and related methods [49],[50], [51] and enabling the development of novel strategies for temporal inference.

References

1. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**, 207-11 (1998).
2. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res* **10**, 1126-37 (2000).
3. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat Biotechnol* **18**, 1001-5 (2000).
4. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* **64**, 3060-71 (2004).
5. Bignell, G.R. *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res* **14**, 287-95 (2004).
6. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-9 (2008).
7. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99-103 (2009).
8. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. in *Science* Vol. 258 818-21 (1992).
9. Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. in *Journal of cell science* (2008).
10. Mitelman, F. Recurrent chromosome aberrations in cancer. in *Mutation research* Vol. 462 247-53 (2000).
11. Albertson, D.G., Collins, C., McCormick, F. & Gray, J.W. Chromosome aberrations in solid tumors. *Nat Genet* **34**, 369-76 (2003).
12. Storchova, Z. & Pellman, D. From polyploidy to aneuploidy, genome instability and cancer. *Nat Rev Mol Cell Biol* **5**, 45-54 (2004).
13. Storchova, Z. & Kuffer, C. The consequences of tetraploidy and aneuploidy. *J Cell Sci* **121**, 3859-66 (2008).
14. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res* **20**, 68-80 (2010).
15. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* (2011).
16. Hicks, J. *et al.* High-resolution ROMA CGH and FISH analysis of aneuploid and diploid breast tumors. *Cold Spring Harb Symp Quant Biol* **70**, 51-63 (2005).
17. Mullighan, C.G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758-64 (2007).

18. Lyng, H. *et al.* GeneCount: genome-wide calculation of absolute tumor DNA copy numbers from array comparative genomic hybridization data. *Genome Biol* **9**, R86 (2008).
19. Hudson, T.J. *et al.* International network of cancer genome projects. *Nature* **464**, 993-8 (2010).
20. Weir, B.A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-8 (2007).
21. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
22. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
23. Stransky, N. *et al.* The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. *Science* (2011).
24. Fujiwara, T. *et al.* Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature* **437**, 1043-7 (2005).
25. Holland, A.J. & Cleveland, D.W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat Rev Mol Cell Biol* **10**, 478-87 (2009).
26. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).
27. Roschke, A.V. *et al.* Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer Res* **63**, 8634-47 (2003).
28. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905 (2010).
29. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-9 (2009).
30. Levanon, K., Crum, C. & Drapkin, R. New insights into the pathogenesis of serous ovarian cancer and its clinical impact. *J Clin Oncol* **26**, 5284-93 (2008).
31. Galipeau, P.C. *et al.* 17p (p53) allelic losses, 4N (G2/tetraploid) populations, and progression to aneuploidy in Barrett's esophagus. *Proc Natl Acad Sci U S A* **93**, 7081-4 (1996).
32. Barrett, M.T. *et al.* Evolution of neoplastic cell lineages in Barrett oesophagus. *Nat Genet* **22**, 106-9 (1999).
33. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
34. Ganem, N.J., Godinho, S.A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460**, 278-82 (2009).
35. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-10 (2009).
36. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005 (2010).
37. Campbell, P.J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A* **105**, 13081-6 (2008).

38. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114-7 (2010).
39. Dutrillaux, B., Gerbault-Seureau, M., Remvikos, Y., Zafrani, B. & Prieur, M. Breast cancer genetic evolution: I. Data from cytogenetics and DNA content. *Breast Cancer Res Treat* **19**, 245-55 (1991).
40. Ganem, N.J., Storchova, Z. & Pellman, D. Tetraploidy, aneuploidy and cancer. in *Curr Opin Genet Dev* Vol. 17 157-62 (2007).
41. Davoli, T., Denchi, E.L. & de Lange, T. Persistent telomere damage induces bypass of mitosis and tetraploidy. *Cell* **141**, 81-93 (2010).
42. Bazeley, P.S. *et al.* A model for random genetic damage directing selection of diploid or aneuploid tumours. *Cell Prolif* **44**, 212-23 (2011).
43. Lin, H. *et al.* Polyploids require Bik1 for kinetochore-microtubule attachment. *J Cell Biol* **155**, 1173-84 (2001).
44. Storchova, Z. *et al.* Genome-wide genetic analysis of polyploidy in yeast. *Nature* **443**, 541-7 (2006).
45. Torres, E.M. *et al.* Identification of aneuploidy-tolerating mutations. *Cell* **143**, 71-83 (2010).
46. Tang, Y.C., Williams, B.R., Siegel, J.J. & Amon, A. Identification of aneuploidy-selective antiproliferation compounds. *Cell* **144**, 499-512 (2011).
47. Cheung, H.W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 12372-7 (2011).
48. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-6 (2010).
49. Hoglund, M., Frigyesi, A., Sall, T., Gisselsson, D. & Mitelman, F. Statistical behavior of complex cancer karyotypes. *Genes Chromosomes Cancer* **42**, 327-41 (2005).
50. Ozery-Flato, M. & Shamir, R. Sorting cancer karyotypes by elementary operations. *J Comput Biol* **16**, 1445-60 (2009).
51. Attolini, C.S. *et al.* A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc Natl Acad Sci U S A* **107**, 17604-9 (2010).


a
Tumor copy−ratio profile by homologue
Allelic copy ratio
Chromosome
b
Summary histogram
Allelic copy
high
balanced
low
Genomic fraction
d
Relationship to purity, ploidy
Fraction cancer nuclei
2N
4N
6N
8N
10N
e
Model−based evaluation
Log−likelihood
peak−fit
karyotype
total
c
Candidate interpretations of copy profile
Allelic copy ratio
Genomic fraction
Absolute somatic copy−numbers

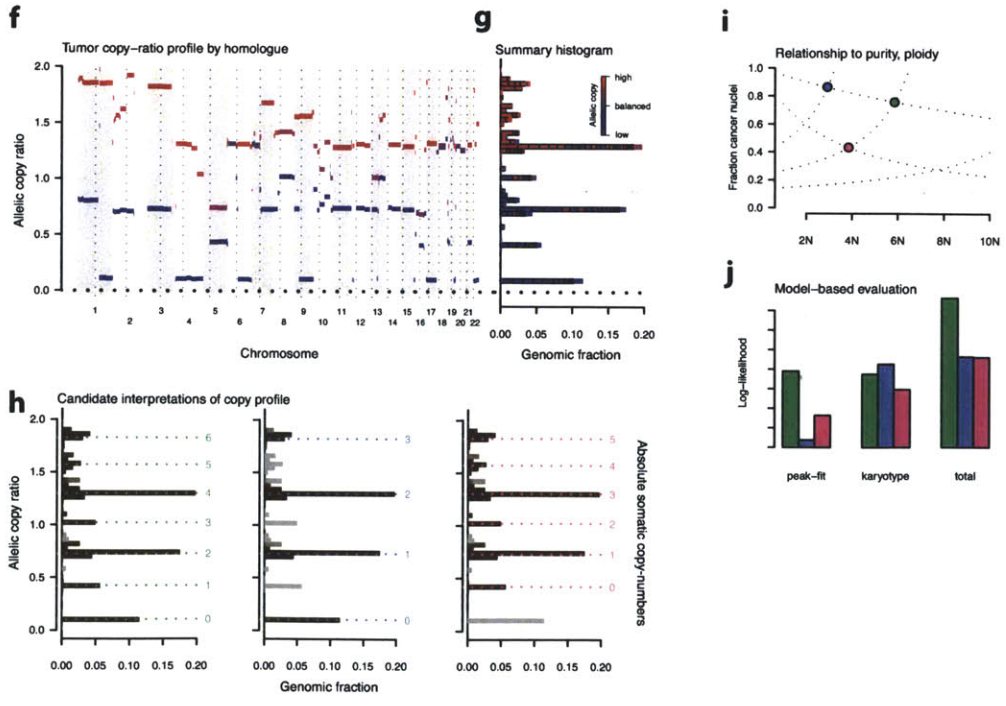

f
Tumor copy−ratio profile by homologue
Allelic copy ratio
Chromosome
g
Summary histogram
Allelic copy
high
balanced
low
Genomic fraction
i
Relationship to purity, ploidy
Fraction cancer nuclei
2N
4N
6N
8N
10N
j
Model−based evaluation
Log−likelihood
peak−fit
karyotype
total
h
Candidate interpretations of copy profile
Allelic copy ratio
Genomic fraction
Absolute somatic copy−numbers

Figure 1 | Example of ABSOLUTE analysis on two tumor samples.
a-e, A lung adenocarcinoma tumor sample processed using ABSOLUTE.
a, Genome-wide view of allelic copy ratios. The copy ratios for both homologues are shown for each genomic segment with locally constant copy number. Color-axis indicates distance between low (blue) and high (red) homologue concentration; segments where these are similar (homologous-allele balance) are purple.
b, Allelic copy-ratio histogram. Allelic copy-ratios shown in (**a**) were binned at 0.04 resolution (*y*-axis); the length of each block corresponds to the (haploid) genomic fraction (*x*-axis) of each corresponding segment in (**a**). Several discrete peaks are visible, each corresponding to either an (unknown) integral copy state in the somatic clone, or to a subclonal alteration.
c, Three potential interpretations of the allelic copy-ratio histogram (**b**) in terms of absolute copy numbers. Horizontal dotted lines indicate the copy ratios corresponding to the indicated absolute somatic copy-numbers.
d, Purity (fraction of tumor nuclei) and cancer-genome ploidy values corresponding to each interpretation in (**c**). Dotted lines denote potential solutions that share either b, the copy-ratio associated with zero somatic copies (from upper-left to lower-right), or δ_τ, the spacing between consecutive integral copy-levels (from lower-left to upper-right). Candidate solutions lie on the indicated grid of $b = 2(1-\alpha)/D$ and $\delta_\tau = \alpha/D$ (eq. 1).
e, The log-likelihood (score) of each solution in terms of the peak-fit of the observed copy-ratios to integer absolute copy-numbers and plausibility of the proposed karyotype. The highest-scoring solution (green) is identified by the combination of peak-fit and karyotype log-likelihood values. This interpretation implies subclonal gain of chromosome 2 (arrows). The peak-fit score alone cannot distinguish between this and an additional solution (blue), in which the arrowed region is closer to an integral copy-state, but the overall peak-fit score is equivalent to that of the first solution.
f-j, An HGS-OvCa sample processed using ABSOLUTE, as in **a-e**. In this sample, the excellent peak-fit score is strongly opposed to a karyotype score indicating a simpler solution, and dominates the final ranking.

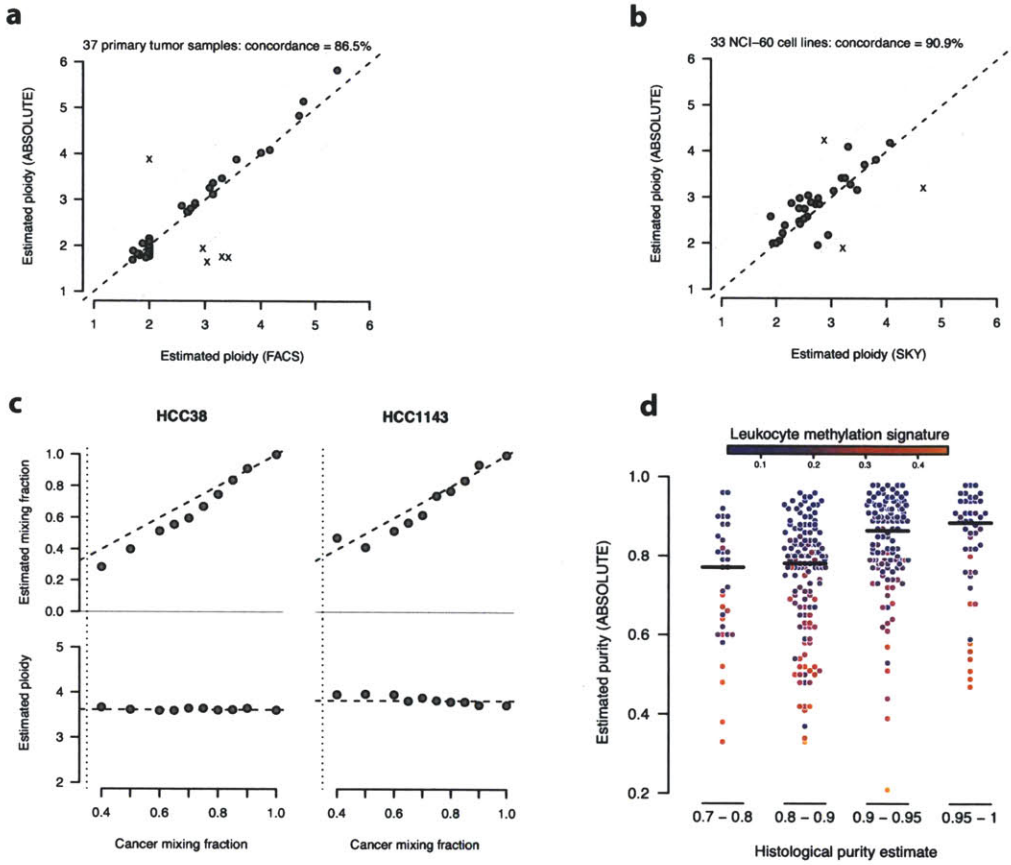

a
37 primary tumor samples: concordance = 86.5%
Estimated ploidy (ABSOLUTE)
Estimated ploidy (FACS)
b
33 NCI–60 cell lines: concordance = 90.9%
Estimated ploidy (ABSOLUTE)
Estimated ploidy (SKY)
c
HCC38
HCC1143
Estimated mixing fraction
Estimated ploidy
Cancer mixing fraction
Cancer mixing fraction
d
Leukocyte methylation signature
Estimated purity (ABSOLUTE)
0.7 – 0.8
0.8 – 0.9
0.9 – 0.95
0.95 – 1
Histological purity estimate
e
called (76.7%)
non–aberrant (9.1%)
insufficient purity (7.3%)
polygenomic (6.9%)
Median called fraction = 0.79
TCGA – HGS–OvCa (485)
RCC (87)
Glioma (27)
TCGA – GBM (192)
Neuroblastoma (25)
Cell lines (364)
HCC (108)
Colorectal Cancer (125)
Medulloblastoma (96)
ALL (235)
MFH (36)
Ovarian Cancer (90)
Prostate Cancer (61)
Esophageal adeno (69)
Other Sarcoma (66)
Liposarcoma (93)
Breast Cancer (134)
Lung squamous (46)
Esophageal squamous (32)
Lung adeno (543)
GIST (36)
MPD (205)
Fraction of samples

Figure 2 | ABSOLUTE method validation and call rate.
a, FACS-based ploidy estimates vs. ABSOLUTE predictions for 37 primary tumor samples. Dashed line indicates *y=x*. The dotted lines indicate a margin for correct prediction of ploidy of +/- 1. Five tumor samples were predicted to have ploidy values that differed by more than one unit from the FACS result (x's).
b, SKY-based ploidy estimates vs. ABSOLUTE predictions for 34 cancer cell-lines. Data are displayed as in **a**.
c, Cancer-normal DNA mixing experiment results for two cell lines. DNA from each cancer cell line was mixed with DNA from the matched B-lymphocyte in varying proportions (*x*-axis). (top) predicted vs. true DNA mixing fractions compared to the *y=x* line (dashed). (bottom) predicted cancer cell-line ploidy vs. mixture purity. Dashed line indicates the median prediction. Ploidy estimates are consistent with previous SKY analysis of these cell lines: http://www.path.cam.ac.uk/~pawefish/cell%20line%20catalogues/breast-cell-lines.htm.
d, Leukocyte methylation signature enrichment in tumors of histologicaly underestimated purity. HGS-OvCa samples are shown grouped according to the indicated histological purity estimates (*x*-axis){Network:2011bq}. Black horizontal lines indicate the median purity of each group, as estimated by ABSOLUTE (*y*-axis). The color of each point corresponds to the degree to which that sample's methylation profile resembled that of purified leukocytes (Methods).
e, ABSOLUTE result types: (i) 'called' -- unique purity / ploidy solution; (ii) 'non-aberrant' -- sample has no detectable somatic copy-number alterations; (iii) 'insufficient purity' -- insufficient fraction of cancer cells; (iv) 'polygenomic' discrete copy-ratio levels could not be determined. See Supplemental figure 5 for examples of each result type.

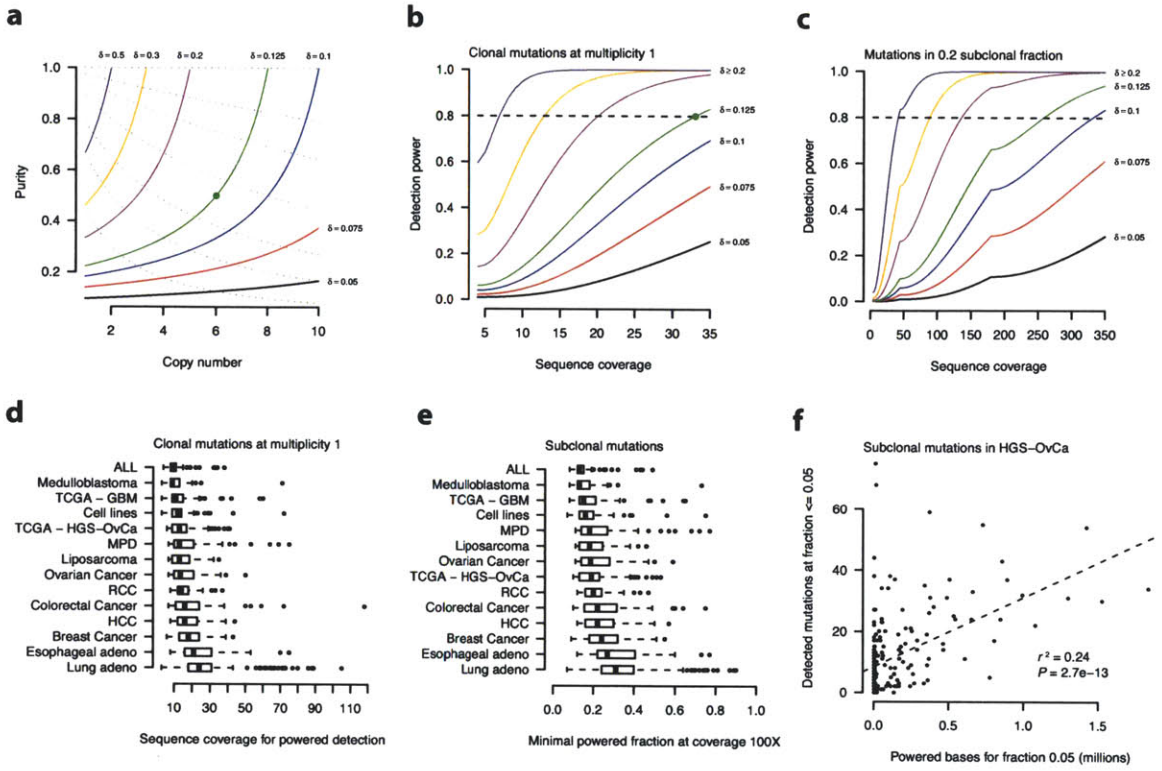

a
δ=0.5 δ=0.3 δ=0.2 δ=0.125 δ=0.1
Purity
Copy number
δ=0.075
δ=0.05
b
Clonal mutations at multiplicity 1
Detection power
Sequence coverage
δ≥0.2
δ=0.125
δ=0.1
δ=0.075
δ=0.05
c
Mutations in 0.2 subclonal fraction
Detection power
Sequence coverage
δ≥0.2
δ=0.125
δ=0.1
δ=0.075
δ=0.05
d
Clonal mutations at multiplicity 1
ALL
Medulloblastoma
TCGA – GBM
Cell lines
TCGA – HGS–OvCa
MPD
Liposarcoma
Ovarian Cancer
RCC
Colorectal Cancer
HCC
Breast Cancer
Esophageal adeno
Lung adeno
Sequence coverage for powered detection
e
Subclonal mutations
ALL
Medulloblastoma
TCGA – GBM
Cell lines
MPD
Liposarcoma
Ovarian Cancer
TCGA – HGS–OvCa
RCC
Colorectal Cancer
HCC
Breast Cancer
Esophageal adeno
Lung adeno
Minimal powered fraction at coverage 100X
f
Subclonal mutations in HGS–OvCa
Detected mutations at fraction <= 0.05
$r^2 = 0.24$
$P = 2.7e{-}13$
Powered bases for fraction 0.05 (millions)

Figure 3 | Effect of tumor purity and ploidy on power for detection of somatic point-mutations by sequencing

a, Combinations of tumor purity and ploidy which imply equal values of δ (eq. 8), the concentration-ratio of molecules present at a single copy per cancer cell (multiplicity 1) in the tumor population. Dotted lines indicate equal values of b (eqn 2).

b, Theoretical power to detect clonal somatic point mutations present at multiplicity 1, as a function of sequence coverage, for various values of δ (shown in **a**). Power was calculated for FPR <= 5×10^{-7}, assuming a uniformly random sequencing error-rate of 10^{-3} (methods, eqn. 7).

a,b, Green dots correspond to an example case of a region amplified to 6 copies, with 1 copy mutated, and purity 50%. Only 1 of 8 alleles at this locus carry the mutation (**a**), and 33 reads are required to detect the mutation with 80% power (**b**).

c, Theoretical power (as in **b**) to detect subclonal somatic point mutations present at multiplicity 1 in 0.2 of cancer cells.

d,e, Values were calculated for each tumor using the purity and ploidy estimates obtained from ABSOLUTE. Samples were considered powered if their detection power (as in **b**, **c**) exceeded 0.8.

f, Tumor purity and local absolute copy-number estimates were used to calculate the number of bases powered for the detection of subclonal mutations present at fraction 0.05 (x-axis, Methods, eq. 7), and to compute the number of subclonal mutations at cancer-cell fraction <= 0.05 in each sample (y-axis). The dashed line, r^2, and P values refer to a linear regression fit of the data points.

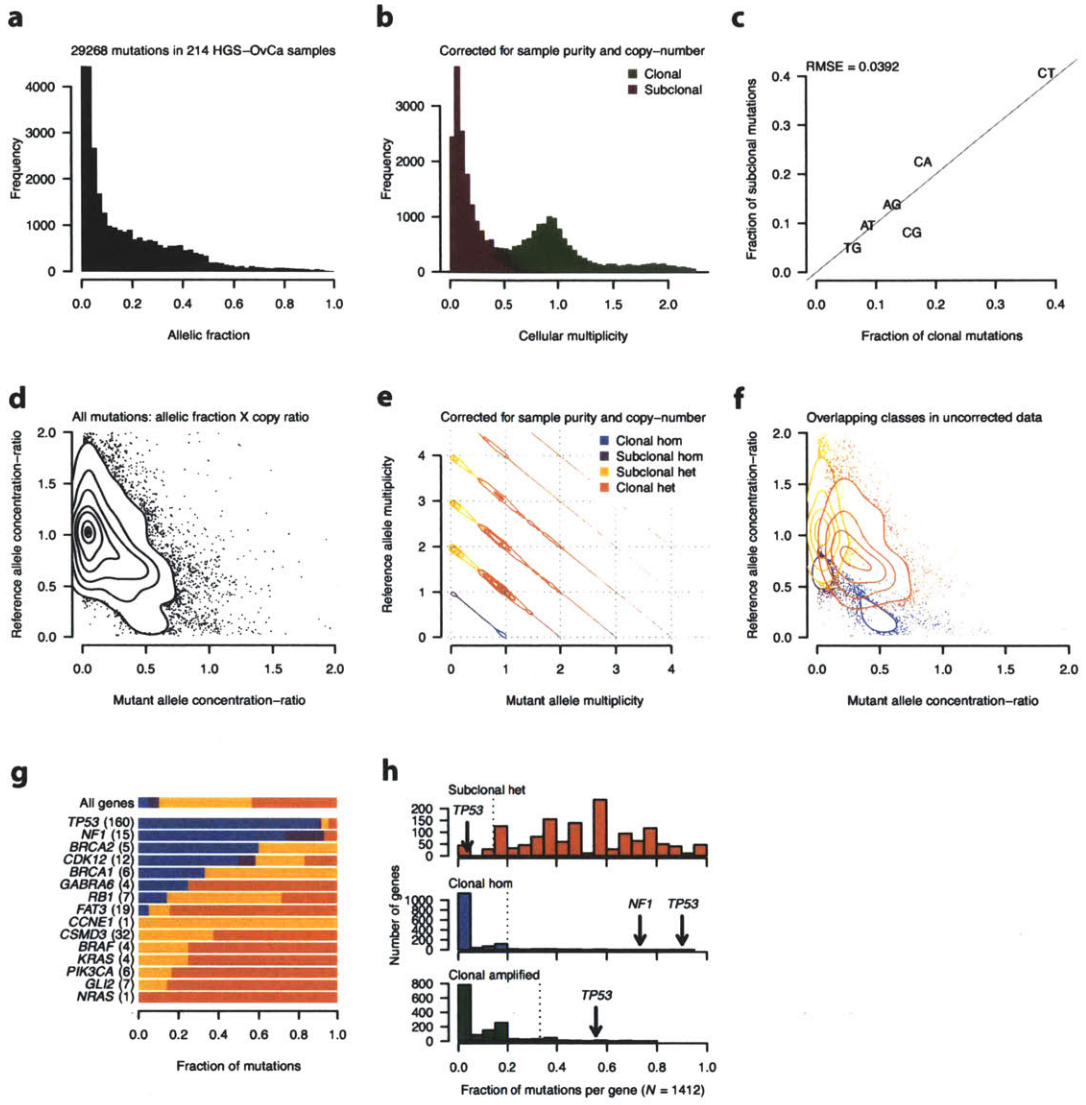

a
29268 mutations in 214 HGS–OvCa samples
Frequency
Allelic fraction
b
Corrected for sample purity and copy–number
Clonal
Subclonal
Frequency
Cellular multiplicity
c
RMSE = 0.0392
CT
CA
AG
AT
CG
TG
Fraction of subclonal mutations
Fraction of clonal mutations
d
All mutations: allelic fraction X copy ratio
Reference allele concentration–ratio
Mutant allele concentration–ratio
e
Corrected for sample purity and copy–number
Clonal hom
Subclonal hom
Subclonal het
Clonal het
Reference allele multiplicity
Mutant allele multiplicity
f
Overlapping classes in uncorrected data
Reference allele concentration–ratio
Mutant allele concentration–ratio
g
All genes
TP53 (160)
NF1 (15)
BRCA2 (5)
CDK12 (12)
BRCA1 (6)
GABRA6 (4)
RB1 (7)
FAT3 (19)
CCNE1 (1)
CSMD3 (32)
BRAF (4)
KRAS (4)
PIK3CA (6)
GLI2 (7)
NRAS (1)
Fraction of mutations
h
Subclonal het
TP53
Clonal hom
NF1
TP53
Clonal amplified
TP53
Number of genes
Fraction of mutations per gene (N = 1412)

Figure 4 | Multiplicity analysis of somatic point-mutations detected by whole-exome sequencing of 214 primary HGS-OvCa tumor samples.
a, Histogram of allelic fraction (alternate / total read-count) values for 29,628 somatic point-mutations detected in 214 primary HGS-OvCa samples[26].
b, Allelic fractions for the mutations shown in (**a**) were converted to point estimates of integral allele-counts per cancer cell (cellular multiplicity; *x*-axis) by correcting for sample purity and local copy-numbers (Methods eq. 11). Subclonal mutations were identified using the model defined in equation 10.
c, The fraction of each of the 6 distinguishable nucleotide substitutions for clonal vs. subclonal point-mutations. The solid grey line indicates $y=x$. RMSE: root mean squared error.
d, Empirical density estimate of allelic concentration-ratios, which are obtained by multiplication of the allelic fraction by the copy-ratio at that locus.
e, Density estimate of allelic multiplicity estimates, as in (**b**), for reference vs. mutant allele. Mutations were classified into the four indicated categories according to their mutant and reference allele multiplicity.
f, The density estimates of allelic concentration-ratios are shown for each of the four mutation classes in **b** are shown superimposed.
g, Mutation classification profiles of genes identified as significantly recurrent in HGS-OvCa, as well as several COSMIC genes with previously observed mutations in these data [26]. Note that only individual point mutations were considered here; the possibility of recessive inactivation via multiple events (compound heterozygosity) was not considered. Color code as in **e**.
h, Histograms of gene classification fractions for 1412 genes having at least 5 mutations. Dashed vertical lines denote the 5^{th} (top) and 95^{th} (other) percentiles of each distribution. No mutations occurring at multiplicity > 1 were observed in *NF1* (not shown).

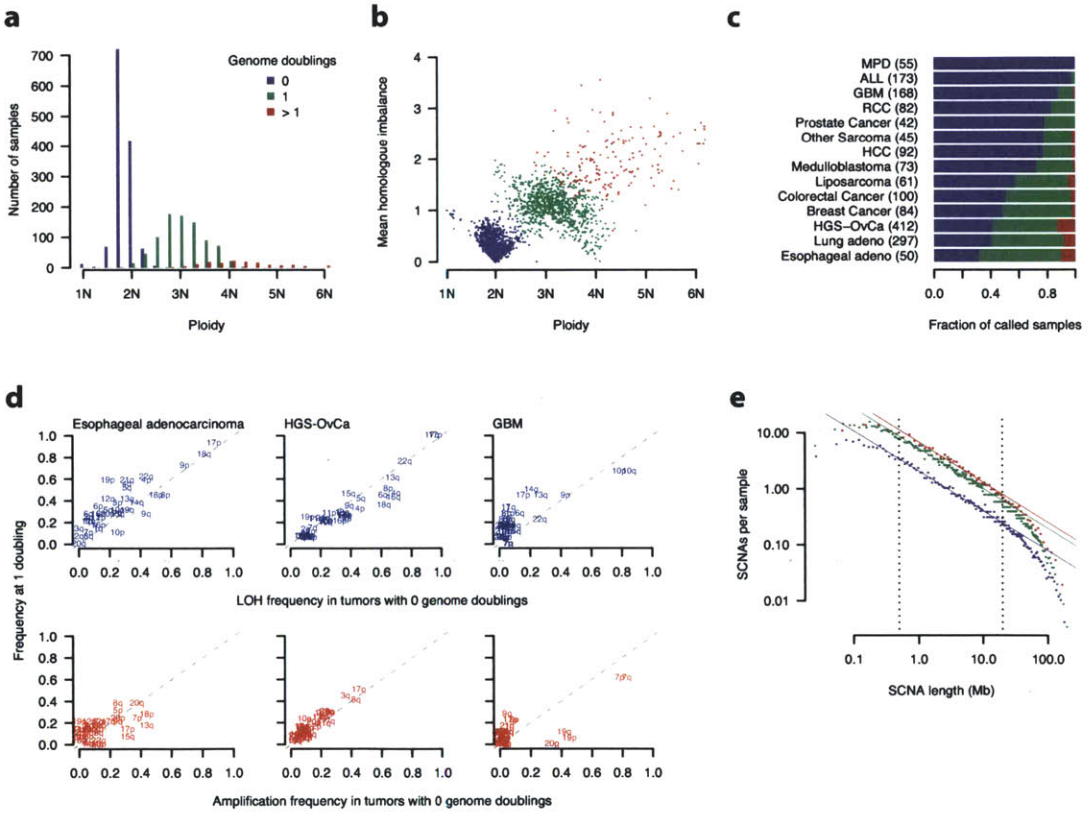

a
Number of samples
Genome doublings
0
1
> 1
Ploidy
b
Mean homologous imbalance
Ploidy
c
MPD (55)
ALL (173)
GBM (168)
RCC (82)
Prostate Cancer (42)
Other Sarcoma (45)
HCC (92)
Medulloblastoma (73)
Liposarcoma (61)
Colorectal Cancer (100)
Breast Cancer (84)
HGS–OvCa (412)
Lung adeno (297)
Esophageal adeno (50)
Fraction of called samples
d
Esophageal adenocarcinoma
HGS-OvCa
GBM
Frequency at 1 doubling
LOH frequency in tumors with 0 genome doublings
Amplification frequency in tumors with 0 genome doublings
e
SCNAs per sample
SCNA length (Mb)

Figure 5 | Incidence and timing of whole genome doubling events in primary cancers.
a, b, Ploidy estimates were obtained from ABSOLUTE. Mean allelic imbalance was calculated as the average difference in absolute allelic copy numbers at every position in the genome. Genome doubling status was inferred from absolute allelic copy-numbers (Methods, Supplemental figure 7).
c, MPD – myeloproliferative disease, ALL – acute lymphoblastic leukemia, GBM - Glioblastoma multiforme, RCC - renal cell carcinoma, HCC - hepatocellular carcinoma, HGS-OvCa - high-grade serous ovarian carcinoma.
d, LOH (loss of heterozygosity) was defined as 0 allelic copies. Amplification was defined as > 1 allelic copy for samples with 0 genome doublings, and as > 2 allelic copies for those with 1 genome doubling. Calls were made based on the modal allelic copy numbers of each chromosome arm. Dashed lines indicate $y=x$.
e, SCNAs, defined as regions differing from the modal absolute copy number of each sample, were binned at adaptive resolution to maintain 250 SCNAs per bin, and renormalized by bin length. The value in each bin was further divided by the number of tumor samples in each genome doubling class, indicated by color as in **a**. Linear regression models were fit independently for each class; solid lines indicate the resulting fits. Only SCNAs $0.5 < x < 20$ Mb were used in the regression (dotted vertical lines)

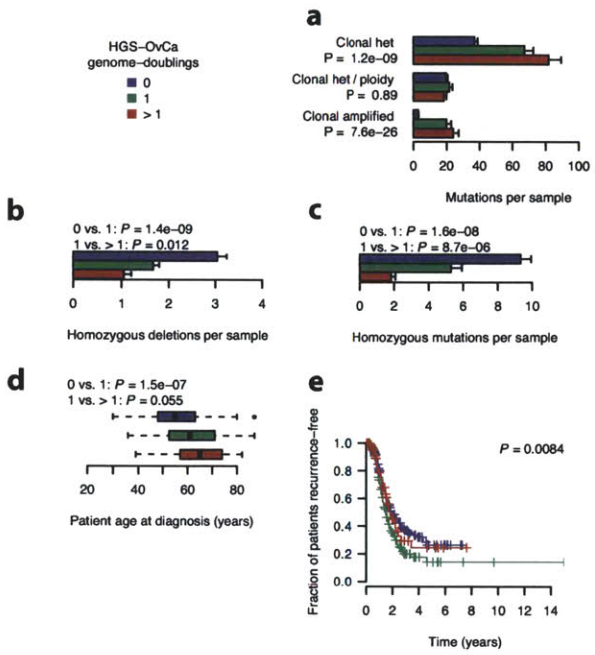

a
HGS–OvCa
genome–doublings
0
1
> 1
Clonal het
P = 1.2e–09
Clonal het / ploidy
P = 0.89
Clonal amplified
P = 7.6e–26
0 20 40 60 80 100
Mutations per sample
b
0 vs. 1: P = 1.4e–09
1 vs. > 1: P = 0.012
0 1 2 3 4
Homozygous deletions per sample
c
0 vs. 1: P = 1.6e–08
1 vs. > 1: P = 8.7e–06
0 2 4 6 8 10
Homozygous mutations per sample
d
0 vs. 1: P = 1.5e–07
1 vs. > 1: P = 0.055
20 40 60 80
Patient age at diagnosis (years)
e
P = 0.0084
Fraction of patients recurrence–free
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10 12 14
Time (years)

Figure 6 | Genetic and clinical associations with genome doubling in primary HGS-OvCa samples.
a-e, Colors correspond to putative genome doubling status, as indicated.
a-c, Number of mutations in indicated classes as a function of genome doublings. *P*-values were calculated with the two-sided Wilcoxin rank-sum test comparing samples with 0 and 1 genome doublings. Error bars indicate standard errors of the means.
d, *P*-values were calculated with the two-sided Wilcoxin rank-sum test.
e, *P*-values were calculated using the log-rank test.

