



TITAN: Inference of copy number architectures in clonal cell populations from tumor whole genome sequence data

Gavin Ha, Andrew Roth, Jaswinder Khattra, et al.

Genome Res. published online July 24, 2014

Access the most recent version at doi:[10.1101/gr.180281.114](https://doi.org/10.1101/gr.180281.114)

P<P	Published online July 24, 2014 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data

Gavin Ha^{1,2}, Andrew Roth^{1,2}, Jaswinder Khattra¹, Julie Ho³, Damian Yap¹, Leah M. Prentice³, Nataliya Melnyk³, Andrew McPherson^{1,2}, Ali Bashashati¹, Emma Laks¹, Justina Biele¹, Jiarui Ding^{1,4}, Alan Le¹, Jamie Rosner¹, Karey Shumansky¹, Marco A. Marra⁵, C Blake Gilks⁶, David G. Huntsman^{3,7}, Jessica N. McAlpine⁸, Samuel Aparicio^{1,7}, and Sohrab P. Shah^{1,4,7,*}

¹Department of Molecular Oncology, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada

²Bioinformatics Training Program, University of British Columbia, 100-570 West 7th Avenue, Vancouver, BC V5Z 4S6, Canada

³Centre for Translational and Applied Genomics, 600 West 10th Avenue, Vancouver, BC V5Z 4E6, Canada

⁴Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, BC V6T 1Z4, Canada

⁵Genome Sciences Centre, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada

⁶Genetic Pathology Evaluation Centre, Vancouver General Hospital, Vancouver, BC V6H 3Z6, Canada

⁷Department of Pathology and Laboratory Medicine, University of British Columbia, 2211 Wesbrook Mall, Vancouver, BC V6T 2B5, Canada

⁸Department of Gynecology and Obstetrics, University of British Columbia, 2775 Laurel Street, Vancouver, BC V5Z 1M9, Canada

Corresponding Author:

Sohrab P Shah, PhD
675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada
email: sshah@bccrc.ca
tel: +1 604 675 8252

Keywords: cancer genomes, tumour evolution, intra-tumour heterogeneity, subclonal events, copy number alterations, loss of heterozygosity, whole genome sequencing, single-cell sequencing, hidden Markov models

Abstract

The evolution of cancer genomes within a single tumour creates mixed cell populations with divergent somatic mutational landscapes. Inference of tumour subpopulations has been disproportionately focused on the assessment of somatic point mutations, whereas computational methods targeting evolutionary dynamics of copy number alterations (CNA) and loss of heterozygosity (LOH) in whole genome sequencing data remain under-developed. We present a novel probabilistic model, TITAN, to infer CNA and LOH events while accounting for mixtures of cell populations, thereby estimating the proportion of cells harbouring each event. We evaluate TITAN on idealized mixtures, simulating clonal populations from whole genome sequences taken from genomically heterogeneous ovarian tumour sites collected from the same patient. In addition, we show in 23 whole genomes of breast tumours that the inference of CNA and LOH using TITAN critically informs population structure and the nature of the evolving cancer genome. Finally, we experimentally validated subclonal predictions using fluorescence *in situ* hybridization (FISH) and single-cell sequencing from an ovarian cancer patient sample, thereby recapitulating the key modelling assumptions of TITAN.

Introduction

Tumour progression follows the principles of clonal evolution (Nowell, 1976). Accumulation of genomic alterations is patterned by phylogenetic branching, creating a substrate for natural selection. Invariably, this leads to the emergence of distinct cell populations (clones) with divergent genotypes and associated phenotypes (Aparicio and Caldas, 2013). Here, we define a clone as a population of cells related by descent from a unitary origin, and as being uniquely identified by the complement of fixed genetic marks comprising its clonal genotype. Genetic marks can consist of somatic mutations such as point mutations, genome rearrangements, copy number alterations (CNA), and loss of heterozygosity (LOH), of which CNA and LOH are the focus of this study. We define the cellular prevalence of a somatic mutation as the proportion of cells harbouring an aberration in the overall (bulk) tumour cell population (Aparicio and Caldas, 2013). Cellular prevalence can be measured approximately through sequencing a bulk sample, or more precisely in independent analysis of single cells (Navin et al., 2011). The dynamics of cellular prevalence of a mutation are reflective of growth (dis)advantages in the presence of treatment or micro-environmental-induced selective pressures, and are thus a useful indicator of the biology underpinning tumour progression.

The clonal evolution theory implies that extant clones are related genetically through a phylogenetic tree. In such population structures, cellular prevalence of a genetic alteration is generally a function of its evolutionary timing: high prevalence mutations are acquired earlier than low prevalence mutations. Thus, ancestral mutations are found at the root of the tree, while descendent mutations are situated towards the leaves. We explored the resulting patterns of alterations acquired after expansion of the ancestral clone, which generates three types of cells in a tumour sample: normal (non-malignant) cells, tumour cells harbouring the alteration, and tumour cells without the alteration. This concept applies to all forms of genomic aberrations including CNA and LOH, despite a disproportionate emphasis on point mutations in the literature (Shah et al., 2009, 2012; Ding et al., 2012; Gerstung et al., 2012; Cibulskis et al., 2013; Landau et al., 2013; Larson and Fridley, 2013; Roth et al., 2014). Indeed, tumours with diverse intra-tumoural patterns of CNA and LOH have been described in breast (Navin et al., 2011; Nik-Zainal et al., 2012), ovarian (Bashashati et al., 2013), renal (Gerlinger et al., 2012, 2014) and brain tumours (Sottoriva et al., 2013).

Whole exome (WES) and genome (WGS) sequencing of a single biopsy are emerging as the dominant experimental designs in large cohort studies of tumour genomic landscapes, with consortia such as the In-

ternational Cancer Genome Consortium (ICGC) poised to generate on the order of 10,000 tumour-normal WGS libraries in the next few years. Characterization of clonal populations from such datasets has been primarily focused on point mutations, which require targeted deep sequencing. Measuring cellular prevalences of CNA and LOH presents unique challenges because these events can span mega-bases, rendering targeted deep sequencing of alleles infeasible. Moreover, heterogeneous mixtures of cells in tumour biopsies present a major limitation in accurate interpretation of WGS data. CNA and LOH events present in only minor cell populations will have diminished statistical signals, and thus are susceptible to false negative detection. Figure 1 depicts the observed read depth (top track) and allelic ratios (middle track) from subclonal deletions (*Deletion I* and *III*) and a high prevalence clonal deletion (*Deletion II*), illustrating the distinct statistical signals arising from differences in cellular prevalence (bottom track). The degree to which CNA and LOH contribute to the inference of evolutionary dynamics cannot be estimated using most current standard approaches. Methods for robust computational models of statistical signals emitted from multiple cell populations within a single tumour sample are therefore under-developed and represent a deficiency in the cancer genomics literature.

We developed a novel probabilistic model called TITAN. The model simultaneously infers CNA and LOH segments from read depth and digital allele ratios at germline heterozygous SNP loci across the genome from tumour WGS data. For each alteration, we assume the event is segregated into the underlying population of three different cell types: normal cells, tumour cells containing the event, and tumour cells without the event (Fig. 2a). We estimate the cellular prevalence of the CNA/LOH with the assumption that co-occurring events will be represented in the same clones, resulting from “punctuated” clonal expansions (Navin et al., 2011; Greaves and Maley, 2012). This motivates a clustering paradigm for statistical inference, allowing for increased power to detect weaker signals in the data across multiple loci, and to distinguish sets of events at different cellular prevalences (Fig. 1). We integrated this approach in a generative, factorial hidden Markov model (HMM) framework. The approach borrows statistical strength across adjacent genomic loci induced by segmental CNA and LOH events spanning multiple contiguous SNPs (Methods).

Our approach is distinct from related methods in the literature. Methods such as APOLLOH (Ha et al., 2012) and Control-FREEC (Boeva et al., 2012) model normal contamination from WGS of tumours, but do

not jointly infer CNA and LOH in a unified statistical approach, nor do they explicitly account for multiple tumour subpopulations. SNP genotyping array-based methods, such as OncoSNP (Yau et al., 2010), analyze CNA while accounting for intra-tumoural heterogeneity in cancer samples but cannot be directly applied to WGS data. Recently developed approaches, ABSOLUTE (Carter et al., 2012) and THetA (Oesper et al., 2013), were designed with the aim of predicting subclonal CNA events specifically for tumour sequencing data. However, neither tool employs a complete model that provides segmentation analysis. Moreover, THetA analyzes subclonal CNA in the absence of allelic ratios, which results in the omission of LOH and allelic imbalance. Finally, OncoSNP-seq (Yau, 2013) accounts for mixed populations in WGS data, but does not model distinct clonal populations in a clustering approach, which is characteristic of punctuated expansions.

We present a rigorous evaluation of TITAN including: 1) single-cell sequencing and fluorescence *in situ* hybridization (FISH) experimental validation of predictions on WGS data from a high grade serous ovarian tumour; 2) systematically engineered *in silico* mixtures with WGS data from multiple intra-patient samples; 3) artificially embedded CNA and LOH events in diploid chromosomes; and 4) 23 triple-negative breast cancer (TNBC) genomes. We compared TITAN with four published methods to demonstrate, with quantitative benchmarking on the ground truth datasets, that TITAN has higher sensitivity to detect low cellular prevalence events without decreasing specificity and has the capacity to accurately estimate cellular prevalences. Application of TITAN to 23 TNBC samples shows that a substantial proportion of clonal diversity is captured in the CNA/LOH dimensions, with low cellular prevalence LOH impacting allele-specific expression inferred from matched RNA-seq data. Finally, results from FISH and single-cell sequencing validation experiments confirm that predicted CNA/LOH events were consistent with our key modelling assumption: distinct cell populations can be identified through inference of CNA/LOH from WGS data. Together, these data show that the cellular prevalence profile of the copy number architecture from WGS provides an effective route to inferring clonal populations in patient tumour samples. Finally, we suggest how the deployment of TITAN in large scale clinical studies can dissect the interplay between clonal evolution, DNA repair deficiencies, and response to therapy.

Results

TITAN is a statistical model for predicting segmental CNA and LOH from matched tumour and normal WGS data. The input to the model is the full set of germline heterozygous SNP loci (identified from the normal sample) and the corresponding read depth and allele ratios at these SNP positions from the tumour. The output is a set of segmental CNA and LOH, clonal cluster memberships, and estimated cellular prevalences (Fig. 2b). The *tumour* and *sample* cellular prevalences are defined as the proportion of the tumour cells and the proportion of the sample (including normal cells) that harbour a CNA/LOH event, respectively. The model is predicated on four main assumptions: 1) joint analysis of allelic ratio and tumour sequence coverage (depth) at approximately one to three million heterozygous germline SNP loci reflect the underlying somatic genotype of the tumour; 2) segmental regions of CNA and LOH span 10s to 1000s of contiguous SNP loci; 3) the observed sequencing signal is an aggregated measure of heterogeneous cellular populations, including normal and tumour subpopulations (Fig. 2a,d); and 4) sets of genetic aberrations observed at similar cellular prevalence possibly co-occurred in the same clone. We incorporated these assumptions into a two-chain factorial HMM (Fig. 2c) (Methods and Supplementary Methods)

In order to evaluate and experimentally validate TITAN predictions, we used the genomes from a set of five synchronously resected pre-treatment high grade serous (HGS) ovarian cancer specimens (DG1136a,c,e,g,i) from the same patient. We obtained Illumina HiSeq 2500 WGS 100bp paired-end data, sequenced at $\sim 30X$, for each tumour sample and the patient's matched normal DNA for a total of six data sets. There were ~ 2.3 million high confidence heterozygous SNPs in the normal genome of DG1136. Across the five tumour samples, there were 2,816 CNA/LOH events (Supplementary Table 1), with a range of event sizes (Supplementary Fig. 1) covering an average of $\sim 1.5M$ SNPs per sample. We used these data to generate two types of benchmarking datasets for which these original events formed the ground truth 'positive set' in the quantitative evaluation. First, we simulated synthetic embedding of sampled CNA/LOH events into diploid chromosomes, and second, we systematically admixed the related but distinct CNA/LOH profiles from the tumour samples in known quantities to simulate mixed tumour populations. Finally, provision over the biological material allowed for experimental validation of TITAN predictions using FISH and single-cell sequencing.

Simulated CNA spike-in experiment demonstrates accurate detection for varying event sizes

We profiled the CNA landscape of DG1136a using two complementary methods: HMMcopy (Ha et al., 2012) and Control-FREEC (Boeva et al., 2012). Both methods identified a large deletion (chr16:46464744-90173515) and an amplification (chr8:97045605-144155272) of interest (Supplementary Fig. 2, Supplementary Table 1a). From these events, we randomly sampled log ratios and allele counts for non-consecutive sets of 10, 100 and 1000 positions. The sampled data were then inserted into diploid heterozygous chromosomes (chr1, 2, 9 and 18) at consecutive SNP positions to simulate segmental CNA and LOH events (Supplementary Methods). In total there were four deletions and four amplifications for each event size. Median genomic sizes of these events were 6.9Kb, 82Kb and 1.2Mb. We also included one deletion and one amplification event spanning 10,000 SNPs (median 12.5Mb) (Supplementary Table 2a). To vary cellular prevalences, we generated spike-in events sampled from two simulated tumour-normal admixtures at 80% and 60% of the original DG1136a dataset, computationally admixed with its matched normal WGS data (Supplementary Methods, Supplementary Fig. 3-6).

