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

Supplementary Material

Gavin Ha[1,2], Andrew Roth[1,2], Jaswinder Khattra[1], Julie Ho[3], Damian Yap[1], Leah M. Prentice[3], Nataliya Melnyk[3], Andrew McPherson[1,2], Ali Bashashati[1], Emma Laks[1], Justina Biele[1], Jiarui Ding[1,4], Alan Le[1], Jamie Rosner[1], Karey Shumansky[1], Marco A. Marra[5], C Blake Gilks[6], David G. Huntsman[3,7], Jessica N. McAlpine[8], Samuel Aparicio[1,7], and Sohrab P. Shah[1,4,7,*]

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

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

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

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

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

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

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

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

Contents

List of Supplementary Figures

List of Supplementary Tables

1 Supplementary Methods

1.1 TITAN analysis workflow

1. First, germline heterozygous SNP positions $\boldsymbol{L} = \{t_i\}_{i=1}^{T}$ are identified from the normal genome using a genotyping tool, such as SAMtools *mpileup* (Li et al., 2009). The analysis uses approximately one to three million loci genome-wide per patient and allows for identification of somatic allelic imbalance events relative to the normal genome.

2. From the tumour genome data, the read counts mapping to the reference base (A allele) and total depth at all positions in $\boldsymbol{L}$ are extracted and represented as $a_{1:T}$ and $N_{1:T}$, respectively.

3. The tumour copy number is normalized for GC content and mappability biases using only the normalization component of HMMcopy (Ha et al., 2012). Briefly, the genome is divided into bins of 1kb and read count is represented as the number of reads overlapping each bin. GC content and mappability bias correction are performed on tumour and normal samples, separately. The corrected read counts for the overlapping 1kb bin at each position of interest $t \in \boldsymbol{L}$, $\bar{N}_t$ and $\bar{N}_t^N$, are used to compute the log ratio, $l_{1:T} = \log\left(\bar{N}_{1:T}/\bar{N}_{1:T}^N\right)$.

4. TITAN analyzes the data $l_{1:T}$, $a_{1:T}$, $N_{1:T}$ to segment the data into regions of CNA/LOH and estimates normal contamination, tumour ploidy and cellular prevalences for Z number of clonal clusters.

5. For range $i = 1\ to\ 5$, run TITAN analysis once for clonal cluster states $Z_i := \{1, ..., i\}$ where $|Z| = i$ is the number of clonal clusters. That is, TITAN is run once for $Z_1 = \in \{1\},\ |Z| = 1$, then independently run again for $Z_2 = \{1, 2\},\ |Z| = 2$, and again for $Z_3 = \{1, 2, 3\},\ |Z| = 3$, etc.

1.2 TITAN probabilistic framework

1.2.1 Representation of mixed populations in heterogeneous tumour WGS data

Copy number data is represented by the log ratio between tumour and normal read depth $l_{1:T}$, which is modeled as Gaussian distributed $l_{1:T} \sim \mathcal{N}\left(l_{1:T}|\mu_{g,z}, \sigma_g^2\right)$ with mean

$$\mu_{k,z} = \log\left(\frac{nc_N + (1-n)\, s_z c_N + (1-n)\,(1-s_z)\, c_{T,g}}{nc_N + (1-n)\,\phi}\right) \tag{1}$$

ϕ is the genome-wide average tumour ploidy, c_N is normal copy number, and $c_{T,g}$ is the copy number of tumour state $g \in \boldsymbol{G}$. The state space of $\boldsymbol{G}$ consists of 21 genotype states and includes the combination of both copy number and allelic imbalance (Supplementary Table 14). Thus, $\mu_{g,z}$ is the parameter representing copy number resulting from the three cell populations, and the overall ploidy of the genome (Fig. 2a). Prior normalization steps lead to diploid baselines; therefore, this formulation allows the model to account for and estimate for average tumour ploidy (Van Loo et al., 2010) or, in the context of WGS data, haploid coverage factor, during inference.

The reference allelic read counts $a_{1:T}$ and tumour depth $N_{1:T}$ is modeled as Binomial distributed $a_{1:T} \sim Bin\left(a_{1:T}|N_{1:T}, \omega_{g,z}\right)$ with probability-of-success parameter

$$\omega_{g,z} = \frac{nc_N r_N + (1-n)\, s_z r_N c_N + (1-n)\,(1-s_z)\, r_{T,g} c_{T,g}}{nc_N + (1-n)\, s_z c_N + (1-n)\,(1-s_z)\, c_{T,g}} \tag{2}$$

r_N and $r_{T,g}$ are reference allelic ratios for normal and tumour (state $g \in \boldsymbol{G}$). This equation can be described as the proportion of reference alleles, considering all cell population types, out of the total alleles (or copies) for a specific locus. Note that for equations 1 and 2, we assume only a finite ($|Z|$) number of clusters, and the presence of only one tumour genotype at any aberrant locus in order to maintain identifiability in the model.

The graphical model is shown in Figure 2 and definitions for states and parameters are described in Supplementary Table 13.

1.2.2 TITAN model assumptions

In this section, we provide the assumptions of TITAN and introduce the concepts of cellular prevalence and clonal clusters more formally. The model is based on four main assumptions:

Assumption 1 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.

Assumption 2 Segmental regions of CNA and LOH span 10s to 1000s of contiguous SNP loci.

Assumption 3 The observed sequencing signal is the sum of the signals from heterogeneous cellular populations, including normal and tumour subpopulations.

Assumption 4 Sets of genetic aberrations are observed at similar cellular prevalence if these events arose from the same clone during punctuated expansion.

For **Assumption 3**, we assume the observed measurements were generated from a composite of three types of cell populations (Yau et al., 2010), which allows for modelling tumours that contain multiple tumour subpopulations (clones). Let s be the proportion of tumour cells that are diploid heterozygous (and therefore normal) at the locus. Then, $(1-s)$ is the *tumour cellular prevalence* or the proportion of the tumour cells containing the event. The relative proportions of the three cell populations are as follows: n, the proportion of sample that are non-malignant cells; $(1-n)\,s$, the proportion of sample that are tumour cells and have normal genotype; and $(1-n)\,(1-s)$, the *sample cellular prevalence* or the proportion of sample that are tumour cells and harbour the CNA or LOH event of interest (Figure 2a). We also assume that, at any aberrant locus, only one tumour genotype is harboured.

For **Assumption 4**, we assume that punctuated clonal expansions likely gave rise to multiple somatic events that will be observed at similar cellular prevalence; therefore, these events can be assigned to one of a finite number of clonal clusters. Because each event in a clonal cluster will have a unique cellular prevalence, we can redefine the parameter s. Let Z be the set of clonal clusters. Then, $(1-s_z)$ is the tumour cellular prevalence at the locus of interest for clonal cluster $z \in Z$. 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). We further assume that there are only a finite ($|Z|$) number of clusters.