Methods

Inference of purity, ploidy, and absolute somatic copy-numbers

ABSOLUTE is based on the assumption that in many specimens, the cancer cell-fraction is grossly monogenomic in terms of copy-number alterations, i.e. most of the cancer cells share a similar copy-number profile. Although clearly, heterogeneity of cancer-cell populations presents a challenge to such analysis, both cytological Mitelman, 2000 #749 and genomic data Albertson, 2003 #821 have supported this assumption, as have reports of similar SCNA profiles obtained from paired primary and metastatic lesions Kuukasjarvi, 1997 #916, Liu, 2009 #812, Campbell, 2010 #867, Navin, 2011 #863. Recent sequencing of single cancer cells has directly demonstrated the presence of cancer-cell populations with nearly identical copy-profiles Navin, 2011 #863.

Additional statistical methods have been reported for integral copy-number interpretation from SNP-array data without experimental ploidy determination Attiyeh, 2009 #767, Greenman, 2010 #759. These methods did not attempt to model tumor purity, however, making them best suited for analysis of cell lines. In addition, several methods have appeared for joint estimation of tumor purity and ploidy without using prior karyotype data, which forced them to either limit the range of possible tumor ploidy values (1.8 - 4.6)N Van Loo, 2010 #818, or else leave the data up to multiple interpretations Popova, 2009 #913, Yau, 2010 #912. As a comparison, the concordance of ploidy estimates from a related method Van Loo, 2010 #818 with the FACS results was 69% compared to 87% for ABSOLUTE; purity of the cell-line mixtures was systematically underestimated, (mean difference of 15%, Supplemental figure 4).

Homologue-specific copy ratios (HSCRs) derived from analysis of microarray or sequencing data are the required input for ABSOLUTE. HSCR values are obtained by analysis of allele-specific copy ratios at polymorphic loci in order to infer germline genotypes, phase of imbalanced heterozygous markers, and copy ratio of both homologues in the cancer clone. The use of HSCRs reduces the ambiguity of observed copy-profiles. For example, the total copy ratio profile of a sample without SCNAs would be equivalent for ploidy values of 1,2,3, etc., however the HSCR profile would rule out odd ploidy values, since these would not be consistent with equal allelic copy-numbers. In order to accurately perform these inferences without a paired-normal sample, we employed a novel method, HAPSEG (S.L.C. et al., manuscript submitted). HAPSEG currently uses an error model tailored to Affymetrix SNP arrays, developed using thousands of array samples, to perform accurate and robust HSCR estimation.

The likelihood of a proposed purity/ploidy solution is obtained by modeling the HSCRs with a Gaussian mixture model, with components centered at the discrete concentration-ratios implied by eq. 1 (eq. (3) below). The model also supports a small fraction of subclonal events which are not restricted to the discrete levels. Candidate solutions are identified by searching for local optima of this likelihood over a large range of purity and ploidy values. This results in a discrete set of candidate solutions with corresponding peak-fit likelihoods (eq. 1, Fig. 1d,h).

Additional information is generally required in order to reliably select the correct tumor purity and ploidy solution from this set of candidates. In a given tumor sample, several combinations of theoretically possible purity, ploidy, and copy number values may map to equivalent copy ratios (fig 1c,h). Furthermore, the presence of subclonal SCNAs may result in a spuriously high ploidy solution with an implausible karyotype receiving a greater peak-fit likelihood by over-discretizing the copy profile, allowing their assignment to

integral copy-levels (fig 1h,i,j).

Previous techniques have dealt with this issue by presenting multiple interpretations of the sample for manual inspection Popova, 2009 #913,Yau, 2010 #912, or by restricting the range of malignant ploidy values Van Loo, 2010 #818. ABSOLUTE models common cancer karyotypes by grouping tumor sets according to similarities in their absolute allelic copy-number profiles (Supplemental figure 1; methods). These models are constructed directly from the tumor data in a boot-strapping fashion, whereby a subset of tumors with relatively unambiguous profiles (eg., e.g. due to high purity values and lack of subclonal SCNAs) is used initialize the models, iteratively allowing more tumors to be called, etc. Previous cytogenetic characterizations of human cancer were used to guide this process Storchova, 2008 #763. These models enable calculation of a karyotype likelihood, for each candidate purity / ploidy solution, reflecting the similarity of the corresponding karyotype to models associated with the specified disease of the input tumor sample (methods, eq. 7).

Integration of the peak-fit and karyotype likelihoods allows for robust and unambiguous identification of the correct purity and ploidy values in many tumor samples (fig 1d,h). The selection of a solution implying a less common karyotype requires greater evidence from the peak-fit of the copy profile. This ensures that the simplest (most common) karyotype adequately explaining the data is preferred over more complex solutions, while preserving the flexibility to identify unexpected karyotypes given sufficient evidence from the peak-fit. Indeed, several unusual karyotypes, including near-haploid (< 1.2N) and hyperaneuploid (> 6N) genomes, were identified using ABSOLUTE (Supplemental figure 2).

Identification and evaluation of candidate tumor purity and ploidy values

We describe identification of candidate tumor purity and ploidy values and calculation of their peak-fit log-likelihood scores using a probabilistic model. These scores quantify the evidence for each solution contributed by explanation of the observed allelic copy-ratios as integral SCNAs. These computations are independent for each sample. The input data consist of N segmented allelic copy-ratios x_i, $i \in \{1, \ldots, N\}$. Each of these is observed with standard error σ_i, and corresponds to a genomic fraction denoted w_i. Each of the x_i is assumed to have arisen from either one of Q integral copy-number states: $\mathbf{Q} = \{0, 1, \ldots, Q-1\}$, or an additional state Z corresponding to sub-clonal copy-number. We refer to the collection of possible copy-states as $\mathbf{S} = \mathbf{Q} \cup Z$. We define $Q+1$ indicators $\mathbf{s}$ for the copy-state of each segment, with $p(s_i)$ representing the probability of segment i having been generated from state $s \in \mathbf{S}$. The integral copy-states of $\mathbf{S}$ are indexed by $q \in \mathbf{Q}$; the non-integral state is denoted by z.

The expected copy-ratio corresponding to each integral allelic copy-number $q(x)$ in a tumor sample is given by:

$$\mu_q = 2[\alpha/D]q(x) + [2(1-\alpha)/D] \tag{2}$$

where D is related to tumor purity and ploidy (α and τ) (Figure 1; the factor of 2 arises because allelic copy-ratios are relative to haploid control concentrations, as opposed to the diploid values assumed by eqn. 1.)

The observed x_i are modeled with a mixture of Q Gaussian components located at $\boldsymbol{\mu} = \{\mu_{q \in \mathbf{Q}}\}$ representing integral copy-states $\mathbf{Q}$ and an additional uniform component Z. The mixture Z allows segments to be assigned non-integral copy values so that occasional subclonal alterations or artifacts do not dramatically

impact the likelihood.

$$
\begin{aligned}
s_i &\sim \text{Multinom}\left(p\left(s_i|w_i,\boldsymbol{\theta}\right)\right) \\
x_i &= \begin{cases} \mu_q + \epsilon_i & \text{if } s_i \in \mathbf{Q} \\ u & \text{if } s_i = Z \end{cases} \\
\epsilon_i &\sim \mathcal{N}\left(0, \sigma_i^2 + \sigma_H^2\right) \\
u &\sim \mathcal{U}(d)
\end{aligned} \tag{3}
$$

$\mathcal{N}$ and $\mathcal{U}$ denote the normal and uniform densities, respectively. The free parameter σ_H represents sample-level noise in excess of the HSCR standard-erorr σ_i, which might represent a moderate number of related clones in the malignant cell population, ongoing genomic instability, or excessive noise due to faulty experimental conditions. The mixture weights $\boldsymbol{\theta} = \{\theta_{s \in \mathbf{S}}\}$ specify the expected genomic fraction allocated to each copy-state. The parameter d represents the domain of the uniform density, corresponding to the range of plausible copy-ratio values (we used $d = 7$).

Some complication arises due to the fact that the data consist of copy-ratios calculated from a segmentation of the genome. For consistent interpretation, the mixture weights $P(s_i|w_i,\boldsymbol{\theta})$ must be calculated for each segment separately, taking into account the variable genomic fraction w_i. This is accomplished by constraining the canonical averages of genomic mass allocated to each copy-state to match those specified by $\boldsymbol{\theta}$:

$$
\forall_{s \in \mathbf{S}}, \left\langle \sum_{i=1}^{N} s_i w_i \right\rangle_{\mathcal{C}} = \theta_s,
$$

where $\langle \cdot \rangle_{\mathcal{C}}$ denotes the average over all configurations $\{s_i\}$, weighted by the function $\mathcal{C} = P(s_i|w_i,\boldsymbol{\lambda})$. This density corresponds to the maximum entropy distribution over $\mathbf{s}$ subject to these constraints:

$$
P\left(s_i|w_i,\boldsymbol{\lambda}\right) = \frac{e^{-\lambda_s s^{\#} w_i}}{\sum_{k \in \mathbf{S}} e^{-\lambda_k k^{\#} w_i}},
$$

where $s^{\#}$ indicates the order of state s in a sequence of copy-states, beginning with 0. Values of the Q Lagrange multipliers $\boldsymbol{\lambda}$ are determined via Nelder-Mead optimization of L_2 loss:

$$
\boldsymbol{\lambda} = \underset{\boldsymbol{\lambda}}{\text{argmin}} \left(\sum_{s \in \mathbf{S}} \left[\left(\sum_{i=1}^{N} \left[w_i P\left(s_i|w_i,\boldsymbol{\lambda}\right) \right] - \theta_s \right)^2 \right] \right)^{\frac{1}{2}}.
$$

This approximation allows for robustness of the peak-fit score to over-segmentation of the data.

The likelihood of a given segment i can then be calculated as:

$$
\mathcal{L}\left(x_i|\boldsymbol{\mu},\sigma_i,\sigma_H,\boldsymbol{\theta},w_i\right) = \sum_{q \in \mathbf{Q}} \left[P\left(q_i|w_i,\boldsymbol{\lambda}\right) \mathcal{N}\left(x_i|\mu_q, \sigma_i^2 + \sigma_H^2\right) \right] + P\left(z_i|w_i,\boldsymbol{\lambda}\right) \mathcal{U}(d),
$$

and the full log-likelihood of the data is then:

$$\sum_{i=1}^{N} \log \mathcal{L}\left(x_i|\boldsymbol{\mu}, \sigma_i, \sigma_H, \boldsymbol{\theta}, w_i\right). \tag{4}$$

We define the parameterization $b = 2(1-\alpha)$, $\delta_\tau = 2(\alpha/D)$, which determines $\boldsymbol{\mu}$ via eq. (2). Candidate purity and ploidy solutions for a tumor sample are identified by optimization of eq. (4) with respect to b and δ_τ. Calculation of eq. (4) requires estimates of $\boldsymbol{\theta}$ and σ_H, which are not known *a priori*. We make an approximation (scale-separation) assuming that locations of the modes of eq. (4) are invariant to moderate fluctuations in these parameters. A provisional likelihood for each x_i may then be calculated by

$$\mathcal{L}_{\mathbf{P}}\left(x_i|\boldsymbol{\mu}, \sigma_i, \sigma\right) = \sum_{q \in \mathbf{Q}} \left[\mathcal{N}\left(x_i|\mu_q, \sigma_i^2 + \sigma^2\right)\right] + \mathcal{U}(d).$$

Candidate purity and ploidy solutions are then identified by optimization of

$$\sum_{i=1}^{N} \log \mathcal{L}_{\mathbf{P}}\left(x_i|\boldsymbol{\mu}, \sigma_i, \sigma_P\right),$$

initiated from all points in a regular lattice spanning the domain of b and δ_τ. The parameter σ_P was set to 0.01 for this study. We verified that the above approximation identified modes equivalent to those obtained through a full Metropolis-Hastings Markov chain Monte Carlo (MCMC) simulation (data not shown). The approximation allows for much simpler computations to be used.

The peak-fit score for each solution is calculated after optimization of $\sigma_H \in [0, 0.04]$:

$$\hat{\sigma}_H = \underset{\sigma_H}{\operatorname{argmax}} \sum_{i=1}^{N} \log \mathcal{L}\left(x_i|\hat{\boldsymbol{\mu}}, \sigma_i, \sigma_H, \boldsymbol{\theta}, w_i\right),$$

with the elements of $\boldsymbol{\theta}$ calculated for each value of σ_H by:

$$\hat{\theta}_q = \sum_{i=1}^{N} w_i \frac{\mathcal{N}\left(x_i|\hat{\mu}_q, \sigma_i^2 + \sigma_H^2\right)}{\mathcal{L}_{\mathbf{P}}\left(x_i|\hat{\boldsymbol{\mu}}, \sigma_i, \sigma_H\right)}; \; \hat{\theta}_z = \sum_{i=1}^{N} w_i \frac{\mathcal{U}(d)}{\mathcal{L}_{\mathbf{P}}\left(x_i|\hat{\boldsymbol{\mu}}, \sigma_i, \sigma_H\right)}.$$

The final calculation of the peak-fit log-likelihood for each mode is obtained by inserting $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}_H$ into eq. (4). Estimates of the copy-state indicators for each segment are calculated as:

$$\hat{\mathbf{q}}_i = P\left(q_i|w_i, \boldsymbol{\lambda}\right) \frac{\mathcal{N}\left(x_i|\hat{\mu}_q, \sigma_i^2 + \hat{\sigma}_H^2\right)}{\mathcal{L}\left(x_i|\hat{\boldsymbol{\mu}}, \sigma_i, \hat{\sigma}_H, \hat{\boldsymbol{\theta}}, w_i\right)},$$

$$\hat{z}_i = P\left(z_i|w_i, \boldsymbol{\lambda}\right) \frac{\mathcal{U}(d)}{\mathcal{L}\left(x_i|\hat{\boldsymbol{\mu}}, \sigma_i, \hat{\sigma}_H, \hat{\boldsymbol{\theta}}, w_i\right)}.$$

Note that each $\hat{\mathbf{q}}_i$ is a vector representing the posterior probability of each $Q \in \mathbf{Q}$ integral copy-states, corresponding to the copy-ratios (locations) $\boldsymbol{\mu}$.

Genome-wide absolute copy-profiles are over-determined with respect to DNA ploidy estimates. An

alternate estimate of ploidy may be calculated as twice the expected absolute allelic copy-number over the genome:

$$\hat{\tau}_g = 2\sum_{i=1}^{N}\left(w_i \sum_{q=0}^{Q-1} q\hat{\mathbf{q}}_{iq}\right). \tag{5}$$

Validation of absolute copy-number predictions was performed by consideration of the expected absolute copy-number, genome-wide. By definition, this quantity ($\hat{\tau}_g$) is an alternate estimate of cancer ploidy. Because $\hat{\tau}_g$ is a weighted average over discrete states identified in the data, it is robust to experimental fluctuations that shift and scale the copy-profile slightly. Validation of this ploidy estimate using FACS data (fig. 2a), demonstrates that the average absolute copy-numbers are correct genome-wide. We adopted this measure of cancer-genome ploidy throughout.

Estimation of ploidy by $\hat{\tau}$ relies on accurate calibration of the input copy-ratios. In contrast, ploidy estimates derived the above expectation over discretized states are more robust to systematic fluctuations (e.g. microarray background, input DNA quantity). For this reason, ABSOLUTE uses $\hat{\tau}_g$ as the final ploidy estimate for each mode. We note that these estimates are generally close to the values of $\hat{\tau}$ obtained by optimization of the peak-fit likelihood (data not shown). Note that, for this computation, the $\hat{\mathbf{q}}_{ij}$ are calculated with $\hat{\theta}_z = 0$, so that the above expectation is over integral states only.

Two additional transformations of the copy-state locations $\boldsymbol{\mu}$ are used when copy-ratios are measured using microarrays. The first of these accounts for the effect of attenuation with an isothermal adsorbtion model [Chiang 2009]:

$$g(x) = \frac{x(1+\hat{\phi})}{(1+x\hat{\phi})},$$

where the value $\hat{\phi}$ parameterizes the attenuation response in a given sample, and is estimated via HAPSEG. The second transformation is a variance stabilization for microarray data adapted from REF Huber:

$$h(x) = \sinh^{-1}\left(\frac{xe^{\sigma_\eta^2}-1}{\sigma_\epsilon}\right),$$

where σ_η and σ_ϵ represent multiplicative and additive noise scales for each microarray, estimated by HAPSEG. This transformation is applied to the marker-level data during estimation of the x_i values, after which their distribution is approximately normal. The normal mixture component specified in (3) then becomes $h(x_i) = h(g(\mu_q)) + \epsilon_i$, and the corresponding likelihood calculations are performed under these transformations.

Karyotype models

In a given tumor sample, several combinations of theoretically possible purity, ploidy, and copy number values may map to equivalent copy ratios (fig 1c,h). Furthermore, the presence of subclonal SCNAs may result in a spuriously high ploidy solution with an implausible karyotype receiving a greater peak-fit likelihood by over-discretizing the copy profile, allowing their assignment to integral copy-levels (fig 1h,i,j).

Previous techniques have dealt with this issue by presenting multiple interpretations of the sample for manual inspection Popova, 2009 #913,Yau, 2010 #912, or by restricting the range of malignant ploidy values Van Loo, 2010 #818. ABSOLUTE models common cancer karyotypes by grouping tumor sets according

to similarities in their absolute allelic copy-number profiles (Supplemental figure 1). These models are constructed directly from the tumor data in a boot-strapping fashion, whereby a subset of tumors with relatively unambiguous profiles (eg., due to high purity values) is used initialize the models, iteratively allowing more tumors to be called, etc. Previous cytogenetic characterizations of human cancer were used to guide this process Storchova, 2008 #763. These models enable calculation of a *karyotype likelihood*, for each candidate purity / ploidy solution, reflecting the similarity of the corresponding karyotype to models associated with the specified disease of the input tumor sample (methods, eq. 7). Integration of the peak-fit and karyotype likelihoods allows for robust and unambiguous identification of the correct purity and ploidy values in many tumor samples (fig 1d,h). The selection of a solution implying a less common karyotype requires greater evidence from the peak-fit of the copy profile.

Prior knowledge of karyotypes characteristic of a particular disease is summarized as a mixture of K multivariate multinomial distributions over the integral allelic copy-states $\mathbf{Q} = [0, 7]$ of each chromosome arm. For a given candidate purity and ploidy solution, the corresponding segmental copy-state indicators for each segment i, $\hat{\mathbf{q}}_{ij}$, are summarized into estimates of the J arm-level allelic copy-numbers, denoted $\hat{\mathbf{C}}$. The karyotype log-likelihood score is calculated as:

$$\mathcal{L}_{\mathbf{K}}\left(\hat{\mathbf{C}}|\mathbf{K}\right) = \log \sum_{i=1}^{K} \left[w_i \prod_{j=1}^{J} \prod_{q \in \mathbf{Q}} \mathbf{K}_{ijq}^{\hat{\mathbf{C}}_{jq}} \right], \tag{6}$$

where w_i denotes the weight of each mixture component. The karyotype models $\mathbf{K}_i$ are $J \times Q$ SCNA probability matrices obtained by clustering arm-level allelic copy-states of modeled copy-profiles using the standard expectation-maximization (EM) algorithm {Dempseter, Laird} for multinomial mixtures. This calculation identifies groups of disease subtypes with similar genomic copy profiles (Supplementary figure 1). Note that copy-states for both homologues of each arm are modeled ($J = 78$). The number of clusters K for each disease was chosen by minimizing the Bayesian information criterion (BIC) complexity penalty: $-2\hat{L}_k + KJ\log(N)$, where $\hat{L}_k$ indicates the sum of $\mathcal{L}_{\mathbf{K}}$ values over the N input samples, computed using K clusters. In order to avoid local minima, the EM algorithm was run 25 times for each value of $K \in [2, 8]$ with randomized starting points and the best model was retained.

The models were constructed in a semi-automated fashion by seeding with relatively unambiguous copy-profiles. As tumors were added, the use of recurrent karyotypes clearly identified the correct solutions of additional samples, etc. For example, LOH of chr17 occurs in nearly 100% of ovarian carcinoma samples [TCGA OV], allowing the model to learn that solutions implying LOH of chr17 are likely to be correct. The qualitative similarity of the resulting ploidy distributions with those obtained from cytogenetic data [Storchova 2008] (Supplemental figure 7a) bolsters the veracity of the models created in this exercise. In total, models for 14 disease types were created. Diseases with fewer than 40 samples called by ABSOLUTE were omitted from this procedure.

Limitations of joint purity/ploidy inference from copy-profile data

Accurate calibration of both the peak-fit and karyotype models to the true level of certainty implied by the data would allow for assignment of probabilities to each candidate solution; we do not believe that the models we have presented here sufficiently capture the complexity of cancer genomes to allow for such

interpretations. We therefore expect that manual review is currently an important part of joint purity/ploidy inference using copy-number data, as was concluded in related studies {Popova, 2009 913},{Yau, 2010 912}.

Even with skilled manual review, analysis with ABSOLUTE may occasionally result in incorrect interpretations, for example genome-doubling without subsequent detectable gains or losses would result in a solution implying half the true ploidy value, which in many cases will correspond to a plausible karyotype model (fig 5d). Alternately, when multiple subclonal SCNAs appear close to the midpoints between adjacent clonal peaks, a solution implying twice the true ploidy may be chosen. We note that samples with no reliably detected SCNAs could not be called in our framework (ploidy 2N or 4N; purity cannot be determined). Such samples were therefore excluded from downstream analysis (see below). Estimation of inference error-rate requires external measurement of sample ploidy. Further validation experiments in diverse tumor types will help to clarify any disease specific caveats.

Fortunately, many cancer samples demonstrated marked SCNAs before and after genome doubling, enabling unambiguous inference for many samples. This aspect of cancer genome evolution was noted previously in breast cancer cytogenetic data {Dutrillaux, 1991 905}. We note that manual review of ABSOLUTE results was performed prior to generation of the FACS validation data or analysis of the NCI60 cell-line ploidy estimates (fig 2a,b).

Identification of samples refractory to purity/ploidy inference

In order to facilitate rapid analysis of many cancer samples, ABSOLUTE attempts to automatically identify copy profiles that cannot be reliably called and to classify them into informative failure categories (fig 2e), which in this study were defined by the following criteria. Define $\hat{\mathbf{m}}$ as the sorted vector of posterior genome-wide copy-state allocations ($\hat{\boldsymbol{\theta}}$), so that $\hat{m}_1$ represents the greatest element of $\hat{\boldsymbol{\theta}}$ (the modal copy-state). This vector is constructed with θ_0 replaced by 0 if $\theta_0 < 0.01$ and $b < 0.15$, so that germline copy-number variants (CNVs) or regions of inherited homozygosity are not confused with small SCNAs implying very pure samples. The categories are then:

1. non-aberrant: $\hat{m}_3 < 0.001$, $\hat{m}_2 < 0.005$, $\hat{\sigma}_H < 0.02$
2. insufficient purity: $\hat{m}_3 < 0.001$, $\hat{m}_2 < 0.005$, $\hat{\sigma}_H \geq 0.02$
3. polygenomic: $\hat{\theta}_z > 0.2$.

These criteria are applied to the top-ranked mode for each sample (combined peak-fit and karyotype scores). Several examples of each outcome are shown in Supplemental figure 5. The above designations led to reasonably good concordance of automated calls with those obtained after manual review. This is an area for potential methodological improvement, likely to be aided by additional validation data.

Cancer cell-line DNA mixing experiment

DNA extracted from two cancer cell lines (HCC38, HCC1143) was mixed with DNA from matched B-lymphocyte cell lines (HCC38BL, HCC1143BL) in various proportions, and hybridized to Affymetrix 250 K Sty SNP arrays. Stock DNA aliquots were created for each cell-line by normalization of DNA concentration to 50ng/μl. Mixing of cancer and matched B-lymphocyte DNA to each required mixing fraction was done by volume.

FACS analysis of primary tumor samples

Formalin-fixed and paraffin-embedded blocks from ovarian serous carcinoma cases were available from tumor-sections corresponding to the frozen blocks from which DNA-aliquots were obtained for SNP-array hybridization. Multiple curls containing at least 70% tumor cell nuclei were cut to an aggregate thickness of 150 microns. Sections were disaggregated and labeled with propidium iodide (DNA stain). FACS was performed to determine ploidy.

Determination of tumor purity via pathology review

Frozen ovarian serous cystadenocarcinoma specimens were collected from multiple hospital tissue banks and maintained frozen in liquid nitrogen vapors. A tissue portion was created with two flanking H&E slides (arbitrarily named top and bottom) as follows: tissues were mounted in optimal cutting temperature media (OCT) and brought to -20°C. A 4μm frozen section (top slide) was cut with a cryostat (Leica Microsystems, Wetzlar, Germany). A specimen for molecular extraction was created by shaving 100 mg of tumor tissue from the tissue face with a scalpel, then a second 4μm frozen section was cut (bottom slide). An H&E stain was conducted on both slide tissue sections using an Autostainer XL with integrated coverslipper (Leica). Digital images of slides were created at 20x resolution using a Scanscope XT (Aperio, Vista, CA, USA). Board-certified pathologists conducted the pathology review remotely via ImageScope software (Aperio). Pathologists initially reviewed each slide at low magnification to determine low power microscopic morphology, then increased magnification to 20x and reviewed 10 representative high power fields on each slide. Diagnosis of ovarian serous cystadenocarcinoma was verified, and tumor purity was determined as the proportion of tumor nuclei present compared to the total nuclei present on the slide. The tumor purity of the extracted specimen was calculated as the average purity score of the top and bottom slides. Quality control included a random review of 10% of slides by a second pathologist to verify consistency of reads.

Leukocyte methylation signature

DNA methylation data for 489 high stage, high grade serous ovarian tumors and eight normal fallopian tube samples was obtained from http://tcga.cancer.gov/dataportal/. In addition, buffy coat samples from two female individuals were obtained. All data were generated with Illumina Infinium HumanMethylation27 arrays, which interrogate 27,578 CpG sites located in proximity to the transcription start sites of 14,475 consensus coding sequencing in the NCBI Database (Genome Build 36). The level of DNA methylation at each probe was summarized with beta values ranging from 0 (unmethylated) to 1 (methylated) **Noushmehr et al [PMID: 20399149]**.

The leukocyte methylation signature was derived as follows. Each probe was ranked by the difference in mean beta value in buffy coat and fallopian tube samples. We retained the 100 top-ranking probes in each direction, designated BC and FT, (buffy coat and fallopian tube enriched, respectively). Let T_{ik} denote the beta value for probe k in tumor sample i. Let B_k denote the average beta value of buffy coat samples for each probe. Let βT_k denote the minimum observed beta value across all tumor samples for the BC probes and the maximum for the FT probes. Denote by f_B the fraction of buffy coat components in the sample, then we have the following equation for each probe: $\beta T_{ik} = \beta B_k f_B + \beta T_k(1 - f_B)$. Solving this equation for f_B gives: $f_B = \beta T_{ik} - \beta T_k / \beta B_k - \beta T_k$. The values of f_B for each of the 200 probes in the signature were

calculated and a kernel density estimate was obtained. The leukocyte signature was then calculated as the mode of this density estimate.

Selection of datasets

We analyzed 2445 Affymetrix 250 K Sty SNP samples from a previous pan cancer survey {Beroukhim 2010} containing 3131 cancer samples. Because our processing of the data required use of the Birdseed algorithm {Korn 2008}, external datasets with no diploid PCR controls could not be used. In addition, cancer types with fewer than 20 samples were excluded. In addition, 680 Affymetrix SNP6.0 samples were taken from the TCGA GBM {} and HGS-OvCa {} studies, as well as 30 cell-line samples, bringing to total sample count to 3155. The complete table of cancer samples analyzed is available as Supplementary file 1. The complete table of ABSOLUTE results is available as Supplementary file 2.

Power calculation for somatic mutation detection in cancer tissue samples

We develop a framework for calculation of statistical power for the detection of mutations. Power to detect a variant depends on the allelic fraction f and local depth of coverage n. To calculate power, we model the idealized scenario in which random sequencing errors occur uniformly with rate ϵ. We calculate a minimum number of supporting reads k such that the probability of observing k or more identical non-reference reads due to sequencing error is less than a defined false-positive rate (FPR):

$$k = \min_m |P(m) \leq \text{FPR},$$

where

$$P(m) = 1 - \sum_{i=0}^{m} \text{Binom}\left(i|n, \epsilon/3\right).$$

Variants with $\geq k$ supporting reads are then considered to be detected. We specified the sequencing error rate $\epsilon = 1 \times 10^{-3}$ and FPR $= 5{\times}10^{-7}$ for all computations in this study. Power is then calculated as:

$$\text{Pow}(n, f) = 1 - \sum_{i=0}^{k-1} \text{Binom}\left(i|n, f\right) + d\ \text{Binom}\left(k|n, f\right), \tag{7}$$

where

$$d = \frac{\text{FPR} - P(k)}{P(k-1) - P(k)}.$$

We consider the case of detecting clonal somatic variants present at a single copy per cancer cell in cancer-tissue derived DNA samples. Given estimates of purity (α) and local absolute copy-number (q_t), the allelic fraction of such variants is:

$$\delta = \frac{\alpha}{2(1-\alpha) + \alpha q_t}. \tag{8}$$

Power is calculated in such cases as $\text{Pow}(n, \delta)$.

In order to simplify the relationship between power and tumor purity/ploidy for presentation in fig. 3, we considered the detection power of the expected locus, over the genome-wide copy average. Power as such is determined by the sample *allelic index* $\delta_\tau = \alpha/D$, which is solely a function of tumor purity/ploidy (eqn.

1). Expected power is obtained by using allelic fraction $f = \delta_\tau$ in eq. (7). This calculation differs only in the substitution of expected genomic copy-number, i.e. ploidy (τ), for the local copy-number q_t in eq: (8).

Power for expected subclonal variants present in fraction s_f of cancer cells is given by $\text{Pow}(n, s_f\delta_\tau)$. This calculation was used for Figure 3c,e. Local copy calculations using $\text{Pow}(n, s_f\delta)$ were used for Figure 3f and Supplemental figure 9.

Detection of somatic point mutations in ovarian carcinoma

We analyzed whole-exome hybrid capture Illumina sequencing (WES)Gnirke, 2009 #848 data from 214 ovarian carcinoma tumor-normal pairs previously analyzed by the TCGA consortium Bell, 2011 #944. We used the program muTect K. Cibulskis, et al., in preparation We have used a newer version of the program muTect than used in previous analysis of this data Bell, 2011 #944. The primary improvement in the new version is a reduction in the prior that somatic mutations be at an allelic fraction of 0.5, allowing greater sensitive at low allelic-fraction mutations, such as clonal events in impure samples, or to subclonal mutations. This procedure resulted in 29,268 somatic mutations.