TITAN was run, from a range of zero to five clonal clusters, on the entire simulated sample (containing 78 events), including the chromosomes without spike-in events. The run with four clonal clusters was selected as optimal based on the *S_Dbw* validity index (Halkidi et al., 2002) (Supplementary Methods). All 54 events with size of 100 SNPs or larger were detected ($\text{TPR} \geq 0.9$); however, only 11 of 24 events with size of 10 SNPs were recalled (Supplementary Table 2b). The global false positive rate (FPR) was 0.04, which was computed by considering all SNP positions in chr1, 2, 9 and 18 where no spike-in data was inserted.

The tumour cellular prevalence estimates for two of the TITAN clonal clusters were 0.52 and 0.36, which were within range of the expected values of 0.52 and 0.39 (Supplementary Methods). For deletions and amplifications respectively, the cellular prevalence for 24 (89%) and 10 (37%) events with 100 SNPs or larger were correctly estimated ($\text{TPR} \geq 0.9$, Supplementary Fig. 3-6, Supplementary Table 2c). Despite the prevalence estimates of many amplifications not matching expected values, the events were still predicted to be subclonal but with a lower prevalence in some instances. On balance, the spike-in experiments demonstrated TITAN is accurate at detecting (sub)clonal events of varying sizes, but illustrates a potential limitation in detection of very small (10 SNP) events and estimation of the true prevalence for amplifications (see Discussion).

TITAN confers improved sensitivity for low prevalence events in simulated tumour subpopulations

To assess the performance of TITAN using benchmarking datasets that are more representative of clonal mixtures, we designed systematic experiments that simulated genomes with multiple tumour subpopulations at known proportions. In silico, we mixed the WGS data from DG1136a,c,e,g,i (Fig. 3a, Supplementary Table 3a). Within each mixture, we defined *clonal* events as CNA/LOH events present in all individual samples, and *subclonal* events as present in only a subset of the samples (Fig. 3b, Supplementary Table 3b-d). The proportion of tumour contribution from each individual sample (Supplementary Table 3a) in the mixture was used to compute the expected cellular prevalence (Supplementary Methods).

We combined DG1136e (67% tumour cellularity) and DG1136g (56% tumour cellularity), at mixture proportion increments of 10% (Methods), resulting in nine ($\sim 30X$) mixtures with two simulated tumour populations at 0.07/0.50, 0.13/0.45, 0.20/0.39, 0.27/0.33, 0.33/0.28, 0.40/0.22, 0.47/0.17, 0.53/0.11, 0.60/0.06 relative ratios (Supplementary Table 3b). Figure 3b illustrates a mixture scenario, which identifies true (sub)clonal events and their expected cellular prevalence. We compared accuracy of detection of CNA and LOH events using TITAN (run once each for a fixed number of clusters ranging from zero to five), APOLLOH (A) (Ha et al., 2012), Control-FREEC (CF) (Boeva et al., 2012) and BIC-seq (B) (Xi et al., 2011). After selecting the optimal number of clusters using the *S_Dbw* validity index (Methods), TITAN's median overall F-measure over the nine mixtures for predicted clonally dominant and subclonal events was 0.90. This was similar to APOLLOH (0.91) and Control-FREEC (0.88), but higher than BIC-seq (0.73) (Supplementary Fig. 7a, Supplementary Table 4a). While the precision for all approaches performed comparably (Fig. 3c), TITAN had higher sensitivity (median 0.91 compared to 0.85 (A), 0.83 (CF), 0.58 (B)), respectively) (Fig. 3d).

TITAN's sensitivity gains could be primarily attributed to improved sensitivity to subclonal events (Fig. 3e,f). Accordingly, we observed improved performance for runs with two or more clusters, but not when run with one cluster (Fig. 3g,h). Over the range of two to five clusters, recall was similar for subclonal events, suggesting TITAN is relatively stable in its predictions when accounting for more than one tumour subpopulation. Despite this stability, we explored the utility of unbiased model selection to choose the optimal number of clusters. We note that three clusters fit the scenario where events may be clonally dominant

(present in both samples) or subclonal, having one of two possible unique cellular prevalences contributing from the individual samples. Using the S_Dbw validity index, three clusters were appropriately selected as the optimal number for the majority of the mixtures (Supplementary Table 3b). The subclonal predictions using the optimal cluster runs consistently out-performed the other methods (Fig. 3e,f). In addition, performance gains were maintained across ranges of event lengths of 10Kb-100Kb, 100kb-1Mb, 1Mb-10Mb, and >10Mb (Supplementary Fig. 8).

Next, we used the DG1136 datasets to generate 10 pairwise (61-69X coverage) and 10 triplet (95-103X coverage) merged combinations of the five intra-tumour samples, mixed at approximately equal proportions (Supplementary Table 3c,d). In the triplet-merged samples (Supplementary Fig. 9, Supplementary Table 4b), TITAN performed comparably for all (clonal and subclonal) amplifications (0.85 median F-measure) relative to APOLLOH (0.85), Control-FREEC (0.85) and BIC-seq (0.60). For deletions and LOH events, TITAN showed statistically significant improvement over the other algorithms (0.91 compared to 0.87 (A), 0.83 (CF), and 0.60 (B); two-sample Wilcoxon rank-sum test $p < 0.001$) and LOH events (0.96 compared to 0.94 (A) and 0.85 (CF), $p < 0.001$). Similar performance was observed in the pairwise-merged samples with comparable F-measure for all amplifications (median 0.87 compared to 0.87 (A), 0.85 (CF), and 0.70 (B)) and statistically significant improvement in F-measure for deletions (0.95 compared to 0.92 (A), 0.87 (CF), and 0.69 (B); $p < 0.005$) and LOH events (0.98 compared to 0.96 (A) and 0.89 (CF)) (Supplementary Fig. 10, Supplementary Table 4c).

As shown for the serial mixture simulation, TITAN was more sensitive for subclonal events than the other methods in both pairwise and triplet merging simulations. Events unique to only one sample (subclonal in the mixture) were predicted with statistically higher recall by TITAN (Fig. 3i,j) (two-sample Wilcoxon rank-sum tests, $p < 0.001$). TITAN was also more sensitive to subclonal deletion and LOH events present in exactly two samples in the triplet-merged mixture ($p < 0.001$), while amplifications were comparable (Fig. 3i). All methods accurately predicted clonally dominant events for each merged simulation. Therefore, while maintaining accuracy of clonal events, TITAN showed clear advantages in detection of the engineered subclonal events. We attribute this increase in sensitivity directly to the consideration of heterogeneous cell populations in the model.

Accurate estimation of cellular prevalence and normal contamination

To evaluate the accuracy of cellular prevalence predictions (one of the key parameters estimated by TITAN), we compared the expected ground truth values for each simulated mixture. For each pairwise mixture, three clonal clusters were expected (one ancestral and two sample-specific), and for triplet mixtures, up to seven clonal clusters (all possible combinations of three samples) were expected. The expected cellular prevalence was computed using the tumour contribution from each individual sample making up the simulated mixture (Supplementary Methods, Supplementary Table 3b-d). Cellular prevalence estimates predicted by TITAN showed high and statistically significant positive correlation (Pearson's $r \geq 0.9$, $p < 0.001$, root mean squared error (RMSE) ≤ 0.11) with the expected tumour cellular prevalence across all samples in the serial (Fig. 4a, Supplementary Table 3b) and merging simulations (Fig. 4b,c, Supplementary Table 3c,d), demonstrating that the model was able to reproduce the engineered clonal structure.

Next, we compared cellular prevalence estimates between TITAN and THetA (Oesper et al., 2013). THetA's estimates also showed statistically significant correlation with expected values (Pearson's $r > 0.86$, $p < 0.001$, Fig. 4d,e); however, the RMSE was lower for TITAN (0.11) compared to THetA (0.18) for the serial mixtures, and similarly for the pairwise mixtures (0.07 compared to 0.12). Due to time complexity limitations, we were only able to run THetA for up to two tumour populations, therefore comparison on the triplet mixtures could not be performed. THetA has super-polynomial time and memory complexity when two or more tumour subpopulations are considered. Moreover, complexity increases with the number of input regions; thus, fewer than 15 segments were required for reasonable runtimes, resulting in lower resolution when inferring subclonal events. Finally, we also used an orthogonal approach (Control-FREEC) to estimate tumour content used to compute the expected cellular prevalence and observed similar correlations and RMSE results for TITAN and THetA (Supplementary Fig. 11, Supplementary Table 5).

Global normal contamination impacts the ability of the model to reconcile the presence of subclonality. We assessed the ability of the model to correctly estimate the global normal contamination in the model. TITAN estimates showed significantly positive correlation for the serial mixtures (Pearson's $r = 0.96$, $p < 0.0001$, RMSE=0.023), pairwise mixtures ($r = 0.86$, $p = 0.0014$, RMSE=0.047), and the triplet mixtures ($r = 0.74$, $p = 0.014$, RMSE=0.048) relative to the expected normal proportion (Supplementary Fig. 12a-c, Supplementary Table 3b-d). TITAN's estimates were considerably more accurate than THetA

for the serial ($r = 0.93$, $p < 0.0002$, RMSE=0.23) and pairwise mixtures ($r = 0.51$, $p = 0.14$, RMSE=0.3) (Supplementary Fig. 12d,e). Therefore, in addition to increased sensitivity for detecting subclonal events, TITAN showed accurate inference of cellular prevalence and normal proportion indicating the model formulation more closely models the mixture of cells generating the data. This not only contributes to increased accuracy in detection, but also adds an interpretive layer to estimating the composition of both tumour and normal cells being sequenced.

Characterization of the subclonal copy number alteration landscape in triple negative breast cancers

Having established quantitative accuracy with benchmarking datasets, we next analyzed a set of 23 triple-negative breast cancers (TNBC) with paired tumour-normal whole genome sequencing (Shah et al., 2012; Ha et al., 2012). We applied TITAN to predict regions of (sub)clonal CNA and LOH in the TNBC genomes (Supplementary Table 6) and profile the patterns of cell population structure inferred from the genome architecture. Six cases were clonally homogeneous (i.e. one clonal cluster) with cellular prevalence between 0.91-0.97, while the remaining 17 cases were more heterogeneous (between two and six number of clonal clusters and cellular prevalence estimates ranging 0.17-0.98 (Supplementary Table 7a). In the 17 heterogeneous cases, the proportion of the genome altered by subclonal events ranged between 0.16 to 0.73, with 12 cases having a higher proportion of subclonal alterations than clonal events (Supplementary Table 7b). This emphasizes the importance of considering mixed populations, which if neglected will lead to vastly different interpretations of the data and preclude inference of evolutionary patterns.

For four of the TNBC cases, whole exome capture data was also available. We applied TITAN to this exome sequencing data to demonstrate that the method can also be used for data for targeted genomic regions. Across the four samples, there were a total of 79,097 (~19,700 per sample) overlapping SNP positions between the exome and WGS data, of which 55,846 (71%) were concordant for predicted copy number between the TITAN results for the two data types (Supplementary Fig. 13, Supplementary Table 8). However, in the WGS data, TITAN appears to resolve the signal into potentially higher number of clonal clusters possibly due to 100 times more SNP loci (Supplementary Table 8).

For 22 of the TNBC cohort, we also analyzed the transcriptomes sequenced via RNA-seq (Shah et al.,

2012; Ha et al., 2012) to assess whether the clonality of LOH predictions influences allele-specific expression. Using the transcriptome allelic ratio (TAR; proportion of reference read counts) as an orthogonal measure, we compared TITAN cellular prevalence estimates to expected allelic imbalance in expression. We considered only RNA-seq positions within deletion LOH segments and estimated the expected baseline TAR as a function of the cellular prevalence $s_z + \frac{(1-s_z)}{2}$. The first term corresponds to the tumour cells having only one copy or allele and the second term represents all other cells without the LOH event (i.e. diploid and heterozygous). Across the 22 TNBC cases and each clonal cluster, the TAR was significantly correlated with the predicted cellular prevalence (Pearson's $r = 0.71$, $p = 1.5 \times 10^{-10}$, Supplementary Fig. 14a). TAR values were observed to be more imbalanced than could be explained by deletion LOH alone. This could be attributed to epigenetic silencing of one or both alleles in cells without LOH. When higher copy number events are considered, such as amplified LOH, this can also result in more imbalance than expected due to stronger representation of the homozygous allele if expression is positively correlated with copy number (Supplementary Fig. 14b). These results indicate that a substantial proportion of mono-allelic expression is associated with coincident subclonal LOH prediction, providing evidence that despite presence in only a minor cell population, these events are impacting the transcriptional program in these tumours.

FISH assays validate the presence of subclonal copy number changes

We performed fluorescence *in situ* hybridization (FISH) assays in two primary tumour HGS ovarian cancer samples, DG1136c and DG1136g, using cryosections from the same tissue blocks used for WGS (Supplementary Methods). We targeted seven TITAN CNA predictions (Supplementary Table 9a): one high prevalence deletion (C-DLOH-1) (Fig. 5), four low prevalence deletions (SC-DLOH-1, 3, 4 and 5) (Supplementary Fig. 15, 16), and one low prevalence gain (SC-GAIN-1, Supplementary Fig. 15) in DG1136g. An additional low prevalence gain (SC-GAIN-2) was assayed in sample DG1136c. Notably, all six low prevalence events were missed (detected as neutral regions) by at least one of the three other comparison methods: HMMCopy, Control-FreeC and THetA (Supplementary Table 10). Each event was scored by counting 100 or more nuclei to obtain a quantitative measure of the prevalence (Supplementary Table 9b-g).