For **Assumption 1** and **Assumption 2**, we assume segmental CNA and LOH events span many contiguous SNP positions. Let G be the genotype states that includes the combination of both copy number and allelic imbalance (Table 14). To capture the shared signals between 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 $Z_{1:T}$ make up the factorial Markov chain for T heterozygous germline SNPs (Figure 2c). The state space is dynamically determined as a function of the number of clonal clusters, resulting in $|G| \times |Z|$ number of state tuples $(g \in G, z \in Z)$ (Table 14).

1.2.3 Hidden state space for joint genotype and clonal cluster

The analysis requires the full set of $\boldsymbol{L} = \{t_i\}|_{i=1}^{T}$ germline heterozygous SNP positions, which generally ranges from one to three million per patient identified from the normal genome (Fig. 2b). The model consists of 21 genotype states $\boldsymbol{G}$ (Supplementary Table 14) and a finite number of clonal clusters $\boldsymbol{Z}$. Each position $t \in \boldsymbol{L}$ can be assigned a state tuple (g, z), for $g \in G$, $z \in Z$, resulting in $21 \times |Z|$ total number of tuples in a factorial combination. The initial state distributions, $\boldsymbol{\pi_G}$ and $\boldsymbol{\pi_Z}$, are conjugate Dirichlet-distributed with hyperparameters $\boldsymbol{\delta_G}$ and $\boldsymbol{\delta_Z}$, respectively.

1.2.4 Joint emission model

At each SNP, copy number data is represented by the log ratio between the tumour and normal read depths $l_{1:T}$, which is modeled as Gaussian distributed $l_{1:T} \sim \mathcal{N}\left(l_{1:T}|\mu_{g,z}, \sigma_g^2\right)$. The reference allelic read counts $a_{1:T}$ and tumour depth $N_{1:T}$ is modeled as Binomial distributed $a_{1:T} \sim Bin\left(a_{1:T}|N_{1:T}, \omega_{g,z}\right)$. The parameters $\mu_{g,z}$ and $\omega_{g,z}$ represent the signals from the three cell populations (Methods). The key insight is that these model parameters are influenced by s_z, rendering the approach more sensitive to events with inherently lower cellular prevalences (Fig. 2d).

A joint emission is used to model $l_{1:T}$, $a_{1:T}$, and $N_{1:T}$ in a multivariate approach. The joint emission is therefore defined as

$$p\left(a_t, N_t, l_t|Z_t = z_t, G_t = g_t, \boldsymbol{\theta}\right) = \begin{cases} Binomial\left(a_t|N_t, \omega_{g,z}\right) \times \mathcal{N}\left(l_t|\mu_{g,z}, \sigma_g^2\right) & g_t > 0, \\ U\left(0, N_t\right) \times \mathcal{N}\left(l_t|0, \Sigma\right) & g_t = 0, z_t = 0 \end{cases} \tag{3}$$

where an outlier state ($g = 0, z = 0$) is represented as a joint uniform and Gaussian distribution with variance Σ is set to a large value (Yau et al., 2010).

The parameters of the emission densities, $\omega_{g,z}$ and $\mu_{g,z}$, were defined above and illustrated as being influenced by cellular prevalence (Fig. 2d). These parameters are functions of the unknown parameters for global normal proportion n and tumour ploidy ϕ, cellular prevalence s_z, and state-specific Gaussian variance

σ_g^2. The prior distributions for these unknown parameters are the following,

$$p\left(s_z|\alpha_z,\beta_z\right) = Beta\left(s_z|\alpha_z,\beta_z\right) \quad (4)$$

$$p\left(n|\alpha_n,\beta_n\right) = Beta\left(n|\alpha_n,\beta_n\right) \quad (5)$$

$$p\left(\phi|\alpha_\phi,\beta_\phi\right) = InverseGamma\left(\phi|\alpha_\phi,\beta_\phi\right) \quad (6)$$

$$p\left(\sigma_g^2|\alpha_g,\beta_g\right) = InverseGamma\left(\sigma_g^2|\alpha_g,\beta_g\right) \quad (7)$$

1.2.5 Genotype and clonal cluster transition model

TITAN employs a non-stationary (heterogeneous) transition model in the HMM, which involves transitioning between both the CNA/LOH genotype and clonal cluster state spaces. Two transition probability matrices, $A_t \in \mathbb{R}^{21\times 21}$ for the genotypes and $T_t \in \mathbb{R}^{|Z|\times|Z|}$ for the clonal clusters, are used to define the joint transition matrix $\boldsymbol{J}_t \in \mathbb{R}^{|K|\times|K|}$ where K is the set of $21 \times |Z|$ number of genotype-clonal cluster state tuples (g, z), $\forall g \in G$ and $\forall z \in Z$.

A_t is the genotype transition probability matrix at position t. Let $A_t(i, j)$ be the probability of transitioning between genotypes states $i \in G$ at position $t-1$ and $j \in G$ at position t. The probability ρ_G, which accounts for the distance d between t and $t-1$ is defined as $\rho_G = 1 - \frac{1}{2}(1 - e^{-d/2*L_G})$, where L_G is a user-defined, expected length of CNA/LOH events (Colella et al., 2007). rho_G is used if transitions are between the same state $(i = j)$, share the same allelic zygosity status $(ZS(i) = ZS(j))$, and share the same copy number $(c_{T,i} = c_{T,j})$,

$$A_t(i,j) = \begin{cases} \rho_G & i = j \; or \\ & (ZS\left(i\right) = ZS\left(j\right) \; and \; c_{T,i} = c_{T,j}) \\ \frac{1-\rho_G}{|K-1|} & otherwise \end{cases} \quad (8)$$

Each row of A_t is then normalized such that $\sum_j A_t\left(i,j\right) = 1, \forall i$.