Inference of point mutation multiplicity

We develop a probabilistic model for inference of the integral multiplicities for both germline and somatic variants, based on knowledge of tumor purity and genome-wide absolute allelic copy-number. Denote the absolute allelic copy-numbers at a mutant locus as q_1 and q_2, with $q_1 \leq q_2$. The possible multiplicities of germline variants are then:

$$\mathbf{g_q} = \{q_1, q_2, q_t\},$$

where $q_t = q_1 + q_2$. Under the assumption that all somatic point-mutations arise uniquely on a single haplotype, the possible multiplicities are:

$$\mathbf{s_q} = \{1, \ldots, q_2\}.$$

Germline mutations are potentially present in both the cancer and normal cell populations, with somatic copy-number alterations affecting the allelic fraction. A heterozygous variant in the germline, with multiplicity g_q in the cancer genome, has allelic fraction:

$$f_{g_q} = \frac{(1-\alpha) + \alpha g_q}{2(1-\alpha) + \alpha q_t}, \tag{9}$$

whereas the allelic fraction of homozygous germline variants is 1 regardless of α. For somatic point mutations, the expected allelic fraction at multiplicity s_q is $f_{s_q} = s_q\delta$, with δ as in eq. (8).

Consider an observed somatic point-mutation of unknown copy $s_q \in \mathbf{s_q}$, observed allelic fraction $\hat{f}$, and with n total reads covering the locus. The complete likelihood of $\hat{f}$ may represented as a mixture of beta distributions corresponding to each element of $\mathbf{s_q}$, plus an additional component S corresponding to subclonal states:

$$\mathcal{L}_\mathbf{m}\left(\hat{f}|n, \mathbf{s_q}, \mathbf{w_q}, w_{s_c}\right) = \sum_{s_q \in \mathbf{s_q}} \left[\frac{w_{s_q}}{q_2}\text{Beta}\left(\hat{f}|nf_{s_q}, n\left(1-f_{s_q}\right)\right)\right] + w_{s_c}S(\hat{f}|n, \lambda), \tag{10}$$

where $w_{s_q} \in \mathbf{w_q}$ specify mixture weights for each state in $\mathbf{s_q}$, and w_{s_c} specifies the subclonal component weight. The subclonal component S is specified by composing a Beta distribution modeling sampling noise with an exponential distribution over subclonal cancer-cell fractions, having a single parameter λ:

$$S(\hat{f}|n,\lambda) = \int_0^1 \text{Beta}\left(\hat{f}|nf, n\left(1-f\right)\right) \text{Exp}(f/\delta|\lambda)\delta^{-1}df$$

Note the change of coordinates in the exponential component using δ; this allows modeling in consistent units of cancer-cell fractions, regardless of tumor purity and local copy-number (note this distribution is renormalized on the unit interval). The probability of a given integral copy-state s_q may then be calculated as:

$$\hat{s}_q = \frac{w_{s_q}}{q_2} \frac{\text{Beta}\left(\hat{f}|nf_{s_q}, n\left(1-f_{s_q}\right)\right)}{\mathcal{L}_{\mathbf{m}}\left(\hat{f}|n, \mathbf{s_q}, \mathbf{w_q}, w_{s_c}\right)}.$$

Similarly, the probability that a given mutation is subclonal is calculated as:

$$\hat{s}_c = w_{s_c} q_2 \frac{S(\hat{f}|n,\lambda)}{\mathcal{L}_{\mathbf{m}}\left(\hat{f}|n, \mathbf{s_q}, \mathbf{w_q}, w_{s_c}\right)}.$$

For the computations in this study, we fixed $\lambda = 25$, w_{s_q} to 0.25, and w_{s_c} to 0.75, which produced a fit to the combined-sample mutation-fraction distribution (fig. 4b). The results presented in fig. 4 were robust to various settings.

We describe a rescaling of somatic variant allele-fractions into point estimates of multiplicity. These estimates may be used to visualize the distribution of somatic variant multiplicities in the cancer genome, along with sampling noise. Consider a mutation with N_a and N_r reads supporting the alternate and reference alleles, respectively. Denote the observed mutant allele-fraction by:

$$\hat{f} = \frac{N_a}{N_a + N_r}.$$

Under the assumption that no systematic bias exists favoring the observation of reads supporting either allele, this quantity estimates the ratio of alternate-allele to the total concentration at this locus:

$$\hat{f} = \frac{C_a}{C_a + C_r} + \epsilon$$

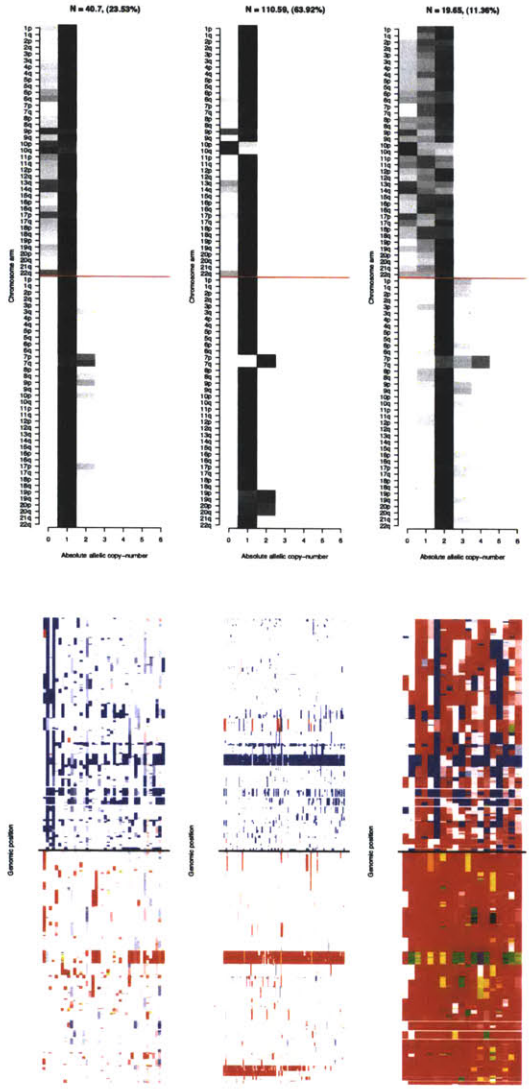
where C_a and C_r denote concentration-ratios of the alternate and reference alleles, respectively, and ϵ represents sampling noise, which is approximately normal for coverage values > 30. Let $\hat{q}_t$ denote the absolute total copy-number at the mutant locus, then $\hat{C} = \hat{b} + \hat{\delta_\tau}\hat{q}_t$ gives the total concentration-ratio. Noting that $C = C_a + C_r$, we can obtain estimates of the variant and reference concentration-ratios as $\hat{C}_a = \hat{f}\hat{C}$, and $\hat{C}_r = \hat{C} - \hat{C}_a$. These estimates can then be rescaled to point estimates of multiplicities for variant and reference alleles:

$$\hat{M}_a = \hat{C}_a/\hat{\delta_\tau},\ \hat{M}_r = \left(\hat{C}_r - \hat{b}\right)/\hat{\delta_\tau} \tag{11}$$

This rescaling projects the reference and variant allelic concentration-ratios onto the lines $y = -x + \hat{q}_t$. This transformation was used to rescale the data in Figure 3b,e.

Simulation of cancer-genome evolution to support genome-doubling inferences

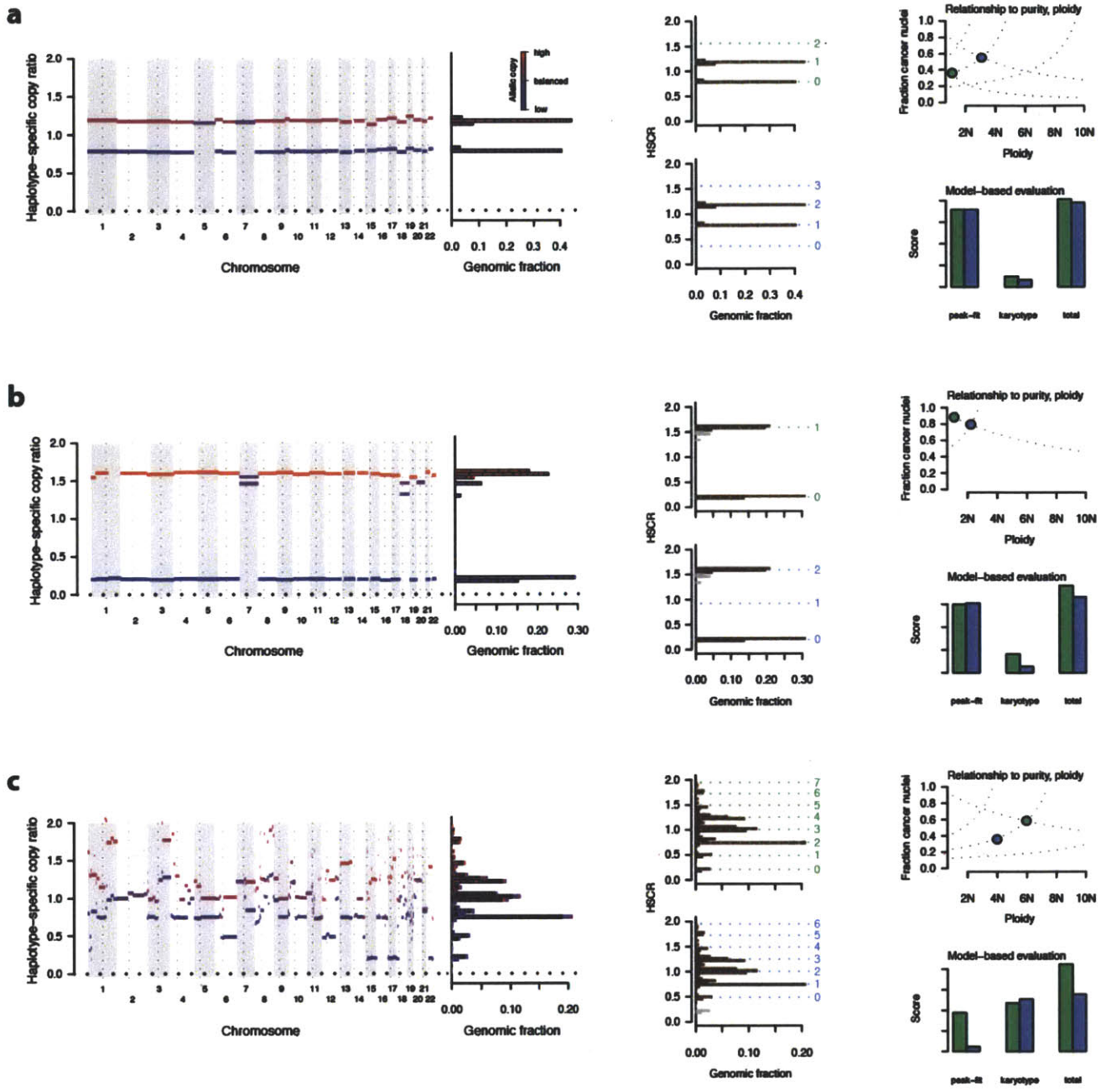
A simple simulation was performed to obtain P-values for the probability that an observed configuration of allelic copy-numbers could be produced from a serial process of independent gains and losses. Genome-wide allelic copy-numbers are summarized at chromosome-arm resolution as integral allelic gains/losses (total of 78 states). We then fix the total number of gains / losses N for the sample, and calculate rates for each arm, which are normalized to probabilities. Simulation of the sample is performed by independently sampling N gains and losses from these probabilities. This was repeated 1000 times for each sample, keeping track of the number of times M that the extent of even high parental-allele copy-number present in the observed sample was attained or exceeded. The P-value is then: $P = M/1000$, if $M > 0$, otherwise $P < 0.001$.



Supplemental Figure 1 | Karyotype models for GBM, constructed by grouping tumor sets according to similarities in their absolute allelic copy-number profiles

(top) multivariate multinomial mixture model with 3 components, representing genomic subtypes.

(bottom) individual GBM samples are shown grouped according to their cluster membership in (top).

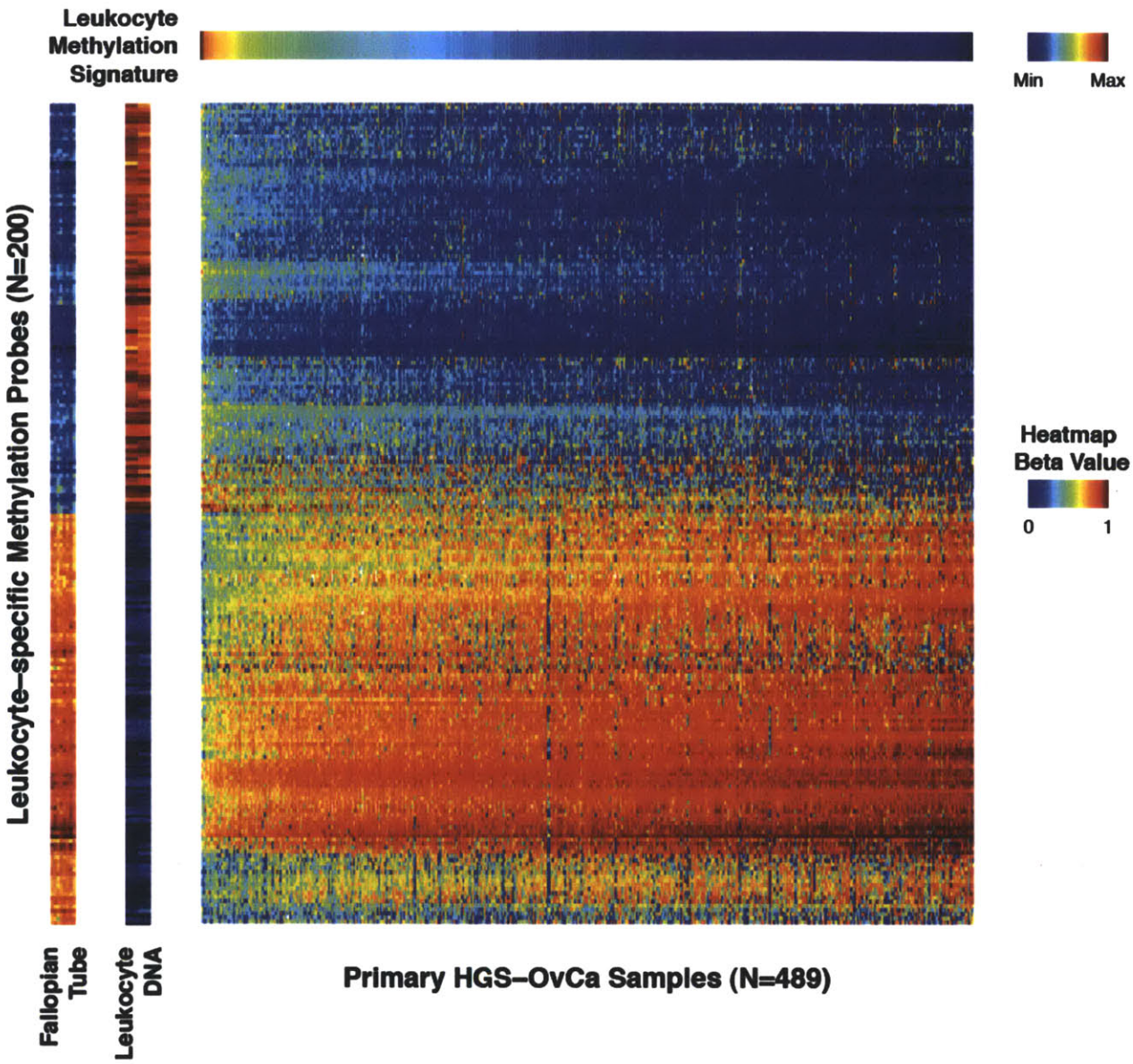


Supplemental Figure 2 | Unusual karyotypes identified using ABSOLUTE.

a-d, Analysis of allelic copy-ratio profiles with ABSOLUTE, as in figure 1.

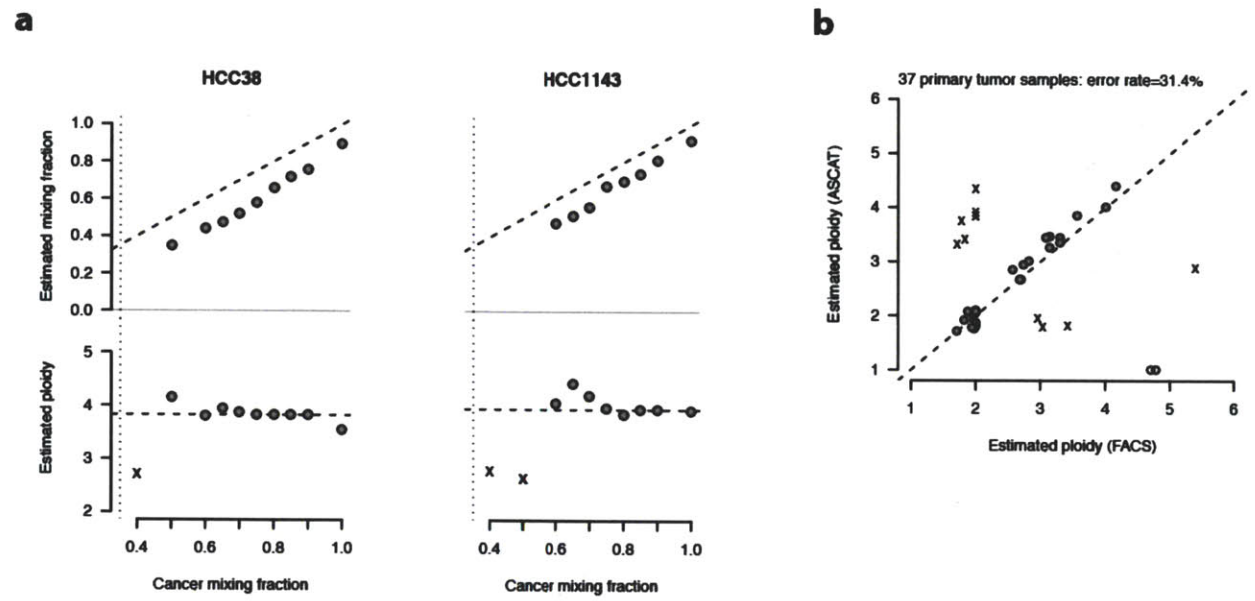
a,b, Near-haploid genomes
c,d, Hyperaneuploid genomes

a, Lung adenocarcinoma sample SM-11ZY. Purity = 0.36, ploidy = 1.12.
b, Glioma sample 'glioma 612'. Purity = 0.88, ploidy = 1.14.
c, HGS-OvCa sample TCGA-25-1320-01A-01D-0452-01. Purity = 0.58, ploidy = 6.03.



Supplemental Figure 3 | Identification of a leukocyte methylation signature in HGS-OvCa samples

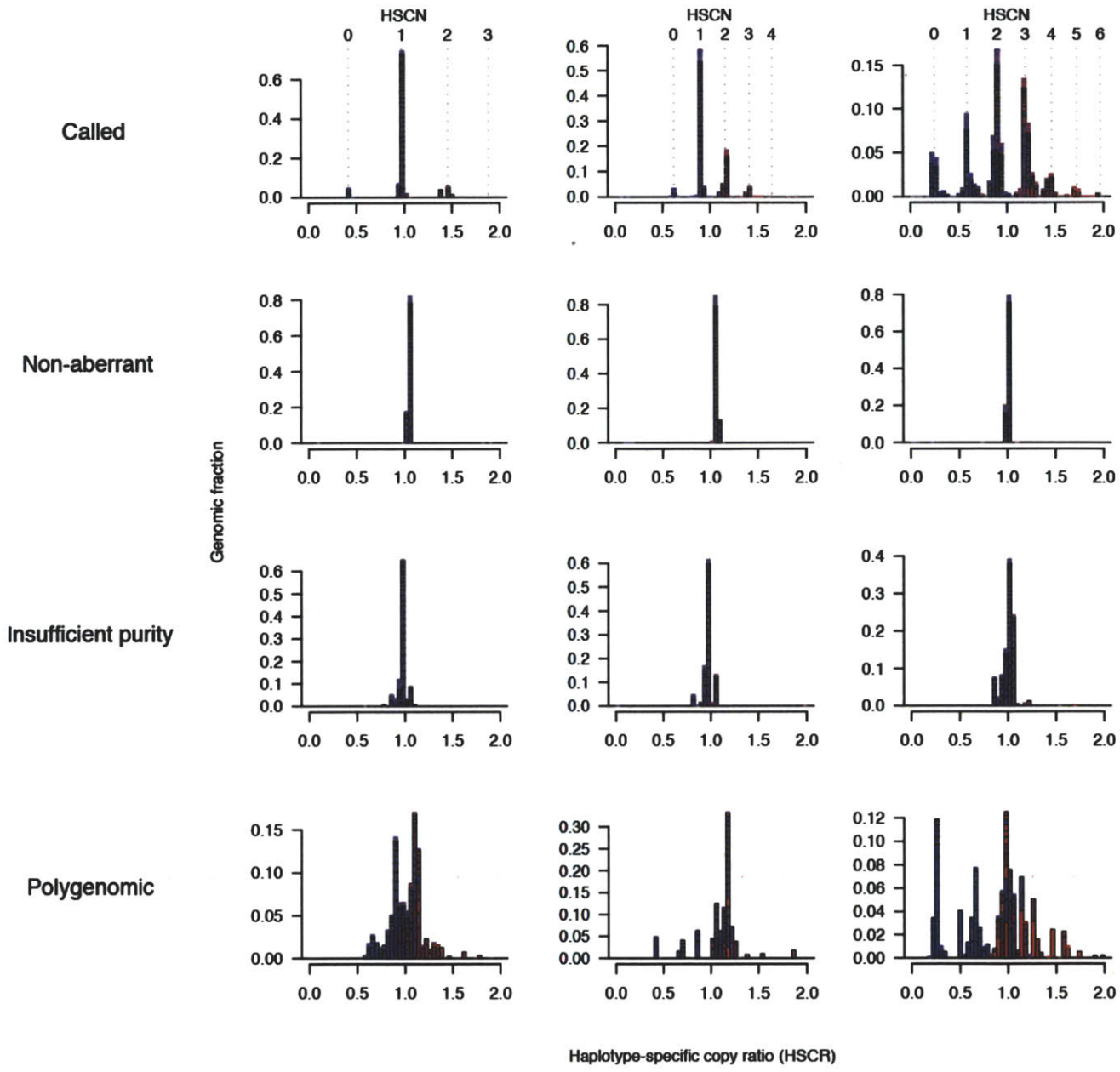
See online methods (x)



Supplemental Figure 4 | Evaluation of the ASCAT {Van Loo, 2010 #818} **method**

a, Cancer-normal cell-line mixing experiment, as in figure 2b. The estimated purity values for the cell-line mixtures obtained by ASCAT were systematically lower than the true values, with a mean difference of 17% and 13% for HCC38 and HCC1143, respectively, and three low-purity mixtures were misinterpreted.

b, Concordance of ASCAT ploidy predictions and FACS analysis, as in figure 2a. The concordance of the ASCAT ploidy estimates with the FACS results was 68.6%, mean absolute difference 0.66. Ploidy of the NCI-60 cell lines could not be evaluated due to the requirement of a matched-normal sample for analysis with ASCAT.



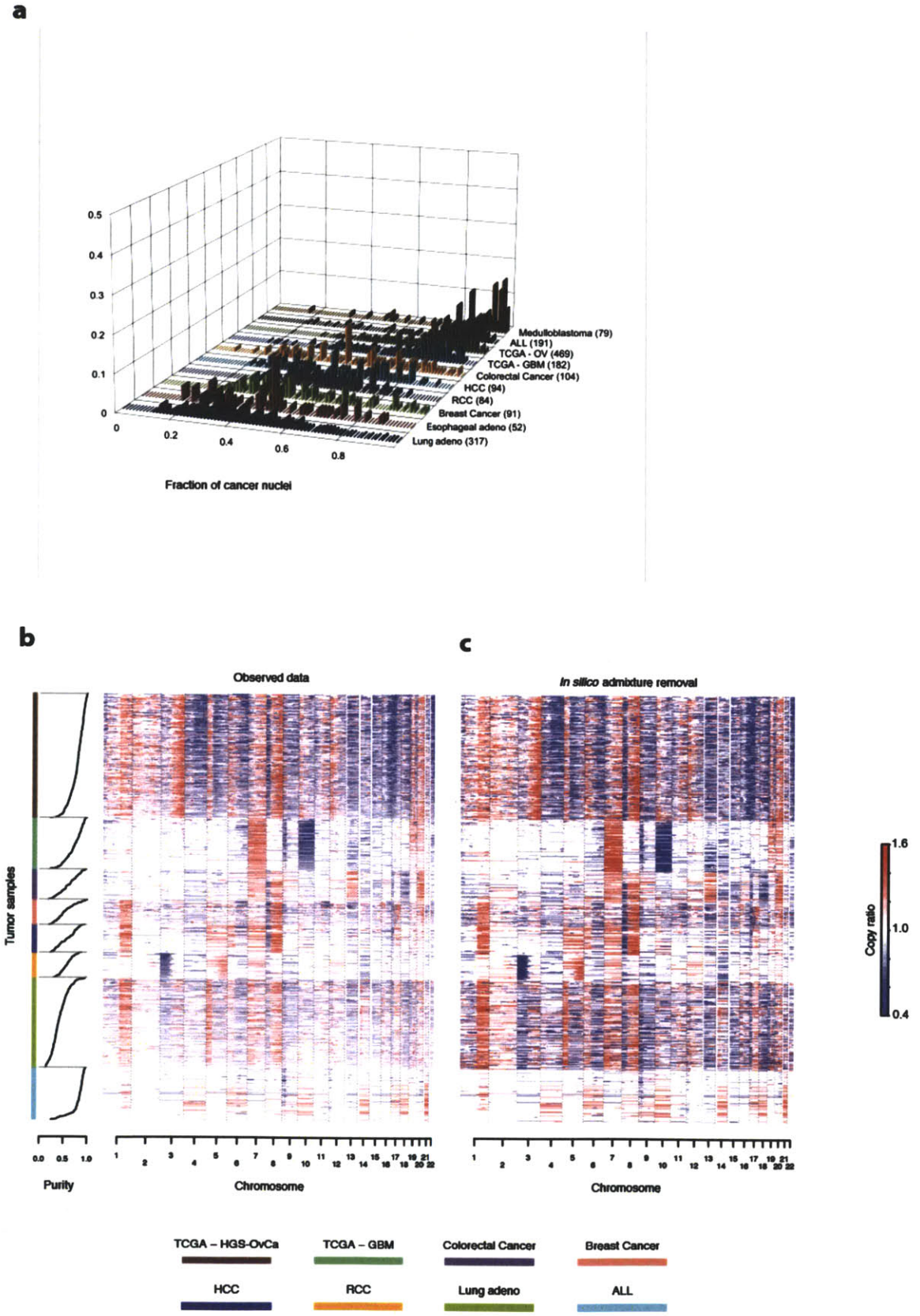
Supplementary figure 5 | Haplotype-specific copy-ratio profiles of called and uncalled cancer samples
a-d, Haplotype-specific copy-ratio (HSCR) profiles for 12 cancer samples. Rows correspond to called samples (**a**), and three distinct failure modes (**b-d**), as indicated (left-hand labels). HSCR profiles are colored according to allelic imbalance, as in figure 1 (**b**,**g**). The distribution of each outcome is shown for each tumor type in figure 2e.

a, Called samples annotated with fit haplotype-specific copy numbers (HSCN; top-axis).

b, The most common failure-mode designation was "non-aberrant" (9.1% of samples), indicating that the copy profile was indistinguishable from that of a normal (diploid) sample.

c, An additional 7.3% of samples failed due to insufficient purity. In these samples, copy aberrations were apparent, but so attenuated as to obscure the pattern of integral copy states.

d, The remaining 6.9% of failed samples were designated as "polygenomic"; although clear SCNAs were observed, they were not consistent with a single dominant copy profile (eqn. 1).


a
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
Fraction of cancer nuclei
Medulloblastoma (79)
ALL (191)
TCGA - OV (469)
TCGA - GBM (182)
Colorectal Cancer (104)
HCC (94)
RCC (84)
Breast Cancer (91)
Esophageal adeno (52)
Lung adeno (317)
b
Observed data
Tumor samples
Purity
Chromosome
c
In silico admixture removal
Chromosome
Copy ratio
1.6
1.0
0.4
TCGA – HGS-OvCa
TCGA – GBM
Colorectal Cancer
Breast Cancer
HCC
RCC
Lung adeno
ALL

Supplemental figure 6 | Distribution of estimated purity in various tumor types, and effect on observed copy-ratios

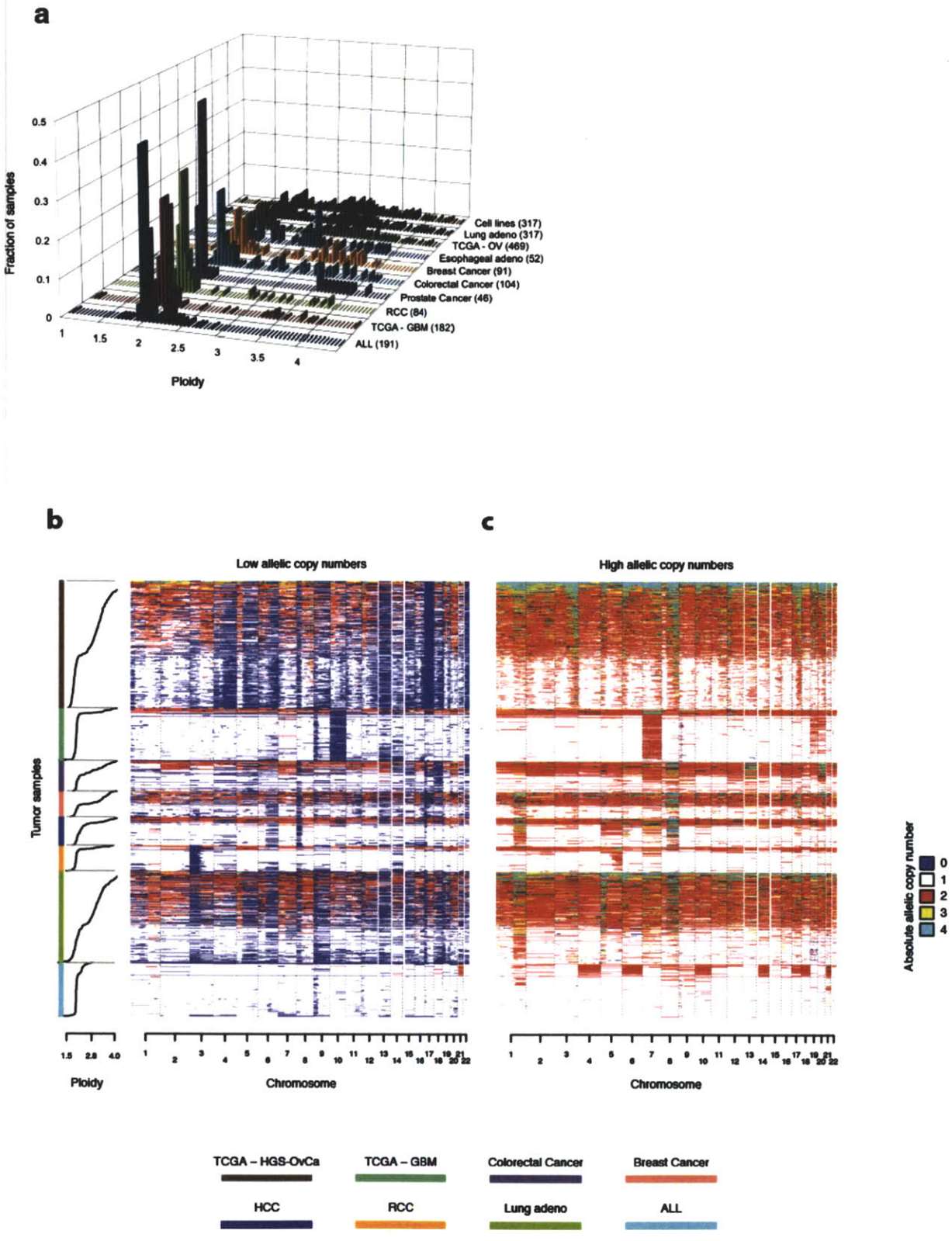
a, Distribution of estimated tumor purity for several datasets. The number of called tumor samples in each group is shown in parentheses. We note that, because heavily contaminated tumors are difficult to call using ABSOLUTE, several of these distributions are biased towards higher purity samples.

b, Data are shown for each indicated sample-type, with samples sorted by tumor purity.

c, Genome wide copy-ratio profiles are shown for several tumor cohorts. The effect of tumor contamination is evident as the 'fading' of copy ratios towards 1 as tumor purity decreases.

d, The data from **c** are shown after '*in silico*' removal of tumor contamination by dividing total absolute copy-number by sample ploidy. This transformation enables more direct comparison between datasets, as samples without copy alterations may be clearly distinguished from heavily contaminated tumors.

c,d, Copy ratios in were median-centered at 1 for plotting.


a
Fraction of samples
0.5
0.4
0.3
0.2
0.1
0
1
1.5
2
2.5
3
3.5
4
Ploidy
Cell lines (317)
Lung adeno (317)
TCGA - OV (469)
Esophageal adeno (52)
Breast Cancer (91)
Colorectal Cancer (104)
Prostate Cancer (46)
RCC (84)
TCGA - GBM (182)
ALL (191)
b
c
Low allelic copy numbers
High allelic copy numbers
Tumor samples
Ploidy
Chromosome
Absolute allelic copy number
0
1
2
3
4
TCGA – HGS-OvCa
TCGA – GBM
Colorectal Cancer
Breast Cancer
HCC
RCC
Lung adeno
ALL

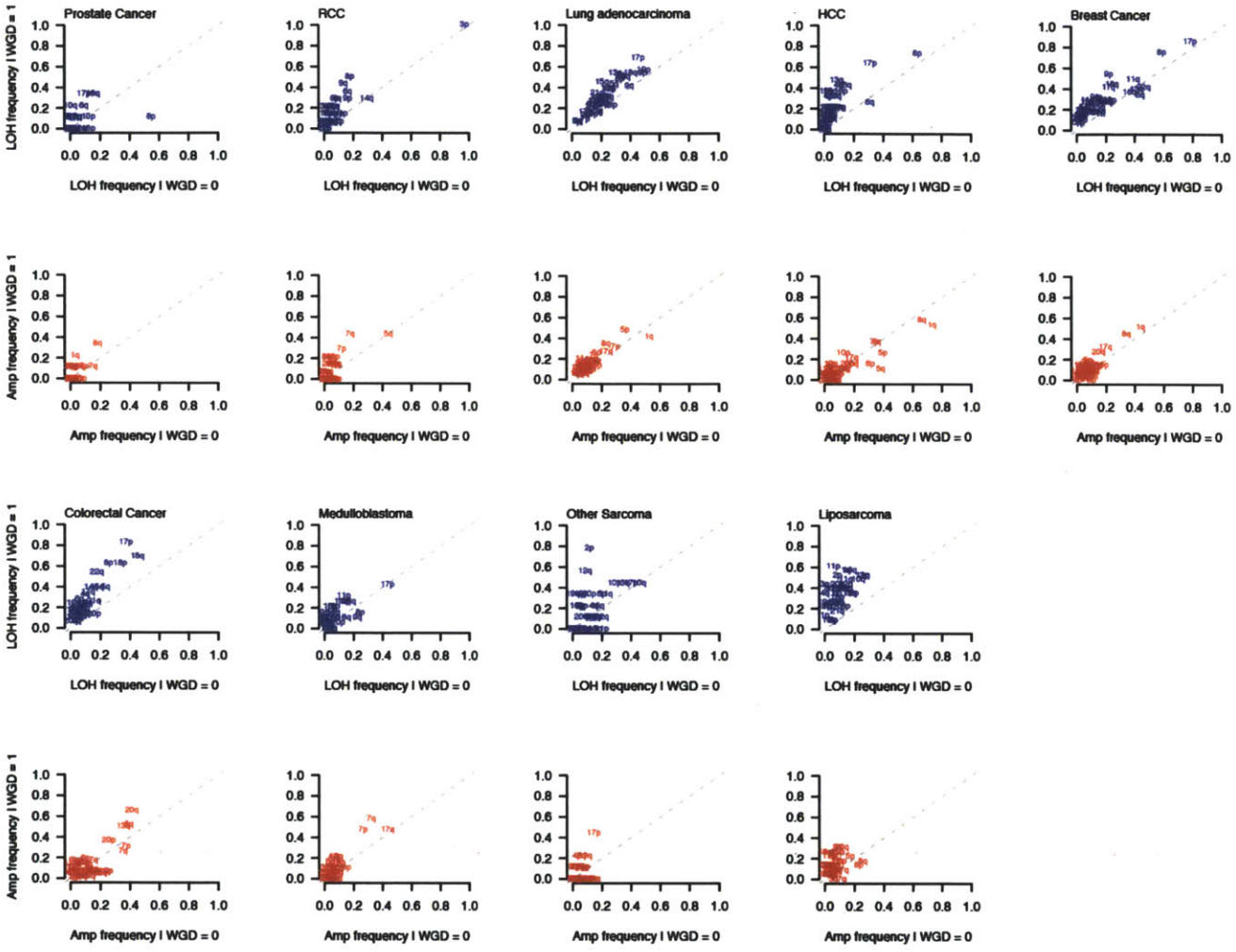
Supplementary figure 7 | | Distribution of estimated ploidy in various tumor types, and correspondence with absolute allelic copy-numbers

a, The number of called tumor samples in each group is shown in parentheses. Because tumors without SCNAs cannot be called using ABSOLUTE, these distributions do not incorporate the prevalence of such samples.

b, Data are shown for each indicated sample-type, with samples sorted by cancer-genome ploidy.