The presence of all seven events was confirmed by FISH (Supplementary Table 9h). Quantitative estimates of prevalence by FISH were broadly consistent with TITAN predictions (Supplementary Table 10).

The highest prevalence event, C-DLOH-1 (chromosome 17 centromere), which represented a clonal deletion at 0.94 TITAN cellular prevalence (and predicted by all other methods), was observed in 0.77 of cells (Fig. 5, Supplementary Table 9h). SC-GAIN-2 (11q13.1), which had TITAN cellular prevalence of 0.62, was present in 0.61 of the nuclei scored by FISH in DG1136c. The remaining five events had predicted cellular prevalence of 0.51 in DG1136g. Two events, SC-DLOH-1 (1p31.1) and SC-DLOH-3 (2p16.1), were both observed at FISH prevalence of 0.48. These events were not predicted by Control-FreeC. SC-GAIN-1 (2p23.2), which was not detected by THetA, had an observed FISH prevalence of 0.35. Neither SC-DLOH-4 (7q35) nor SC-DLOH-5 (21q21.1), which were observed at FISH prevalences of 0.26 and 0.36, were predicted by HMMcopy, Control-FreeC nor THetA (Supplementary Table 10). Thus, experimental re-validation using cytogenetic techniques on cryopreserved patient material confirmed low and high prevalence gains and deletions predicted by TITAN, emphasizing the enhanced sensitivity conferred by modelling the presence of multiple populations.

Validation of TITAN predictions using single-cell sequencing confirm the presence multiple tumour populations

We further validated the CNA predictions from DG1136g using single-cell sequencing of targeted positions. The nuclei were isolated and sorted from disaggregated frozen tissue blocks and sequenced using multiplex PCR reactions and Fluidigm access array technology (Supplementary Methods). Two sets of events, Set1 and Set2 (Supplementary Table 11a, 12a), each included one high prevalence clonal LOH event, two subclonal deletions, two heterozygous diploid regions (Supplementary Fig. 17). For each set, 42 single cells were sorted, followed by library construction and sequencing; statistical analysis was then carried out independently for the two sets (Supplementary Methods).

This experiment focused on LOH events because confirmation of homozygosity (the absence of one allele) in single-cell sequencing is generally unambiguous. For statistical robustness, we interrogated multiple SNPs within each prediction of LOH (10-11 SNPs) and heterozygous (2-3 SNPs), negative control regions. We also selected previously validated somatic point mutations (SNVs), including a homozygous SNV in *TP53*, from this tumour. Because it is widely accepted that *TP53* mutation is a tumour-initiating event in HGS ovarian cancer (Ahmed et al., 2010; Cancer Genome Atlas Research Network, 2011; Bashashati et al.,

2013), this mutation was expected to be present in all tumour cells. *TP53*, along with the other SNVs, were used as markers to distinguish tumour and contaminating normal nuclei in this experiment (Supplementary Methods). This resulted in 14 tumour and 14 normal nuclei for Set1 (Fig. 6a, Supplementary Table 11b), and 9 tumour and 9 normal nuclei for Set2 (Fig. 6b, Supplementary Table 12b). The remaining nuclei contained insufficient read coverage for analysis.

For predicted clonal LOH events, we expected to observe homozygous signals for SNPs in all tumour nuclei. By contrast, for subclonal events, we expected homozygous SNPs to be present in only a sub-fraction of tumour nuclei. We used two statistical tests (Methods) to determine if an LOH event in a nucleus was present across the set of positions in the event. This involved controlling for expected allele drop out frequency (from unequal amplification of alleles) inferred from the normal nuclei at the predicted heterozygous loci. Over the set of positions in an event, we classified each nucleus as heterozygous or homozygous (or unknown if statistically inconclusive). As expected, for each of the normal nuclei in both Set1 and Set2, all LOH events were classified as heterozygous, independently confirming the initial grouping of cell types using mutations. In addition, the four negative control heterozygous events HET1, HET3 (Fig. 6a,c), HET4, HET5 (Fig. 6b,d) were each classified as heterozygous in all tumour nuclei for which sufficient coverage was obtained. By contrast, for the predicted clonal LOH events C-DLOH-1 (Fig. 6a,c) and C-NLOH-1 (Fig. 6b,d), all tumour nuclei were classified as homozygous, confirming the LOH predictions were clonally dominant. For each of the predicted subclonal deletion events (SC-DLOH-1, 3, 4 and 5), the tumour nuclei were divided into two groups with homozygous and heterozygous status, respectively. The proportions of tumour nuclei with homozygous status in these events were 0.54 (7/13 for SC-DLOH-1), 0.71 (10/14 for SC-DLOH-3), 0.50 (4/8 for SC-DLOH-4), and 0.50 (4/8 for SC-DLOH-5), which were generally consistent with the TITAN cellular prevalence estimate of 0.51 (Supplementary Table 3e). Therefore, in two independently executed single-cell sequencing experiments, we were able to relate our predictions back to the key modelling assumptions of TITAN and confirm the presence of the three cell types (Fig. 2a): 1) a population of normal cells, 2) a population of tumour cells harbouring the CNA/LOH event, and 3) a population of cells without the CNA/LOH event.

Discussion

TITAN is a novel algorithm that jointly analyzes both the tumour read depth and digital allele read counts for segmentation of subclonal CNA and LOH in whole genome sequencing of tumours. The advantages of TITAN are three-fold. First, the proper deconvolution of signals in the sequencing reads using the proposed sampling model allows for improved performance for predicting CNA and LOH (Fig. 3c,d). Second, the algorithm is more sensitive to subclonal events, which generally have more diluted signals, demonstrated by the serial and merging mixture experiments from a HGS ovarian cancer dataset (Fig. 3e-j). Third, estimation of tumour cellular prevalence and normal proportion is a powerful feature which enables inference of evolutionary clonal dynamics of the tumour's genome at CNA and LOH scales (Fig. 4, Supplementary Fig. 12). Importantly, using deep-sequencing of single nuclei, we were able to successfully confirm the presence of 6 (sub)clonal LOH and 2 subclonal copy number gain events. FISH experiments confirmed TITAN was able to detect low prevalence events that were not predicted by three other methods. In summary, results from experiments over a broad range of synthetic data, patient tumour data, and experimental validation have established that, through appropriate statistical modelling, the composition of cell populations in source tumour samples can be accurately identified by inference of CNA and LOH events in WGS data, and that consideration of mixed population leads to enhanced accuracy.

We demonstrated single-cell analysis as a viable experimental validation for unambiguously confirming the presence of TITAN-predicted (sub)clonal LOH events at the resolution of individual cells. The quantification of cellular prevalence was still challenging due to the small numbers of nuclei, and inference of patterns of clonal evolution was limited by the analysis of only three LOH events in each cell. Scaling up the numbers of tumour nuclei and LOH events will enable the statistical analysis of mutual exclusion and co-occurrence of events that are unattainable from single bulk tumour biopsies, and construction of clonal evolutionary phylogenies (Potter et al., 2013). More conventional FISH experiments also corroborated TITAN predictions, including low prevalence deletions and gains. Interestingly, for two events, prevalences determined by FISH were lower (0.26 and 0.36) than TITAN cellular prevalence estimates. While TITAN may have over-clustered them into a higher prevalence clonal cluster, the presence of these predicted CNAs was still confirmed by FISH. Moreover, the lack of prediction of these events by any of the other competing methods illustrates TITAN's enhanced sensitivity. These results also suggest that ~ 0.30 prevalence may fall

below the model's ability to accurately resolve the clustering for samples sequenced at $\sim 30X$. Thus, higher sequence coverage may be required for greater accuracy in cellular prevalence estimates.

TITAN has several limitations due to its specific modelling assumptions. TITAN does not model more than one aberrated genotype at the same locus, but instead assumes that clones harbouring the subclonal event coexist with tumour population(s) that have a normal (diploid heterozygous) genotype. An example of this limitation is shown in Figure 1, in which the event at chr2q23.3-q24.1 is likely an aggregated signal from a tumour population with a hemizygous deletion and another with a copy neutral LOH (subsequent duplication). In particular, it is difficult to distinguish among coexisting clones that harbour amplifications of variable copies. In order to model multiple tumour genotypes, the mixture representation model (Fig. 2a) will need to be re-formulated, yet given only tumour read depth and allele counts, the joint analysis of these signals may yield multiple solutions.

The primary aim of TITAN is the improved sensitivity for detecting subclonal CNA/LOH events, and we have demonstrated this performance in benchmarking mixtures up to $\sim 90X$ sequencing coverage (Fig. 3i, 4c); however, this is limited to resolving events in major clones at detectable prevalence. The full enumeration of clonal cell populations, including minor clones, is a limitation of TITAN and remains a difficult problem in the analysis of single tumour biopsies. This was demonstrated with FISH, which revealed the presence of rare, isolated nuclei harbouring genotypes not detectable by TITAN. Solutions to address these limitation will require integration of additional data, such as somatic SNVs (Carter et al., 2012; Roth et al., 2014; Fischer et al., 2014), genomic rearrangement breakpoints, increased coverage of sequencing (Nik-Zainal et al., 2012), multi-patient biopsies (Bashashati et al., 2013; Gerlinger et al., 2014), and ultimately, single-cell analysis (Navin et al., 2011; Potter et al., 2013).

TITAN estimates parameters using the EM algorithm; however, this approach may return locally optimal solutions influenced by the initializations prior to inference. Informed initializations of the key parameters, normal proportion and tumour ploidy, using orthogonal sources provided by histopathology may improve EM-converged parameter values and overall predictions. Also, we determined the optimal number of clonal clusters which best represented the expected number of clonal groups in the sample by employing a modified internal evaluation measure $S_{D_{bw}}$. A more robust solution would be to integrate the model selection directly into the framework, such as representing the clonal cluster groupings using a phylogenetic tree to

relate inferred clones into their ancestral lineages.

Our results indicate that neglecting to model WGS data as a composite of different tumour populations with diverse somatic genomes results in an incomplete representation of a tumour's CNA and LOH landscapes. This is exemplified by the analysis of the TNBC genomes, which suggested that the somatic copy number architecture in tumours may have evolved substantially. It was recently shown that the inference of subclonal mutations impacts the outcome of chronic lymphocytic leukemia under treatment-based selective pressures (Landau et al., 2013). Whether similar outcome correlations are detectable when CNA and LOH are the aberrations measured for evolutionary patterns remains an open question. We anticipate that the TITAN framework will provide a robust analytical route for studying the degree to which clonal evolution, driven from the CNA and LOH perspective, influences treatment sensitivity.

A downstream clinical application of TITAN will be the identification of tumours with DNA repair defects that would make them susceptible to genotoxic drugs. Compromised homologous recombination (HR) due to loss of *BRCA1/2* is a dominant molecular feature of HGS ovarian cancer (Cancer Genome Atlas Research Network, 2011) and other epithelial malignancies (Yang et al., 2013). HR defects lead to accrual of genomic structural changes through un-repaired double strand genomic breaks (Lord and Ashworth, 2012). HR-deficient cells carry the genomic footprints of progressive accumulation of structural genomic aberrations including regions of CNA and LOH. Tumours exhibiting subclonal CNA and LOH events are likely to have acquired HR defects and would be good candidates for treatment with platinum-based drugs or PARP inhibitors (Wang et al., 2012; Mukhopadhyay et al., 2012). Large scale whole genome sequencing studies of epithelial cancers coupled with reliable homologous recombination deficiency assays will be an appropriate way to test this hypothesis.

In conclusion, the TITAN statistical framework represents a significant advance in the field of copy number and LOH analysis for tumour genome sequencing data. The implications for clonal evolution of point mutations in clinical trajectories are well-documented; we suggest that TITAN will enable the execution of complementary studies to investigate the role of genome architecture in driving the evolutionary selection of clonal cell populations.

Methods

The TITAN statistical model

To model tumours containing multiple tumour subpopulations, we assume the observed measurements were generated from a composite of three types of cell populations (Yau et al., 2010) with relative proportions as follows: n : the proportion of non-malignant cells; $(1 - n) s_z$: the proportion of tumour cells with normal genotype; and $(1 - n) (1 - s_z)$: the *sample cellular prevalence* or the proportion of tumour cells harbouring the CNA or LOH event of interest (Fig. 2a). s_z is the proportion of tumour cells that is diploid heterozygous (and therefore normal) at the locus. Thus, $(1 - s_z)$ is the *tumour cellular prevalence* or the proportion of the tumour population containing the event. We assume multiple somatic events share similar cellular prevalence and thus can be assigned to one of a finite number of clonal clusters, $z \in Z$. This allows for sufficient data points to robustly infer the model parameters by borrowing statistical strength. The simultaneous inference and clustering of each data point to $z \in Z$ is the primary distinguishing feature over related work (Yau et al., 2010; Van Loo et al., 2010; Carter et al., 2012; Oesper et al., 2013; Yau, 2013).