T_t is the clonal cluster transition probability matrix at position t. Let $T_t(m, n)$ be the transition probability from clonal cluster $m \in Z$ at position $t-1$ to cluster $n \in Z$ at position t. Higher probabilities are used when transitioning to the same clonal cluster $(m = n)$. This is represented using $\rho_Z = 1 - \frac{1}{2}(1 - e^{-d/2*L_Z})$,

where L_Z is the user-defined, expected length of clonal cluster segments.

$$T_t(m,n) = \begin{cases} \rho_Z & m = n \\ 1 - \rho_Z & otherwise \end{cases} \tag{9}$$

1.2.6 Learning and inference

We use the expectation maximization (EM) algorithm to estimate the model parameters

$$\boldsymbol{\theta} = \{n, s_{1:|Z|}, \phi, (\sigma^2)_{1:21}, \boldsymbol{\pi_G}, \boldsymbol{\pi_Z}\} \tag{10}$$

given all the data $\boldsymbol{\mathcal{D}} = \{l_{1:T}, a_{1:T}, N_{1:T}\}$. For the expectation step, we use the forwards-backwards algorithm to compute the joint-posterior marginal probabilities,

$$p\left(G_t = k, Z_t = z | \boldsymbol{\mathcal{D}}, \boldsymbol{\theta}\right) = \gamma\left(G_t = k, Z_t = z\right) \tag{11}$$

The resulting expectation of the complete log-likelihood at EM iteration n is

$$
\begin{aligned}
Q^{(n)} &= \mathbb{E}_{G|\mathcal{D},\theta^{(n-1)}}\left[\log p\left(\boldsymbol{Z}, \boldsymbol{\mathcal{D}}|\theta\right)\right] && (12)\\
&= \sum_{g=1}^{G} p\left(G_0 = g|\boldsymbol{\mathcal{D}}, \boldsymbol{\theta^{(n-1)}}\right) \log Multinomial\left(G_0|\boldsymbol{\pi_G}\right) && (13)\\
&\quad + \sum_{z=1}^{Z} p\left(Z_0 = z|\boldsymbol{\mathcal{D}}, \boldsymbol{\theta^{(n-1)}}\right) \log Multinomial\left(Z_0|\boldsymbol{\pi_Z}\right)\\
&\quad + \sum_{t=1}^{T}\left\{\sum_{i=1}^{G}\sum_{j=1}^{G} p\left(G_t = j, G_{t-1} = i|\boldsymbol{\mathcal{D}}, \boldsymbol{\theta^{(n-1)}}\right) \log A_t\left(i,j\right)\right\}\\
&\quad + \sum_{t=1}^{T}\left\{\sum_{m=1}^{Z}\sum_{n=1}^{Z} p\left(Z_t = m, Z_{t-1} = n|\boldsymbol{\mathcal{D}}, \boldsymbol{\theta^{(n-1)}}\right) \log T_t\left(m,n\right)\right\}\\
&\quad + \sum_{z=1}^{Z}\sum_{t=1}^{T}\left\{\sum_{g=1}^{G} p\left(G_t = g, Z_t = z|\boldsymbol{\mathcal{D}}, \boldsymbol{\theta^{(n-1)}}\right)\left\{\log Binomial\left(a_t|\omega_{g,z}, N_t^T\right) + \log\mathcal{N}\left(l_t|\mu_{g,z}, \sigma^2\right)\right\}\right\}\\
&\quad + \sum_{z=1}^{Z}\log Beta\left(s_z|\alpha_z, \beta_z\right) + \sum_{g=1}^{G}\log InvGamma(\sigma_g|\alpha_g, \beta_g)\\
&\quad + \log Beta\left(n|\alpha_n, \beta_n\right) + \log InvGamma\left(\phi|\alpha_\phi, \beta_\phi\right)\\
&\quad + \log Dirichlet\left(\boldsymbol{\pi_G}|\boldsymbol{\delta^G}\right) + \log Dirichlet\left(\boldsymbol{\pi_Z}|\boldsymbol{\delta^z}\right)
\end{aligned}
$$

In the maximization step, we estimate the set of unknown parameters $\boldsymbol{\theta}$ using coordinate descent until convergence criteria is met. If the likelihood does not increase by more than 2%, then the coordinate descent is stopped.

The EM algorithm alternates between the E- and M-steps until a specified number of iterations or until convergence criteria is met. The result are the converged parameters $\hat{theta}$. The Viterbi algorithm is then applied to find the optimal genotype and clonal cluster state paths for the HMM using parameters $\hat{theta}$.

1.2.7 Choosing the optimal number of clonal clusters

Recall that for each $i = 1\ to\ 5$, TITAN is run once for the setting of $Z_i := \{1, ..., i\}$ for $|Z_i| = i$. To determine the run with the optimal number of initialized clusters $|Z_i|$, we used an internal validation scoring approach called the S_Dbw validity index (Halkidi et al., 2002). S_Dbw penalizes over-fitting due to increasing number of clusters by minimizing within cluster variances ($scat$) and maximizing density-based

cluster separation ($Dens$),

$$S\_Dbw(|Z_i|) = 25 * Dens(|c_T| * |Z_i|) + scat(|c_T| * |Z_i|)$$

where $Dens$ and $scat$ are defined in Halkidi et al. (2002) and $|c_T|$ is the number of copy levels. This was applied to our runs by defining the copy number log ratio $l_{1:T}$ as the *internal data* and the resulting joint states of (c_T, z), for $c_T \in \{0 \dots 5\}$ and $z \in \{1 \dots |Z_i|\}$, are the clusters in the internal validation. The computation of the S_Dbw index is based on using $|c_T| * |Z_i|$ number of internal evaluation clusters. For instance, when TITAN is run with $Z_2 = \{1, 2\}$, the number of clusters in the S_Dbw internal evaluation is $|c_T| * 2 = 10$. An S_Dbw index value is computed for each run of TITAN using a fixed number of clonal clusters $|Z_i|$ and the $optimalIndex = \arg\min_i \{S\_Dbw\,(|Z_i|)\}$ is chosen as the optimal run.

We acknowledge that a more robust solution could be to integrate the model selection directly into the framework, ideally as a phylogenetic tree to relate inferred clones into their ancestral lineages. Currently inferring phylogenies of clones directly from the WGS data is beyond the scope of this contribution and would require significant development at the level of mathematical first principles.

1.3 TITAN code availability

TITAN is implemented in an R package, called `TitanCNA`, which is available through Bioconductor. The functionality implemented in R consists of the component for GC and mappability bias correction, which uses a wrapper for the HMMcopy (Ha et al., 2012)); and the HMM component that performs segmentation and inference of subclonal copy number. The forwards-backwards and Viterbi algorithms in the HMM are implemented in C, and are interfaced as dynamic function objects within R. The time and memory complexity is $\mathcal{O}\left(K^2T\right)$ and $\mathcal{O}\left(KT\right)$, respectively, where K is the number of joint states (g, z) and T is the number of positions. Because TITAN models a range of clonal clusters using a joint state space of the clusters and genotypes, K scales based on the specified number of clusters. Instructions on software usage can be accessed at `http://compbio.bccrc.ca/software/titan/`.