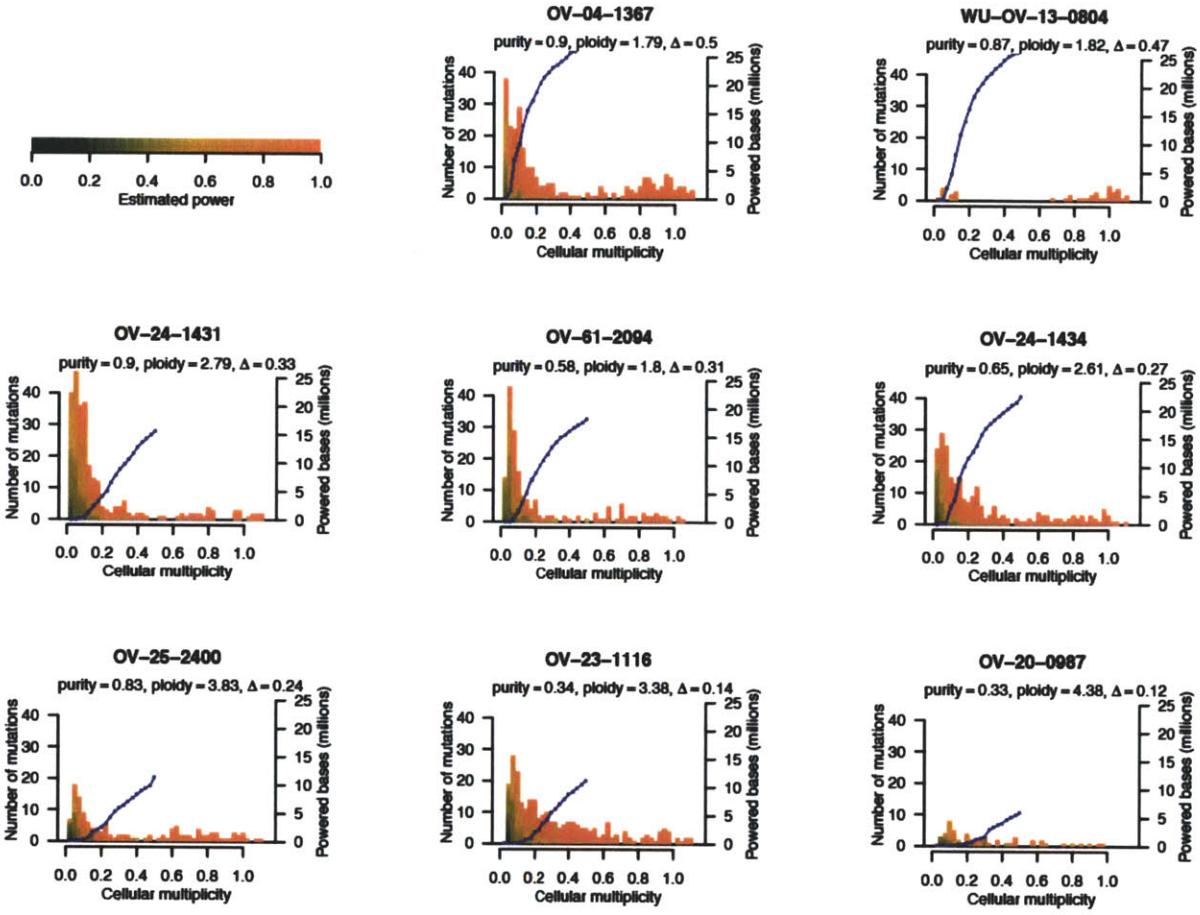
c,d, Absolute copy-numbers of genomic segments for each sample are partitioned into low (**c**) and high (**d**) allelic copy values.

b,c,d, Genome doubling events are discernable as inflections in the ploidy distributions of each cancer (**b**), corresponding with an increase in genome-wide allelic copy-numbers (**c,d**).



Supplementary figure 8 | Incidence of chromosome arm-level SCNAs in non genome doubled (WGD=0) vs. genome doubled (WGD=1) samples from 10 cancer types.

As in figure 3g. LOH (loss of heterozygosity) was defined as 0 allelic copies. Amplification was defined as > 1 allelic copy for samples with 0 genome doublings, and as > 2 allelic copies for those with 1 genome doubling. Calls were made based on the modal allelic copy numbers of each chromosome arm. Dashed lines indicate *y=x*.



Supplemental figure 9 | Power for detection of subclonal somatic mutations in 8 HGS-OvCa samples.

Histograms of cellular multiplicity point-estimates (as in supplemental figure 9b) are shown for individual tumor samples, in order of decreasing Δ, which determines the expected detection power for each sample (figure 5). Individual point mutations are colored according to their estimated detection power, calculated using their observed allelic fraction, coverage, tumor purity, and local absolute copy-number (Methods, eqn. X).

Coverage values and absolute copy-numbers were used to calculate theoretical detection power at each base for subclonal fractions between 0.025 and 0.5 (Methods, eqn. X). The number of bases with power >= 0.8 are shown for each fraction (blue curve, right axis).

Chapter 5

—

Discussion and Future Directions

Scott L. Carter

CIN signature (Chapter 2)

Subsequent studies have validated the prognostic ability of the CIN signature in additional tumor types, including colon cancer [1] and several other cancers. A recent report described the use of the CIN signature (chapter 2) [2] for the construction of a "CINSARC" gene expression signature, conopensisting of 67 genes. This genelist was obtained by intersecting three *t* tests for gene expression values in 183 sarcoma samples, comparing expression in tumors with many vs. few CGH imbalances, grade 3 vs. grade 2, and high vs. low expression of the CIN signature. High expression of the resulting "CINSARC" genes was significantly predictive of poor clinical outcome in a validation set of 127 sarcoma profiles [3].

In 2009, Julian Downward and colleagues[1] examined the relationship between genes described in the CIN-signature (Chapter 2) with those identified as being related to the response of primary tumors and model systems to treatment with taxanes[4]. Although microtubule-stabilizing (MTS) agents such as taxanes are important chemotherapeutics, their mechanism of action is incompletely understood. The specific mechanisms by which cell death is induced following MTS exposure may depend on a competent mitotic spindle checkpoint[5]. However, the inactivation of this checkpoint is likely to be a prerequisite for the development of CIN [6], [7], [5]. The karyotypic complexity of cancer cell lines has been associated with altered cytotoxic response profiles in vitro [8], suggesting that the poor success of treating CIN cancers with MTS agents may be partially due to processes required for the generation and maintenance of the CIN phenotype, rather than the ability of CIN cancers to rapidly evolve resistance mechanisms to cytotoxic therapeutics.

Downward and colleagues derived a MTS expression signature by analysis from expression profiling of 4 cancer cell lines and an ovarian cancer xenograft model treated with taxanes or epothilone. Notably, 27 non cell-cycle regulated genes from the CIN70 signature[2] were significantly repressed following MTS treatment, but not by treatment

[1] Including Dr. Zoltan Szallasi, the senior author of the CIN-signature paper.

with the cytotoxic agent 5-FU[4]. In addition, these genes were observed to have lower expression across several data sets generated from tumors prior to and following treatment via MTS exposure[4]. The authors identified 50 "MTS-repressed" genes by intersecting genes down-regulated following MTS exposure in several datasets with the CIN70 genes[4]. Silencing 22/50 of these genes by shRNAi caused significantly impaired cell viability and/or promoted increases in sub-G1 DNA content in the HCT-116 cell line, and silencing all of them produced the same effect in two additional cell lines. Taken together, these results suggest that DNA damage and spindle checkpoint genes, as identified in the CIN signature[2] (Chapter 2), may play a causal role in mediating cell-death following MTS exposure.

In addition to its utility in cancer prognosis, subsequent experimental investigation of specific genes comprising the "CIN signature" has confirmed roles for several such genes as functionally related to CIN. CIN scores were computed for 10,152 genes. Over-expression of several of the top 70 scoring CIN genes had been previously reported to cause aneuploidy (PTTG1), chromosomal instability (AURKB, CCNB1, CCNB2, FLJ10036), or polyploidy (TPX2, AURKA) (references and the full list of the top 70 genes are given in Appendix A of this thesis).

Recently, high levels of the Mps1 protein, (formerly TTK, CIN-gene #23) were reported to contribute to the ability of breast cancer cells to tolerate aneuploidy, although over-expression of Mps1 was insufficient to induce aneuploidy or transformation of immortalized human cells [9]. Inhibition of this gene by shRNAi resulted in increased apoptosis in the human cancer cell-line T47D, and reduced growth in breast cancer xenografts [9]. These results provide functional validation for an additional CIN gene and provide another demonstration of non-oncogene addiction in human cancer cells [10], [11].

Another gene, *FOXM1* (CIN-gene #3) has since been shown to be an important master transcriptional regulator, disrupted in human cancer, and has been recently identified to play a role in high-grade serous ovarian carcinoma[12]. Previous reports had described

the chromatin modifying polycomb-group (PcG) gene *EZH2* (CIN-gene #37) as over-expressed in breast cancer. Interestingly, recurrent mutations in this gene have recently been reported in myelodysplastic syndrome (MDS), showing patterns of homozygosity, frame-shift, and truncation consistent with tumor suppressor function [13], [14].

In 2006, shortly after the CIN signature paper appeared, David Pellman and colleagues reported results from a synthetic lethality screen for yeast genes essential to the viability of polyploid strains [15]. Only 39 out of 3,740 mutations screened exhibited ploidy-specific lethality. Almost all of these mutations affected genomic stability by impairing homologous recombination, sister chromatid cohesion, or mitotic spindle function. Several of these genes had high CIN-scores, such as RAD54L (#75), BUB1 (#421), and RAD51 (#487), although a rigorous mapping to yeast/human gene-homologs was not performed.

Although aneuploidy is found in approximately 90% of human cancers, it produces a strongly deleterious phenotype in normal eukaryotic cells[16]. In 2007, Angelika Amon and colleagues addressed this paradox by investigating haploid yeast strains for which single chromosome gains of each individual autosome had been constructed[17]. The authors used proliferation assays in various conditions and gene expression microarrays to characterize the phenotypic consequences on these induced aneuploidies. They showed that aneuploid strains shared a number of phenotypes, including defects in cell cycle progression, increased glucose uptake, and increased sensitivity to conditions interfering with protein synthesis and protein folding[17]. To investigate the general effects of aneuploidy, the authors down-weighted genes on the disomic chromosome present in each strain and clustered the profiles, revlealing that several disomic strains, and all of the strains bearing multiple extra chromosomes exhibited a gene expression signature characteristic of the yeast environmental stress response (ESR) [18].

Notably, the above growth and gene-expression phenotypes were observed only with gain of yeast chromosomes bearing transcribed genes (and not with artificial chromosome constructs); this indicates that the observed phenotypes were likely the result of protein

imbalances, and not due to mitotic defects due to purely geometric differences in the aneuploid strains relative to wild-type. However, the gene expression profile shared by aneuploid strains grown under phosphate-limiting chemostat conditions was also observed in strains with artificial chromosomes[17], indicating that replication and maintenance of extra DNA has distinct phenotypic consequences.

The gene annotations most over-represented in the 397 aneuploidy genes co-regulated in the aneuploid strains were ribosome biogenesis, and carbohydrate metabolism, which bear no obvious resemblance to the annotations associated with the CIN signature [2]. This suggests that in human cancer, aneuploidy does not arise randomly, but rather is due to positive selection, in which case it occurs with proliferative advantages. This is supported by the patterns of chromosomal SCNAs in human cancer, where gain and loss of particular arms are highly recurrent within particular diseases and cell-lineages [19].

Because virtually all oncogenic pathways are expected to result in a cellular phenotype with greater proliferative capacity, it is difficult to disentangle gene expression 'signatures' related to specific oncogenic mechanisms from the downstream induction of normal cell-cycle transcriptional programs [20]. This caveat applies to all such "gene expression signatures" related to the cancer phenotype, whether they seek to describe the induction of specific known oncogenes [21], [22], or cellular states related to cancer development [23], [24], [25], [26], [2], [27].

In the case of the CIN signature, many genes were annotated in functional categories related to replication and segregation of chromosomes. However, many such genes are also by their nature cell-cycle regulated [28]. An important aspect of the CIN-signature work involved dissecting out the prognostic ability of genes specifically associated with aneuploidy (as ascertained by regionally coordinated expression deviation[2]), vs. those with expression patterns driven by the cell-cycle (in the HeLa cancer cell-line [20])[2].

[2] One might wonder whether the use of a cancer cell-line to identify genes regulated by the cell-cycle might result in the identification of genes with such expression patterns due

This procedure revealed that many genes with high CIN-scores did not appear to be regulated by the cell cycle (and that these genes were highly prognostic). To further examine this relationship further, it might be informative to compare the genes correlated with each of the core cell-cycle and CIN genes across large and diverse collections of cancer gene expression profiles. Although these gene-sets will almost certainly overlap, clustering or PCA analysis of each signature's expression "neighborhood" might reveal evidence for finer correlation structure, given sufficient data.

The difficulty of de-coupling the downstream effectors of cellular oncogenes from the normal functioning of physiological processes was further highlighted in a recent paper describing the regulation of 13 such signatures by the MYC oncogene in human cancer cell-lines [29]. The c-Myc oncoprotein was recently demonstrated to function as a post-initiation regulatory mechanism, controlling pause-release by the RNA polymerase II (Pol II) transcription initiation apparatus [30]. Furthermore the expression of this protein was linked to the production of a large population of actively transcribed genes in embryonic stem cells [30]. These results suggest that MYC over-expression may orchestrate the unregulated proliferation characteristic of stem cells by directly regulating the transcription of thousands of target genes. Notably, MYC is the most frequently amplified gene in human cancer [19].

Further dissection of chromosomal instability as a cellular phenotype might be aided by screening for genes with CIN-associated phenotypes, such as progression of the cell-cycle through the mitotic-spindle checkpoint in the presence of nocodazole, a microtubule poison, as previously described [6]. Such efforts could be integrated with other high-throughput screening efforts on cell-lines with partially defined karyotypes[3]

---

to the cancer phenotype itself, rather than as a normal constituent of the cell cycle program.

[3] The use of ABSOLUTE to further resolve these karyotypes and annotate them with respect to genome doubling status may be used to interpret screening data in these cell lines.

In 2011, Zoltan Szallasi and colleagues demonstrated excellent correlation between the CIN score and a structural chromosomal complexity score based on SNP array data, as well as a measure of nuclear DNA mass (ploidy)[31]. In addition, they showed that breast cancers with CIN signature values greater than the 75th percentile showed improved prognosis relative to the 50-75% percentile, suggesting that extreme excess transcription of the CIN-genes was associated with decreased fitness, possibly due to the difficulty of maintaining an extreme mutator phenotype[31].

HAPSEG (Chapter 3)

In addition to its utility for HSCR estimation, we exploit unequal copy-numbers in parental homologues to achieve direct haplotype phasing, which is of general interest beyond cancer genome analysis. We demonstrate the validity of our phasing estimates by comparison with phasing methods based on reference chromosome panels.

We believe that HAPSEG has potential application to areas other than cancer biology. Specifically, the capacity of HAPSEG to produce definitive long-range haplotype phasing from bi-allelic measurements will be of immediate interest. For example, we are currently using HAPSEG on whole-genome sequencing data in order to generate densely-typed reference panels for imputation of rare variants using whole-exome sequencing.

Specifically, because the reference panel can be genotyped at greater density than the input genotypes, information about untyped SNPs can be filled in using the reference panels, a technique termed *imputation* in the genetics literature. Several statistical algorithms have been described implementing imputation within this framework [32], [33], [34], [35], [36], (reviewed in[37]). Importantly, all of these techniques assume that the phase of the heterozygous genotypes is unknown, requiring them to average over statistical configurations of the haploid genotypes.

Long-range haplotyping provided by HAPSEG in aneuploid cancer samples provides a means of obtaining segmental regions of definitive phase. This information has the potential to increase the sensitivity of imputation in these regions, and in adjacent unphased regions, allowing for inference of lower minor allele frequency (MAF)

variants. This may present an attractive strategy by which to undertake a pan-cancer analysis of preferential somatic alteration of specific alleles in cancer development. Identification of recurrent alterations of specific variants would be of interest because they would be contribute evidence of positive selection at these loci during cancer evolution, possibly identifying novel mechanisms of interaction with germline risk and evolution of the cancer phenotype.

The common risk locus at 8q24 was previously reported to show ~1.5 fold preferential amplification in colorectal cancer [38], and later to functionally interact with MYC by affecting a binding site for the Wnt-regulated transcription factor TCF4, with the risk allele showing stronger binding. Although genome-wide ChIP assay revealed this element as the strongest TCF4 binding site within 1 Mb of *MYC* [39], "an unambiguous correlation between rs6983267 genotype and MYC expression was not detected" [39]. At the same time, Matthew Freedman and colleagues reported that this locus displayed enhancer activity and long-range interaction with MYC in colorectal cancer [40].

Unfortunately, the unbiased genome-wide detection of novel selected variants at effect sizes of ~20-fold or less will likely require the analysis of several thousand cancer samples [41][4]. Early studies using smaller sample sizes have not yielded definitive evidence for the moderate prevalence of allelic selection in cancer evolution [42], [43]. We hope to probe deeper into lower effect-size variants by using HAPSEG, as described above, on several thousand cancer samples produced in collaboration with The Cancer Genome Atlas.

ABSOLUTE (Chapter 4)

As of 2011, characterization of tumor copy-profiles and point-mutations using ABSOLUTE has become an essential methodological component of our cancer genome analysis pipeline at the Broad institute, and provides a key foundation for virtually every aspect of our continued analytical developments. This will impact future efforts in The

[4] Identification of such variants requires large sample-sizes because only *heterozygous* SNPs with somatic copy alteration are informative.

Cancer Genome Atlas (TCGA) project, in which we are characterizing 500 tumor/normal pairs in >20 tumor-types, as well as all of our other cancer genome projects. In addition, several insights into the evolution of aneuploidy and genome doubling in human cancer are discussed below.

Aneuploidy precedes genome doubling in many human cancers

Analysis of modeled absolute allelic copy-number profiles across human cancers produced by ABSOLUTE allowed for refinement in the understanding of cancer genome evolution. The profiles were consistent with a period of chromosomal instability early in the development of many tumors, followed by the emergence of a stable aneuploid clone, as previously described [44]. Our data further indicate that genome doublings occur in a subset of cancer cells harboring the characteristic arm-level SCNAs of the corresponding tumor type. The genomes of these cancers were therefore shaped by selection at broad resolution prior to doubling and clonal outgrowth.

Our model whereby aneuploidy precedes genome doubling is broadly consistent with an earlier interpretation of primary breast cancer FACS/SKY profiles [45], and has recently been recapitulated in studies of macro-dissected and ploidy-sorted cell-populations [46], and single cell sequencing [47], of primary breast tumors. However, most other descriptions of genome doubling in cancer typically have referenced tetraploidization itself as an initiating oncogenic event [48],[49],[50], [51], [52], [53], an interpretation inconsistent with our data. A notable exception occurred in prostate cancer samples, a subset of which were modeled as near-tetraploid, with few broad SCNAs; further validation is required to confirm these predictions.

Annotation of cell-line collections using ABSOLUTE will inform future genetic screening of cell-lines, allowing integration of shRNAi sensitivity profiles with specific aneuploidies, genome-doubling status, and point-mutation multiplicities. Such annotations may be further used to screen for mutations that are specific to aneuploidy or genome-doubling status.

The relationship between aneuploidy and genome doubling

Polyploidy has previously been proposed to buffer the deleterious effects of aneuploidy [54]. Deleterious phenotypes caused by gain of individual yeast chromosomes were more severe in haploid yeast strains than in diploids, consistent with the doubled baseline DNA content buffering proteomic imbalances due to aneuploidy [17], [54], [55]. This may provide an explanation for the frequent genome doublings observed in human cancers, especially in those of epithelial lineage (Chapter 4).

Buffering of aneuploidy may also partly explain the rarity of near-haploid karyotypes in human cancer. These genomes virtually always display absence of focal alterations and loss of all autosomes except for those with the greatest amplification frequency (eg. Chromosomes 7 and 8 in lung adenocarcinoma and GBM; Chapter 4). Such configurations may represent islands of selective advantage, since the paths leading to them by concerted loss of individual chromosomes likely involve intermediates with severe relative imbalances in the genome; virtually no karyotypes were observed with ploidy values between 1.25 and 1.6 (Chapter 4). In addition, the random loss of any chromosome from this configuration would likely be lethal; a state which we expect would be incompatible with ongoing chromosomal instability.

Buffering of genomic imbalances by genome doubling would not explain why only some cancers pass through genome doubling during their evolution. High-grade serous ovarian carcinoma (HGSOC) may be an especially instructive disease in which to consider this question[56],[12], because the rates of specific chromosome-arm deletions and amplifications are nearly identical in the genome-doubled cases (~55%) vs. non-doubled cases (45%) (Chapter 4). Thus, aneuploidy itself is unlikely to influence somatic evolution with regard to genome doubling in these cancers.

In 2009, David Pellman and colleagues demonstrated a mechanistic link between abnormal centrosome content and chromosomal instability in human cells [57]. Human cells normally contain two centrosomes, which localize to either end of the nascent mitotic spindle during mitosis where they are critical in the nucleation of microtubule

polymerization forming the spindle scaffold. The authors showed that cells with doubled centrosome content could sometimes cluster their four centrosomes into a bipolar orientation, resulting in a reasonably normal mitosis, although with a greater rate of chromosome mis-segregation[57]. However, when this clustering did not occur, a multi-polar mitosis ensued, resulting in catastrophic errors and total lethality. Because genome doubling by cytokinesis failure [48], [58], [49], [59], or mitotic bypass [52] doubles the centrosome number, these results predict that the rate of chromosomal instability would be greater in the resulting cells.

In 2010, Titia de Lange and colleagues demonstrated that critical erosion of telomeric DNA provokes a persistent DNA damage response in mouse embryonic fibroblasts (MEFs), leading to cytokinesis failure, resulting in polyploidy. We observed a significant association between genome doubling and increased age at diagnosis, consistent with a role for telomeric erosion in HGSOC genome doubling. Alternately, this association could be related to menopausal status and estrogen signaling.

Genome doubled cancers have greater incidences of focal SCNAs than non-doubled cases (Chapter 4), and this increase did not appear to be explainable solely by the increased DNA content subjected to equivalent mutation rates in the two sample classes. This raises the possibility that the same mechanism which produced the excess focal SCNAs in the doubled genomes caused bypass of cytokinesis in these cells, leading to the genome doubling as previously described[52], but independent of telomere status. Deeper analysis of the absolute allelic copy-numbers of focal alterations in genome-doubled cancers may shed light on whether they occurred pre- or post- doubling, at least.

In the final days before the defense of this thesis on 08/29/2011, additional insights into the relationship between aneuploidy and cancer evolution were published. Angelika Amon and colleagues reported that multiple budding yeast strains bearing additional copies of single yeast chromosomes showed an increase in various forms of genetic instability, thus demonstrating that aneuploidy itself drives a mutator phenotype [60]. By showing that aneuploidy enhances genetic recombination and defective DNA damage

repair, the authors provide a mechanistic link between aneuploidy itself and further genomic instability [61]. Notably, they found that loss of yeast artificial chromosomes (YACs)[5] was increased in 9/13 of aneuploid strains tested, relative to wild type[60], an observation which is consistent with our results in primary human cancers, where individual chromosomal losses were predicted to exceed gains (Chapter 4).

In the same issue of *Science*, a separate group demonstrated that the STAG2 gene is homozygously deleted in a small number of glioblastoma, Ewing's sarcoma, and melanoma cell lines [62]. Because the STAG2 gene is located on the human X chromosome, only a single genetic hit is required to recessively inactivate the gene in either male patients or female patients in which the remaining allele is silenced by the X-inactivation. The *STAG2* gene encodes a subunit of the cohesin complex, which regulates the separation of sister chromatids during cell division. The authors showed that targeted correction of the endogenous mutant allele of *STAG2* in human glioblastoma cells restored sister chromatid cohesion [62].

Additional future applications of ABSOLUTE

Clustering of absolute allelic copy-profiles at chromosome arm resolution produced 3 clusters of GBM samples (Supplemental figure 1, chapter 4) [6]. Integration with additional genomic features resolved from sequencing data will allow further characterization of genomic heterogeneity in diverse cancer types. In addition, see the discussion section of Chapter 4.

[5] Because these do not contain transcribed genes, they are not subject to selection for beneficial protein products, and allow for measurement of the rates of chromosomal gain and loss.

[6] The models used for clustering may not have been sensitive enough to resolve small clusters of rare sample types, given the available data. For example several near-haploid samples are observed as outliers in the first cluster (Supplemental figure 1, chapter 4).

Analysis with ABSOLUTE provides information that helps distinguish somatic from germline mutations in unpaired DNA samples[7]. Differences in the effect of tumor purity on the allelic-fractions of somatic vs. germline mutations can be exploited to distinguish the two classes of events in unpaired tumor DNAs.

Promising results were obtained in preliminary analyses using ABSOLUTE to probabilistically classify variants as somatic or germline by modeling the observed allelic fraction of each variant compared to the expected fraction given by each possible germline and somatic multiplicity. Receiver operating curve AUC values of ~0.85 for somatic classification recall as verified by analysis of paired-normal DNA (not shown). Note that this refers exclusively to the clonal variants. Subclonal somatic alterations rarely occur with plausible germline allelic fractions, making them much easier to distinguish in the data[8].

Further integration of this information with computational prediction of functional consequences [63],[64] and statistical analysis of recurrence patterns would likely result in adequate power to distinguish recurrent somatic alterations in particular genes. We note that heavily contaminated tumor samples sequenced at high coverage would be ideal for such analysis. The sensitivity of this approach can be increased by joint analysis of multiple tumor samples of differing purity, perhaps from different sectors of a tumor specimen, although it will be important to account for subclonal variation.

---

[7] This might be the case when considering analysis of DNA derived from paraffin embedded tumor specimens; such samples are available from many clinical trials, and are thus richly annotated with valuable retrospective data.

[8] Although an important exception occurs in the case of very pure tumor samples or cancer cell-lines.

References

1. Mayhew CN, Carter SL, Fox SR, Sexton CR, Reed CA, et al. (2007) RB loss abrogates cell cycle control and genome integrity to promote liver tumorigenesis. Gastroenterology 133: 976–984. doi:10.1053/j.gastro.2007.06.025

2. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z (2006) A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 38: 1043–1048. doi:10.1038/ng1861

3. Chibon F, Lagarde P, Salas S, Pérot G, Brouste V, et al. (2010) Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. Nature Medicine 16: 781–787.

4. Swanton C, Nicke B, Schuett M (2009) Chromosomal instability determines taxane response. Proceedings of the national academy of science

5. Amon A (1999) The spindle checkpoint. Current opinion in genetics & development.

6. Lengauer C, Kinzler KW, Vogelstein B (1997) Genetic instability in colorectal cancers. Nature 386: 623–627. doi:10.1038/386623a0

7. Lengauer C, Kinzler K, Vogelstein B (1998) Genetic instabilities in human cancers. Nature.

8. Roschke A, Lababidi S, Tonon G (2005) Karyotypic "state" as a potential determinant for anticancer drug discovery. Proceedings of the ....

9. Daniel J, Coulter J, Woo J-H, Wilsbach K, Gabrielson E (2011) High levels of the Mps1 checkpoint protein are protective of aneuploidy in breast cancer cells. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1007645108

10. Solimini NL, Luo J, Elledge SJ (2007) Non-Oncogene Addiction and the Stress Phenotype of Cancer Cells. Cell 130: 986–988. doi:10.1016/j.cell.2007.09.007

11. Luo J, Solimini NL, Elledge SJ (2009) Principles of cancer therapy: oncogene and non-oncogene addiction. Cell 136: 823–837. doi:10.1016/j.cell.2009.02.024

12. Network TCGAR, Network TCGAR, sites DWGATS, Medicine GSCBCO, Broad Institute, et al. (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615. doi:10.1038/nature10166

13. Ernst T, Chase AJ, Score J, Hidalgo-Curtis CE, Bryant C, et al. (2010) Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nat Genet 42: 722–726. doi:10.1038/ng.621

14. Nikoloski G, Langemeijer SMC, Kuiper RP, Knops R, Massop M, et al. (2010) Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. Nat Genet 42: 665–667. doi:10.1038/ng.620

15. Storchová Z, Breneman A, Cande J, Dunn J, Burbank K, et al. (2006) Genome-wide genetic analysis of polyploidy in yeast. Nature 443: 541–547. doi:10.1038/nature05178

16. Storchova Z, Pellman D (2004) From polyploidy to aneuploidy, genome instability and cancer. Nature Reviews Molecular Cell Biology 5: 45–54.

17. Torres EM, Sokolsky T, Tucker CM, Chan LY, Boselli M, et al. (2007) Effects of Aneuploidy on Cellular Physiology and Cell Division in Haploid Yeast. Science 317: 916–924. doi:10.1126/science.1142210

18. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell 11: 4241–4257.

19. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905. doi:10.1038/nature08822

20. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, et al. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell 13: 1977–2000. doi:10.1091/mbc.02-02-0030.