The inputs to the model are quantities readily extracted from WGS sequence data (Methods, Fig. 2b). The analysis requires the genome-wide set of T germline heterozygous SNP positions derived from the normal genome, which generally ranges from one to three million per patient. At each SNP, copy number data from the tumour genome is represented by the log ratio between the tumour and normal read depths $l_{1:T}$. We assume $l_{1:T}$ is Gaussian distributed: $l_{1:T} \sim \mathcal{N}(l_{1:T} | \mu_{g,z}, \sigma_g^2)$. We assume the reference allelic read counts from the tumour $a_{1:T}$ are Binomial distributed $a_{1:T} \sim \text{Bin}(a_{1:T} | N_{1:T}, \omega_{g,z})$, where $N_{1:T}$ represents the sequencing depth at each position. The cluster-specific parameters $\mu_{g,z}$ and $\omega_{g,z}$ are functions of s_z (Fig. 2d), and therefore represent the signals from the three types cell populations. This formulation enables TITAN to be more sensitive to events with lower cellular prevalences.

Segmental CNA and LOH events span many contiguous SNP positions, thereby inducing spatial correlation along the chromosome. To capitalize on expected shared signals from adjacent positions, TITAN was implemented as a two-factor hidden Markov model (HMM) where the hidden genotypes $G_{1:T}$ and the hidden clonal cluster memberships $Z_{1:T}$ comprise the two chains (Fig. 2c). The state space is dynamically expanded as a function of clonal cluster membership, resulting in $|G| \times |Z|$ number of state tuples

($g \in G, z \in Z$) (Supplementary Table 14). The HMM is fit to the data using expectation maximization (EM) as described in the Supplementary Methods.

The final output of TITAN is a list of segment boundaries that represent CNA and LOH events with accompanying estimates of the genotype, the cellular prevalence ($1 - s_z$) and clonal population cluster membership for each event. In addition, estimation of global parameters n the normal proportion and ϕ the estimated ploidy are output. The parameters of the probabilistic graphical model (Fig. 2c) are defined in Supplementary Table 13 and full mathematical details are described in Supplementary Methods.

Analysis workflow

The analysis workflow of TITAN for tumour whole genome sequencing data is shown in Figure 2b. First, germline heterozygous SNP positions $\mathbf{L} = \{t_i\}_{i=1}^T$ is identified from the normal genome using SAMtools *mpileup* (Li et al., 2009). The analysis focuses on ~ 1 -3 million loci genome-wide per patient and allows for identification of somatic allelic imbalance events (Ha et al., 2012). From the tumour genome data, the read counts mapping to the reference base (A allele) and total depth at all positions in \mathbf{L} are extracted and represented as $a_{1:T}$ and $N_{1:T}$, respectively.

The tumour copy number is normalized for GC content and mappability biases using only the normalization component of HMMcopy (<http://bioconductor.org/packages/2.11/bioc/html/HMMcopy.html>). Briefly, the genome is divided into bins of 1kb and read count is represented as the number of reads overlapping each bin. Loess curve fitting and correction was performed on tumour and normal samples, separately. The corrected read counts for the overlapping 1kb bin at each position of interest $t \in \mathbf{L}$, \bar{N}_t and \bar{N}_t^N , is used to compute the log ratio, $l_{1:T} = \log(\bar{N}_{1:T}/\bar{N}_{1:T}^N)$.

TITAN jointly analyzes the data $l_{1:T}$, $a_{1:T}$, $N_{1:T}$ to segment the data into regions of CNA/LOH and estimate normal contamination, tumour ploidy and cellular prevalences for Z number of clonal clusters. For a range of $i = 1$ to 5, TITAN is run once for the set of clonal cluster states $Z_i := \{1, \dots, i\}$, where $|Z_i| = i$ is the number of clonal clusters. The optimal number of clusters i is then chosen using the minimum *S_Dbw* validity index (Supplementary Methods).

In-silico mixture experiments simulating multiple tumour subpopulations

Five intra-patient samples from patient DG1136 was used to simulate multiple cellular populations by mixing combinations of samples at known proportions. For the predefined serial mixture experiment, nine whole genome mixtures at $\sim 30X$ coverage were generated by sampling reads from DG1136e and DG1136g at mixing proportions of 10% increments (0.1e/0.9g, 0.2e/0.8g, ..., 0.8e/0.2g, 0.9e/0.1g). The expected relative tumour content contribution from the two samples were computed for each mixture based on tumour cellularity of 67% and 56%, respectively, as a consensus estimates by APOLLOH and the pathological review (Supplementary Table 3b). HMMcopy and APOLLOH (Ha et al., 2012) results from the individual samples were used as ground truth CNA and LOH events, respectively with default parameters were used (<http://compbio.bccrc.ca/software/>). For the merging of two or three samples at approximately equal proportions, five intratumour samples were merged together to generate ten pairs at $\sim 60X$ coverage (Supplementary Table 3c) and ten triplets at $\sim 90X$ coverage (Supplementary Table 3d) coverage for each combination. This was done using SAMtools (Li et al., 2009) merge command.

Precision, recall, and F-measure were computed based copy number status at heterozygous germline SNP positions from the individual samples (prior to mixing) predicted by APOLLOH/HMMcopy. Performance was calculated for deletions, gains, and LOH independently, and averaged together for overall assessment shown in Figure 3c-j. The number of SNPs for ground truth deletion, amplification and LOH events used to calculate performance metrics are given in Supplementary Table 1, 3, 4. See Supplementary Methods for more details.

Statistical tests for single-cell sequencing experiments

Two statistical tests were used to determine if an event in a nucleus was statistically significant for LOH. First, we addressed allelic drop-out, which is the preferential amplification of one allele at a heterozygous locus, leading to a homozygous signal that may be mistaken for LOH. Using the expected allelic drop-out rate (*DOR*) of 0.28 (Set1) and 0.48 (Set2) determined from the normal nuclei (Supplementary Table 11c, 12c), we applied a one-tailed binomial test in which the null hypothesis asserts that the proportion of homozygous positions is not greater than the expected *DOR*. The second test examined whether the allelic ratio distribution (Figure 6a,b) across the positions for an event showed statistically significant difference

compared to the expected heterozygous allelic ratio (HAR) as determined from the normal nuclei. Finally, the maximum of the (Benjamini & Hochberg (FDR) adjusted) p-values between the two tests was used to determine if an event was statistically significant ($FDR < 0.05$) for LOH status, or heterozygous (HET) otherwise; events were designated as unknown or ambiguous for non-significant FDR and absence of a heterozygous position (Supplementary Table 11e,f and 12e,f).

Additional methods on mathematical details of the TITAN model, the inference algorithm and software as well as experimental protocols for generating the validation data are provided in the Supplementary Methods.

Data access

The ovarian cancer genome sequence data, including the single-cell data, have been submitted to the European Genome-phenome Archive (EGA; <https://www.ebi.ac.uk/ega/>) under accession EGAS00001000547. TITAN is available at <http://compbio.bccrc.ca/software/titan/> and can be downloaded from Bioconductor under the R package, `TitanCNA`.

Acknowledgements

We wish to thank the British Columbia Cancer Foundation for research funding support. In addition, this work was funded by the Canadian Institutes for Health Research (CIHR), Genome Canada/Genome British Columbia, Canadian Cancer Society Research Institute, and the Terry Fox Research Institute grants to SPS and SA. SPS is supported by the Michael Smith Foundation for Health Research and is the Canada Research Chair (CRC) in Computational Cancer Genomics. SA is the CRC in Molecular Oncology. GH is supported by the Natural Sciences and Engineering Research Council of Canada. We thank Dr. Sarah Mullaly for critical reading of the manuscript.

Author's contributions

SPS oversaw the project. SPS, SA, and GH conceived and wrote the manuscript. GH, SPS and AR: algorithm design. GH implemented the software and carried out all analytical experiments. AR, AM, LMP, AB,

JD, AL, JR and KS: data analysis and discussions. JK, DY, EL and JB: single-cell sequencing experiment. JH, NM, LMP: FISH assays. CBG, DGH, JNM: sample preparation and histopathological review. MAM: genome sequencing.

Figures

Figure 1: Detection of subclonal deletions in whole genome sequencing data of a triple negative breast cancer genome. Copy number is represented as the log ratio of tumour and normal read depth. Discrete copy number status shown is predicted as either a hemizygous deletion (HEMD; green), copy neutral (NEUT; blue), or gain/amplification (AMP; red). Allelic ratios are computed as the proportion of reads matching the reference genome. The LOH status shown is one of heterozygous (HET; grey), LOH (green), copy neutral LOH (NLOH; blue), or allele-specific gain/amplification (ASCNA; red). Subclonal deletions are observed to have a weaker log ratio signal that is closer to 0 and shows less spreading in allelic ratios (Deletion I) compared to clonal deletions (Deletion II); the sample cellular prevalence estimates (proportion of sample) for ‘Deletion I’ indicates it is in a subclonal cluster ‘Z2’. ‘Deletion I’ and ‘Deletion III’ are clustered into the same subclonal cluster because they share similar signals, and therefore, the same cellular prevalence in the data. ‘Deletion II’ is present in all tumour cells, indicated by being in the clonal cluster ‘Z1’. Tumour cellularity of 84% (normal contamination of 16%) is denoted with a black horizontal line. The average tumour ploidy (haploid coverage factor) was estimated as 1.66 by genome-wide analysis. The log ratio and symmetric allelic ratio ($\max(\text{reference reads}, \text{variant reads})/\text{depth}$) Gaussian kernel densities are shown for all deletions in chr2.

Figure 2: Description of the TITAN probabilistic framework

(a) Representation of the aggregate copy number signal from mixed populations in a heterogeneous tumour sample. c is the aggregate signal that is composed of 3 components: normal population (white circles), tumour populations with the deletion (green decagons) and without the event (blue decagons). n is the normal proportion; s_z is the tumour proportion for the z^{th} clonal cluster that *does not* contain the event; c_{norm} and c_{DEL} are normal and tumour copy numbers. Therefore, $(1 - s_z)$ corresponds to the proportion of tumour harbouring the event, also defined as the tumour cellular prevalence of the z^{th} clonal cluster.

(b) Analysis workflow for TITAN. Three inputs are required: 1) Heterozygous positions identified in the normal DNA predicted by genotyping tools such as SAMtools mpileup (Li et al., 2009); 2) Reference counts a and read depth N are extracted at these positions from aligned reads in the tumour DNA sequence data; 3) The tumour and normal read depths, N and N_N , are normalized independently to correct GC content and mappability biases; log ratios $l = \log(N/N_N)$ of the corrected read counts are computed. The output are the optimal sequence of CNA/LOH genotypes and clonal cluster memberships at each position. Model parameter for normal contamination n , tumour cellular prevalence s_z , and tumour ploidy ϕ are estimated.

(c) Probabilistic graphical model of TITAN. Shaded nodes are known or observed quantities; open nodes are random variables of unknown quantities. Arrows represent conditional dependence between random variables. Full details and definitions are in Methods and Supplementary Table 13. **(d)** Parameter trace of $\omega_{g,z}$ and $\mu_{g,z}$ when cellular prevalence varies. s_1 and s_2 is shown as the tumour cellular prevalence (i.e. transformed using $1 - s_z$). n is normal proportion and ϕ is average tumour ploidy. Each CNA/LOH genotype is shown (Supplementary Table 14) with the associated integer copy number in parentheses.

Figure 3: Performance of TITAN in serial and merging simulations using real intratumoural samples from a HGS ovarian carcinoma.

(a) Patient DG1136 had biopsies synchronously resected from four sites in the primary tumour of the right ovary and one site from the left pelvic sidewall metastasis.

(b) Illustration demonstrating the expected proportions in a simulation of two tumour subpopulations. The tumour content of Sample *a* (80%) and Sample *b* (70%) inform the sample cellular prevalence in the merged Sample *a* + *b*. Events found in all samples of the mixture represent simulated clonal events. For example, the (green) deletion is present in 75% of the merged sample (or 100% of tumour cells) given that the normal proportion is 25%. Events present in a subset of samples in the mixture simulate subclonal events such as for the (red) gain unique to Sample *a* which is present in 40% of the merged sample or 53% of the tumour cells.

(c-f) Performance of the serial mixture experiment between TITAN, APOLLOH (Ha et al., 2012) (which includes HMMcopy), Control-FREEC (Boeva et al., 2012), and BIC-seq (Xi et al., 2011). The mixture proportion includes 0.1:0.9, 0.2:0.8, ..., 0.9:0.1 relative ratios of DG1136e:DG1136g. Precision **(c)** and recall **(d)** are shown for subclonal and clonal events averaged across gains, deletions, and LOH events. Recall performance for truth events found uniquely in Sample *e* **(e)** or Sample *g* **(f)** are shown. ‘Mixture Proportion’ is defined as the ideal mixing fractions (e.g. 10%, 20%, etc.); expected tumour ‘cellular prevalence’ is defined as the expected tumour contribution, at a given mixture proportion, from each individual sample making up the mixture. The expected tumour cellular prevalence shown was computed by adjusting the mixture proportion for tumour content of 67% and 56% for DG1136e and DG1136g, respectively. Ground truth events were identified in the individual samples of the mixture using APOLLOH/HMMcopy and expected tumour cellular prevalence values are shown in Supplementary Table 3b.