1.4 Biospecimen collection of intratumoural ovarian carcinoma samples

Ethical approval was obtained from the University of British Columbia (UBC) Ethics Board. Women undergoing debulking surgery (primary or recurrent) for carcinoma of ovarian/peritoneal/fallopian tube origin were approached for informed consent for the banking of tumour tissue. Patient DG1136 was chosen as a high-grade serous carcinoma where more than one sample was surgically extracted from four sites in the primary tumour on the right ovary and left pelvic sidewall prior to treatment. The peripheral blood lymphocyte sample was collected prior to surgery. Tissue sections were subject to expert histopathological review (GT) to assess the presence of invasive tumour, pre-malignant or benign changes, lymphocytic infiltration, necrosis and tumour cellularity. Genomic DNA was extracted from fresh frozen tumour tissue and patient matched peripheral blood lymphocytes as previously described (Bashashati et al., 2013). Constructed libraries were sequenced on the Illumina HiSeq 2000, according to Illumina protocols, generating 100bp paired-end reads. The amount of sequence generated ranged from 94 to 110 gigabases total for an estimated coverage of sequencing between 29X and 35X (Supplementary Table 3a). The sequenced reads were aligned to the reference genome (build GRCh37, hg19) using BWA (Li and Durbin, 2009).

1.5 FISH validation of subclonal events in ovarian carcinoma samples

BACs were directly labelled with Spectrum Green, Spectrum Orange, or Alexa 647 using a Nick Translation Kit (Abbott Molecular, Illinois, USA) and chromosomal locations were validated using normal metaphases from blood (results not shown). Specific BAC and control probe identifiers are listed in the corresponding figures.

FISH on frozen tissue sections was performed according to the Frozen Tissue Prep for FISH protocol and the Paraffin Pretreatment protocol from Abbott Molecular, with several changes. Please refer to `https://www.abbottmolecular.com/contactus/fishtechsupport/keyproductinformation/specimen/frozen-tissue-prep-for-fish.html` and `https://www.abbottmolecular.com/contactus/fishtechsupport/keyproductinformation/vysisproducts/paraffin-pretreat.html`.

Briefly, 5m tissue cryosections were mounted onto positively charged slides. The slides were fixed in 4% formaldehyde in PBS at 4C for 15 minutes, washed briefly with PBS, and allowed to dry at room

temperature for at least 24 hours. The slides were then pretreated in 0.2N HCl for 20 minutes, followed by 10 mM citric acid buffer for 45 minutes at 80C. After washing twice with 2X SSC, the slides were digested in a pepsin solution at 37C, washed twice with 2X SSC, and dehydrated in an ethanol series. Probes were co-denatured with the tissues at 73C for 5 minutes and hybridized at 37C for 16 to 18 hours. After hybridization, slides were washed with 2X SSC/0.3% NP40 at 73C for 2 minutes, dehydrated in an ethanol series, and counterstained with 4,6-diamidino-2-phenylindole (DAPI).

Slides were scored manually using an oil immersion 63x objective and z-stack images were captured using Metasystems software (MetaSystems Group Inc., Belmont, MA, USA). To avoid bias in counting, tumour nuclei were selected using the DAPI filter based on size and morphology, then the number of different probe signals were counted by switching to the respective filters. To obtain prevalence estimates from FISH, at least 100 nuclei were scored using a combination of manually looking down the microscope and with images. Next, for each nuclei, the Event Ratio between event and control was computed. The final FISH count prevalence used is the proportion of nuclei having ratio < 1 (deletion), equal to 1 (neutral), and > 1 (gain). For event SC-DLOH-5 (chr21:chr21:22060503-22231762), no suitable control could be used because chr21 is deleted in a subset of cells; therefore, the raw cell count proportion is used as the prevalence for this event.

1.6 TNBC sample collection and sequencing

The genome and transcriptome sequencing files for triple negative breast cancers can be downloaded at the European Genome-phenome Archive (EGA; `https://www.ebi.ac.uk/ega/`) under the accession EGAS00001000132. Details of ethical consent, biospecimen collection, histopathlogical review, library preparation, and sequencing are described in (Shah et al., 2012). Mutations originally identified using JointSNVMix (Roth et al., 2012) and MutationSeq (Ding et al., 2012) for 16 genomes sequenced using ABi/SOLiD. Targeted deep amplicon sequencing data was generated on a selection of these positions and determined to be somatic using the Binomial exact test (Shah et al., 2012).

1.6.1 Application of TITAN to TNBC whole exome-capture sequencing data

Four TNBC samples (SA030, SA052, SA065, SA073) previously analyzed (Shah et al., 2012; Ha et al., 2012) contained both whole genome and exome sequencing data. TITAN was applied to these samples

with a modification in the normalization procedure for GC-content bias correction. The loess curve-fitting correction method was applied to only positions overlapping exons in the hg18 build, which was downloaded in BED format from UCSC. All other TITAN parameter settings were initialized to default values as per usage for other WGS samples in this manuscript. Performance of the TITAN results on these samples were computed by comparing the copy number concordance with the TITAN results for the corresponding WGS data at all overlapping SNP positions. A 'match' for a deletion or amplification at an overlapping position if both were less than 2 or both greater than 2, respectively.

1.7 Spike-in simulation experiment

Generating and inserting simulated data In a controlled experiment, we simulated a sample with pre-defined CNA events using real data from WGS data for sample DG1136a. First, we identified a large deletion (chr16:46464744-90173515) and a large amplification (chr8:97045605-144155272) from the CNA profiles predicted by both APOLLOH and Control-FREEC (Supplementary Fig. 2). From the heterozygous SNP positions within these events, we randomly sampled allelic ratio and log ratio data for non-consecutive sets of 10, 100, 1000 SNPs for each event; we sampled four replicates, resulting in eight sets of loci for each set size. We also sampled one set of 10000 SNPs for each event (Supplementary Table 2a). Next, we identified four whole chromosomes with diploid heterozygous status (chr1, 2, 9 and 18) for which we randomly inserted the 26 sampled loci sets, each as a contiguous spike-in CNA event, without overlaps (Supplementary Fig. 3-6). The median length of the spike-in events were 6.9kb, 82.5kb, 1.2Mb, and 12.5Mb, respectively.

Generating subclonal events with expected prevalence To assess the performance of cellular prevalence estimates, we also generated two additional tumour-normal admixtures at 80% and 60% of the original DG1136a sample mixed with the matched normal. These samples provide allelic ratio and log ratio data which simulate varying cellular prevalence from the expected tumour content of 0.52 and 0.39 based on the original tumour content of 0.65 for DG1136a. The samples were generated by randomly sampling the desired proportion of reads from the tumour and normal, separately for each chromosome, then merging all the reads together. For the spike-in, similar to before, 26 sampled loci sets taken from the same large events

were used for each admixture. The spike-in events were inserted into the same copy neutral chromosomes (chr1, 2, 9 and 18) (Supplementary Fig. 3-6).