21. Huang E, Ishida S, Pittman J, Dressman H (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. Nature.

22. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2005) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439: 353–357. doi:10.1038/nature04296

23. Ramaswamy S, Ross KN, Lander ES, Golub TR (2002) A molecular signature of metastasis in primary solid tumors. Nat Genet 33: 49–54. doi:10.1038/ng1060

24. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al. (2004) Gene Expression Signature of Fibroblast Serum Response Predicts Human Cancer Progression: Similarities between Tumors and Wounds. PLoS Biology 2: e7. doi:10.1371/journal.pbio.0020007.sd002

25. Kho AT (2004) Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. Genes & development 18: 629–640. doi:10.1101/gad.1182504

26. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proceedings of the National Academy of Sciences of the United States of America 101: 9309–9314. doi:10.1073/pnas.0401994101

27. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, et al. (2008) An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 40: 499–507. doi:10.1038/ng.127

28. Pollack J (2006) Chromosome instability leaves its mark. Nat Genet.

29. Wolfer A, Wittner BS, Irimia D, Flavin RJ, Lupien M, et al. (2010) MYC regulation of a "poor-prognosis" metastatic cancer cell state. Proceedings of the National Academy of Sciences 107: 3698–3703. doi:10.1073/pnas.0914203107

30. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, et al. (2010) c-Myc Regulates Transcriptional Pause Release. Cell 141: 432–445. doi:10.1016/j.cell.2010.03.030

31. Birkbak NJ, Eklund AC, Li Q, McClelland SE, Endesfelder D, et al. (2011) Paradoxical Relationship between Chromosomal Instability and Survival Outcome in Cancer. Cancer Research 71: 3447–3452. doi:10.1158/0008-5472.CAN-10-3667

32. Li Y (2006) Li: Mach 1.0: rapid haplotype reconstruction and... - Google Scholar. Am J Hum Genet.

33. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81: 559–575. doi:10.1086/519795

34. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906–913. doi:10.1038/ng2088

35. Howie B, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.

36. Browning BL, Yu Z (2009) Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. The American Journal of Human Genetics 85: 847–861. doi:10.1016/j.ajhg.2009.11.004

37. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nature Publishing Group 11: 499–511. doi:10.1038/nrg2796

38. Tuupanen S, Niittymaki I, Nousiainen K, Vanharanta S, Mecklin JP, et al. (2008) Allelic Imbalance at rs6983267 Suggests Selection of the Risk Allele in Somatic Colorectal Tumor Evolution. Cancer Research 68: 14–17. doi:10.1158/0008-5472.CAN-07-5766

39. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, et al. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 41: 885–890. doi:10.1038/ng.406

40. Pomerantz M, Ahmadiyeh N, Jia L, Herman P, Verzi M, et al. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41: 882–884.

41. Dewal N, Freedman ML, LaFramboise T, Pe'er I (2010) Power to detect selective allelic amplification in genome-wide scans of tumor data. Bioinformatics 26: 518–528. doi:10.1093/bioinformatics/btp694

42. LaFramboise T, Dewal N, Wilkins K, Pe'er I, Freedman ML (2010) Allelic Selection of Amplicons in Glioblastoma Revealed by Combining Somatic and Germline Analysis. PLoS Genetics 6: e1001086. doi:10.1371/journal.pgen.1001086.t001

43. van Loo P, Nordgard S (2010) Allele-specific copy number analysis of tumors. Proceedings of the national academy of science

44. Albertson D, Collins C, McCormick F, Gray J (2003) Chromosome aberrations in solid tumors. Nat Genet 34: 369–376.

45. Dutrillaux B, Gerbault-Seureau M, Remvikos Y, Zafrani B, Prieur M (1991) Breast cancer genetic evolution: I. Data from cytogenetics and DNA content. Breast Cancer Res. Treat. 19: 245–255.

46. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, et al. (2010) Inferring tumor progression from genomic heterogeneity. Genome research 20: 68.

47. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90–94. doi:10.1038/nature09807

48. Fujiwara T, Bandi M, Nitta M, Ivanova E, Bronson R, et al. (2005) Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. Nature 437: 1043–1047.

49. Ganem NJ, Pellman D (2007) Limiting the Proliferation of Polyploid Cells. Cell 131: 437–440. doi:10.1016/j.cell.2007.10.024

50. Storchova Z, Kuffer C (2008) The consequences of tetraploidy and aneuploidy. Journal of Cell Science 121: 3859–3866. doi:10.1242/jcs.039537

51. Holland A, Cleveland D (2009) Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nature Reviews Molecular Cell Biology 10: 478–487.

52. Davoli T, Denchi EL, Lange T de (2010) Persistent Telomere Damage Induces Bypass of Mitosis and Tetraploidy. Cell 141: 81–93. doi:10.1016/j.cell.2010.01.031

53. Bazeley PS, Nestor Kalinoski AL, Ways JA, Liu ST, Ramdath RS, et al. (2011) A model for random genetic damage directing selection of diploid or aneuploid tumours. Cell Proliferation 44: 212–223. doi:10.1111/j.1365-2184.2011.00746.x

54. Torres EM, Williams BR, Amon A (2008) Aneuploidy: Cells Losing Their Balance. Genetics 179: 737–746. doi:10.1534/genetics.108.090878

55. Williams BR, Prabhu VR, Hunter KE, Glazier CM, Whittaker CA, et al. (2008) Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. Science 322: 703–709. doi:10.1126/science.1160058

56. Levanon K, Crum C, Drapkin R (2008) New Insights Into the Pathogenesis of Serous Ovarian Cancer and Its Clinical Impact. Journal of Clinical Oncology 26: 5284–5293. doi:10.1200/JCO.2008.18.1107

57. Ganem NJ, Godinho SA, Pellman D (2009) A mechanism linking extra centrosomes to chromosomal instability. Nature 460: 278–282. doi:10.1038/nature08136

58. Shi Q, King R (2005) Chromosome nondisjunction yields tetraploid rather than aneuploid cells in human cell lines. Nature.

59. Pellman D (2007) Cell biology: aneuploidy and cancer. Nature 446: 38–39. doi:10.1038/446038a

60. Sheltzer JM, Blank HM, Pfau SJ, Tange Y, George BM, et al. (2011) Aneuploidy Drives Genomic Instability in Yeast. Science 333: 1026–1030. doi:10.1126/science.1206412

61. Kolodner RD, Cleveland DW, Putnam CD (2011) Aneuploidy Drives a Mutator Phenotype in Cancer. Science 333: 942–943. doi:10.1126/science.1211154

62. Solomon DA, Kim T, Diaz-Martinez LA, Fair J, Elkahloun AG, et al. (2011) Mutational Inactivation of STAG2 Causes Aneuploidy in Human Cancer. Science 333: 1039–1043. doi:10.1126/science.1203619

63. Sunyaev S, Ramensky V, Koch I (2001) Prediction of deleterious human alleles. Human molecular genetics

64. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nature methods 7: 248–249. doi:10.1038/nmeth0410-248

Appendix A

—

Supplementary Material for Chapter 2

Scott L. Carter, Aron C Eklund, Isaac S Kohane, Lyndsay N Harris, and Zoltan Szallasi

Figure S1. Total functional aneuploidy (tFA) is correlated with the total number of chromosomal aberrations in the NCI60 cell line panel as measured by SKY analysis. The number of structural aberrations (NSA) is equal to the total number of insertions, deletions, translocations and homologously staining regions, present in at least two metaphases[1]. A) The tFA of each cell line is correlated with its NSA ($r = 0.43$). Each point represents a cell line. The significance of this correlation was verified by permutation testing (10,000 random permutations of cell-line labels, $p = 0.0009$). B) The Pearson correlations of individual genes to tFA and to NSA are correlated ($r = 0.61$). For each gene, the correlation between tFA and the expression level of the gene is plotted on the horizontal axis, and the correlation between NSA and the expression level of the gene is plotted on the vertical axis. Each point represents a gene.

While the trend in these plots is clear, a perfect correlation between tFA and NSA cannot be expected. Structural chromosomal aberrations are directly observed by SKY analysis in only a few cells, and those results are averaged and extrapolated for the entire cell population. Furthermore, the tFA metric was conceived to take both copy-number aberrations and structural aberrations into account. In accordance with these considerations the above outlined results confirm that tFA and NSA are correlated, though not equivalent.

Figure S2. SNPchip aCGH data from the NCI60 cell line panel confirms that FA is an estimate of regional chromosomal aberrations and that tFA is an estimate of the total level of chromosomal aberrations. Mean copy number (MCN) was determined for each sub-band using SNPchips (see supplementary methods). A) MCN and FA are correlated across cell lines for most sub-bands. The cumulative distribution of the MCN/FA correlations is plotted with each point representing a sub-band. B) MCN and FA are correlated across sub-bands for most cell lines. The cumulative distribution of correlations is plotted with each point representing a cell line. C) tFA is correlated with the sum of the magnitudes of MCN deviations ($r=0.52$). The significance of this correlation was verified by permutation testing (10,000 random permutations of cell-line labels, $p = 0.0004$). D) The Pearson correlations of individual genes to total MCN-based aneuploidy and to tFA are correlated ($r = 0.66$). For each gene, the correlation between tFA and the expression level of the gene is plotted on the horizontal axis, and the correlation between total MCN-based aneuploidy and the expression level of the gene is plotted on the vertical axis. Each point represents a gene.

Although the observed correlation is significant, a perfect correlation between expression-based tFA and SNPchip-based total aneuploidy is unlikely for several reasons. First, while there is a significant ploidy regulation of gene expression in aneuploid tumors[2-4], the mean correlation between DNA copy number and gene expression levels is around 0.2-0.3[4]. Second, the distribution of cell-line based correlation (panel B) appears to be bi-modal, perhaps indicating that failed hybridizations for a subset of the cell-lines might be contaminating the analysis.

Figure S3. Total functional aneuploidy (tFA) is a significant predictor of clinical outcome in four breast, lung, and brain cancer datasets[5-8] of 18 datasets evaluated. Samples are categorized as having poor (good) prognosis if the level of tFA is above (below) the mean tFA across all samples in that cohort. Data sets are identified by the first author of the publication listed in the references of the main text. Significance was assessed using the log-rank test comparing survival curves corresponding to the two

prognosis classes.

Figure S4. Correlation between gene expression profiles and tFA is conserved in diverse human cancer datasets. The heat map shows normalized ranks of correlations for 10,151 genes (columns), ordered according to the final integrated ranking (see methods). Datasets (rows) are ordered by the mean difference between the normalized correlation ranks and the ranks of the integrated ranking. For each gene (vertical bar) the color indicates the ranking of the Pearson correlation coefficient between the gene and tFA across the data set. Yellow indicates highest level of correlation to tFA in a given dataset; blue indicates lowest (negative) correlation. The three datasets used to derive the integrated ranking are marked with asterisks. This figure is also available as an image file for better viewing.

Figure S5. Visualization of the CIN25 signature in eight datasets. Heat maps showing normalized expression profiles of the 25-gene CIN signature. Samples are ordered by total expression of genes in the CIN25 signature. Categorical prediction of outcome is determined by above average (bad) or below average (good) total CIN expression (the two categories are divided by the black line). Red and blue indicate over and under-expression of a gene, respectively.

Figure S6. The prognostic ability of cell-cycle regulated genes is dependent on CIN score.

A, The 764 genes with cell-cycle scores exceeding the false discovery rate (FDR) threshold (vertical blue line) were divided into four groups based on combined thresholds of CIN score and cell-cycle score. A high cell-cycle score threshold (vertical black dashed line) was determined by removing genes having the highest cell-cycle score from the CC_highCIN group until the average cell-cycle score in this group was equal to that of the CC_lowCIN group.

B, The prognostic ability of the four groups was compared across 18 independent cohorts. Each bar represents the log-rank test statistic comparing the survival curves corresponding to the two prognosis classes generated using the indicated gene group for outcome prediction. While the highCIN_midCC group (yellow) significantly risk-stratified nine of the eighteen cohorts, the lowCIN_CC group (turquoise, F5A) only did so for five (F5B). Because these gene-sets were constructed to have equal average cell-cycle scores, we conclude that the difference in performance was due to the difference in CIN-score between the groups.

Although the highCIN_highCC group (orange) significantly risk-stratified eleven of the eighteen cohorts, we noted that this group contained higher CIN-score genes than the highCIN_midCC group. Furthermore, we noted that the CIN70_lowCC group was able to risk-stratify ten of the eighteen cohorts (Figure 5B), suggesting that cell-cycle score is relevant for prediction primarily due to the fact that many of the genes with the highest cell-cycle scores also had very high CIN-scores.

The seven genes in the lowCIN_highCC group (green) significantly risk-stratified five of the eighteen cohorts, whereas the the lowCIN_midCC group (salmon) only did so for two. This might indicate that the predictive ability of the lowCIN_highCC group is indeed due to detection of increased proliferation, as the lowCIN_midCC group is likely to include a higher number of false-positive cell-cycle genes.

Figure S7. Multivariate analysis of the CIN25 and proliferation signatures revealed that CIN25 was generally more relevant for risk-stratification of cancer cohorts.

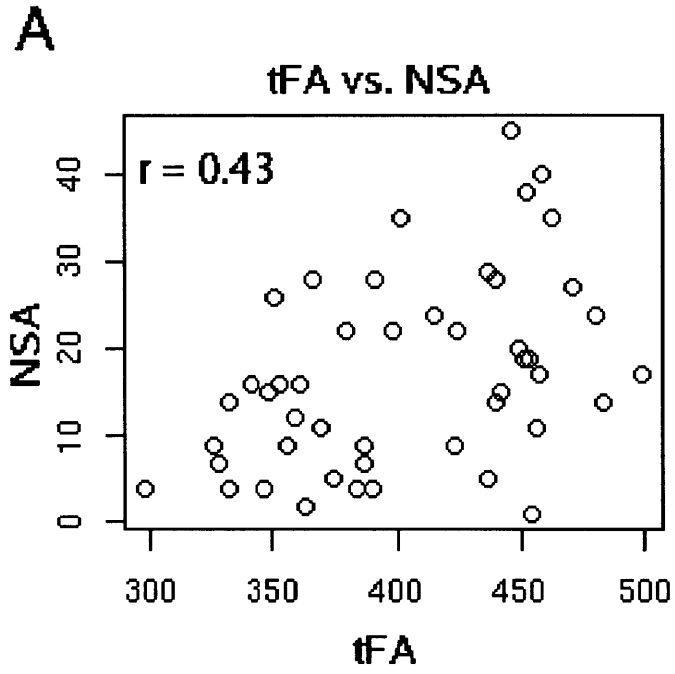
Multivariate analysis was performed using a Cox proportional hazards model built from the CIN25 and proliferation signatures. Shown are bars representing the -log of *P* values reflecting the significance of each signature's contribution to the model.

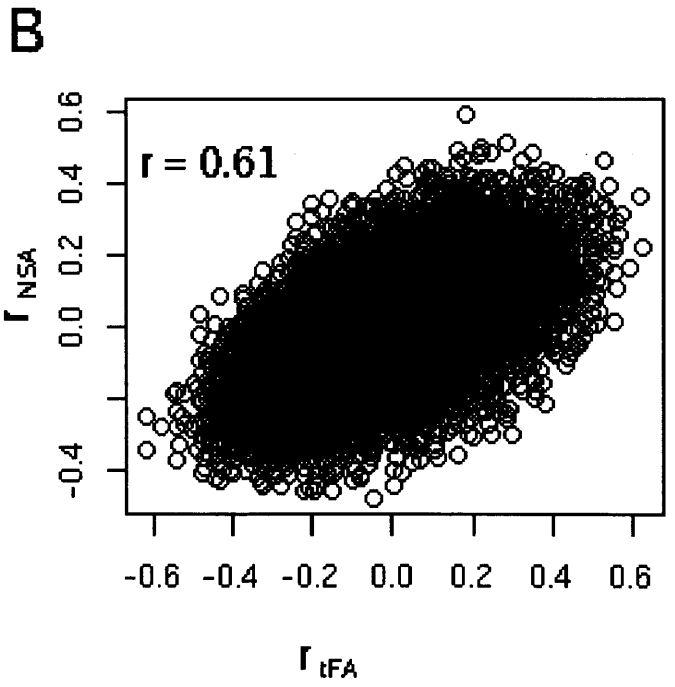
In five of eighteen cohorts evaluated, the CIN25 signature was significant in the Cox model given the proliferation signature. The proliferation signature was not significant given the CIN25 signature in any of the cohorts. The combined signatures significantly risk-stratified eleven of the eighteen cohorts (data not shown).

Figure S8. Removal of proliferation-associated genes from the CIN signature does not impair its predictive ability for clinical outcome. A) The correlation coefficient between each gene and the population doubling time of NCI60 cell lines was calculated across the cell line panel. The cumulative distribution of these correlations is shown for the entire gene list, for the CIN25 signature and for the set of 8 genes of the CIN signature without significant correlation with the population doubling time (CIN_no_dt). The cumulative distribution of correlations after permuting the cell line labels is shown in pink. B) Patients in each clinical data set were divided into two groups based on the average log expression of the CIN25 signature or the CIN_no_dt gene list. Each bar represents the log-rank test statistic comparing the survival curves corresponding to the two prognosis classes generated using the indicated gene group for outcome prediction. The CIN signature and CIN_no_dt signatures show similar power to predict clinical outcome. C) Cumulative distribution of the CIN score for the genes contained in the proliferation signature[9]. D) The proliferation signature genes were divided into two sets based on their CIN scores. Patients in each clinical data set were divided into two groups based on the average log expression of the proliferation signature genes with a high CIN score (prolif_highCIN) or the proliferation signature genes with a low CIN score (prolif_lowCIN). Each bar represents the log-rank test statistic comparing the survival curves corresponding to the two prognosis classes generated using the indicated gene group for outcome prediction. Proliferation signature genes with a low CIN score produced successful classification of 5 data sets (at a significance level of $p=0.05$) as opposed to the 10 data sets correctly classified by the full proliferation signature. These results suggest that the ability of the proliferation signature genes to predict clinical outcome is dependent in part on its inclusion of high CIN-score genes. Furthermore, the ability of CIN genes to predict the clinical outcome of various solid tumors is mainly due to the fact that they were derived to characterize the level of chromosomal aberrations and not due to their association with the proliferation rate of the cells.

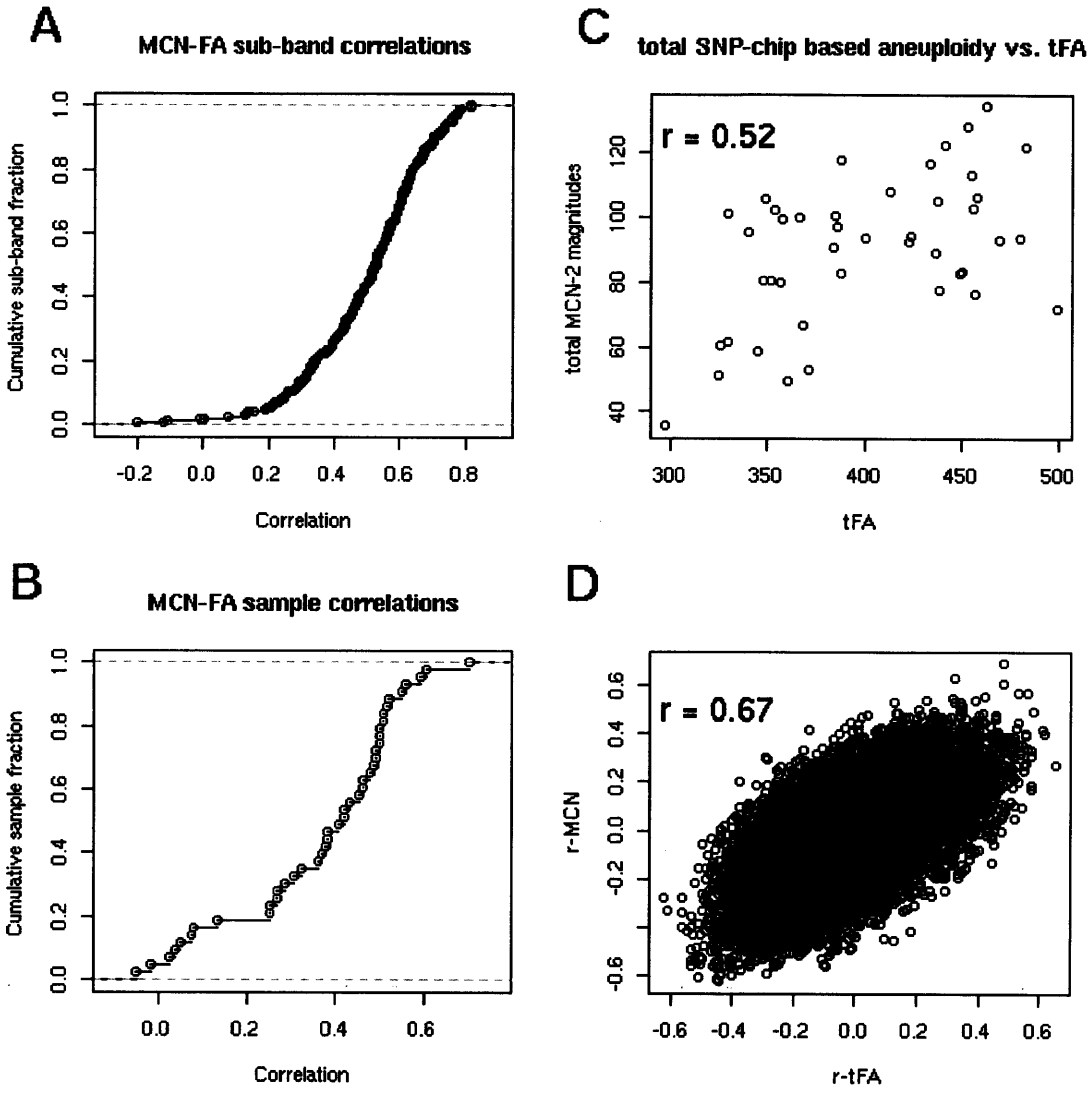
Figure S9. The CIN signature forms a coherent module such that its genes are more correlated with each other than with tFA. For each cancer dataset, we show the distribution of pair-wise correlations of genes in the CIN signature (red), as well as the distribution of correlations between the CIN genes and the tFA measure (black). The distributions of pair-wise correlations of those CIN genes without significant correlation with the doubling time in the NCI60 panel are shown in green. In most of the datasets the internal coherence of the CIN module exceeds the level of correlation to tFA.

Figure S10. The CIN signature does not generate significant predictions of clinical outcome for six of eighteen datasets evaluated. See legend for Figure 2.

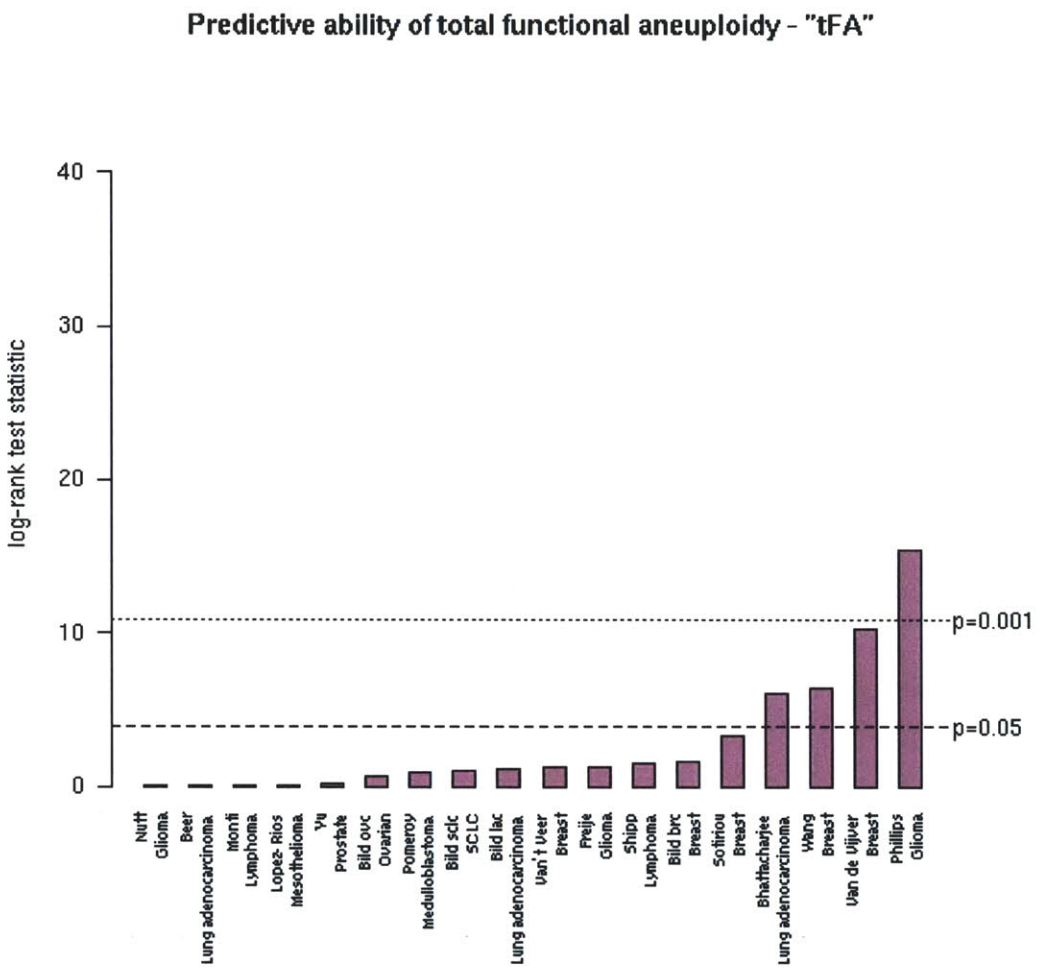

A
tFA vs. NSA
r = 0.43
NSA
0
10
20
30
40
300
350
400
450
500
tFA

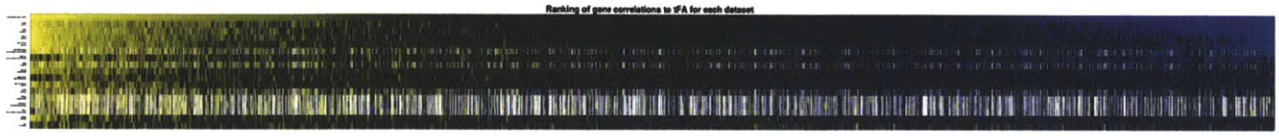

B
r = 0.61
r NSA
-0.4
0.0
0.2
0.4
0.6
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
r tFA

S2


A
MCN-FA sub-band correlations
Cumulative sub-band fraction
Correlation
B
MCN-FA sample correlations
Cumulative sample fraction
Correlation
C
total SNP-chip based aneuploidy vs. tFA
r = 0.52
total MCN-2 magnitudes
tFA
D
r = 0.67
r-MCN
r-tFA

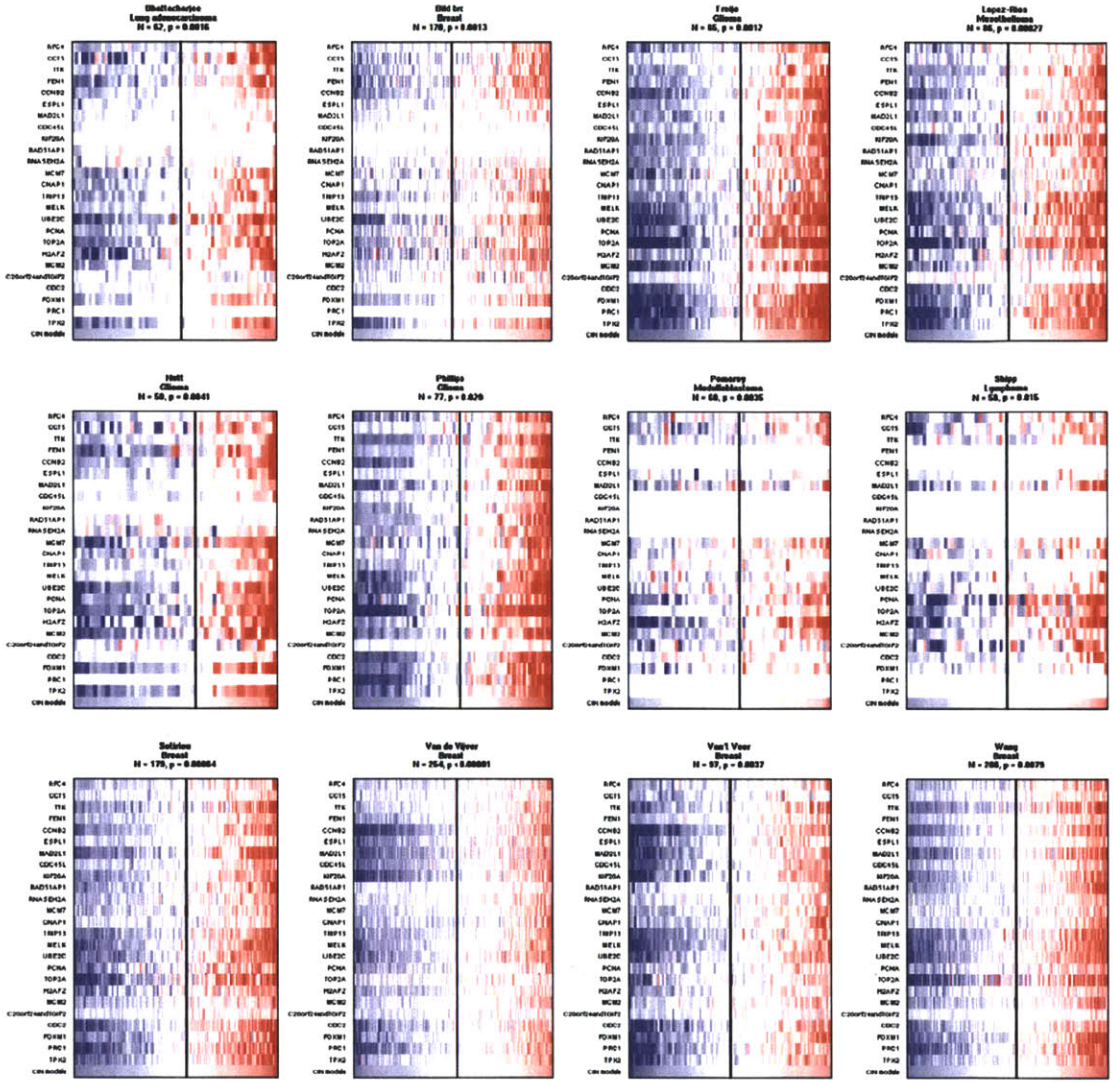
S3


Predictive ability of total functional aneuploidy - "tFA"
log-rank test statistic
40
30
20
10
0
p=0.001
p=0.05
Nutt
Glioma
Beer
Lung adenocarcinoma
Monfi
Lymphoma
Lopez- Rios
Mesothelioma
Yu
Prostate
Bild ovc
Ovarian
Pomeroy
Medulloblastoma
Bild scle
SCLC
Bild lac
Lung adenocarcinoma
Van 't Veer
Breast
Pleije
Glioma
Shipp
Lymphoma
Bild brc
Breast
Sotiriou
Breast
Bhattacharjee
Lung adenocarcinoma
Wang
Breast
Van de Vijver
Breast
Phillips
Glioma