(g-h) Serial mixture performance for TITAN runs initialized with number of clusters ranging from 1 to 5. Recall performance for events found uniquely in DG1136e **(g)** or DG1136g **(h)** represent events that are subclonal within the simulated mixture. Average recall across deletions, gains, and LOH events are shown. The one cluster run represents the scenario where only one tumour population exist.

(i-j) Comparison of recall performance distributions across 10 paired **(i)** and 10 triplet **(j)** merging simulations for TITAN (T), APOLLOH/HMMcopy (A), and Control-FREEC (CF). Performance is shown for simulated subclonal events, which were present uniquely in exactly one (Subclonal 1) and exactly two (Subclonal 2) samples making up the mixture, and in contrast, clonally dominant events were present in all samples of the mixture (Clonal).

Figure 4: Performance of TITAN tumour cellular prevalence estimates for serial (30X) and pairwise (60X)/triplet (90X) merging simulations of intra-tumour samples from a HGS ovarian carcinoma. Pearson correlation coefficients (*r*) and root mean squared error (RMSE) were computed for TITAN **(a-c)** and THetA (Oesper et al., 2013) **(d-e)**. Correlation and RMSE were computed by comparing the cellular prevalences of the predicted clusters with the prevalence of the expected clusters across the mixture samples. Each data point represents an expected clonal cluster with a unique tumour cellular prevalence. Ground truth and expected tumour cellular prevalence values were computed from the tumour contribution from each individual sample making up the simulated mixture (Supplementary Table 3b-d).

Figure 5: Fluorescence in-situ hybridization (FISH) validation of TITAN predictions for chromosomes 1 and 17 in DG1136g. **(a)** Subclonal hemizygous deletion, SC-DLOH-1, in chromosome 1 was validated using BAC probe RP11-795A13 (orange, chr1:69851036-70025173). Control probe for copy neutral regions was RP11-159J14 (green, chr1:69454844-69606688). FISH imaging shows tumour cells with a deletion (green arrow) and diploid (white arrow) at this region. **(b)** Clonal deletion, C-DLOH-1, in chromosome 17 was validated using the centromeric probe, CEP 17. The BAC probes RP11-147K16 (orange, chr17: 3294803-3452243) and RP11-982O5 (blue, chr17:55475584-55662513) were used as controls. The majority of cells were observed to harbour the deletion. FISH count prevalence was computed as the proportion of nuclei with event:control count ratio that is < 1 (deletion) or > 1 (gain) (Supplementary Table 9h). FISH imaging is shown at 63X magnification. Copy number predictions are shown using log ratios (normalized tumour depth/normal depth). Copy neutral (blue), hemizygous deletion (green), and copy gain (red) predictions are shown. Cellular prevalence estimates for clonal cluster 1 (Z1) and cluster 2 (Z2) predicted by TITAN are shown; tumour cellularity is indicated with the black horizontal line.

Figure 6: Single-cell validation of subclonal deletions in DG1136g using deep DNA sequencing of individual nuclei. **(a,b)** The 28 nuclei for Set1 and 18 nuclei for Set2 were designated as tumour and normal cell type using the status of mutations. The mutant allele ratio ($variant\ reads/depth$) for mutations and symmetric allele ratio ($max(reference\ reads, variant\ reads)/depth$) for SNP positions are shown for Set1 **(a)** and Set2 **(b)** events. Low coverage positions are shaded in grey. **(c,d)** The LOH status for each event for Set1 **(c)** and Set2 **(d)** were determined using the binomial test for drop-out and Wilcoxon rank sum test for allelic ratios. *TP53* mutation status is shown. The LOH status for each heterozygous (HET) and LOH (C-DLOH, SC-DLOH) event is shown. ‘Tumour’ nuclei having the LOH event (green) or not having the event (blue) to illustrate the original 3 component mixture model (Fig. 2a). Normal nuclei are designated ‘Normal’ (white). Unknown events (grey) were inconclusive for HET or LOH status. See Supplementary Methods for details.

References

- Ahmed A. A, Etemadmoghadam D, Temple J, Lynch A. G, Riad M, Sharma R, Stewart C, Fereday S, Caldas C, Defazio A, *et al.*, 2010. Driver mutations in tp53 are ubiquitous in high grade serous carcinoma of the ovary. *J Pathol*, **221**(1):49–56.
- Aparicio S and Caldas C, 2013. The implications of clonal genome evolution for cancer medicine. *N Engl J Med*, **368**(9):842–851.
- Bashashati A, Ha G, Tone A, Ding J, Prentice L. M, Roth A, Rosner J, Shumansky K, Kalloger S, Senz J, *et al.*, 2013. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol*, **231**(1):21–34.
- Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, and Barillot E, 2012. Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**(3):423–425.
- Cancer Genome Atlas Research Network, 2011. Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353):609–615.
- Carter S. L, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird P. W, Onofrio R. C, Winckler W, Weir B. A, *et al.*, 2012. Absolute quantification of somatic dna alterations in human cancer. *Nature Biotechnology*, **30**(5):413–421.
- Cibulskis K, Lawrence M. S, Carter S. L, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander E. S, and Getz G, *et al.*, 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, **31**(3):213–219.
- Ding L, Ley T. J, Larson D. E, Miller C. A, Koboldt D. C, Welch J. S, Ritchey J. K, Young M. A, Lamprecht T, McLellan M. D, *et al.*, 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**(7382):506–510.
- Fischer A, Vázquez-García I, Illingworth C. J. R, and Mustonen V, 2014. High-definition reconstruction of clonal composition in cancer. *Cell reports*, **7**(5):1740–1752.
- Gerlinger M, Horswell S, Larkin J, Rowan A. J, Salm M. P, Varela I, Fisher R, McGranahan N, Matthews N, Santos C. R, *et al.*, 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, **46**(3):225–233.
- Gerlinger M, Rowan A. J, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, *et al.*, 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*, **366**(10):883–892.
- Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, and Beerenwinkel N, 2012. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, **3**:811.
- Greaves M and Maley C. C, 2012. Clonal evolution in cancer. *Nature*, **481**(7381):306–313.

- Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K, *et al.*, 2012. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research*, **22**(10):1995–2007.
- Halkidi M, Batistakis Y, and Vazirgiannis M, 2002. Clustering validity checking methods: part ii. *SIGMOD Rec.*, **31**(3):19–27.
- Landau D. A, Carter S. L, Stojanov P, McKenna A, Stevenson K, Lawrence M. S, Sougnez C, Stewart C, Sivachenko A, Wang L, *et al.*, 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, **152**(4):714–726.
- Larson N. B and Fridley B. L, 2013. Purbayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics (Oxford, England)*, **29**(15):1888–1889.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and Subgroup . G. P. D. P, *et al.*, 2009. The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16):2078–2079.
- Lord C. J and Ashworth A, 2012. The dna damage response and cancer therapy. *Nature*, **481**(7381):287–294.
- Mukhopadhyay A, Plummer E. R, Elattar A, Soohoo S, Uzir B, Quinn J. E, McCluggage W. G, Maxwell P, Aneke H, Curtin N. J, *et al.*, 2012. Clinicopathological features of homologous recombination-deficient epithelial ovarian cancers: sensitivity to parp inhibitors, platinum, and survival. *Cancer Research*, **72**(22):5675–5682.
- Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, *et al.*, 2011. Tumour evolution inferred by single-cell sequencing. *Nature*, **472**(7341):90–94.
- Nik-Zainal S, Van Loo P, Wedge D. C, Alexandrov L. B, Greenman C. D, Lau K. W, Raine K, Jones D, Marshall J, Ramakrishna M, *et al.*, 2012. The life history of 21 breast cancers. *Cell*, **149**(5):994–1007.
- Nowell P. C, 1976. The clonal evolution of tumor cell populations. *Science*, **194**(4260):23–8.
- Oesper L, Mahmoody A, and Raphael B. J, 2013. Theta: Inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome biology*, **14**(7):R80.
- Potter N. E, Ermini L, Papaemmanuil E, Cazzaniga G, Vijayaraghavan G, Titley I, Ford A, Campbell P, Kearney L, and Greaves M, *et al.*, 2013. Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Research*, **23**(12):2115–2125.
- Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, and Shah S. P, *et al.*, 2014. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, **11**(4):396–398.
- Shah S. P, Morin R. D, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, *et al.*, 2009. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**(7265):809–813.

- Shah S. P, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, *et al.*, 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**(7403):395–399.
- Sottoriva A, Spiteri I, Piccirillo S. G. M, Touloumis A, Collins V. P, Marioni J. C, Curtis C, Watts C, and Tavaré S, 2013. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci*, **110**(10):4009–4014.
- Van Loo P, Nordgard S. H, Lingjærde O. C, Russnes H. G, Rye I. H, Sun W, Weigman V. J, Marynen P, Zetterberg A, Naume B, *et al.*, 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci*, **107**(39):16910–16915.
- Wang Z. C, Birkbak N. J, Culhane A. C, Drapkin R, Fatima A, Tian R, Schwede M, Alsop K, Daniels K. E, Piao H, *et al.*, 2012. Profiles of genomic instability in high-grade serous ovarian cancer predict treatment outcome. *Clinical Cancer Research*, **18**(20):5806–5815.
- Xi R, Hadjipanayis A. G, Luquette L. J, Kim T.-M, Lee E, Zhang J, Johnson M. D, Muzny D. M, Wheeler D. A, Gibbs R. A, *et al.*, 2011. Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proc Natl Acad Sci*, **108**(46):E1128–E1136.
- Yang L, Luquette L. J, Gehlenborg N, Xi R, Haseley P. S, Hsieh C.-H, Zhang C, Ren X, Protopopov A, Chin L, *et al.*, 2013. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, **153**(4):919–929.
- Yau C, 2013. Oncosnp-seq: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics*, **29**(19):2482–2484.
- Yau C, Mouradov D, Jorissen R. N, Colella S, Mirza G, Steers G, Harris A, Ragoussis J, Sieber O, and Holmes C. C, *et al.*, 2010. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol*, **11**(9).