Application of TITAN and performance assessment TITAN was run on all chromosomes including the ones with spike-in events. The parameters were estimated as per usual for a genome-wide sample. TITAN was run once for each clonal cluster setting between 1 and 5 with cluster 4 being selected as the optimal number of clusters based on the S_Dbw validity index. Predicted cellular prevalence estimates for two of the clusters were 0.52 and 0.36, which is within range of the expected values (Supplementary Fig. 3-6).

Performance was computed by comparing the TITAN predicted copy number status of the SNPs within each spike-in event. For each event, the true positive rate (TPR) for copy number was computed as the proportion of positions with copy number status less than 2 for matching deletions or greater than 2 for matching gains (Supplementary Table 2b). We also computed the size-based TPR as the proportion of matching positions out of all positions in the spike-in events with the same size (same number of SNPs) (Supplementary Table 2c). The TPR for cellular prevalence estimates of spike-in events was computed using a matching criteria of ± 0.05 of the expected prevalence (0.52 and 0.39). For copy number and cellular prevalence, a spike-in event is considered a true positive prediction when TPR> 0.9. The false positive rate (FPR) was computed on a global scale, considering all positions in chr1, 2, 9 and 18 where no spike-in data is found.

1.8 Mixture simulation experiments using intra-tumour samples from an ovarian carcinoma

Five intra-patient ovarian carcinoma samples from patient DG1136 were used to simulate multiple cellular populations by mixing combinations of samples at known proportions. Four of these samples (DG1136a, c, e, g) were obtained from regional, adjacent biopsies of the same tumour in the right ovary and the fifth sample (DG1136i) was taken from the metastasis in the left pelvic sidewall (Figure 3a). The tumour content (cellularity) for each of the five individual samples were determined as the consensus (average) between pathology immunohistochemical and APOLLOH-predicted estimates. Using the tumour content estimates and the relative sequence coverage, the proportion of tumour from each sample contributing to a simulated

mixture can be computed (Supplementary Table 3a). The contributing tumour proportions also represent the expected, simulated subclonal cellular prevalence in the mixture (Supplementary Table 3b-d). Two types of mixture simulations were used.

1.8.1 Generating serial mixtures

For the predefined serial mixture experiment, nine whole genome mixtures at ∼30X coverage were generated by sampling reads from DG1136e and DG1136g at mixing proportions of 10% increments (0.1 DG1136e + 0.9 DG1136g, 0.2 DG1136e + 0.8 DG1136g, ..., 0.8 DG1136e + 0.2 DG1136g, 0.9 DG1136e + 0.1 DG1136g). Because the two individual samples contain normal contamination, the expected mixture proportions were adjusted based on 67% and 56% normal estimates, respectively. This resulted in the relative tumour content contribution of 0.07 DG1136e/0.50 DG1136g, 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 for the mixtures (Supplementary Table 3b). Therefore, this becomes the expected sample prevalence for a mixture.

To formalize this, let the mixture proportion bep_e for DG1136e and p_g for DG1136g ,and tumour content be t_e for DG1136e and t_g for DG1136g. Then, the expected *sample* cellular prevalence for events contributing uniquely from DG1136e in the simulated mixture is $s_{e,sample} = p_e * t_e$; the *sample* cellular prevalence for events unique to DG1136g is $s_{g,sample} = p_g * t_g$. The *tumour* cellular prevalence are then computed as $s_{e,tumour} = s_{e,sample} / (s_{e,sample} + s_{g,sample})$ and $s_{g,tumour} = s_{g,sample} / (s_{e,sample} + s_{g,sample})$ for events in the mixture that are contributed uniquely from DG1136e and DG1136g, respectively.

1.8.2 Generating merged mixtures

For the merging of two or three samples at approximately equal proportions, five intratumour samples were merged together to generate ten pairs at ∼60X coverage (Supplementary Table 3c) and ten triplets at ∼90X coverage (Supplementary Table 3d) coverage for each combination. This was done using SAMtools (Li et al., 2009) merge command. The expected cellular prevalence for each mixture, once again, was computed based on tumour contributions from individual samples making up the mixture while also adjusting for sequencing coverage. Therefore, the *sample* expected cellular prevalence for the merged mixture of Sample a with c_a coverage and t_a tumour content and Sample b with c_b coverage and t_b tumour content is computed

as $s_{a,sample} = p_a * t_a$ and $s_{b,sample} = p_b * t_b$ for events uniquely in a and b, respectively, where the mixture proportions are computed as $p_a = c_a / (c_a + c_b)$ and $p_b = c_b / (c_a + c_b)$. The *tumour* cellular prevalence is $s_{a,tumour} = s_{a,sample} / (s_{a,sample} + s_{b,sample})$ and $s_{b,tumour} = s_{b,sample} / (s_{a,sample} + s_{b,sample})$ for events in the mixture that are contributed uniquely from Sample a and b, respectively

1.8.3 Computing performance metrics

HMMcopy and APOLLOH (Ha et al., 2012) results from the individual samples were used as ground truth CNA and LOH events, respectively. Default parameters were used for APOLLOH/HMMcopy (`http://compbio.bccrc.ca/software/`). The truth set consists of CNA/LOH status at all germline heterozygous (HET) SNP positions included in the APOLLOH results for each of the five samples. The rationale for using all germline HET positions in the evaluation is that it represents a genome-wide assessment, such that larger events are given more weight because they span more loci. Spurious, potentially false, ground truth events that span fewer loci, and thus perhaps less confident, are weighted less. Furthermore, every evaluation examines the same set of positions, providing a more comparable performance metric across methods and alleviating the complexity in varying precision of boundaries between approaches.

Precision, recall, and F-measure performance for TITAN, APOLLOH, Control-FREEC (Boeva et al., 2012), and BIC-seq (Xi et al., 2011) analysis on the simulated mixture samples were computed based on the CNA/LOH status of predicted segments overlapping the ground truth SNP loci. Precision was computed as the proportion of SNP loci in which the predicted CNA/LOH status from the overlapping segment matched the ground truth status at all CNA/LOH predicted SNP loci. Recall was computed as the proportion of SNP loci in which the predicted CNA/LOH status from the overlapping segment matched the ground truth status at all CNA/LOH truth SNP loci. The performance was computed for deletions, gains, and LOH status, independently, and averaged together when evaluating for overall assessment. True deletion and amplification loci were determined if both predictions and ground truth were < 2 and > 2, respectively. True LOH loci were determined as presence or absence in predictions, matching the ground truth. Ground truth subclonal events that are a mixture of two different tumour genotypes (non-diploid-heterozygous) were excluded from the performance because these events were uncommon, could possibly lead to identifiability issues, and all tools were only capable or designed to return a single prediction genotype. For evaluating

size-based performance, ground truth events from each individual sample (not the simulated mixtures) were grouped into ranges of length 10kb-100kb, 100kb-1Mb, 1Mb-10Mb, and greater than 10Mb; precision, recall, and F-measure were computed similarly for each size group.