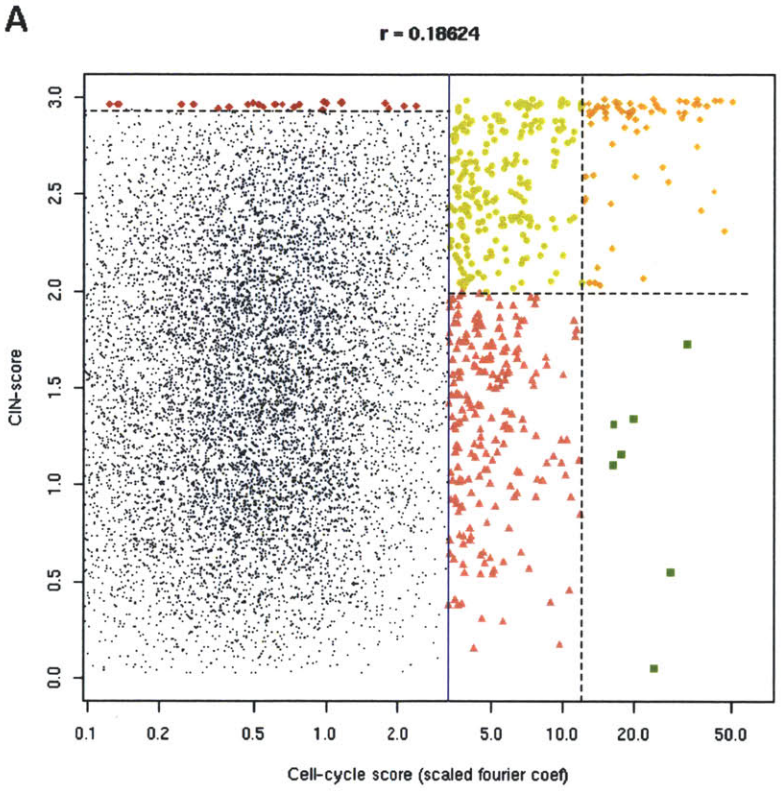


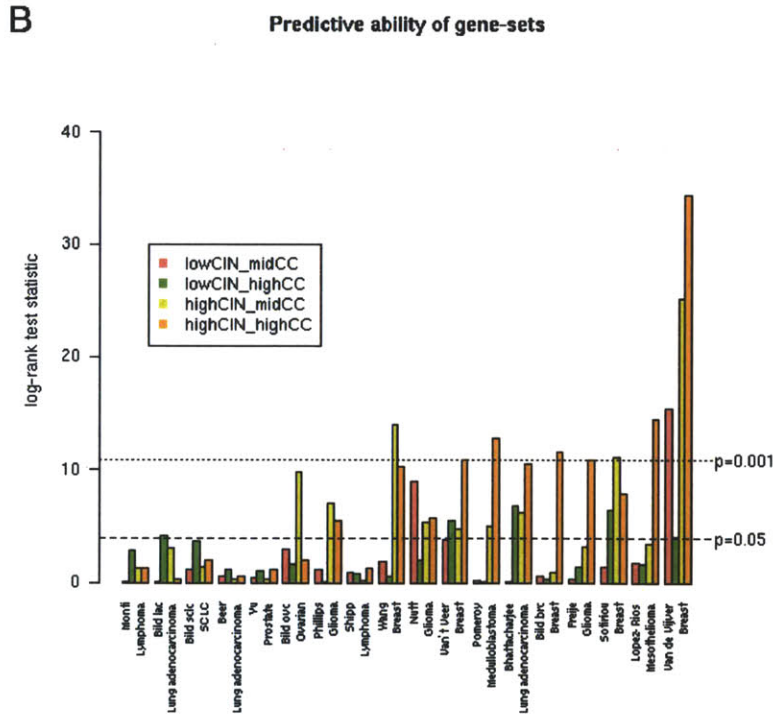
S4

S5

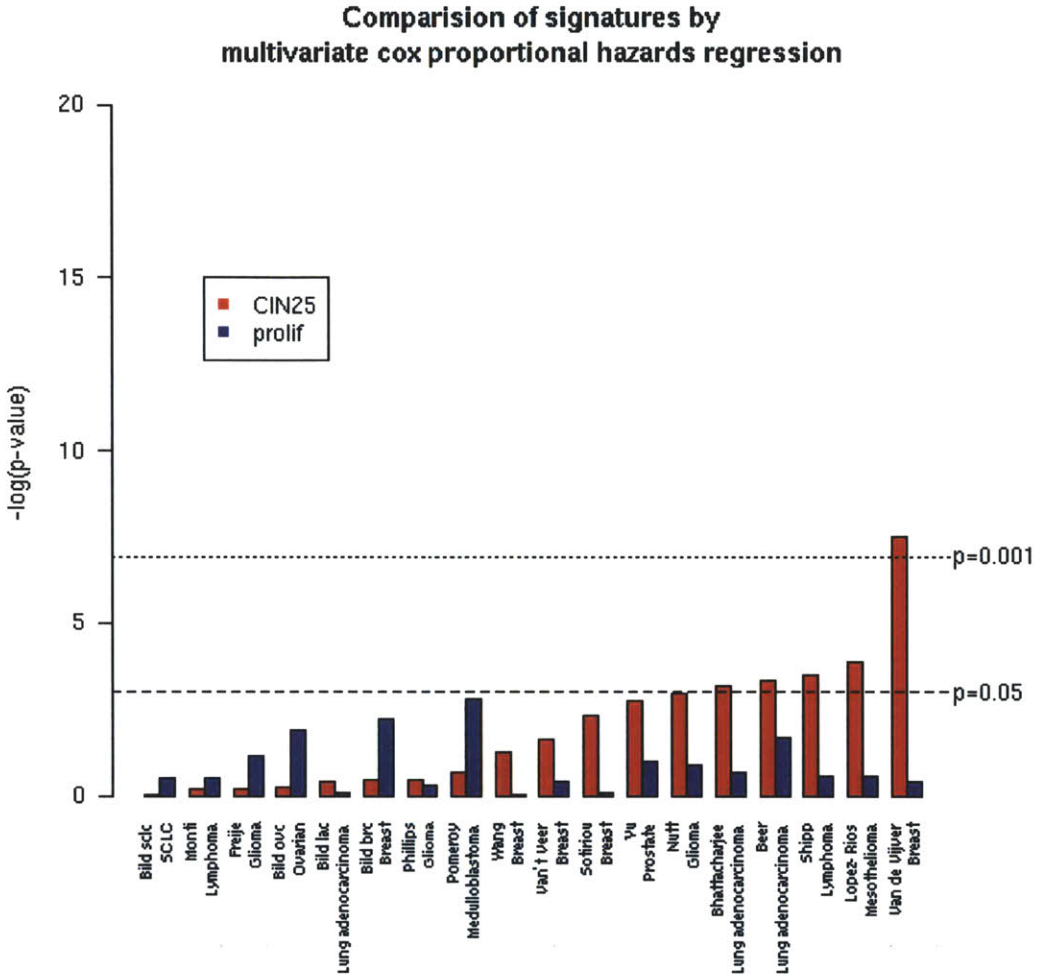

Bhattacharjee
Lung adenocarcinoma
N = 57, p = 0.0016
Bild Inc
Breast
N = 178, p = 0.0013
Freije
Glioma
N = 85, p = 0.0012
Lopez-Ibor
Mesothelioma
N = 86, p = 0.00027
Natt
Glioma
N = 50, p = 0.0041
Phillips
Glioma
N = 77, p = 0.029
Pomeroy
Medulloblastoma
N = 60, p = 0.0035
Shipp
Lymphoma
N = 58, p = 0.015
Sotiriou
Breast
N = 175, p = 0.00054
Van de Vijver
Breast
N = 254, p < 0.00001
Van't Veer
Breast
N = 97, p = 0.0037
Wang
Breast
N = 286, p = 0.0079

S6

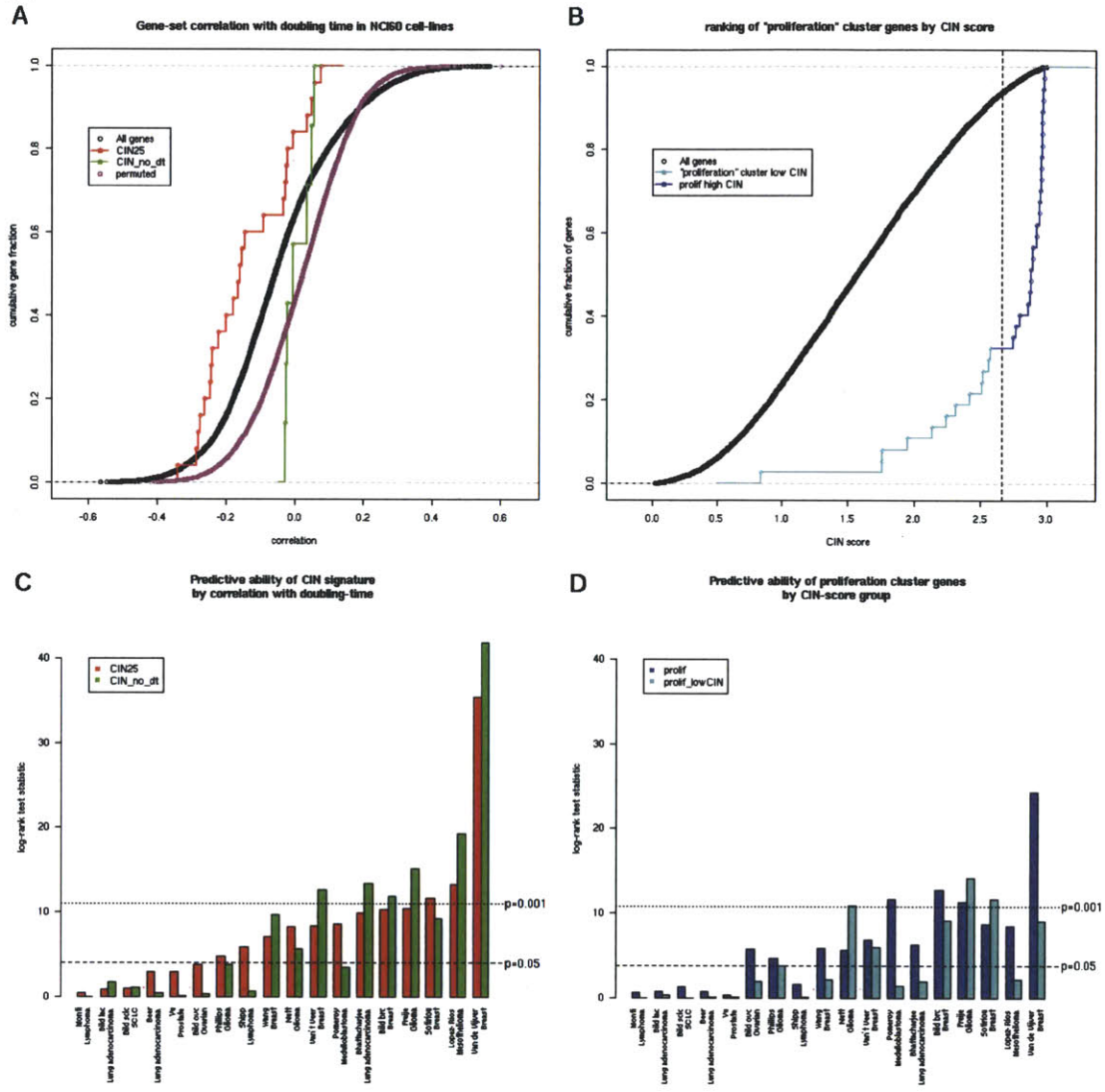

A
r = 0.18624
CIN-score
Cell-cycle score (scaled fourier coef)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.1
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0


B
Predictive ability of gene-sets
log-rank test statistic
lowCIN_midCC
lowCIN_highCC
highCIN_midCC
highCIN_highCC
p=0.001
p=0.05
0
10
20
30
40
Ivshina
Lymphoma
Bild lac
Lung adenocarcinoma
Bild scl
SCLC
Beer
Lung adenocarcinoma
Yu
Prostate
Bild ovc
Ovarian
Phillips
Glioma
Shipp
Lymphoma
Wang
Breast
Natt
Glioma
Van't Veer
Breast
Pomeroy
Medulloblastoma
Bhattacharjee
Lung adenocarcinoma
Bild brc
Breast
Freije
Glioma
Sotiriou
Breast
Lopez-Rios
Mesothelioma
Van de Vijver
Breast

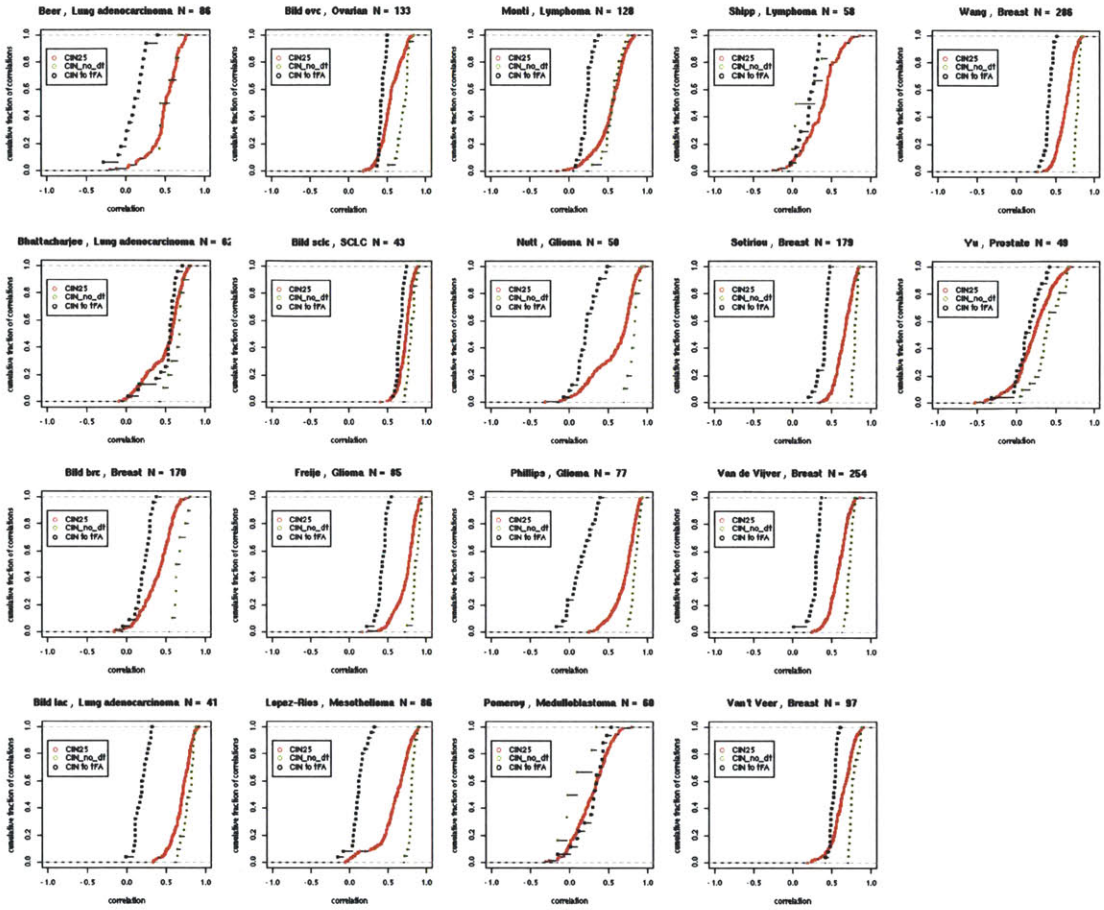
S7


Comparision of signatures by
multivariate cox proportional hazards regression
-log(p-value)
20
15
10
5
0
CIN25
prolif
p=0.001
p=0.05
Bild scic SCLC
Monti Lymphoma
Freije Glioma
Bild ovc Ovarian
Bild lac Lung adenocarcinoma
Bild brc Breast
Phillips Glioma
Pomeroy Medulloblastoma
Wang Breast
Van t Veer Breast
Sotiriou Breast
Yu Prostate
Nutt Glioma
Bhattacharjee Lung adenocarcinoma
Beer Lung adenocarcinoma
Shipp Lymphoma
Lopez-Rios Mesothelioma
Van de Vijver Breast

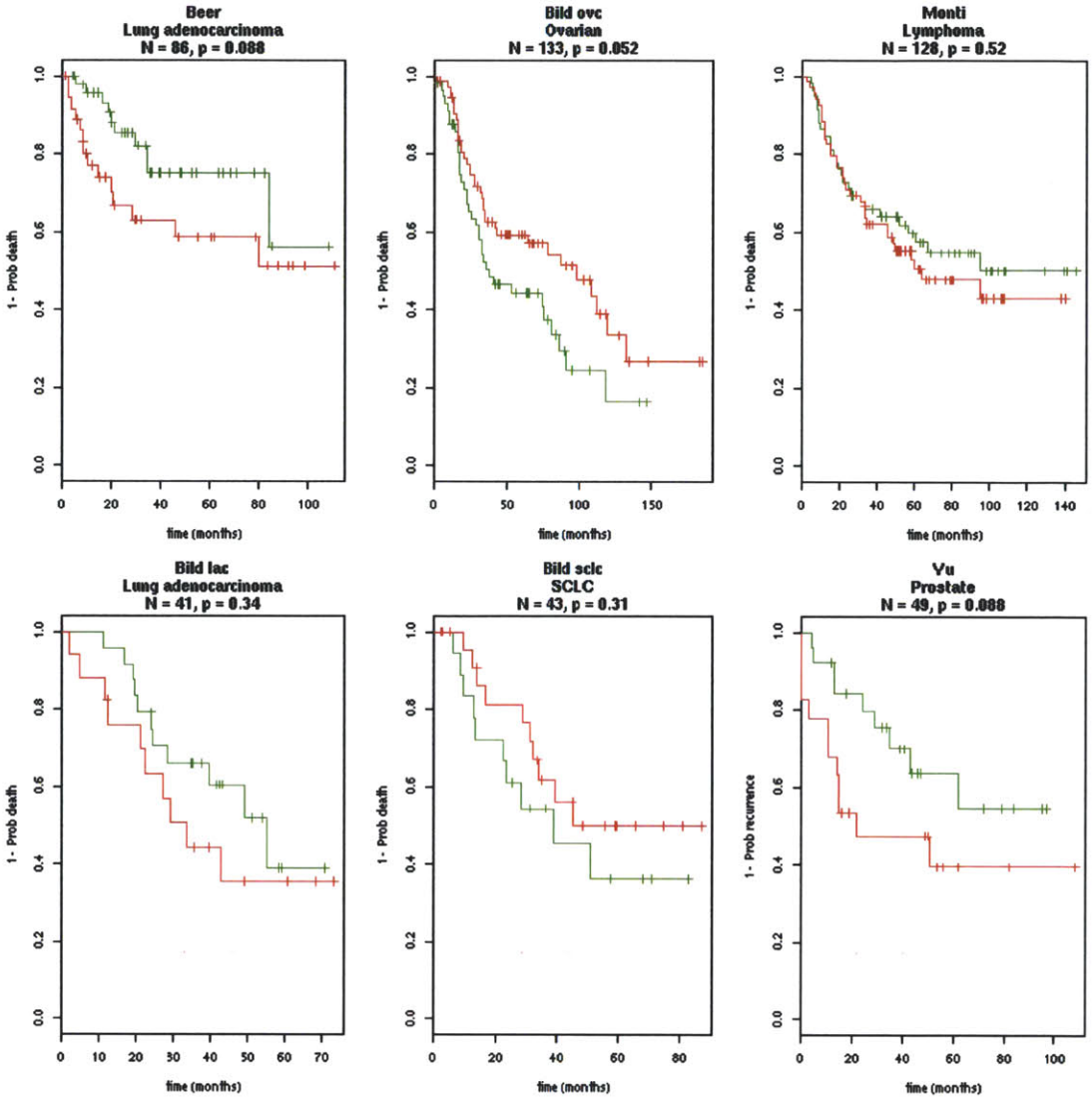
S8


A
Gene-set correlation with doubling time in NCI60 cell-lines
All genes
CIN25
CIN_no_dt
permuted
cumulative gene fraction
correlation
B
ranking of "proliferation" cluster genes by CIN score
All genes
"proliferation" cluster low CIN
prolif high CIN
cumulative fraction of genes
CIN score
C
Predictive ability of CIN signature
by correlation with doubling-time
CIN25
CIN_no_dt
log-rank test statistic
p=0.001
p=0.05
D
Predictive ability of proliferation cluster genes
by CIN-score group
prolif
prolif_lowCIN
log-rank test statistic
p=0.001
p=0.05

S9


Beer , Lung adenocarcinoma N = 86
Bild ovc , Ovarian N = 133
Monti , Lymphoma N = 128
Shipp , Lymphoma N = 58
Wang , Breast N = 286
Bhattacharjee , Lung adenocarcinoma N = 62
Bild sclc , SCLC N = 43
Nutt , Glioma N = 50
Setiriou , Breast N = 179
Yu , Prostate N = 49
Bild brc , Breast N = 170
Freije , Glioma N = 85
Phillips , Glioma N = 77
Van de Vijver , Breast N = 254
Bild lac , Lung adenocarcinoma N = 41
Lopez-Rios , Mesothelioma N = 86
Pomeroy , Medulloblastoma N = 60
Van't Veer , Breast N = 97
CIN25
CIN_no_df
CIN to tFA
cumulative fraction of correlations
correlation

S10


Beer
Lung adenocarcinoma
N = 86, p = 0.088
1 - Prob death
time (months)
Bild ovc
Ovarian
N = 133, p = 0.052
1 - Prob death
time (months)
Menti
Lymphoma
N = 128, p = 0.52
1 - Prob death
time (months)
Bild lac
Lung adenocarcinoma
N = 41, p = 0.34
1 - Prob death
time (months)
Bild sclc
SCLC
N = 43, p = 0.31
1 - Prob death
time (months)
Yu
Prostate
N = 49, p = 0.088
1 - Prob recurrence
time (months)

References mentioned in the supplementary figure legends:

1. Roschke, A.V., Tonon, G., Gehlhaus, K.S., McTyre, N., Bussey, K.J., Lababidi, S., Scudiero, D.A., Weinstein, J.N. & Kirsch, I.R. Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer Res* **63**, 8634-47 (2003).
2. Wolf, M. et al. High-resolution analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression. *Neoplasia* **6**, 240-7 (2004).
3. Tonon, G. et al. High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A* **102**, 9625-30 (2005).
4. Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L. & Brown, P.O. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* **99**, 12963-8 (2002).
5. Bhattacharjee, A. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**, 13790-5 (2001).
6. Wang, Y. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).
7. van de Vijver, M.J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).
8. Phillips, H.S. et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157-73 (2006).
9. Whitfield, M.L., George, L.K., Grant, G.D. & Perou, C.M. Common markers of proliferation. *Nat Rev Cancer* **6**, 99-106 (2006).

Supplementary Table 1

The top 70 genes with the highest levels of consistent correlation with tFA in 3 cancer associated data sets were determined as described in the supplementary methods. The genes are listed in descending order in terms of their correlation coefficients with tFA. The 28 genes with known functional association with CIN are underlined and the appropriate references are listed.

1. TPX2	is required for normal spindle morphology and centrosome integrity[1] its overexpression provokes polyploidization[2] it is required for targeting Aurora-A kinase to the spindle apparatus[3]
2. PRC1	critical to the formation of the central spindle[4] contributes to the correct formation of the spindle during the metaphase[5]
3. FOXM1	is a key transcriptional regulator of genes involved in chromosomal segregation and cytokinesis[6]
4. CDC2	play important regulatory roles in cell cycle control. plays an important role in the proper timing of mitotic spindle formation[7]
5. C20orf24/TGIF2	
6. MCM2	involved in the initiation of eukaryotic genome replication
7. H2AFZ	this particular histone is required for embryonic development
8. TOP2A	plays an essential catalytic role in chromosome segregation[8]
9. PCNA	is involved in the RAD6-dependent DNA repair pathway
10. UBE2C	is required for the destruction of mitotic cyclins
11. MELK	is a potential regulator of the G2/M progression and may act

	antagonistically to the CDC25B phosphatase
12. TRIP13	
13. CNAP1	is responsible for targeting CNAP1, and possibly condensin, to mitotic chromosomes.[9]
14. MCM7	involved in the initiation of eukaryotic genome replication
15. RNASEH2A	shows increased activity during DNA replication
16. RAD51AP1	
17. KIF20A	is required for cytokinesis[10] Relocation of the Aurora B/INCENP/survivin passenger protein complex requires KIF20A[11]
18. CDC45L	required to the initiation of DNA replication
19. MAD2L1	is a component of the mitotic spindle assembly checkpoint that prevents the onset of anaphase until all chromosomes are properly aligned at the metaphase plate p31(comet, formerly known as Cmt2) counteracts the function of Mad2 and is required for the silencing of the spindle checkpoint[12] reconstitution of Mad2 in Brca1deficient cells partially restored the spindle checkpoint[13] binds directly to CDC20[14]
20. ESPL1	is required for sister chromatid separation during mitosis in human cells [15]
21. CCNB2	is a member of the cyclin family and its overexpression results in CIN[16]
22. FEN1	causes repeat instability and aberrant DNA repair
23. TTK	is required for centrosome duplication and for the normal progression of mitosis[17]

	is a critical regulator of genetic stability, it is required for the spindle assembly checkpoint[18]
24. CCT5	
25. RFC4	Mutation in this replication factor leads to aberrant checkpoint control in response to damage to chromosomes[19] or DNA damage.
26. ATAD2	
27. ch-TOG	has roles during mitosis in focusing MT minus ends at spindle poles, maintaining centrosome integrity, and contributing to spindle bipolarity[20]
28. NUP205	
29. CDC20	plays a role in chromosome separation[21]
30 CKS2	binds to the catalytic subunit of the cyclin dependent kinases
31. RRM2	nzyme essential for the production of deoxyribonucleotides prior to DNA synthesis in S phase
32. ELAVL1	
33. CCNB1	is a member of the cyclin family and its overexpression results in CIN[16]
34. RRM1	essential for the production of deoxyribonucleotides prior to DNA synthesis in S phase
35. AURKB	its aberrant expression leads to chromosomal instability.[22]
36. MSH6	
37. EZH2	deregulated expression of EZH2 is associated with loss of differentiation and development of poorly differentiated breast cancer in humans

38. CTPS	is important in the biosynthesis of phospholipids and nucleic acids, and plays a key role in cell growth
39. DKC1	The DKC1 protein binds to the box H + ACA small nucleolar RNAs and the RNA component of telomerase.
40. OIP5	
41. CDCA8	binds Survivin and INCENP in vitro, it is required for stability of the bipolar mitotic spindle[23].
42. PTTG1	its overexpression causes aneuploidy[24].
43. C10 or f3	
44. H2AFX	its aberrant expression leads to chromosomal instability[22]
45. CMAS	
46. BRRN1	is required for the conversion of interphase chromatin into condensed chromosomes[25]
47. MCM10	is required for proper chromosome condensation.
48. LSM4	
49. MTB	it is subunit of the condensin II complex implicated in chromosome assembly and segregation[26]
50. ASF1B	
51. ZWINT	is essential for mitotic checkpoint signaling[27].
52. TOPK	
53. FLJ10036	its aberrant expression leads to chromosomal instability[22]

54. CDCA3	cell cycle regulated gene
55. ECT2	plays a critical role in cytokinesis[28]
56. CDC6	a regulator at the early steps of DNA replication
57. UNG	
58. MTCH2	
59. RAD21	Scc1 is essential for the association of kinetochores with microtubules.[29]
60. ACTL6A	
61.	
62. SFRS2	
63. HDGF	
64. NXT1	
65. NEK2	is a kinetochore-associated protein kinase essential for faithful chromosome segregation[30]
66. DHCR7	
67. STK6	(Aurora-A) plays a role in the stabilization at the spindle pole during chromosome segregation. its overexpression leads to tetraploidization[31]
68. NDUFAB1	
69. KIAA0286	
70. KIF4A	it is a kinesin playing essential roles in the organization of central

spindles[32]

1. Garrett, S., Auer, K., Compton, D.A. & Kapoor, T.M. hTPX2 is required for normal spindle morphology and centrosome integrity during vertebrate cell division. *Curr Biol* **12**, 2055-9 (2002).
2. Heidebrecht, H.J., Adam-Klages, S., Szczepanowski, M., Pollmann, M., Buck, F., Endl, E., Kruse, M.L., Rudolph, P. & Parwaresch, R. repp86: A human protein associated in the progression of mitosis. *Mol Cancer Res* **1**, 271-9 (2003).
3. Kufer, T.A., Sillje, H.H., Korner, R., Gruss, O.J., Meraldi, P. & Nigg, E.A. Human TPX2 is required for targeting Aurora-A kinase to the spindle. *J Cell Biol* **158**, 617-23 (2002).
4. Mollinari, C., Kleman, J.P., Jiang, W., Schoehn, G., Hunter, T. & Margolis, R.L. PRC1 is a microtubule binding and bundling protein essential to maintain the mitotic spindle midzone. *J Cell Biol* **157**, 1175-86 (2002).
5. Ban, R., Irino, Y., Fukami, K. & Tanaka, H. Human mitotic spindle-associated protein PRC1 inhibits MgcRacGAP activity toward Cdc42 during the metaphase. *J Biol Chem* **279**, 16394-402 (2004).
6. Costa, R.H. FoxM1 dances with mitosis. *Nat Cell Biol* **7**, 108-10 (2005).
7. Kramer, A., Mailand, N., Lukas, C., Syljuasen, R.G., Wilkinson, C.J., Nigg, E.A., Bartek, J. & Lukas, J. Centrosome-associated Chk1 prevents premature activation of cyclin-B-Cdk1 kinase. *Nat Cell Biol* **6**, 884-91 (2004).
8. Carpenter, A.J. & Porter, A.C. Construction, characterization, and complementation of a conditional-lethal DNA topoisomerase IIalpha mutant human cell line. *Mol Biol Cell* **15**, 5700-11 (2004).
9. Ball, A.R., Jr., Schmiesing, J.A., Zhou, C., Gregson, H.C., Okada, Y., Doi, T. & Yokomori, K. Identification of a chromosome-targeting domain in the human condensin subunit CNAP1/hCAP-D2/Eg7. *Mol Cell Biol* **22**, 5769-81 (2002).
10. Hill, E., Clarke, M. & Barr, F.A. The Rab6-binding kinesin, Rab6-KIFL, is required for cytokinesis. *Embo J* **19**, 5711-9 (2000).
11. Gruneberg, U., Neef, R., Honda, R., Nigg, E.A. & Barr, F.A. Relocation of Aurora B from centromeres to the central spindle at the metaphase to anaphase transition requires MKlp2. *J Cell Biol* **166**, 167-72 (2004).
12. Xia, G., Luo, X., Habu, T., Rizo, J., Matsumoto, T. & Yu, H. Conformation-specific binding of p31(comet) antagonizes the function of Mad2 in the spindle checkpoint. *Embo J* **23**, 3133-43 (2004).
13. Wang, R.H., Yu, H. & Deng, C.X. A requirement for breast-cancer-associated

gene 1 (BRCA1) in the spindle checkpoint. *Proc Natl Acad Sci U S A* **101**, 17108-13 (2004).

14. Luo, X., Tang, Z., Xia, G., Wassmann, K., Matsumoto, T., Rizo, J. & Yu, H. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nat Struct Mol Biol* **11**, 338-45 (2004).
15. Waizenegger, I., Gimenez-Abian, J.F., Wernic, D. & Peters, J.M. Regulation of human separase by securin binding and autocleavage. *Curr Biol* **12**, 1368-78 (2002).
16. Sarafan-Vasseur, N., Lamy, A., Bourguignon, J., Pessot, F.L., Hieter, P., Sesboue, R., Bastard, C., Frebourg, T. & Flaman, J.M. Overexpression of B-type cyclins alters chromosomal segregation. *Oncogene* **21**, 2051-7 (2002).
17. Fisk, H.A., Mattison, C.P. & Winey, M. Human Mps1 protein kinase is required for centrosome duplication and normal mitotic progression. *Proc Natl Acad Sci U S A* **100**, 14875-80 (2003).
18. Fisk, H.A., Mattison, C.P. & Winey, M. A field guide to the Mps1 family of protein kinases. *Cell Cycle* **3**, 439-42 (2004).
19. Krause, S.A., Loupart, M.L., Vass, S., Schoenfelder, S., Harrison, S. & Heck, M.M. Loss of cell cycle checkpoint control in Drosophila Rfc4 mutants. *Mol Cell Biol* **21**, 5156-68 (2001).
20. Cassimeris, L. & Morabito, J. TOGp, the human homolog of XMAP215/Dis1, is required for centrosome integrity, spindle pole organization, and bipolar spindle assembly. *Mol Biol Cell* **15**, 1580-90 (2004).
21. Acquaviva, C., Herzog, F., Kraft, C. & Pines, J. The anaphase promoting complex/cyclosome is recruited to centromeres by the spindle assembly checkpoint. *Nat Cell Biol* **6**, 892-8 (2004).
22. Gollin, S.M. Mechanisms leading to chromosomal instability. *Semin Cancer Biol* **15**, 33-42 (2005).
23. Gassmann, R., Carvalho, A., Henzing, A.J., Ruchaud, S., Hudson, D.F., Honda, R., Nigg, E.A., Gerloff, D.L. & Earnshaw, W.C. Borealin: a novel chromosomal passenger required for stability of the bipolar mitotic spindle. *J Cell Biol* **166**, 179-91 (2004).
24. Yu, R., Lu, W., Chen, J., McCabe, C.J. & Melmed, S. Overexpressed pituitary tumor-transforming gene causes aneuploidy in live human cells. *Endocrinology* **144**, 4991-8 (2003).
25. Oliveira, R.A., Coelho, P.A. & Sunkel, C.E. The condensin I subunit Barren/CAP-H is essential for the structural integrity of centromeric heterochromatin during mitosis. *Mol Cell Biol* **25**, 8971-84 (2005).

26. Ono, T., Losada, A., Hirano, M., Myers, M.P., Neuwald, A.F. & Hirano, T. Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell* **115**, 109-21 (2003).
27. Wang, H. et al. Human Zwint-1 specifies localization of Zeste White 10 to kinetochores and is essential for mitotic checkpoint signaling. *J Biol Chem* **279**, 54590-8 (2004).
28. Saito, S., Tatsumoto, T., Lorenzi, M.V., Chedid, M., Kapoor, V., Sakata, H., Rubin, J. & Miki, T. Rho exchange factor ECT2 is induced by growth factors and regulates cytokinesis through the N-terminal cell cycle regulator-related domains. *J Cell Biochem* **90**, 819-36 (2003).
29. Hoque, M.T. & Ishikawa, F. Cohesin defects lead to premature sister chromatid separation, kinetochore dysfunction, and spindle-assembly checkpoint activation. *J Biol Chem* **277**, 42306-14 (2002).
30. Lou, Y., Yao, J., Zereshki, A., Dou, Z., Ahmed, K., Wang, H., Hu, J., Wang, Y. & Yao, X. NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checkpoint signaling. *J Biol Chem* **279**, 20049-57 (2004).
31. Meraldi, P., Honda, R. & Nigg, E.A. Aurora-A overexpression reveals tetraploidization as a major route to centrosome amplification in p53-/- cells. *Embo J* **21**, 483-92 (2002).
32. Kurasawa, Y., Earnshaw, W.C., Mochizuki, Y., Dohmae, N. & Todokoro, K. Essential roles of KIF4 and its binding partner PRC1 in organized central spindle midzone formation. *Embo J* **23**, 3237-48 (2004).
33. Oei, S.L., Keil, C. & Ziegler, M. Poly(ADP-ribosylation) and genomic stability. *Biochem Cell Biol* **83**, 263-9 (2005).