Chromosome 2

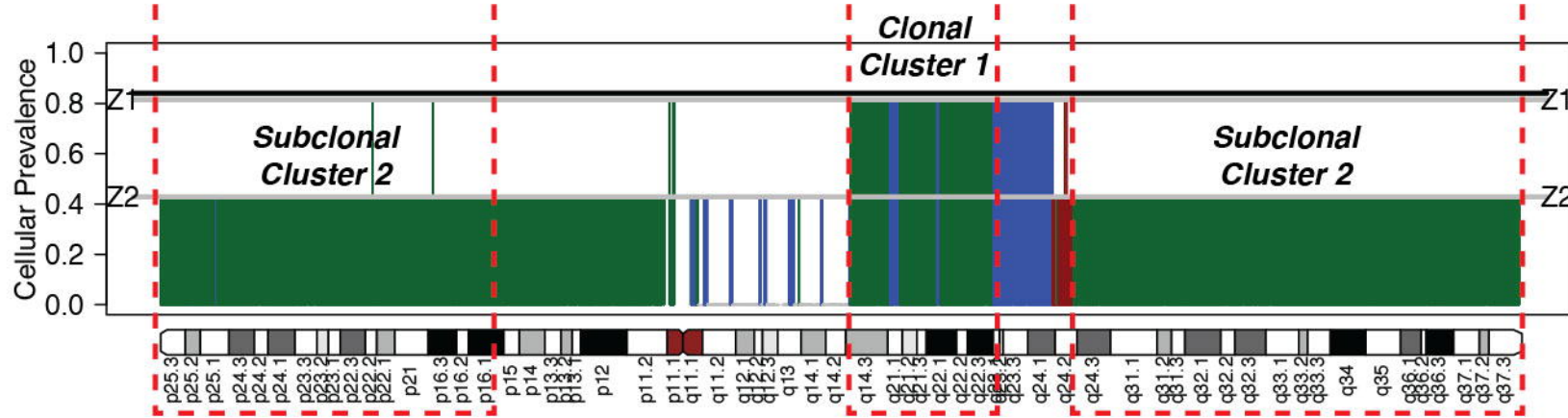
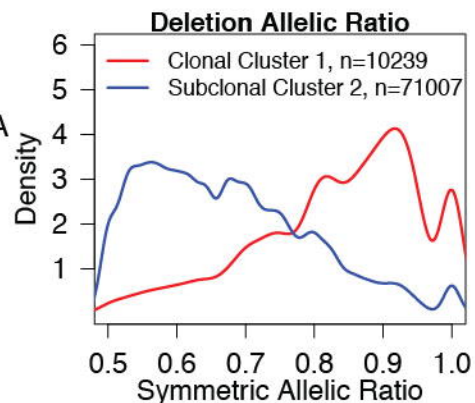
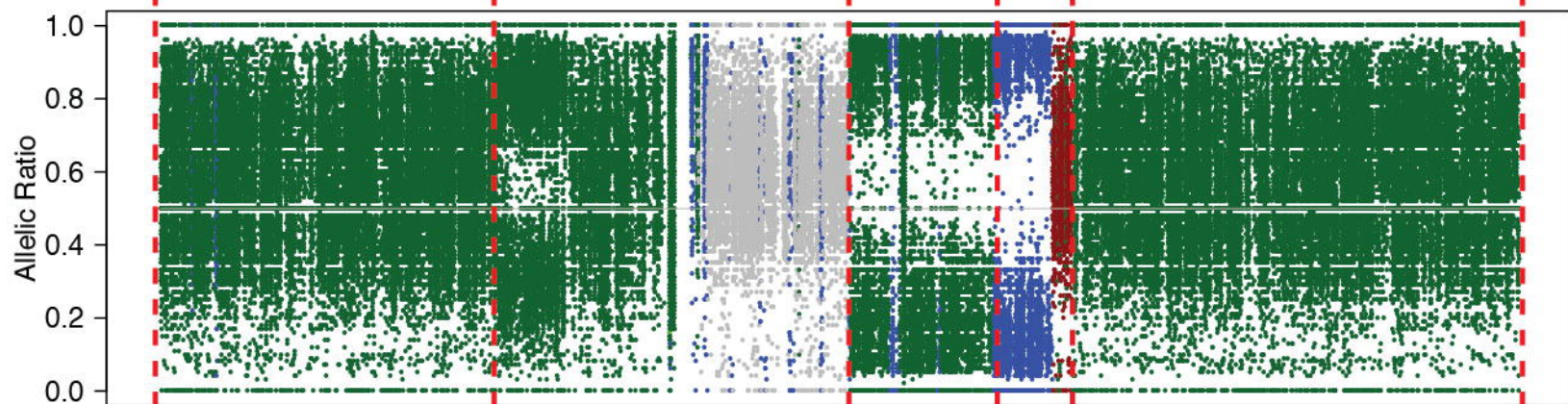
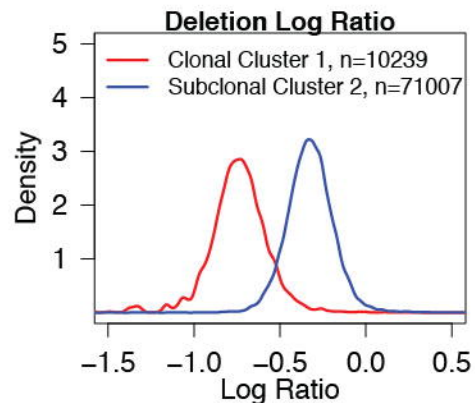
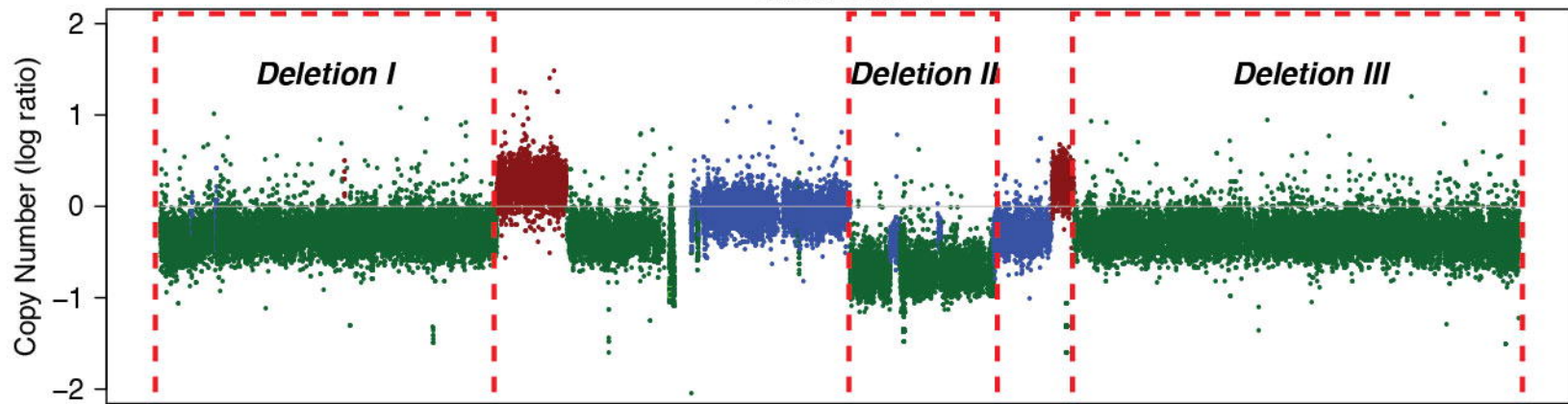


Figure 1

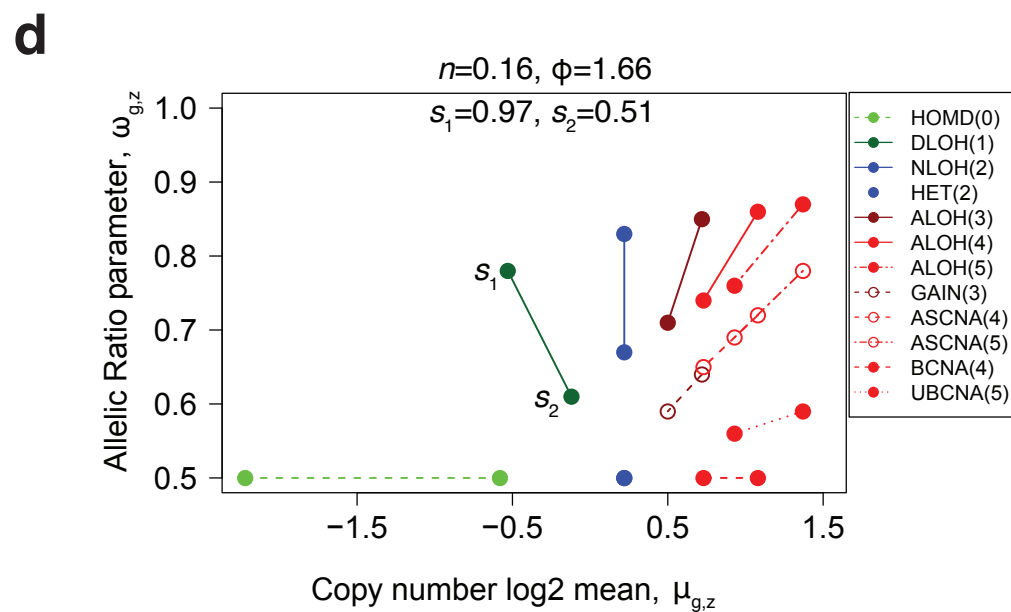
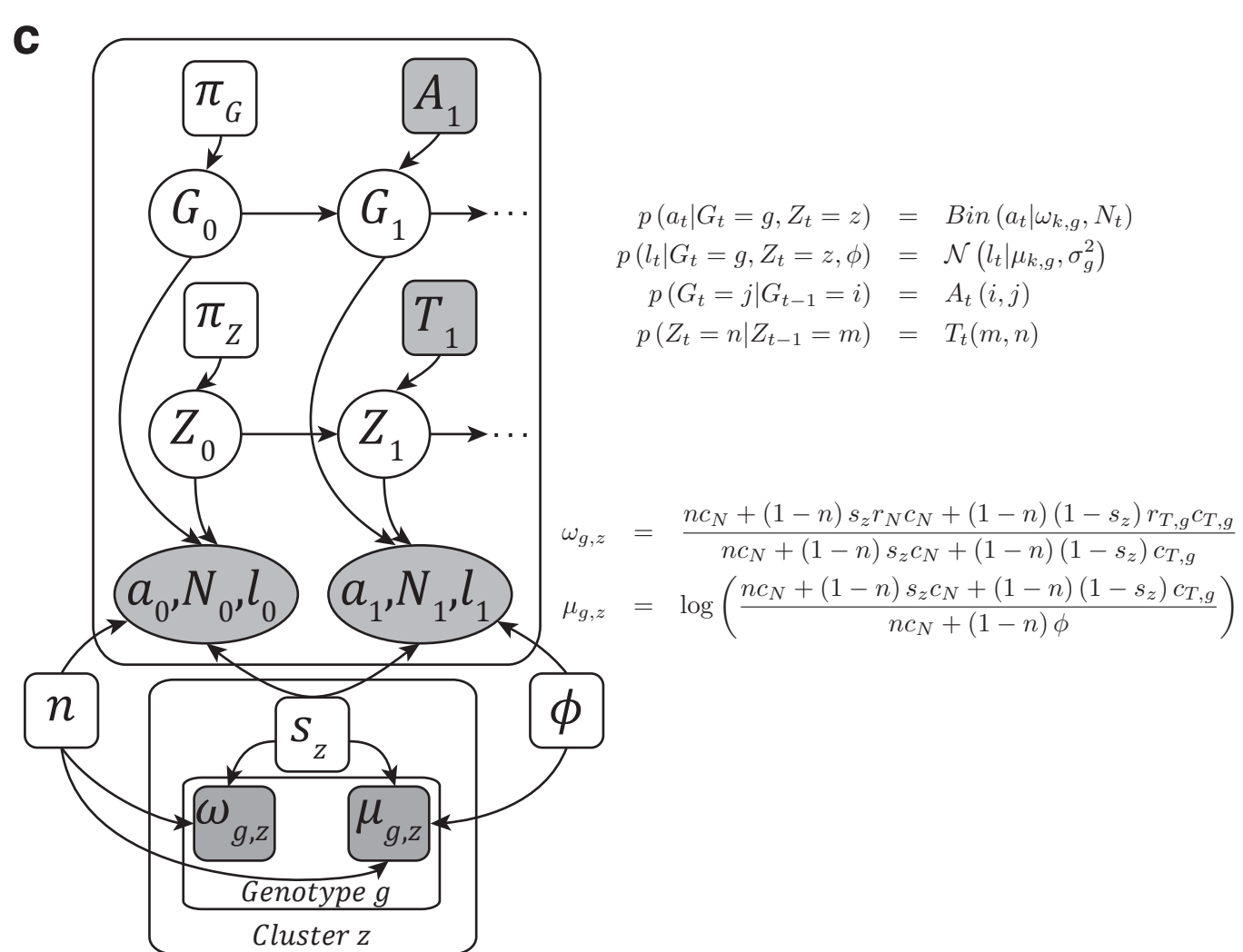
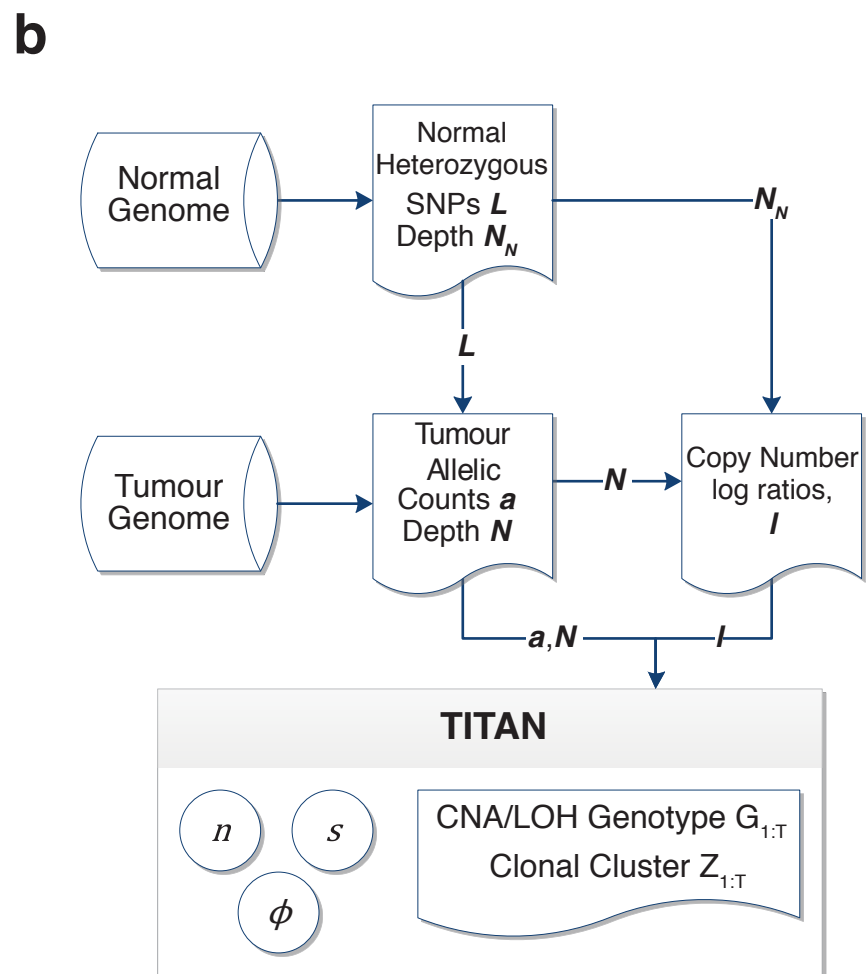
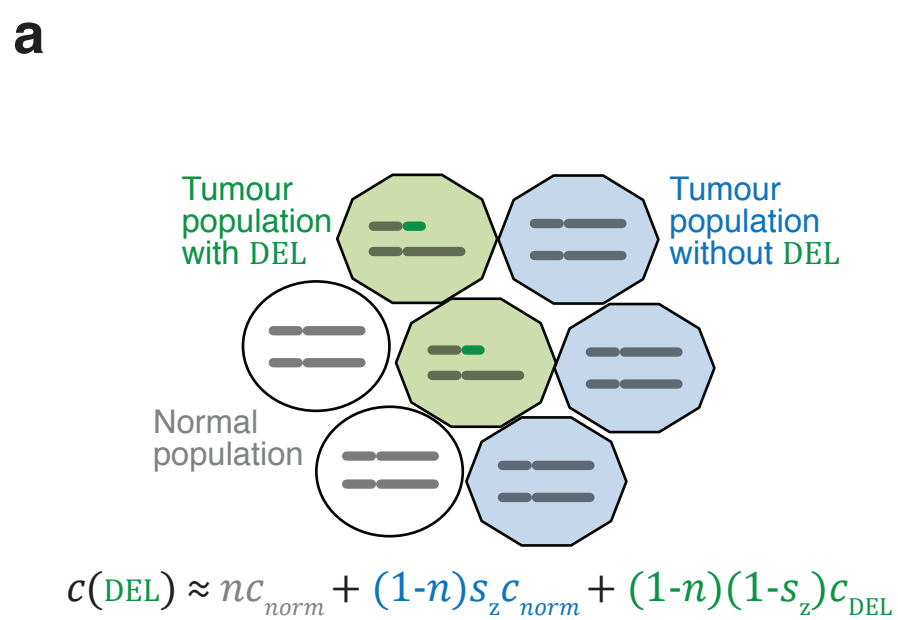