For evaluation of cellular prevalence, the proportion of tumour contribution from each individual sample making up the simulated was used to compute the expected cellular prevalence (Supplementary Table 3b-d). Figure 3b illustrates a mixture scenario, and identifies true (sub)clonal events and expected cellular prevalence. For example, Sample A has 80% tumour content and Sample B has 70%. If these samples were mixed at equal proportions to generate Mixture X, then X will have a tumour subclone (population) of 40% contributing from Sample A (0.5 * 80%), another tumour subclonal of 35% from Sample B (0.5 * 70%), and normal population (0.5 * (20%+30%)). Clonally dominant events in X are considered as those that are present in both Sample A and B, and will have expected cellular prevalence of 40%+35%=70%. Subclonal events in X are those that are present in exactly one of the samples but not both. For example, if a GAIN event was only found in Sample A, then the expected cellular prevalence of this GAIN in X is 40%. Because the individual samples may have different sequence coverage, we have also adjusted for this.

The number of expected clonal clusters with unique cellular prevalence is taken as the permutation of the number of simulated tumour populations. For the serial and pairwise analysis, three possible clonal clusters exist; for the triplet simulation, up to seven clusters may be present. The correlation analysis used a sample size based on the expected number of clusters across all mixtures in each experiment (Fig. 4, Supplementary Fig. 11).

1.8.4 Usage details of other copy number prediction software

APOLLOH Input data consisted of read counts at heterozygous germline SNP positions identified using SAMtools `mpileup`. HMMcopy was used to generate input copy number data; settings included `width=1000` and `quality=0` for `readCounter` during bin read count generation, and `param$mu <- log(c(1, 1.4, 2, 2.7, 3, 4.5) / 2, 2)` for the R function `HMMsegment` during segmentation. Default configuration parameters, as is provided in `apolloh_K18_params_Illumina_stromalRatio_Hyper10k_min10max200.mat` downloaded from `http://compbio.bccrc.ca/software/apolloh/`, were used for APOLLOH to predict regions

of LOH, allele-specific amplification (ASCNA), and heterozygous (HET).

Control-FREEC (version 6.0) was applied to the mixtures using the following parameters: `ploidy=2`, `contaminationAdjustment=TRUE`, `sex=XX`, `uniqueMatch=TRUE`, `window=1000`. Mappability was corrected using the input file '`out100m2_hg19.gem`' (2 mismatches) and the SNPs used for B-allele frequency (LOH) analysis was provided in `hg19_snp137.SingleDiNucl.1based.txt` (downloaded from `http://bioinfo-out.curie.fr/projects/freec/tutorial.html`). The output file with the extension "`.CNVs`" and containing the inferred segments were used to evaluate the performance of Control-FREEC.

BIC-seq BIC-seq (version 1.1.2) was used with bin size of 1kb and λ set to 10, while default settings were used for all other parameters. The output file with extension "`.bicseg`" containing the resulting segments were used for copy number performance evaluation. Copy number loss and gain were determined as segments having "`log2.copyRatio`" $< \log_2(0.75)$ with "`log10.pvalue`" $< \log_{10}(0.0001)$ and "`log2.copyRatio`" $> \log_2(1.25)$ with "`log10.pvalue`" $< log_{10}(0.0001)$, respectively.

THetA We also compared the results to THetA (Oesper et al., 2013), which is a post-segmentation software for estimating cellular prevalence. Due to limitations in runtime and memory, we ran THetA for one normal and one tumour population ($n = 2$) using conservatively large BIC-seq segments ($\lambda = 200$), and subsequently filtered for regions larger than 5 Mbp. The default parameter settings, such as heuristics, were used. Then, we ran THetA for one normal and two tumour populations ($n = 3$) using the $n = 2$ results, filtering down to only the 15 largest non-diploid or full-chromosome sized segments, and changing any zero lower bound copy number heuristics to 1. All other parameters were set to default values.

1.9 Comparison of cellular prevalence with RNA-seq for TNBC

RNA-seq data for TNBC was obtained from the study by (Ha et al., 2012) and aligned as previously described (Shah et al., 2012) using BWA v0.5.5 to the human genome reference (NCBI build 36, hg18) and a database of known exon-exon junctions obtained from different annotation databases (Ensembl, RefSeq, AceView). Allele counts were extracted for all germline SNP positions using filters for base phred qualities

(> 5), mapping qualities (> 30), and depth threshold (> 10).

1.10 Validation using targeted deep amplicon DNA sequencing of single-cell nuclei

Predicted copy number deletions were validated using sequencing of DNA isolated from nuclei of individual cells in sample DG1136g. Deletion events can be more easily confirmed in single-cell sequencing by observing homozygosity, or the absence of one allele, as compared to copy number gains. Somatic point mutations were used to help distinguish tumour and normal nuclei. Two sets of events, Set1 and Set2, were selected for validation from DG1136g, each included one clonal deletion, two subclonal deletions, two heterozygous diploid regions, and a set of previously validated somatic mutations. For each set, 42 nuclei were sorted separately, and library construction and sequencing were carried out.

1.10.1 Selection of positions for validation of deletion events

Deletions in single cells were interrogated at heterozygous germline SNP loci. To improve the likelihood of observing a true deletion, multiple loci overlapping a deletion were used; this also allowed for distinguishing signals from random allele drop-out during sequencing. For deletions and diploid regions, 10-11 and 2-3 positions were selected, respectively. These regions were HET-1 (chr1:56977819-68910999), HET-3 (chr2:82870237-86078478), C-DLOH-1 (chr17:17415217-21074153), SC-DLOH-1 (chr1:70539053-117275764) and SC-DLOH-3 (chr2:31374733-80861750) for Set1, and HET-4 (chr7:138768839-141135114), HET-5 (chr21:19359230-19674681), C-NLOH-1 (chr17:55290843-62185764), SC-DLOH-4 (chr7:143777995-153688808), SC-DLOH-5 (chr21:22084693-25770230) for Set2, where 'C' represents clonally dominant and 'SC' represents subclonal. Additional criteria for selecting these positions were as follows: 1) SNP positions overlapped Affymetrix SNP6.0 array loci. These positions were likely also found within populations-based studies used in the array design; 2) Positions were equally spaced across the deletion region; 3) 500bp flanking regions to left and right of chosen positions did not contain any germline variants (heterozygous or homozygous). This helps with primer design and leads to more optimal primer amplification.