Appendix B

—

Markov chain monte carlo simulations for tumor purity/ploidy inference

Scott L. Carter and Gad Getz

Genomic DNA extracted from primary solid tumor samples is derived from a mixture of cancer cells having abnormal genomes, and various other genomically normal cells associated with the tumor. The relative concentration of a given DNA locus in this mixture is directly related to it's copy number in the cancer cell population, the total amount of DNA present in each cancer cell (ploidy), and the percentage of normal cells present. Because the fraction of tumor cells and the tumor ploidy in a given sample are generally unknown, much of copy number analysis to date has been restricted to considering relative changes within a given sample. Substantial variation and bi-modality exists in the ploidy of several common solid tumor types, making direct comparisons of lesions across tumor-sets difficult. Here we consider the problem of jointly inferring the sample purity, tumor ploidy, and absolute copy-number levels in DNA samples isolated from primary solid tumors. We adopt a Bayesian approach throughout, which allows consistent treatment of the many sources of inherent uncertainty, and which naturally accommodates multiple possible solutions.

1 Basic considerations

We presume that the raw data have been normalized appropriately and fitted to a piecewise-constant regression function (ie., a segmentation of the genome), which makes use of the assumption that nearby loci in the (normal) genome are likely to have the same CN level. This is the only point in our analysis in which genomic locations are used. A substantial amount of technical noise is typically removed from the data by this procedure.

We express the relation between the observed log copy number ratio X and the absolute copy number q as $X = \Delta q + b$. The quantities b and Δ are directly related to the unknown purity: α, and ploidy: τ of the tumor sample by:

$$b = \frac{2(1-\alpha)}{\alpha\tau + 2(1-\alpha)}, \Delta = \frac{\alpha}{\alpha\tau + 2(1-\alpha)} \tag{1}$$

Where $\alpha \in [0, 1]$, with $\alpha = 1$ corresponding to a 'pure' sample, consisting only of cancer cells. The cancer-cell ploidy τ is parameterized such that genomically normal (diploid) cells correspond to $\tau = 2$. A description of this derivation can be found in Chapter 4 of this thesis (eqn. 1).

This derivation assumes that each cancer cell present in the sample is genomically identical. While this is almost certainly not the case in a realistic tumor sample, we consider that the majority of genomic loci will be present at the same absolute copy number level in each cancer cell. We furthermore assume that each non-cancer cell present in the sample is genomically normal. While this has been somewhat controversial, recent studies support the validity of this assertion [Ref Polyak NatGen '08].

2 Graphical model

The observed data consists of N segments having relative CN log-ratio X and mass W, where $\sum_{i=1}^{N} W_i = 1$. The individual segments X are presumed to be generated from an unknown number of cluster centers μ_c, corresponding to discrete genomic copy-number levels, with noise σ constant in

a given sample (for now). The values of the μ_c are determined by the absolute cluster copy-number Z_c, the sample purity α, and the tumor ploidy τ (eq. 1).

All of the information about the values of α and τ is contained in the cluster-center positions μ_c. Furthermore, the 'weight' of the evidence from a given cluster is independent of the cluster mass W_c. For example, CN-level 0 (homozygous deletion) may be present at genomic locations corresponding to only 25 probes (out of 2.5×10^5, say). However, if this hypothetical genome has copy numbers 0-6 present, then these 25 probes ($\frac{1}{10000}th$ of the raw data) represent $\frac{1}{7}th$ of the evidence for α and τ.

Accurate inference on α and τ therefore requires that the individual segments be grouped into the correct number of CN-level clusters, and that the cluster positions μ_c are correctly determined. Unfortunately, it is generally impossible to know these quantities in advance. It is therefore necessary to integrate over uncertainty regarding groupings of observed segments into CN-level clusters. This is accomplished by jointly sampling C, μ_c and $W_c = \sum W_i 1\{c_i = C\}$ from a Dirichlet process:

$$P(C, \mu_c, W_c | X, \Theta) = DP(X, W, \gamma, \Theta)$$

Where γ is the concentration parameter of the DP. The fact that the data points X have differing associated masses W renders the DP somewhat non-standard; the details are left to Section 3.

With a clustering of segments: $\{C, \mu_c, W_c\}$ in hand, our task is to fit them to Q unknown absolute CN-levels $Z_c \in 0, 1, \ldots, Q-1$. It is crucial in this formulation to apply a complexity penalty over the number of occupied levels (Q). Currently, we use a penalty: $N * -\log(span)$, where N is the number of segments in a sample and *span* is the difference between the maximum and minimum genome level indices. Also, assignment of total-copy values of 0 (i.e., homozygous deletion) is severely penalized. Note that the model presented in Chapter 4 of this thesis is free from this defect, as it optimizes the weights of each mixture component to fit the data. This procedure makes extra levels irrelevant, since they receive mixing weights of 0.

$$\begin{aligned} P(Z_c, \alpha, \tau | C, \mu_c, \Theta) &= \frac{P(\mu_c | Z_c, \alpha, \tau, \Theta) P(Z_c, \alpha, \tau | \Theta)}{P(\mu_c | \Theta)} \\ &= \frac{P(\mu_c | Z_c, \alpha, \tau) P(Z_c | \Theta) P(\alpha | \cdot) P(\tau | \cdot)}{P(\mu_c | \Theta)} \\ &= \frac{P(\alpha | \cdot) P(\tau | \cdot) \prod^C P(\mu_c | Z_c, \alpha, \tau)}{P(\mu_c | \Theta)} \end{aligned}$$

We use the notation $P(\alpha|\cdot)$ to emphasize the dependence on a prior which we wish to avoid specifying at this time. Assignment of the Z_c to CN levels $\{0 : Q-1\}$ is complicated by the requirement that the configuration must by definition satisfy: $\sum^C Z_c W_c = \tau$ (i.e., the ploidy of the cancer genome must equal the sum of it's parts). Furthermore, we constrain the dispersion of mass around τ such that the standard deviation scales with $\tau\kappa_\tau$. This constraint imposes a structural prior reflecting our biological intuition that higher-ploidy genomes are likely to populate a greater number of CN levels. The constraints $\mathbf{C_Z}$ on a micro-configuration of segment-clusters $\{Z_c\}$, conditional on an unknown κ_τ, are therefore:

$$\mathbf{C_Z} = \delta\left(\sum^C z_c w_c - \tau\right) \delta\left(\left(\sum^C z_c w_c - \tau\right)^2 - \tau\kappa_\tau\right) \quad (2)$$

These constraints imply dependence between assignments of segment-clusters to absolute CN-levels, making direct Gibbs sampling intractable. Re-parameterization to an alternate ensemble in which the mass center and dispersion are distributionally constrained by τ would simplify inference.

The maximum-entropy density over Z_c subject to the moment-constraints given in (2) is:

$$P(Z_c|W_c, \Lambda_\tau) = \frac{e^{-\lambda_1 W_c Z_c - \lambda_2 W_c (Z_c - \tau)^2}}{\sum_{Z_c=0}^{Q-1} e^{-\lambda_1 W_c Z_c - \lambda_2 W_c (Z_c - \tau)^2}}$$

Values of the Lagrange multipliers $\Lambda_\tau = \{\lambda_1, \lambda_2\}$ are readily determined via Nelder-Mead optimization of L2 loss, such that:

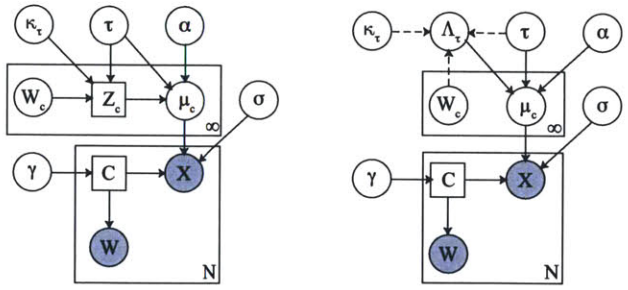
$$\left\langle \sum^{C} Z_c W_c \right\rangle_{\mathcal{Q}} = \sum^{C} W_c \langle Z_c \rangle_{\mathcal{Q}} = \tau$$

$$\sum^{C} W_c \left\langle (Z_c - \tau)^2 \right\rangle_{\mathcal{Q}} = \tau \kappa_\tau$$

Where $\langle \cdot \rangle_{\mathcal{Q}}$ denotes the average over all configurations $\{Z_c\}$, weighted by the function $\mathcal{Q} = P(Z_c|W_c, \Lambda_\tau)$.

We may then write:

$$\begin{aligned} P(\alpha, \tau|\mu_c, \Theta) &= \sum_{\{Z_c\}} \mathbf{C_Z} P(Z_c, \alpha, \tau|\mu_c, \Theta) \\ &= \sum_{\{Z_c\}} \left(\frac{P(\alpha|\cdot)P(\tau|\cdot)}{P(\mu_c|\Theta)} \mathbf{C_Z} \prod^{C} (\mu_c|z_c, \tau, \alpha) \right) \\ &\approx \frac{P(\alpha|\cdot)P(\tau|\cdot)}{P(\mu_c|\Theta)} \prod^{C} \left(\sum_{Z_c=0}^{Q-1} P(z_c|w_c, \Lambda_\tau) P(\mu_c|z_c, \tau, \alpha) \right) \end{aligned}$$



(a) Microcannonical model

(b) Variational approximation

Figure 1: Graphical models with plate notation summarizing the conditional dependencies in the model. Blue circles denote observed data. Dashed lines denote deterministic relationships.

This approximation allows us to replace the sum of products with a product of sums in the factorized likelihood, allowing independent summation over possible assignments of each segment-cluster to a Z_c level. The Z_c may therefore be marginalized out of the likelihood:

$$
\begin{aligned}
P(\mu_c|Z_c, W_c, \Psi) &\approx \sum_{Z_c=0}^{Q-1} P(\mu_c|Z_c, \hat{\sigma}_c, \alpha, \tau) P(Z_c|W_c, \Lambda_\tau) \\
&= P(\mu_c|\Psi)
\end{aligned}
$$

For ease of notation, we define: $\Psi = \{\alpha, \tau, \hat{\sigma}_c, \Lambda_\tau\}$. Where $\hat{\sigma}_c$ is represents the standard error of a given cluster-center:

$$
\hat{\sigma}_c = \sqrt{\frac{\sum_i \frac{\left(X_{[c_i=c]} - \mu_c\right)^2 + \sigma^2_{x[c_i=c]}}{\sigma^2_{x[c_i=c]}}}{\sum_i \frac{1}{\sigma^2_{x[c_i=c]}}}}
$$

The data-liklihood may then be written as:

$$
P(X|\Theta) = \prod_{i=1}^{N} P\left(x_i|C, \mu_c, \sigma\right) \prod^{C} P(\mu_c|\Psi)
$$

Where a two-dimensional grid over α and τ allows easy calculation of:

$$
P(\alpha, \tau|\mu_c) = P(\alpha|\cdot) P(\tau|\cdot) \prod^{C} P(\mu_c|\Psi)
$$

$P(\mu_c)$ is then readily obtained via:

$$
\begin{aligned}
P(\mu_c) &\approx \int_\alpha \int_\tau P(\alpha|\cdot) P(\tau|\cdot) \prod^{C} P(\mu_c|\Psi) d\tau d\alpha \\
&\approx \sum^{\alpha_{grid}} \sum^{\tau_{grid}} P(\alpha|\cdot) P(\tau|\cdot) \prod^{C} P(\mu_c|\Psi)
\end{aligned}
$$

We use metropolis moves to jump between modes of the posterior, based on the equations for α and τ. i.e., we propose jumps like: $b' \leftarrow b + 1$, $\Delta' \leftarrow \Delta$, etc.

3 Dirichlet process

The standard Dirichlet process (described in [Antoniak 1974] and [West 1995]) is a method for drawing a sample from an infinite-dimensional space consisting of probability distributions over clusterings of a finite number N of data points.

Cluster assignments c_i are sampled for each datum conditional on the assignments of all other data-points $\mathbf{c}_{-i}$:

$$
P(c_i = c|\mathbf{c}_{-i}) \propto \begin{cases} \frac{\sum_{j \neq i} 1_{[c_j=c]}}{N-1+\gamma} \int F(x_i|\Theta) dH_{\mathbf{c}_{-i}}(\Theta) & \text{if } c = c_j \text{ for some } j \neq i; \\ \frac{\gamma}{N-1+\gamma} \int F(x_i|\Theta) dG_0(\Theta) & \text{otherwise.} \end{cases}
$$

Where G_0 is the base measure of Θ, and $H_{\mathbf{c}_{-i}}(\Theta)$ gives the posterior distribution of Θ given G_0 and all observations for which $j \neq i$ and $c_j = c$. In general, draws of Θ from the DP are not distinct, and the number of distinct values grows as $\mathcal{O}(\log N)$.

In our case, the individual data points X correspond to genomic segments of various sizes W. We wish to construct a DP in such a way as to be equivalent to the case where the data are observations of the individual probes, but where probes in the same segment always fall in the same cluster. A corollary of this requirement is that each sample genome should have the same expected number of CN clusters *under any genome segmentation* (marginal on X and γ). We therefore modify the clustering probabilities:

$$P(c_i = c|\mathbf{c}_{-i}) \propto \begin{cases} \frac{\sum_{j \neq i} W_{[c_j = c]}}{1 - W_i + \gamma} \int F(x_i|\Theta) dH_{\mathbf{c}_{-i}}(\Theta) & \text{if } c = c_j \text{ for some } j \neq i; \\ \frac{\gamma}{1 - W_i + \gamma} \int F(x_i|\Theta) dG_0(\Theta) & \text{otherwise.} \end{cases} \tag{3}$$

The concentration parameter γ plays an important role in determining the number of distinct clusters k. We wish to specify the prior distribution on the expected number of clusters k_0 in a given sample to be independent of the segment sizes W. Because we generally have strong biological intuition for reasonable values of k, we can specify a relatively tight prior on k_0. This will be important for the development of a more flexible error model.

In order handle uncertainty in γ, we must be able to calculate $P(k|\gamma, N, W)$, the probability (marginal on X) that the N segments of mass $W_{[i \ldots N]}$ form k distinct clusters, given concentration parameter γ. This calculation differs from that of the standard DP in which the W all equal 1. Our task is to enumerate the multiplicities of a given configuration: $\{k, N, W\}$. Normalization by the total multiplicities over $k \in \{1 \ldots N\}$ then gives the sought probability. Following [Antoniak 1974], we define a sequence of polynomials:

$$\begin{aligned} M_1(x) &= (x + 0) \\ M_2(x) &= (x + w_1)\, M_1(x) = (x + 0)\,(x + w_1) \\ &\vdots \\ M_n(x) &= \left(x + \sum_{i=1}^{N-1} w_i\right) M_{n-1}(x) = x(x + w_1) \ldots \left(x + \sum_{i=1}^{N-1} w_i\right) \end{aligned}$$

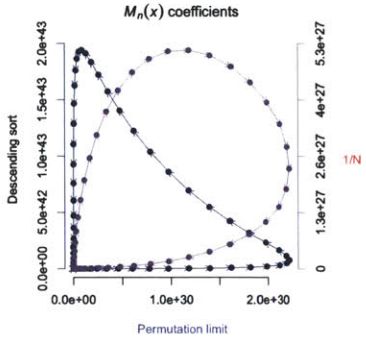
$M_n(x)$ is therefore a polynomial of degree N in x:

$$M_n(x) = {}_n m_1 x + {}_n m_2 x^2 + \cdots + {}_n m_n x^n$$

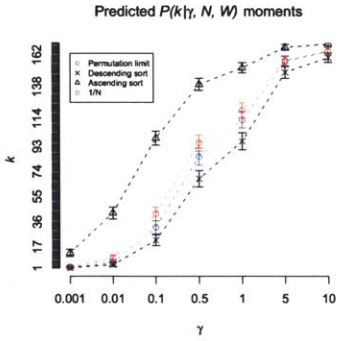
In the standard DP where all $W_{[1 \ldots N]} = 1$, the coefficients ${}_n m_{[1 \ldots n]}$ are the unsigned Stirling numbers of the first kind, and

$$P(k|\gamma, N) = \frac{{}_n m_k \gamma^k}{M_n(\gamma)} \text{[Antoniak 1974].} \tag{4}$$

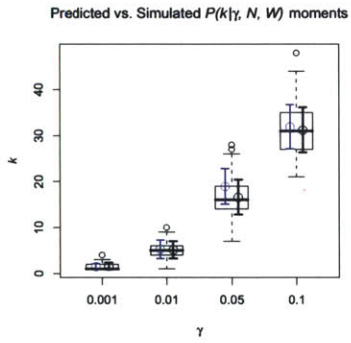
Unfortunately, our modification of the $M_n(x)$ polynomials to take into account the different segment masses has the problem that the values of the coefficients are clearly dependant on a given ordering of the $W_{[1 \ldots N]}$. Because the DP is infinitely exchangeable, all permutations of W lead to the same equilibrium distribution. We therefore compute the average $M_n(x)$ coefficients under successive permutations of W, until approximate convergence is reached.



(a) Coefficients of $M_n(x)$ generated from a descending sort of W vs. the permutation limit of W (dark blue), and from uniform values of $W_{1\ldots N} = 1/N$ vs. the permutation limit of W (purple). Arrows indicate the ordering of the points corresponding to ${}_nm_1 \ldots {}_nm_n$.



(b) Expectation and standard deviation of $P(k|\gamma, N, W)$ obtained from eq. 4 using various coefficients, evaluated under several values of γ.



(c) Expectation and standard deviation of $P(k|\gamma, N, W)$ obtained from eq. 4 with permutation limit coefficients (blue), and from direct simulation of the DP, marginal on X (black), with boxplots showing values of k obtained over 250 MCMC iterations, following a burn-in of 250 iterations.

Figure 2: ${}_nm_k$ coefficients, predicted moments, and comparison with simulation results.

This procedure allows us to calculate $P(k|\gamma, N, W)$ using eq.(4), by plugging in the coefficients obtained from the permutation limit. The predicted moments of this distribution are in good agreement with results obtained from simulation of the DP (marginal on X) (figure 2(c)). We may then consider sampling γ from $P(\gamma|k, N, W) \propto P(\gamma)P(k|\gamma, N, W)$, which can be accomplished using a one-dimensional grid over $\log\gamma$. However, we prefer to specify our prior information in units of k, as this quantity has obvious biological meaning.

Specifically, we may construct a prior on the 1st moment of k, marginal on N and W as follows:

$$E[k|\gamma, N, W] = k_0 \sim \text{Scaled-Inv-}\chi^2(\nu_{k_0}, \pi_{k_0})$$

Where $k_{0\theta} = \{\nu_{k_0}, \pi_{k_0}\}$ correspond to equivalent sample size and location parameters, respectively.

Posterior samples of k_0, conditional on observing S values of k (in S samples, say) are obtained by sampling:

$$k_0|\bar{k}, S \sim \sqrt{\frac{\nu_{k_0}\pi_{k_0}^2 + S\bar{k}^2}{\chi^2(\nu_{k_0} + S)}}$$

Values of the expectation $E[k|\gamma, N, W]$ may be pre-computed for a given $\{N, W\}$ over a grid on $\log\gamma$ to form a map $f : \gamma \to k_0$ such that $f^{-1}(k_0)$ gives the required value of γ for the DP (figure 2(c)).

3.1 Sampling considerations

We specify a prior distribution over μ_c:

$$G_0(\mu_c) = \int_\alpha \int_\tau P(\alpha, \tau|\mu_{\mathbf{c}}{}^{(l-1)})\text{Comb}(\mu_c, \alpha, \tau, Q)d\alpha, d\tau$$

Where $\mu_{\mathbf{c}}{}^{(l-1)}$ denotes the set of cluster means obtained from the previous MCMC iteration. Integration over the conditional posterior of α and τ is well worth the extra computational effort required for a single MCMC iteration, as doing so allows rapid convergence of the simulation (on the order of 100 iterations). Figure 3 illustrates the DP on an example sample.

$$\text{Comb}(\mu_c, \alpha, \tau, Q) = \frac{1}{Q}\sum_{q=0}^{Q-1}\delta(\mu_c - f(\alpha, \tau, q))$$

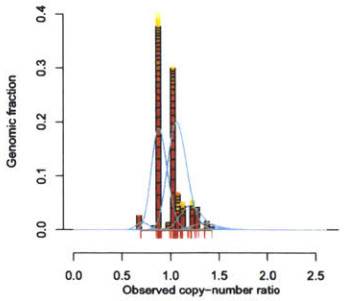
Where $f(\alpha, \tau, q)$ is given in eq. 1.

$$H_{\mathbf{c}_{-i}}(\hat{\mu_c}) = \int_{\mu_c} \mathcal{N}(\hat{\mu_c}|G_0(\mu_c), \sigma)\, dG_0(\mu_c)$$

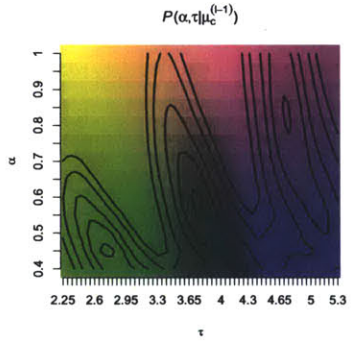
$$F(x_i, \Theta) = \mathcal{N}(x_i|\mu_c, \sigma)$$

$$\hat{\mu_c} = \frac{\sum_{j\neq i}\left(\frac{X_{[c_j=c]}}{\sigma^2_{x[c_j=c]}}\right)}{\sum_{j\neq i}\left(\frac{1}{\sigma^2_{x[c_j=c]}}\right)}$$

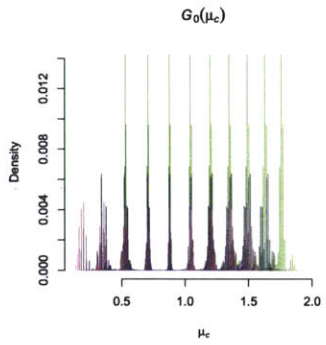
Where $\sigma_{x[i]}$ is the [not yet defined] technical noise estimate of a given segment i.



(a) Clustering of observed segments into CN-level clusters by the Dirichlet process. Individual segments are drawn as rectangles, with height proportional to segment size W, and colored according to σ_x. Five clusters are represented as cyan log-gaussians.



(b) Posterior distribution computed using the clusters sampled from the DP (a). Contours are drawn at (0.99, 0.95, 0.9, 0.8, 0.7) times the log-density at the mode.



(c) $G_0(\mu_c)$ computed from the posterior density in (b). The color of each stick indicates the position on the corresponding (α, τ) grid (b).

Figure 3: Example sample from the DP at equilibrium. A tri-modal posterior density on (α, τ) corresponding to unit shifts in ploidy (b) gives rise to a prior $G_0(\mu_c)$ (c) in which the 'comb teeth' from the three modes overlap at unit shifts in absolute copy-number.

4 Model extension and refinement

The full hierarchical graphical model is illustrated in figure 4.

4.1 Error model

Thus far our discussion has presumed that segments are present in equal copy numbers in each tumor cell in a given sample. This assumption is not realistic, and many segments exist at non-clonal copy numbers in most tumors, resulting in segments that appear much further from the predicted copy-number levels than technical noise (σ_x) can explain. Failure to account for this source of variation can result in dramatic failure of the model, as extraneous copy-number levels are instantiated to fit single non-clonal segments. This situation would often result in a value of Δ equal to twice the true value, corresponding to lower purity α and higher ploidy τ.

In order to deal with this situation, we allow segments to be fit by one of two mixture models, indicated by Z_σ. Segments for which $Z_\sigma = 1$ are fit with variance $\sigma_x + S$, where S is a scale of global noise due to heterogeneity: σ_H. The scale S is proportional to Δ, and is therefore a function of α and τ.

We modify our DP clustering probabilities (eq. 3), taking into account that the only parameter the cluster-members share is the mean ($\Theta = \{\mu_c\}$):

$$P(c_i = c|\mathbf{c}_{-i}) \propto \begin{cases} \frac{\sum_{j \neq i} W_{[c_j=c]}}{1-W_i+\gamma} \int_{\mu_c} \int_{Z_{\sigma_i}} F(x_i|\mu_c, Z_{\sigma_i}, \sigma_{x_i}) P(Z_{\sigma_i}|X_i) dZ_{\sigma_i} dH_{\mathbf{c}_{-i}}(\mu_c) & \text{if } c = c_j \text{ for some } j \neq i; \\ \frac{\gamma}{1-W_i+\gamma} \int_{\mu_c} \int_{Z_{\sigma_i}} F(x_i|\mu_c, Z_{\sigma_i}, \sigma_{x_i}) P(Z_{\sigma_i}|X_i) dZ_{\sigma_i} dG_0(\mu_c) & \text{otherwise.} \end{cases} \tag{5}$$

$$P(Z_{\sigma_i} = 0|X_i) \propto \int_{\mu_c} \mathcal{N}(X_i|\mu_c, \sigma_{x_i}) dG_0(\mu_c) P(Z_{\sigma_i} = 0|\theta_{Z_\sigma})$$

$$P(Z_{\sigma_i} = 1|X_i) \propto \int_{\mu_c} \mathcal{N}(X_i|\mu_c, \sigma_{x_i} + S) dG_0(\mu_c) P(Z_{\sigma_i} = 1|\theta_{Z_\sigma})$$

Application of Bayes' rule, along with conditional independence of the $\{Z_{\sigma_{c_{-i}}}\}$ gives:

$$\begin{aligned} H_{c_{-i}}(\mu_c) &\propto \int_{\{Z_{\sigma_{c_{-i}}}\}} P(\{X_{c_{-i}}\}|\mu_c, \{Z_{\sigma_{c_{-i}}}\}, \sigma) P(\{Z_{\sigma_{c_{-i}}}\}|\{X_{c_{-i}}\}) G_0(\mu_c) d\{Z_{\sigma_{c_{-i}}}\} \\ &= \prod_{j \in \{c_{-i}\}} \int_{Z_{\sigma_j}} P(X_j|\mu_c, Z_{\sigma_j}, \sigma_j) P(Z_{\sigma_j}|X_j) dZ_{\sigma_j} G_0(\mu_c) \end{aligned}$$

This likelihood can be calculated over a 1-dimensional grid on μ_c, and empirically normalized for each Z_σ mixture. In the implementation described here, the DP location prior G_0 is a finite comb of $Q(= 9)$ delta functions corresponding to a single position on (α, τ).

5 MCMC sampling

Given the graphical model of figure 4 and the observed data X, σ_x, W, our task is to generate samples from the posterior distribution.

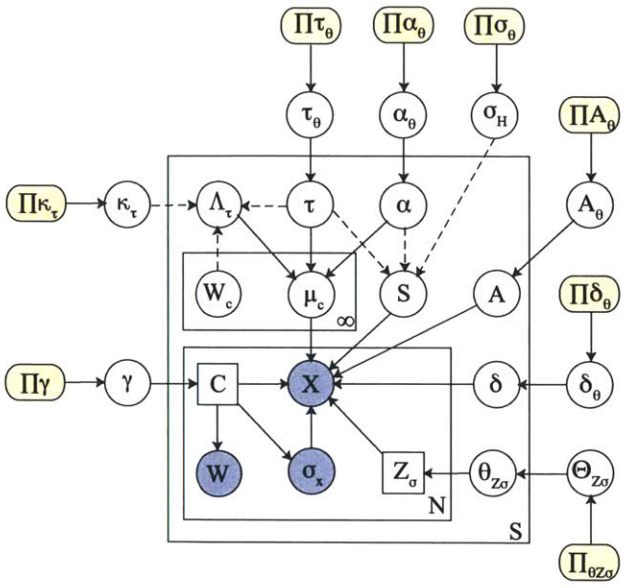


Figure 4: Graphical model for multiple samples, represented by the plate S. Beige ovals denote model hyper-parameters. A and δ account for small variability in attenuation and (log) offset, respectively, in a given sample. The hyper-parameters for those nodes greatly reduces the effective number of extra parameters for these effects. Adding these nodes was important to allow for realistic amounts of uncertainty regarding α and τ.

The algorithm proceeds as follows: For each sample

Choose a metropolis jump from 1 of 5 possible move types:

1. $\tau' \leftarrow \mathcal{N}(\tau, 0.025), \alpha' \leftarrow \mathcal{N}(\alpha, 0.025)$ $\tau' \leftarrow \frac{\tau}{2}, \alpha' \leftarrow \frac{2\alpha}{\alpha+1}$
2. $\tau' \leftarrow \tau + 1, \alpha' \leftarrow \frac{2\alpha}{2-\alpha}$
3. $\tau' \leftarrow \mathcal{N}(\tau, 0.025), \delta' \leftarrow \delta + \log \frac{2+\alpha(\tau-2)}{2+\alpha(\tau'-2)}$
4. Stay still.

Sample parameters:

1. Sample θ'_{Z_σ}
2. Compute $S' = \Delta' * \sigma_H * tooth_{2-1}$
3. Sample $\{C', \mu'_c, W'_c\} \sim DP(X, \sigma_X, C, W, \gamma, \theta_{Z_\sigma}), \{Z_\sigma\}$ are marginalized over.
4. Optimize $\Lambda'_\tau | \tau', W'_c$
5. Marginalize over $\{Z_c\}$ to obtain $P(\mu'_c | \alpha', \tau', W'_c, \Lambda'_\tau)$
6. Compute $\sum \log P(X_i | \sigma_X, C'_i, \mu'_{C_i}, Z'_{\sigma_i}, S', \delta', A')$
7. The log-likelihoods of the remaining sample-level parameters are straight-forward to compute.
8. Accept the new configuration with probability given by the standard Metropolis-Hastings ratio.

9. Repeat for all samples.

10. Gibbs sample the top level parameters from their conditional posteriors given all sample-level parameter values and the model hyper-parameters. (straight-forward)

6 Validation results

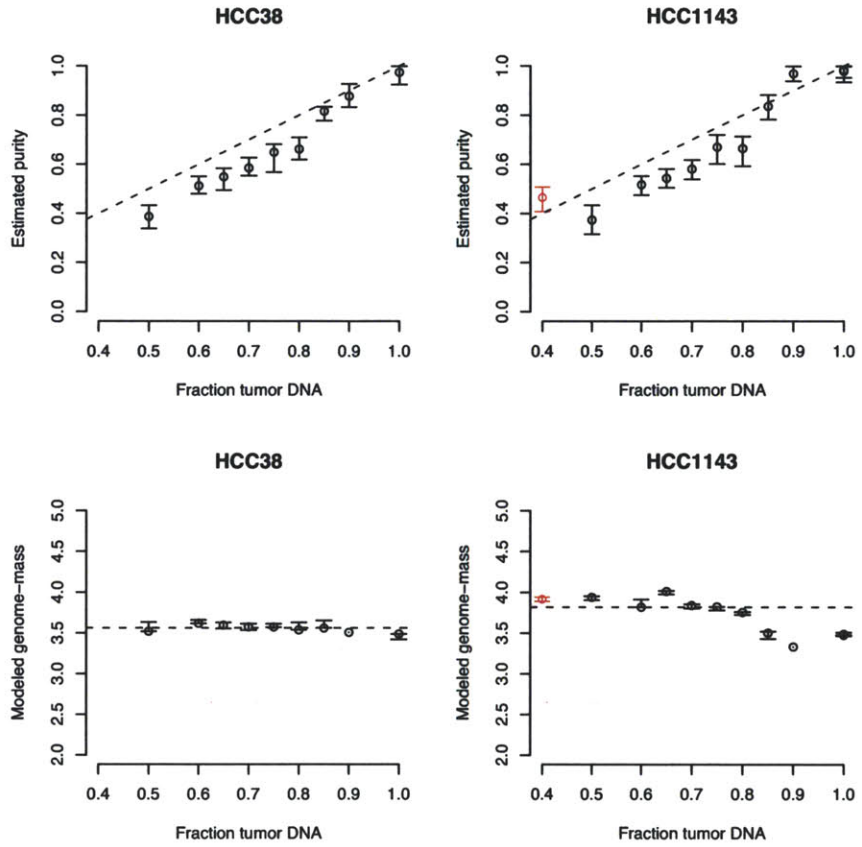


Figure 5: Results of the sampling algorithm for purity/ploidy inference in two cell-line mixing experiments. These results are highly consistent with the results shown in Figure 2 of Chapter 4; further details of the experiment are available in Chapter 4.