Figure 2

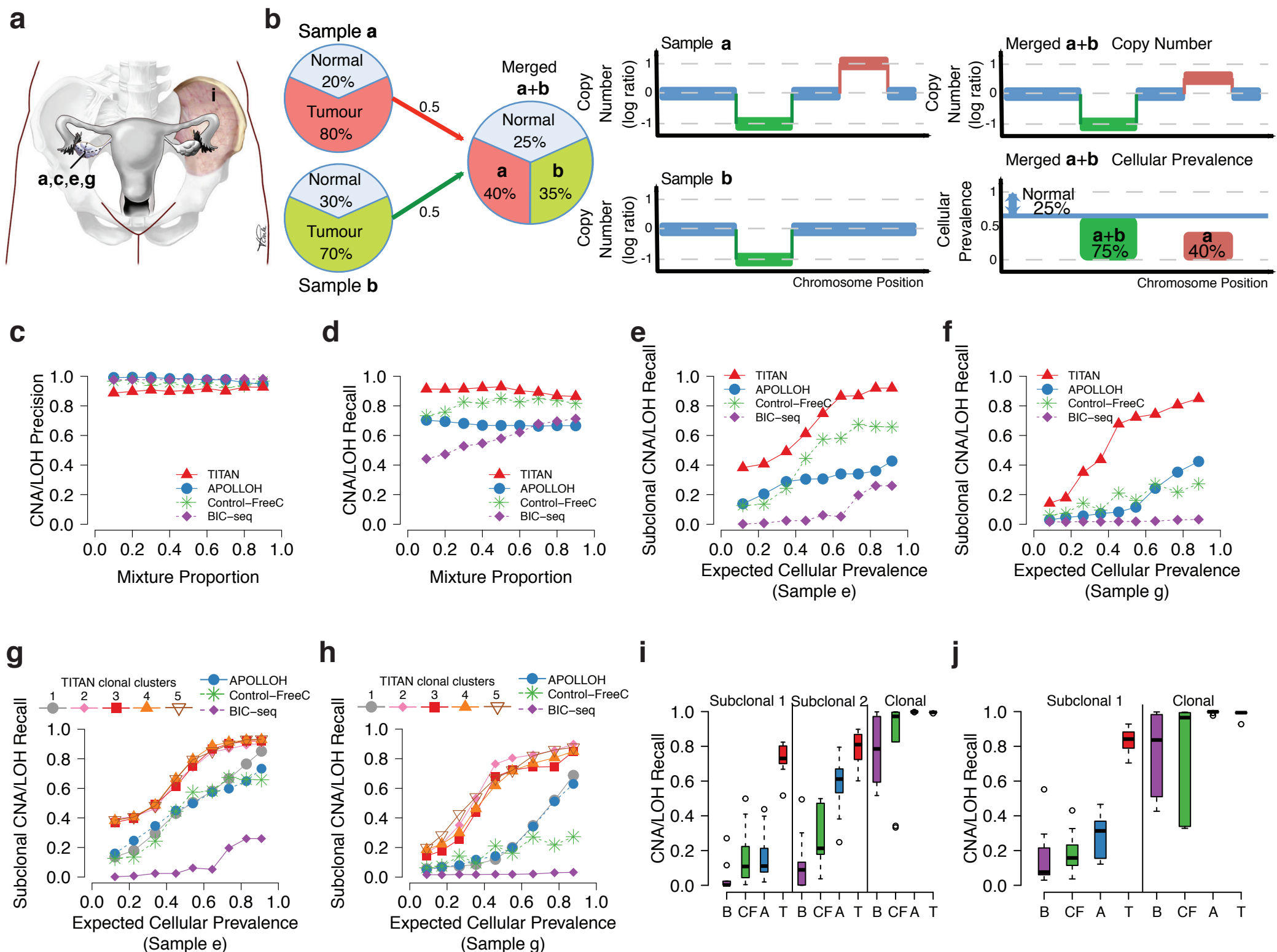
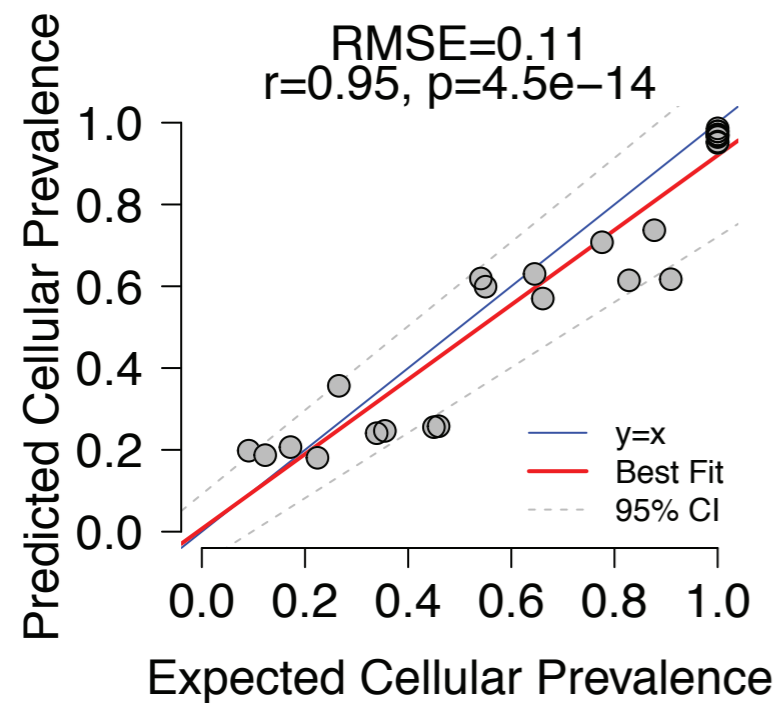
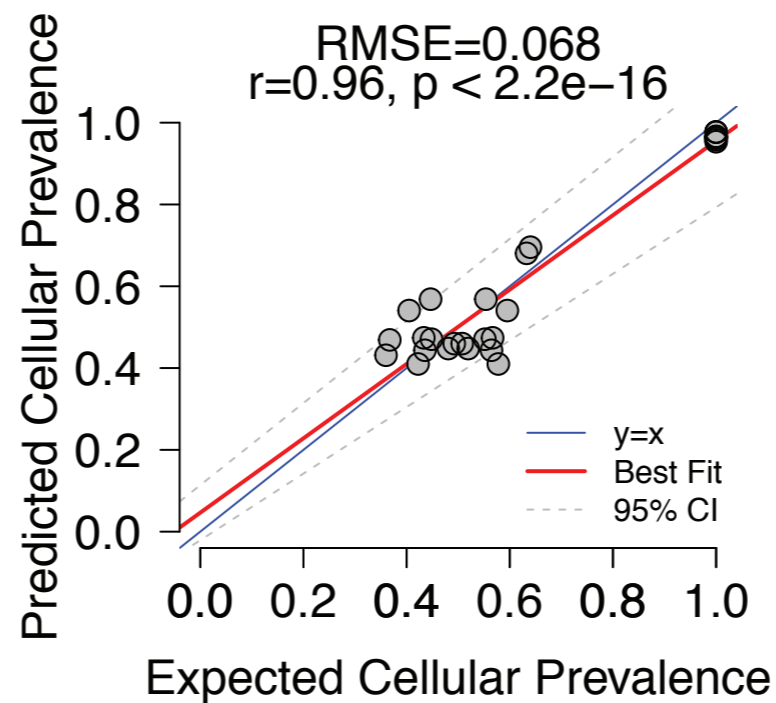
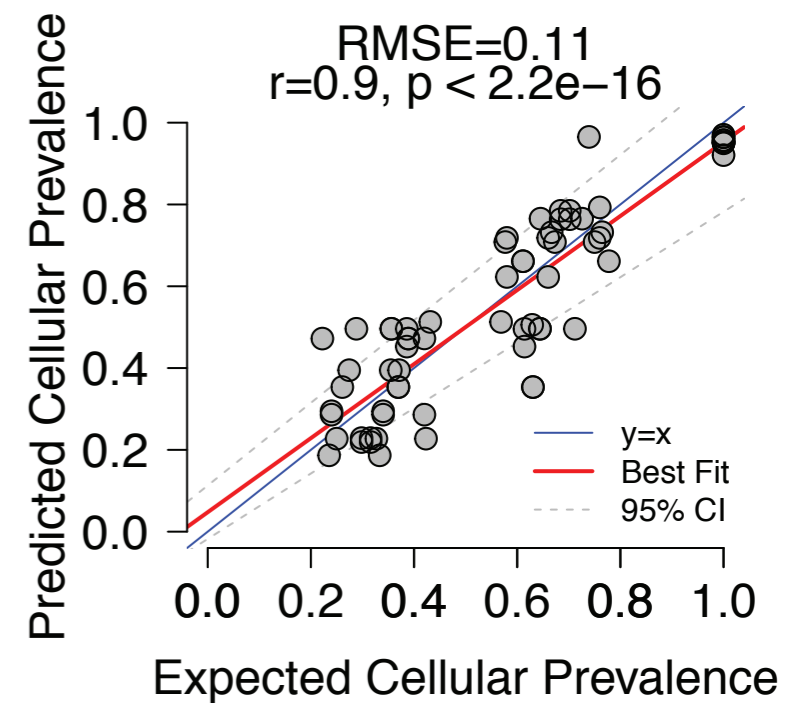
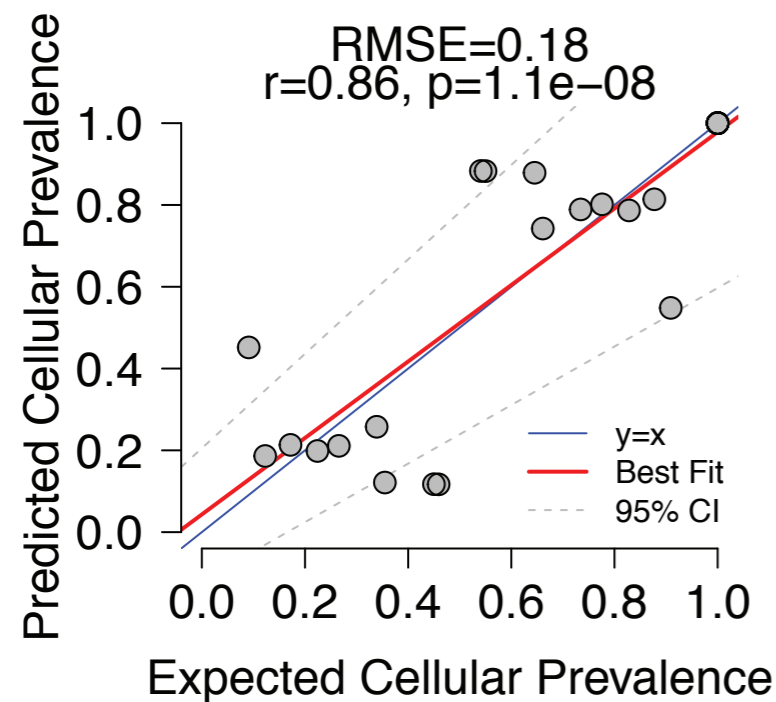
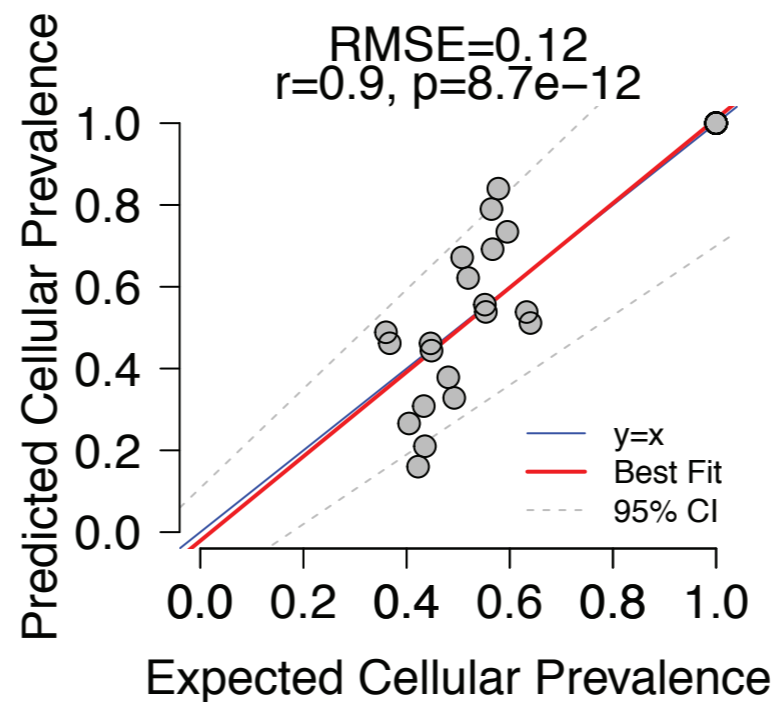