Mutations were chosen from a list of previously validated SNVs via AmpliCrazy primer design platform sequenced on a MiSeq. Clonally dominant mutations (TP53, CSMD1, ARID1B, RFC3) were selected to later help distinguish tumour and normal cells. In particular, TP53 was validated as a clonally dominant

homozygous mutation (containing only the variant allele). Additional clonally dominant mutations were selected for Set1 (FGD5) and Set2 (GABRA5, GALNT16, LRRC36, SPTB) (Supplementary Table 11a, 12a). We also included mutations that were found within subclonal deletions for Set1 (ABCA4, DENND2C, SULT6B1) and Set2 (MUC3A, XRCC2) to investigate biallelic inactivation in tumour cells.

1.10.2 Single-cell sequencing of nuclei DNA for ovarian cancer sample DG1136g

Nuclei preparation and sorting Single cell nuclei were prepared using a sodium citrate lysis buffer containing Triton X-100 detergent. Solid tissue samples were first subjected to mechanical homogenization using a laboratory paddle-blender. The resulting cell lysates were passed twice through a 70-micron filter to remove larger cell debris. Aliquots of freshly prepared nuclei were visually inspected and enumerated using a dual counting chamber hemocytometer (Improved Neubauer, Hausser Scientific, PA) with Trypan blue stain. Single nuclei were flow sorted into individual wells of microtitre plates using propidium iodide staining and a FACSAria II sorter (BD Biosciences, San Jose, CA).

Genomic DNA (gDNA), which refers to the bulk tumour DNA and can contain stromal DNA, is a potential source of contamination in the nuclei buffer during preparation. Included in each set were control nuclei samples with the absence of DNA templates, called non-template control (NTC) cells. These samples were used as the background control because any signal present will be from gDNA contamination as well as various amplicon and primer artefacts.

Multiplex and singleplex PCRs Somatic coding SNVs catalogued and validated in bulk tissue genome sequencing experiments were picked for mutation-spanning PCR primers design using Primer3. Common sequences were appended to the 5' ends of the gene-specific primers to enable downstream barcoded adaptor attachment using a PCR approach. Multiplex (24) PCRs were performed using an ABI7900HT machine and SYBR GreenER qPCR Supermix reagent (Life Technologies, Burlington, ON). The 24-plex reaction products from each nucleus were used as input template to perform 48 singleplex PCRs using 48 by 48 Access Array IFCs according to the manufacturer's protocol (Fluidigm Corporation, San Francisco, CA). Flow sorting plate wells without nuclei and 10 ng gDNA aliquots were used for negative and positive control reactions, respectively.

Nuclei-specific amplicon barcoding and nucleotide sequencing Pooled singleplex PCR products from each nucleus were assigned unique molecular barcodes and adapted for MiSeq flow-cell NGS sequencing chemistry using a PCR step. Barcoded amplicon libraries were pooled and purified by conventional preparative agarose gel electrophoresis. Library quality and quantitation was performed using a 2100 Bioanalyzer with DNA 1000 chips (Agilent Technologies, Santa Clara, CA) and a Qubit 2.0 Fluorometer (Life Technologies, Burlington, ON). Next-generation DNA sequencing was conducted using a MiSeq sequencer according to the maufacturer's protocols (Illumina Inc., San Diego, CA).

1.10.3 Analysis of single-cell sequencing data

Initial analysis of sequenced reads Paired end FASTQ files from the MiSeq sequencer were aligned to human genome build 37 downloaded from the NCBI using the `mem` command from the bwa 0.7.5a package. Allelic count data was extracted from the BAM files using a custom Python script which filtered out positions with base or mapping qualities below 10.

For each position, both mutation SNVs and SNPs, one-tailed binomial exact tests were independently applied to the reference and variant alleles in order to determine the presence or absence while accounting for sequencing errors and gDNA contamination. The error and contamination variant ratio was computed for each position by looking at the mean variant allelic ratio (variant reads divided by depth) for the flanking bases of the amplicon at that position from the NTC samples. This parameter encapsulated both the sequencing bias of the amplicon and the presence of gDNA contamination. The one-tailed binomial exact test was used to estimate whether the variant allelic ratio of the position was greater than expected. Similarly, the test was applied to the reference allelic ratio (reference reads divided by depth) for the same position. A *present* status was used for statistically significant test (Benjamini and Hochberg adjusted p-value < 0.05) and *absent* otherwise for the reference and variant alleles. Positions with fewer than a depth of 50 were considered *low_coverage*. Positions with *low_coverage* in $\geq 50\%$ of all nuclei in a set were also removed.

Distinguishing tumour and normal nuclei First, the nuclei were filtered for global low coverage if fewer than 10 positions had sufficient coverage (≥ 50 reads); these nuclei were excluded from the analysis. Next, normal nuclei were determined conservatively based on *absent TP53* variant allele status, and *absent* or

low_coverage variant allele status for all other mutations. While SNP positions for the regions of interest should be heterozygous in normal cells, we do not use these in the criteria due to allelic drop-out. For the remaining nuclei, each were classified as tumour if it had a *present* TP53 variant allele status but *absent* TP53 reference allele status; however, if TP53 was *low_coverage*, then at least one mutation with *present* variant allele status sufficed for tumour designation. All remaining nuclei were classified as *Unknown* because the data was ambiguous for determining normal or tumour.

The 42 nuclei in Set1 were divided into 14 with global low coverage, 14 normal, and 14 tumour; Set2 were divided into 23 with global low coverage, 9 normal, 9 tumour, and 1 *Unknown*.

Calculating the expected allelic drop-out rate and heterozygous allelic ratio Allelic drop-out refers to the preferential amplification of one allele for a heterozygous position, and this can be mistaken for the homozygous signal arising from loss of heterozygosity. As a result, approximately 10 positions were selected to assess the LOH status in individual nuclei for predicted deletion events (from the bulk WGS sample). The expected drop-out rate was computed as the proportion of (sufficient coverage) positions with *present* status for one of reference or variant but not both (XOR) out of all positions from every normal nuclei. Drop-out rates (DOR) for Set1 and Set2 were 0.28 and 0.48, respectively.