7 R source code

```
H_model_met_gibbs = function( sample.seg.dat,  sample.HSCN_params,
    sample_GRID, Q, genome_coverage, gen_model, data_params, init_model,
     gibbs_params, metrop_params, sample_DP_coefs, sample_k_0_maps, PLOT
    , DEBUG=FALSE )
```

```
{
   S = length(sample.seg.dat)
   comb_Q = Q
   N_iter = gibbs_params$N_iter

  ### initialize
   gen_point = list(S)
   first_lambda_guess = list(S)
   est = list()
   est$MCMC = list()

   est$MCMC$N_accept = array( 0, dim=c(N_iter+1) )

   est$assign_Z_CN = list(S)
   est$assign = list(S)
   est$assign_mu_c = list(S)
   est$assign_W_c = list(S)
   est$assign_Z_sigma = list(S)

   est$Z_sigma = list(S)
   est$seg_logliks = list()
   est$prop_seg_logliks = list()
   est$cluster_Z_CN = list(S)
   est$cluster_W = list(S)

   est$c_mu = list(S)
   est$c_stderr = list(S)

   for( s in 1:S )
   {
      alpha_GRID = sample_GRID[[s]]$alpha_GRID
      tau_GRID = sample_GRID[[s]]$tau_GRID
      N_alpha = length(alpha_GRID)
      N_tau = length(tau_GRID)
      N_grid = N_alpha * N_tau

      if( !any(is.na(gen_model)) )
      {
         gen_point[[s]] = c( which.min( abs( alpha_GRID - gen_model$
             alpha[s]) ),
                             which.min( abs( tau_GRID - gen_model$tau[s]) )
                                      )
      }
      else{ gen_point[[s]] = NA }

      est$cluster_Z_CN[[s]] = list()
      est$cluster_W[[s]] = list()
      est$Z_sigma[[s]] = list()

      P_s = length( sample.seg.dat[[s]][,"copy_num"] )
      est$assign[[s]] = array( NA, dim=c(P_s, N_iter+1) )
      est$assign_mu_c[[s]] = array( NA, dim=c(P_s, N_iter+1) )
      est$assign_W_c[[s]] = array( NA, dim=c(P_s, N_iter+1) )
      est$assign[[s]][,1] = c(1:P_s)    ### converges fast!!!
```

```
    est$assign_Z_CN[[s]] = array( NA, dim=c(P_s, Q, (N_iter+1) ) )
    est$assign_Z_sigma[[s]] = array( NA, dim=c( (N_iter+1), P_s, 2 ) )

    est$seg_logliks[[s]] = array( NA, dim=c(P_s, (N_iter+1)) )
    est$prop_seg_logliks[[s]] = array( NA, dim=c(P_s, (N_iter+1)) )

    first_lambda_guess[[s]] = c(0,0)
}

K_tau_hat = array( dim=S )
sample_sigma_H_hat = array( dim=c(S,2) )

est$alpha = array( dim=c(S,N_iter+1) )
est$tau = array( dim=c(S,N_iter+1) )
est$alpha[,1] = 0.5
est$tau[,1] = 2.5

est$prop_type = array( NA, dim=c(S,N_iter+1) )
est$MH_correction = array( NA, dim=c(S,N_iter+1) )

est$alpha_theta = array( dim=c( 2, N_iter+1 ) )
est$tau_theta = array( dim=c( 2, N_iter+1 ) )
est$sigma_H = array( dim=c( N_iter+1) )
est$K_tau = array( dim=c(N_iter+1) )
est$k_0= array( dim=c(N_iter+1) )
est$loglik_D = array( -Inf, dim=c(S,N_iter+1) )

est$prop_loglik_D = array( NA, dim=c(S,N_iter+1) )

est$DP_loglik = array( NA, dim=c(S,N_iter+1) )
est$Z_sigma_loglik = array( NA, dim=c(S,N_iter+1) )
est$cluster_Z_CN_loglik = array( NA, dim=c(S,N_iter+1) )

est$prop_DP_loglik = array( NA, dim=c(S,N_iter+1) )
est$prop_Z_sigma_loglik = array( NA, dim=c(S,N_iter+1) )
est$prop_cluster_Z_CN_loglik = array( NA, dim=c(S,N_iter+1) )

est$k = array( dim=c(S,N_iter+1) )
est$theta_Z_sigma = array( dim=c(S,2,N_iter+1) )
est$Theta_Z_sigma = array( dim=c(2,N_iter+1) )

est$delta_theta = array( dim=c(2,N_iter+1))
est$delta = array( dim=c(S, N_iter+1))

est$AT_theta = array( dim=c(2,N_iter+1))
est$AT = array( dim=c(S, N_iter+1))

est$epsilon_theta = array( dim=c(2,N_iter+1))
est$epsilon = array( NA, dim=c(S, N_iter+1))
est$Gmass = array( NA, dim=c(S, N_iter+1))
est$Lambda = array( NA, dim=c(S, LAMBDA_D, N_iter+1))

est$alpha_theta[,1] = init_model$alpha_theta
est$tau_theta[,1] = init_model$tau_theta
```

```
    est$delta_theta[,1] = init_model$delta_theta
    est$delta[,1] = init_model$delta_theta[1]
    est$epsilon[,1] = init_model$epsilon_theta[1]
    est$epsilon_theta[,1] = init_model$epsilon_theta
    est$AT_theta[,1] = init_model$AT_theta
    est$AT[,1] = init_model$AT_theta[1]

    est$K_tau[1] = init_model$K_tau
    est$k_0[1] = init_model$k_0
    est$sigma_H[1] = init_model$sigma_H

    ## init Z_sigma
    for( s in 1:S )
    {
       est$theta_Z_sigma[s,,1] = init_model$theta_Z_sigma
    }
    est$Theta_Z_sigma[,1] = init_model$Theta_Z_sigma

##########

    if( !is.na(gibbs_params$random_seed) ) { set.seed(gibbs_params$random
       _seed) }
    proposals = rep(0,4)

    for( i in 1:N_iter )
    {
       accepted = 0
       for( s in 1:S )
       {
        ####   Sample-specific constants
          seg.dat = sample.seg.dat[[s]]
          obs = list()

          obs$HSCN_params = sample.HSCN_params[[s]]
          obs$d = tx_data( sample.HSCN_params[[s]], seg.dat[,"copy_num"]
              )

          obs$seg_sigma = seg.dat[,"seg_sigma"]

          obs$W = seg.dat[,"W"]
          obs$alpha_GRID = sample_GRID[[s]]$alpha_GRID
          obs$tau_GRID = sample_GRID[[s]]$tau_GRID
          obs$sample_k_0_map = sample_k_0_maps[[s]]
          obs$sample_DP_coefs = sample_DP_coefs[[s]]
        ####

       ## get a proposal on alpha/tau
          new_conf = get_alpha_tau_proposal( est, s, i, conf, obs, gibbs_
              params, metrop_params )

       ## sample rest of model given proposal
       ## includes loglik calculation
          new_conf = sample_conf_point( new_conf, est, s, i, Q, obs,
              gibbs_params, metrop_params )
```

```
            W = new_conf$cluster_W
            Z_CN = new_conf$cluster_Z_CN
            Gmass = sum( (c(1:Q)-1) * W %*% Z_CN )

            if( data_params$DATA_TYPE == "allele" ){ Gmass = Gmass * 2 }

            msg = paste("alpha: ", round(new_conf$alpha,4), ", tau: ",
                round(new_conf$tau,4), ", Gmass: ", round(Gmass,4), ",
                loglik: ", round(new_conf$loglik,4), sep="" )
            print(msg)

            cur_loglik = est$loglik_D[s,i]
            MH_ratio = exp(new_conf$loglik - cur_loglik) * new_conf$MH_
                correction
            log_accept_prob = min( (new_conf$loglik - cur_loglik) + log(new
                _conf$MH_correction) , 0 )

            if( DEBUG == TRUE )
            {
                est$prop_loglik_D[s,(i+1)] = new_conf$loglik
                est$prop_DP_loglik[s,(i+1)] = new_conf$cluster_res$loglik
                est$prop_seg_logliks[[s]][,(i+1)] = new_conf$seg_logliks
                est$prop_Z_sigma_loglik[s,(i+1)] = new_conf$Z_sigma_loglik
                est$prop_cluster_Z_CN_loglik[s,(i+1)] = new_conf$cluster_Z_
                    CN_loglik
            }

            if( log(runif( 1,0,1 )) < log_accept_prob )    ## accept move
            {
                chr="O"
                est = accept_move( new_conf, est, s, i )
                accepted = accepted + 1
            }
            else
            {
                chr="X"
                est = reject_move( est, s, i )
            }
            cat(chr)
        K_tau_hat[s] = calc_K_tau_hat( est$cluster_W[[s]][[(i+1)]], est$
            cluster_Z_CN[[s]][[(i+1)]], est$tau[s,(i+1)], Q, gibbs_params
            $zero_p )

cat( length(unique( est$assign[[s]][,(i+1)])) );  cat(", ")
        }
cat("\n")

        est$alpha_theta[,(i+1)] = alpha_theta_posterior( est$alpha[,(i+1)
            ], gibbs_params$pi_alpha_theta )
        est$tau_theta[,(i+1)] = alpha_theta_posterior( est$tau[,(i+1)],
            gibbs_params$pi_tau_theta )

        est$K_tau[(i+1)] = K_tau_post( K_tau_hat, gibbs_params$K_tau_theta
```

```
        )

    est$k_0[(i+1)] = H_k_0_post( est$k[,(i+1)], gibbs_params$k_0_theta
        )

    est$sigma_H[(i+1)] = est$sigma_H[i]

    stz = apply( matrix(est$theta_Z_sigma[,,(i+1)],nrow=S), 2, sum )
    est$Theta_Z_sigma[1,(i+1)] = rbeta( 1, gibbs_params$pi_theta_Z_
        sigma[1] + stz[1], gibbs_params$pi_theta_Z_sigma[2] + stz[2] )
    est$Theta_Z_sigma[2,(i+1)] = 1 - est$Theta_Z_sigma[1,(i+1)]

    HP = gibbs_params$pi_delta_theta
    res = Norm_Inv_Chi_sq_post( est$delta[,(i+1)], HP[1], HP[2], HP
        [3], HP[4] )
    est$delta_theta[1,(i+1)] = res$mu
    est$delta_theta[2,(i+1)] = res$sigma

    HP = gibbs_params$pi_AT_theta
    res = Norm_Inv_Chi_sq_post( est$AT[,(i+1)], HP[1], HP[2], HP[3],
        HP[4] )
    est$AT_theta[1,(i+1)] = res$mu
    est$AT_theta[2,(i+1)] = res$sigma

    HP = gibbs_params$pi_epsilon_theta
    res = Norm_Inv_Chi_sq_post( est$epsilon[,(i+1)], HP[1], HP[2], HP
        [3], HP[4] )
    est$epsilon_theta[1,(i+1)] = res$mu
    est$epsilon_theta[2,(i+1)] = res$sigma


      print( paste("k_0:␣", round(est$k_0[i+1],3), sep="" ) )
      print( paste("sigma_H:␣", round(est$sigma_H[(i+1)],3), sep="" )
          )
      print( paste("K_tau:␣", round(est$K_tau[(i+1)],3), sep="") )
      print( paste("Theta_Z_sigma:␣", round(est$Theta_Z_sigma[,(i+1)
          ],3), sep="") )
      print( paste("delta_theta:␣", round(est$delta_theta[,(i+1)],3),
          sep="") )
      print( paste("AT_theta:␣", round(est$AT_theta[,(i+1)],3), sep=""
          ) )
      print( paste("epsilon_theta:␣", round(est$epsilon_theta[,(i+1)
          ],3), sep="") )
      print( paste("loglik:␣", round( sum(est$loglik_D[,(i+1)],na.rm=T
          ), 3), sep="") )
      print( paste(accepted, "␣moves␣accepted", sep="") )
}

print(est$k_0)
print(est$sigma_H)
print(est$K_tau)
print( apply(est$loglik_D, 2, sum, na.rm=T ) )

print("Move␣acceptances:␣" )
```

```
    print( est$MCMC$N_accept )
    print( sum(est$MCMC$N_accept[1:i]) )

    return( est )
}


get_span_penalty = function( Z_CN, k, eps=1e-5 )
{
    Z = apply( Z_CN, 2, sum, na.rm=T )
    Z = Z / sum(Z)
    Q = length(Z)
    i = 1
    while( Z[i] < eps ) { i = i + 1 }
    j = 1
    while( rev(Z)[j] < eps ) { j = j + 1}
    Nused = Q - (j-1 + i-1)

    ## BIC penalizes high-complexity solutions less.

   ## sez gaddy
#    H = k * -log(Nused)

   ## More BIC-like penalty:
    H = Nused * -log(k)

    return(H)
}

sample_DP = function( obs, s, i, est, pi_c_mu, pi_c_sigma )
{
    ### get clusters from DP
    conf = list()

    k_0_to_gamma_map = function( k_0, k_0_map )
    {
        ix = which.min( abs(k_0 - k_0_map[,1]) )
        return( k_0_map[ix,2] )
    }

    use_gamma = k_0_to_gamma_map( est$k_0[i], obs$sample_k_0_map )

    cluster_res = DP_post( obs$d, use_gamma, obs$seg_sigma, est$assign[[s
        ]][,i], obs$W, pi_c_sigma, pi_c_mu, obs$sample_DP_coefs )

    conf$assign = cluster_res$assign
    conf$cluster_W = cluster_res$c_W

    clusters = unique( conf$assign )
    K = length(clusters)
    for( j in 1:K )
    {
        cc = which( conf$assign == clusters[j] )
        conf$assign_mu_c[cc] = cluster_res$means[j]
```

```
        conf$assign_sigma_c[cc] = cluster_res$sigmas[j]
        conf$assign_W_c[cc] = cluster_res$c_W[j]
    }
    conf$k = K

    conf$DP_loglik = cluster_res$loglik
    conf$c_mu = cluster_res$means
    conf$c_stderr = cluster_res$stderrs

    return( conf )
}

sample_conf_point = function( conf, est, s, i, Q, obs, gibbs_params,
    metrop_params )
{
 ## Marginalize over Z_CN
   W_C = conf$cluster_W

    if( data_params$DATA_TYPE == "total" )
    {

        if( LAMBDA_D == 1 )
        {
            resid = Inf
            while( resid > 0.01 )
            {
                lambda_res = get_Lambda( conf$tau, W_C, est$K_tau[i], Q,
                    gibbs_params$zero_p, first_lambda_guess=NA )
                resid = lambda_res$resid
                cat(".")
            }

            m3 = -lambda_res$Lambda * ( (c(1:Q)-1) %*% t(W_C) )
        }

        if( LAMBDA_D == 2 )
        {
            lambda_res = get_Lambda( conf$tau, W_C, est$K_tau[i], Q, gibbs_
                params$zero_p, first_lambda_guess=est$Lambda[s,,(i-1)] )

            m31 = -lambda_res$Lambda[1] * ( (c(1:Q)-1) %*% t(W_C) )
            ff = c( gibbs_params$zero_p, rep(0, (Q-1) ) )
            m32 = -lambda_res$Lambda[2] * ( log( (c(1:Q)-1) + ff) - log(
                conf$tau))^2 %*% t(W_C)
            m3 = m31 + m32
        }
    }

    if( data_params$DATA_TYPE == "allele" )
    {
        use_tau = conf$tau / 2
       W_C = W_C

        resid = Inf
```

```
        while( resid > 0.01 )
        {
            lambda_res = get_Lambda( use_tau, W_C, est$K_tau[i], Q, gibbs_
                params$zero_p, first_lambda_guess=NA )
            resid = lambda_res$resid
            cat(".")
        }

        m3 = -lambda_res$Lambda[1] * ( (c(1:Q)-1) %*% t(W_C) )
    }
    conf$Lambda = lambda_res$Lambda

    if( lambda_res$resid > 0.01 )
    {
        cat("!!!!")
        print( lambda_res$resid )
        conf$Lambda = NA
    }

### Q x K
    loc_comb = get_gridpoint_comb( conf, Q, obs$HSCN_params )

    res = get_SS_comb_LL( conf, loc_comb, est$sigma_H[i] )  ### updates Z
        _sigma
    m1 = res$SS_LL
    conf$Z_sigma = res$Z_sigma

  ## Sample theta_Z_sigma
    sz = apply( conf$Z_sigma, 2, sum)
    P_s = dim( conf$Z_sigma )[1]

    conf$theta_Z_sigma = est$theta_Z_sigma[s,,i]

  ## marginzalize over Z_CN
    res = get_seg_MM_soft_assign( m1, m3, Q, lambda_res$norm_term )

    conf$cluster_Z_CN_loglik = res$loglik
    conf$cluster_Z_CN = res$soft_assign

    if( !is.finite( conf$cluster_Z_CN_loglik ) ) { stop() }

    conf = calc_conf_loglik( conf, est, s, i, obs )
    conf$Gmass = sample_Gmass( conf )

    return( conf )
}

sample_Gmass = function( conf )
{
    CN = rep( NA, conf$k )
    for( k in 1:conf$k )
    {
        CN[k] = which( rmultinom( 1, 1, conf$cluster_Z_CN[k,] ) == 1 ) - 1
    }
```

```
    Gmass = sum( CN * conf$cluster_W )
    return(Gmass)
}

theta_Z_sigma_gibbs_post = function( pi_theta_Z_sigma, Z_sigma )
{
    theta_Z_sigma = c(0,0)

    s1 = pi_theta_Z_sigma[1] + sum( Z_sigma[,1] )
    s2 = pi_theta_Z_sigma[2] + sum( Z_sigma[,2] )

    theta_Z_sigma[1] = rbeta( 1, s1, s2 )
    theta_Z_sigma[2] = 1 - theta_Z_sigma[1]

    return( theta_Z_sigma )
}


calc_conf_loglik = function( conf, est, s, i, obs)
{
    ### Sample loglik calculation
    k = conf$k
    p = conf$theta_Z_sigma[1]
    x = sum(conf$Z_sigma[,1] * conf$cluster_W ) * k

    Z_sigma_loglik = x*log(p) + (k-x) * log(1-p)    ## log of un-
        normalized binomial density

    S = dim( est$theta_Z_sigma )[1]
    theta_Z_sigma_loglik = 0

    delta_loglik = dnorm( conf$delta, est$delta_theta[1,i], est$delta_
        theta[2,i], log=TRUE )
    AT_loglik = dnorm( conf$AT, est$AT_theta[1,i], est$AT_theta[2,i], log
        =TRUE )
    epsilon_loglik = dnorm( conf$epsilon, est$epsilon_theta[1,i], est$
        epsilon_theta[2,i], log=TRUE )

    conf$Z_sigma_loglik = Z_sigma_loglik

    alpha_loglik = dnorm( conf$alpha, est$alpha_theta[1,i], est$alpha_
        theta[2,i], log=T)
    tau_loglik = dnorm( conf$tau, est$tau_theta[1,i], est$tau_theta[2,i],
         log=T )

    H = get_span_penalty( conf$cluster_Z_CN, conf$k )

    conf$loglik = H +
                  conf$cluster_Z_CN_loglik +
                  conf$DP_loglik  +
                  Z_sigma_loglik +
                  theta_Z_sigma_loglik +
                  delta_loglik +
                  AT_loglik +
```

```
                    epsilon_loglik +
                    alpha_loglik +
                    tau_loglik


  if( is.nan( conf$loglik ) )
  {
     dump.frames( dumpto="Hmet.framedump", to.file=TRUE )
  }

   return( conf )
}


get_gridpoint_comb = function( conf, Q, HSCN_params )
{
   res = get_b_and_delta( conf$alpha, conf$tau )
   means = get_means( Q, res$delta, res$b, conf$delta, conf$epsilon,
      HSCN_params )

   comb_X = means

   Delta = res$delta
   t2 = means[2]
   comb_S = log( (Delta + t2) / t2, 2 )

   return( list("X"=comb_X, "S"=comb_S) )
}


accept_move = function( conf, est, s, i )
{
   est$alpha[s,(i+1)] = conf$alpha
   est$tau[s,(i+1)] = conf$tau
   est$delta[s,(i+1)] = conf$delta
   est$AT[s,(i+1)] = conf$AT
   est$epsilon[s,(i+1)] = conf$epsilon
   est$Gmass[s,(i+1)] = conf$Gmass
   est$Lambda[s,,(i+1)] = conf$Lambda

   est$Z_sigma[[s]][[(i+1)]] = conf$Z_sigma
   est$theta_Z_sigma[s,,(i+1)] = conf$theta_Z_sigma

   est$assign[[s]][,(i+1)] = conf$assign
   est$cluster_W[[s]][[(i+1)]] = conf$cluster_W
   est$cluster_Z_CN[[s]][[(i+1)]] = conf$cluster_Z_CN

   est$assign_mu_c[[s]][,(i+1)] = conf$assign_mu_c
   est$assign_W_c[[s]][,(i+1)] = conf$assign_W_c
   est$k[s,(i+1)] = conf$k

   est$c_mu[[s]] = conf$c_mu
   est$c_stderr[[s]] = conf$c_stderr
```

```
    est$loglik_D[s,(i+1)] = conf$loglik

    est$MCMC$N_accept[i] = est$MCMC$N_accept[i] + 1

    est$DP_loglik[s,(i+1)] = conf$DP_loglik
    est$Z_sigma_loglik[s,(i+1)] = conf$Z_sigma_loglik
    est$cluster_Z_CN_loglik[s,(i+1)] = conf$cluster_Z_CN_loglik

    est$prop_type[s,(i+1)] = conf$type
    est$MH_correction[s,(i+1)] = conf$MH_correction

    clusters = unique( conf$assign )
    K = length(clusters)
    for( j in 1:K )
    {
        cc = which( conf$assign == clusters[j] )
        est$assign_Z_sigma[[s]][(i+1),cc,1] = conf$Z_sigma[j,1]
        est$assign_Z_sigma[[s]][(i+1),cc,2] = conf$Z_sigma[j,2]
    }

    return( est )
}

reject_move = function( est, s, i )
{
    est$alpha[s,(i+1)] = est$alpha[s,i]
    est$tau[s,(i+1)] = est$tau[s,i]
    est$delta[s,(i+1)] = est$delta[s,i]
    est$AT[s,(i+1)] = est$AT[s,i]

    est$epsilon[s,(i+1)] = est$epsilon[s,i]
    est$Gmass[s,(i+1)] = est$Gmass[s,i]
    est$Lambda[s,,(i+1)] = est$Lambda[s,,i]

    est$Z_sigma[[s]][[(i+1)]] = est$Z_sigma[[s]][[i]]
    est$theta_Z_sigma[s,,(i+1)] = est$theta_Z_sigma[s,,i]

    est$assign[[s]][,(i+1)] = est$assign[[s]][,i]

    est$cluster_Z_CN[[s]][[(i+1)]] = est$cluster_Z_CN[[s]][[i]]
    est$cluster_W[[s]][[(i+1)]] = est$cluster_W[[s]][[i]]

    est$assign_mu_c[[s]][,(i+1)] = est$assign_mu_c[[s]][,i]
    est$assign_W_c[[s]][,(i+1)] = est$assign_W_c[[s]][,i]
    est$assign_Z_CN[[s]][,,(i+1)] = est$assign_Z_CN[[s]][,,i]
    est$k[s,(i+1)] = est$k[s,i]

    est$loglik_D[s,(i+1)] = est$loglik_D[s,i]

    est$DP_loglik[s,(i+1)] = est$DP_loglik[s,i]
#   est$seg_logliks[[s]][,(i+1)] = est$seg_logliks[[s]][,i]
    est$Z_sigma_loglik[s,(i+1)] = est$Z_sigma_loglik[s,i]
    est$cluster_Z_CN_loglik[s,(i+1)] = est$cluster_Z_CN_loglik[s,i]
```

```
    est$assign_Z_sigma[[s]][(i+1),,] = est$assign_Z_sigma[[s]][i,,]

    return( est )
}

calc_sigma_H_hat = function( conf, obs )
{
    P_s = dim( conf$Z_sigma )[1]
    DF =  P_s
    sigma_H_hat = array( NA, dim=P_s )
    aX = array( NA, dim=P_s )

    Z_sigma = conf$Z_sigma
    assign = conf$assign
    comb_post_set = conf$comb_post_set

    K = conf$k
    clusters = unique( conf$assign )

    for( k in 1:K )
    {
        cc = which( assign == clusters[k] )
        comb_post = comb_post_set[[k]]

 ## integrate over mu_c
        xx = sapply( obs$d[cc], "-", as.vector(comb_post$X) )
        sq_dist = apply( xx^2 * as.vector(comb_post$Y), 2, sum )

        Si_sq = sum( as.vector(comb_post$S)^2 * as.vector(comb_post$Y) )
        xx = sum(sq_dist * Z_sigma[cc,2] / Si_sq) / sum(Z_sigma[cc,2])
        sigma_H_hat[cc] = sqrt(xx)
    }

    sigma_H_hat[ is.nan(sigma_H_hat) ] = 0
    sigma_H_hat[ sigma_H_hat < 0 ] = 0

    sigma_H = mean( sigma_H_hat )

    return( list("DF"=DF, "sigma_H_hat"=sigma_H) )
}

sigma_H_post = function( sample_sigma_H_hat, pi_sigma_H )
{
    DF = sum( sample_sigma_H_hat[,1] )
    sigma_H_hat = mean( sample_sigma_H_hat[,2] )

    post = Inv_Chi_sq_post( DF, sigma_H_hat, pi_sigma_H[1], pi_sigma_H[2]
        )

    return(post)
}


calc_K_tau_hat = function( W, Z_CN, tau, Q, zero_p )
```

```
{
  ## this is over super-segs
  ## Calculate sufficient statistic for K_tau
   P_s = length(W)

   E_Gmass_dist = 0
   ff = c( zero_p, rep(0, (Q-1) ) )

   for( p in 1:P_s )
   {
      if( data_params$DATA_TYPE == "total" )
      {
         E_Gmass_dist = E_Gmass_dist + W[p] * sum( Z_CN[p,] * ( log( (c
             (1:Q)-1) + ff) - log(tau))^2, na.rm=TRUE )
      }
      if( data_params$DATA_TYPE == "allele" )
      {
         E_Gmass_dist = E_Gmass_dist + W[p] * sum( Z_CN[p,] * ((c(1:Q)
             -1) - tau/2)^2, na.rm=TRUE )
      }
   }
   K_tau_hat = E_Gmass_dist / tau

   return( K_tau_hat )
}

DP_post = function( data, gamma, sigma_x, assign_init, W, pi_sigma, pi_
    mu, stirling_coef )
{
   N = length(data)
   assign = assign_init
   clusters = array(NA, dim=N+1)
   cluster_stats = list(N+1)

   cluster_prior = list("mu"=pi_mu$mu_0, "sigma"=pi_sigma$sigma_0)

   for( j in 1:(N+1) )
   {
      cluster_stats[[j]] = 0
      clusters[j] = sum(assign == j)
   }

   c_lik = array(0, dim=N+1)
   dirty = array(TRUE, dim=N+1)
   dirty[ clusters == 0 ] = FALSE

   for( n in 1:N )
   {
      c_lik = rep(0, length(c_lik) )
      skip_count = 1

      NZ_clusters = which( clusters != 0 )
      for( j in 1:length(NZ_clusters) )
      {
```

```
        c_ix = NZ_clusters[j]
        c_dat = (assign == c_ix)

        if( dirty[c_ix] == TRUE )
        {
            res = get_cluster_stats( c_dat, data, sigma_x, pi_mu, pi_
                sigma )
            cluster_stats[[c_ix]] = res
            dirty[c_ix] = FALSE
        }

      ## if the current point is the only thing in the cluster...
       # This seems to work empirically (as well as theoretically)
        if( sum(c_dat) == 1 & c_dat[n] == TRUE )
        {
            skip_count = skip_count + 1    ## at most = 2
            next
        }

### all snps are equal (almost, should use n_probes instead of length
    ...)
        const = sum(W[c_dat]) / (1 - W[n] + gamma)

        if( c_dat[n] == FALSE )
        {
            c_res = cluster_stats[[c_ix]]
        }
        else
        {
            ## remove the current point from the cluster
            ## and re-compute cluster stats

            c_dat[n] = FALSE
            c_res = get_cluster_stats( c_dat, data, sigma_x, pi_mu, pi_
                sigma )
        }

        res = cluster_prob( c_res, data[n], sigma_x[n] )
        c_lik[c_ix] = const * res$cluster_prob
    }

    ## prob to open a new cluster
      const = (gamma) / (1- W[n] + gamma)
      cluster_prior$mu = data[n]
      cluster_prior$stderr = 0
      res = cluster_prob( cluster_prior, data[n], sigma_x[n] )
      h_lik = skip_count * const * res$cluster_prob

      new_c_ix = min( which(clusters==0) )
      c_lik[new_c_ix] = h_lik

      c_lik[ is.nan(c_lik)] = 0
      if( sum( c_lik ) > 0 )
      {
```

```
      res = which( rmultinom( 1, 1, c_lik ) == 1 )
    }
    else
    {
      print("DP: all c_lik==0!!!")
      res = new_c_ix
    }

  ## if a point has been assigned to a different cluster..
    if( assign[n] != res )
    {
      clusters[ assign[n] ] = clusters[ assign[n] ] - 1
      if( clusters[ assign[n] ] > 0 )
      {
        dirty[assign[n] ] = TRUE
      }
      else
      {
        cluster_stats[[ assign[n] ]] = 0
      }
      dirty[res] = TRUE
      clusters[res] = clusters[res] + 1
      assign[n] = res
    }
}

NZ_clusters = unique(assign)

c_mu = rep(0, length(NZ_clusters) )
c_stderr = rep(0, length(NZ_clusters) )
c_W = rep(0, length(NZ_clusters) )

for( j in 1:length(NZ_clusters) )
{
    c_ix = NZ_clusters[j]
    c_dat = (assign == c_ix)

    if( dirty[c_ix] == TRUE )
    {
      res = get_cluster_stats( c_dat, data, sigma_x, pi_mu, pi_sigma
          )
      cluster_stats[[c_ix]] = res
      dirty[c_ix] = FALSE
    }

    c_res = cluster_stats[[c_ix]]
    c_mu[j] = c_res$mu
    c_stderr[j] = c_res$stderr
    c_W[j] = sum( W[c_dat] )
}

DP_loglik = calc_DP_loglik( N, length(NZ_clusters), gamma, stirling_
    coef )
```

```
    return( list("means"=c_mu, "stderrs"=c_stderr, "c_W"=c_W, "assign"=
        assign, "loglik"=DP_loglik, "K"=length(NZ_clusters) ) )
}

calc_DP_loglik = function( N, K, alpha, coefs )
{
  ## actually returns the log of the normalized probability

  ## Escobar and West 1995 for DP loglik. No correction for segment
      sizes!
#    loglik = log(stirling_coef[K]) + lgamma(N-1) + K*alpha + lgamma(
    alpha) - lgamma(alpha + N)

  ## Carter and Getz 2008 with seglen correction.
    loglik = log(coefs[K]) + K*log(alpha) - log( eval_M_polynom(alpha,
        coefs) )


    return(loglik)
}

eval_M_polynom = function( alpha, coefs )
{
    pows = c(1:length(coefs))
    res = sum( coefs * alpha^pows )

    return(res)
}


mod_DP_conc_post = function( k, N, coefs, gamma_theta )
{
    ### Escobar and West 1995
#    eta = rbeta( 1, alpha_DP, N )
#    new_alpha = rgamma( 1, a_alpha_DP[1] - log(eta), a_alpha_DP[2] + J -
    1 )

    N_gamma = 25
    gamma_GRID = exp( seq( log(0.0005), log(0.05), length.out=N_gamma )
        )

    for( i in 1:N_gamma )
    {
       grid_res[i] = calc_DP_loglik( N, k, gamma_GRID[i], coefs ) +
                     dgamma( gamma_GRID[i], gamma_theta[1], gamma_theta
                         [2], log=T )
    }

    mx = max(grid_res)
    nf = sum( exp(grid_res - mx) )
    grid_density = exp(grid_res-mx) / nf

    res = which( rmultinom( 1, 1, grid_density ) )
    new_gamma = grid_density[res]
```

```
    return(new_gamma)
}


cluster_prob = function( c_res, x, sigma_x )
{
    n1 = dnorm( x, c_res$mu, sqrt(sigma_x^2 + c_res$stderr^2) )
#    n1 = dnorm( x, c_res$mu, sigma_x )

    return( list("cluster_prob"=n1) )
}

c_sigma_hat = function( c_data, c_mean, c_var )
{
    c_SD = sqrt( sum( ((c_data-c_mean)^2 + c_var) / c_var) / sum( 1/c_var
        ) )
    return(c_SD)
}

get_cluster_stats = function( c_ix, data, sigma_x, pi_mu, pi_sigma )
{
    c_var = sigma_x[c_ix]^2
    c_dat = data[c_ix]

    N = length(c_dat)

    mu_hat = sum( c_dat / c_var  ) / sum( 1/c_var )
    sigma_hat = c_sigma_hat( c_dat, mu_hat, c_var )


# new parameters
    y_bar = mu_hat
    s = sigma_hat

# priors
    mu_0 = pi_mu$mu_0
    k_0 = pi_mu$k_0
    sigma_0 = pi_sigma$sigma_0
    v_0 = pi_sigma$v_0

    mu_n = mu_0 * k_0 / (k_0 + N) + y_bar * N / (k_0 + N)
    k_n = k_0 + N
    v_n = v_0 + N

    if( N > 1 )
    {
        v_n_sigma_sq = v_0 * sigma_0^2 + (N -1)*s^2 + (y_bar - mu_0)^2 *(k
            _0 * N)/(k_0 + N)
        sigma_n = sqrt(v_n_sigma_sq / v_n )
    }
    else
    {
        sigma_n = sqrt(c_var[1])
```

```
    }


# posterior samples
    post_mu = rnorm( 1, mu_n, sqrt( sigma_n^2 / k_n)   )
    dd = rchisq( 1, v_n )
    post_sigma = sqrt(   v_n * sigma_n^2 / dd   )

    return( list("mu"=post_mu, "stderr"=post_sigma/sqrt(k_n)) )
#    return( list("mu"=post_mu, "stderr"=post_sigma) )
}
```