Figure 3

Serial Mixture

Pairwise Merge Mixture

Triple Merge Mixture

a**b****c****d****e****Figure 4**

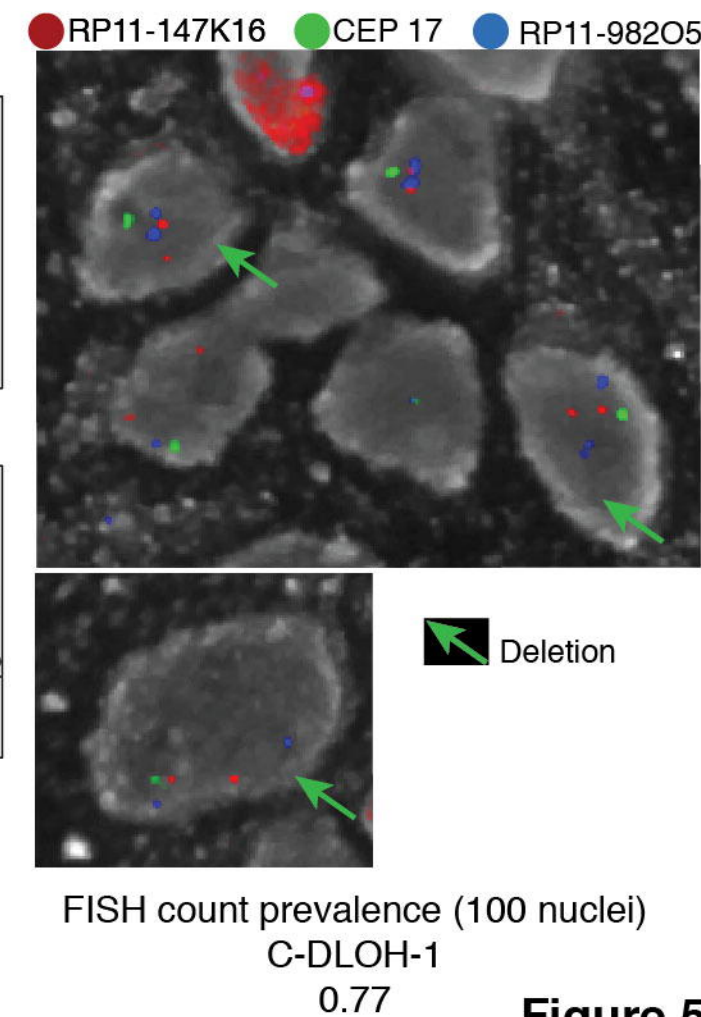
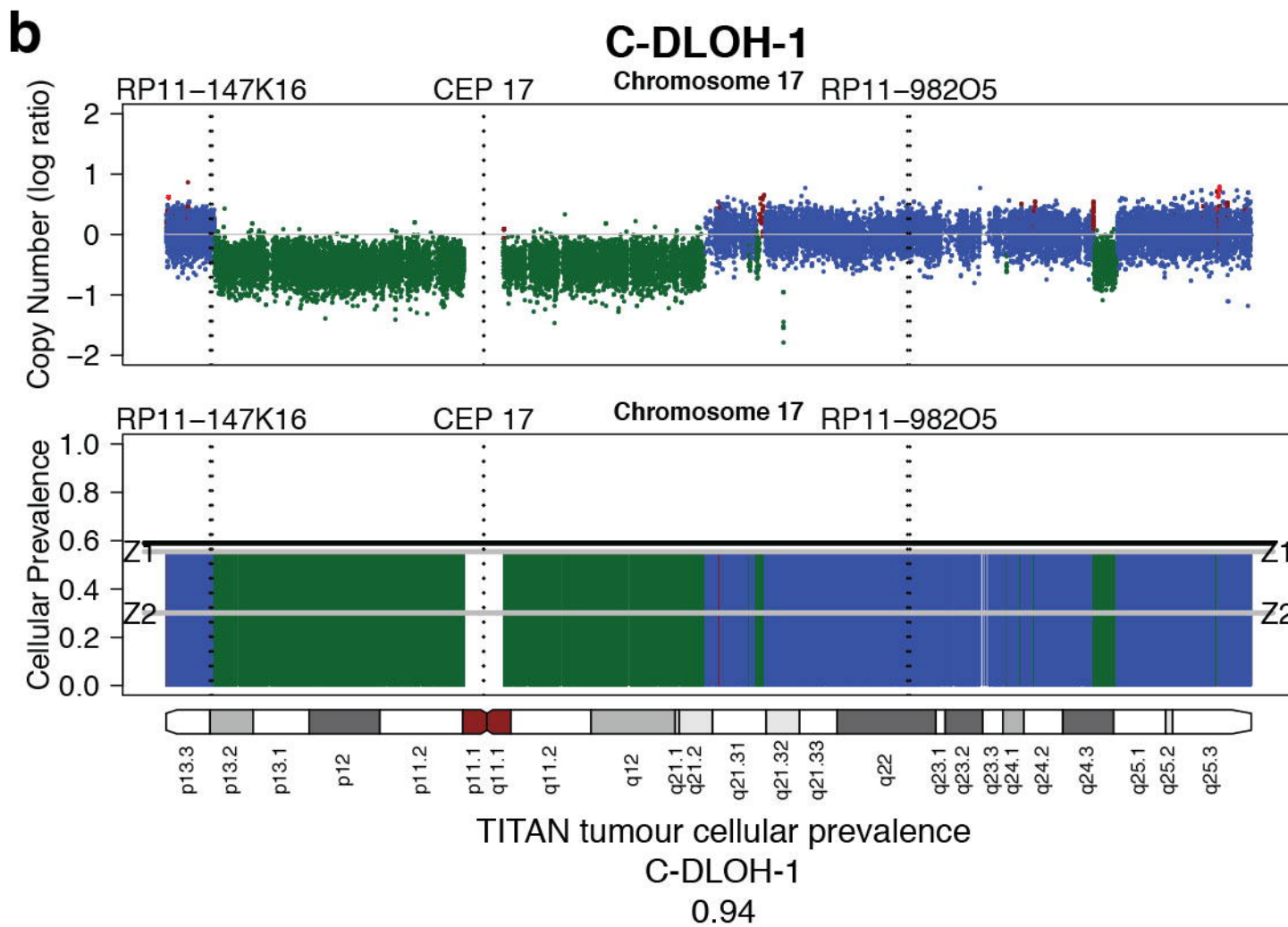
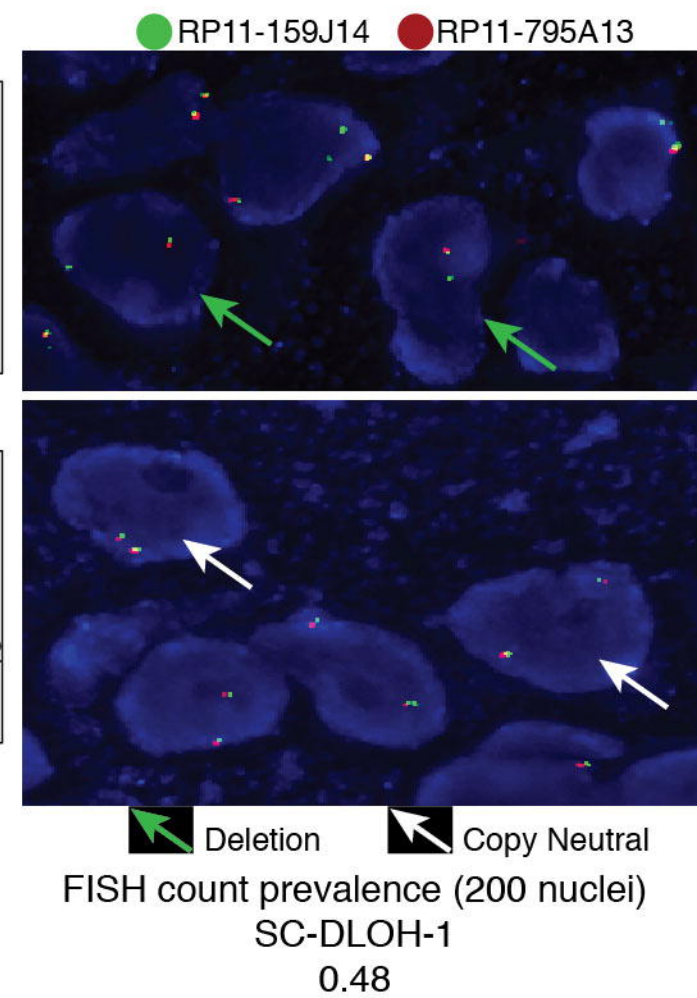
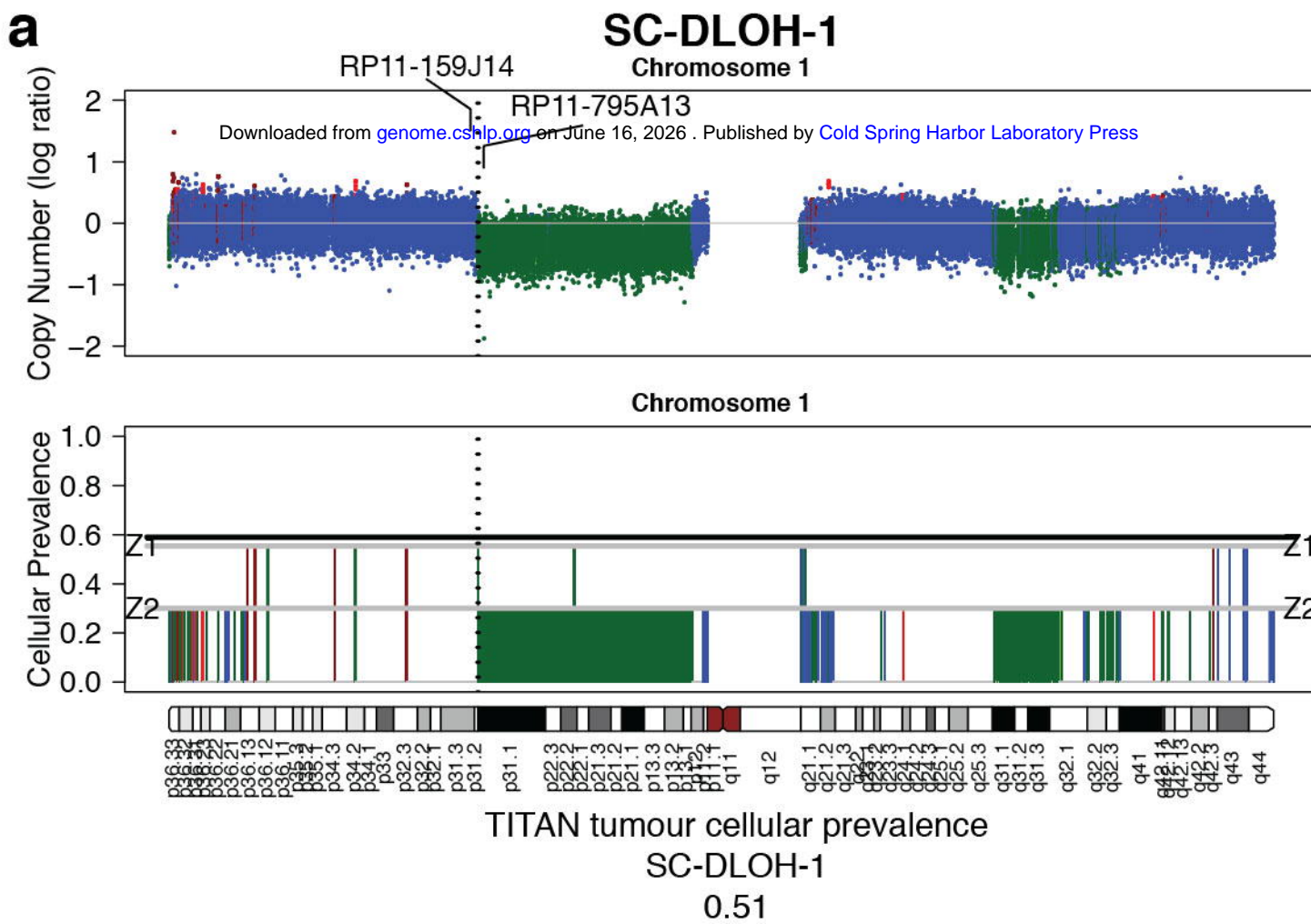


Figure 5