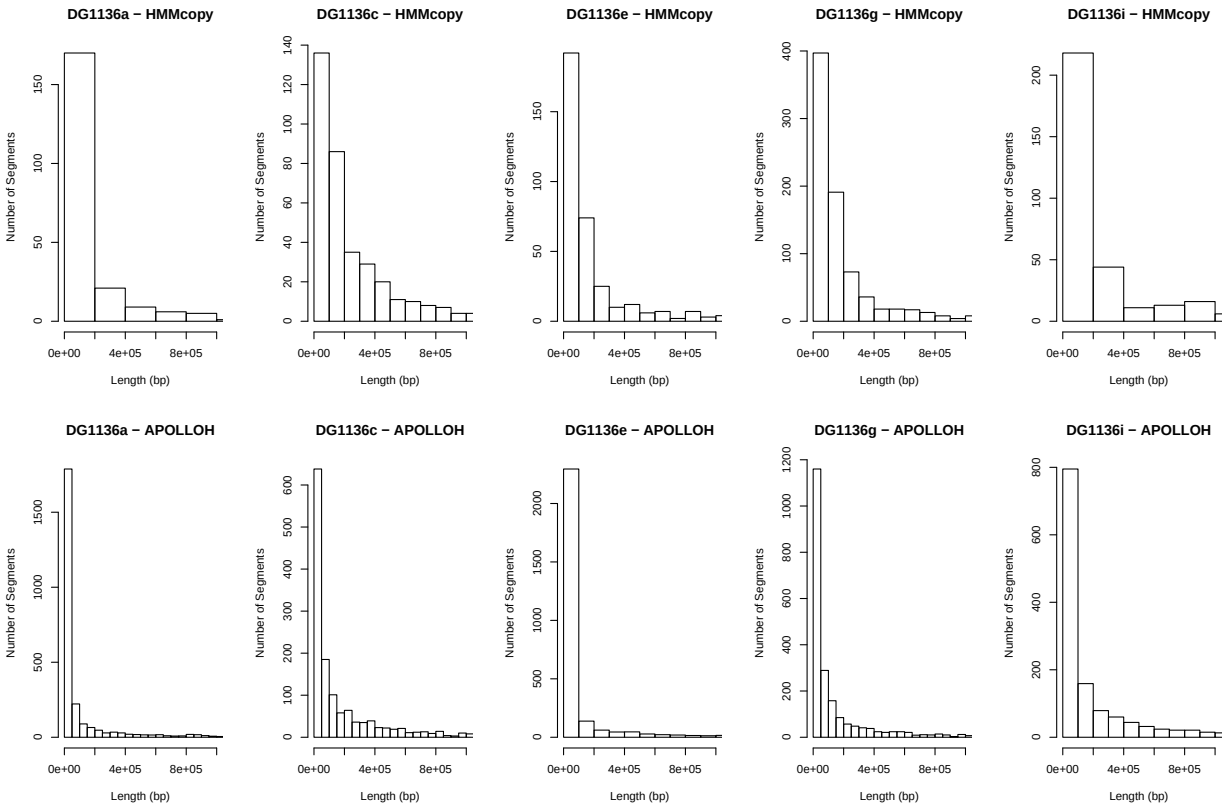
The expected allelic ratio for a heterozygous position is subject to gDNA contamination that can deviate this value away from the theoretical 0.5 ratio. Therefore, to account for this artefact, the expected allelic ratio was computed as the median across all (sufficient coverage) heterozygous positions, determined by having both reference and variant *present* status, from every normal nuclei. The expected heterozygous allelic (HAR) ratio for Set1 and Set2 were 0.57 and 0.68, respectively.

Two statistical tests to determine LOH event status To determine the LOH status of an event across all SNP positions within the event, two statistical tests were applied to each event. First, the event is assessed for being a true LOH and not due to allelic drop-out. We used a one-tailed binomial test in which the null hypothesis asserts that the ratio of homozygous:heterozygous positions is not greater than the drop-out rate. The drop-out rate was used as the expected ratio (probability of success); number of homozygous positions, determined by *present* reference XOR variant status, is the number of successes; and the total number of is the number of trials. The second analysis is a one-sample Wilcoxon signed rank test that was used to examine

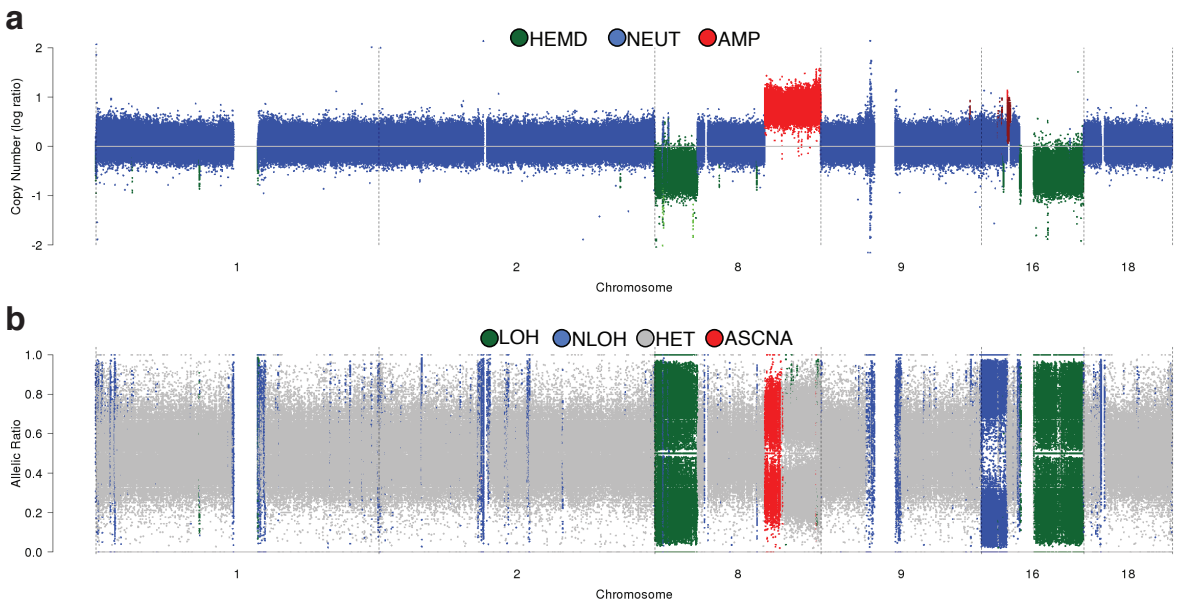
whether the allelic ratio distribution across the positions within the event was significantly different than the expected HAR. In particular, a one-tailed Wilcoxon test was used to assess if the symmetric allelic ratio, $SAR = \left(\frac{\max(ref\ reads, variant\ reads)}{depth}\right)$, distribution is greater than HAR. These two tests were applied to deletion and diploid heterozygous events for each type of test, separately. The p-values were adjusted using Benjamini & Hochberg correction across all events and all tumour or normal nuclei, separately.

Because the second test did not account for drop-out, both tests were combined by taking the maximum adjusted p-value to generate the final p-value representing the event. This conservatively ensured that a statistically significant final p-value (< 0.05 for both Set1 and Set2) indicated an LOH event that was supported by a homozygous allelic ratio and not due to allelic drop-out. The event was designated as heterozygous (HET) if the final p-value was not statistically significant and unknown (UNK) if the final p-value was not statistically but did not contain at least one heterozygous position (*present* status for both reference and variant). The cellular prevalence for each event was then computed based on nuclei that had the event status of LOH or HET.

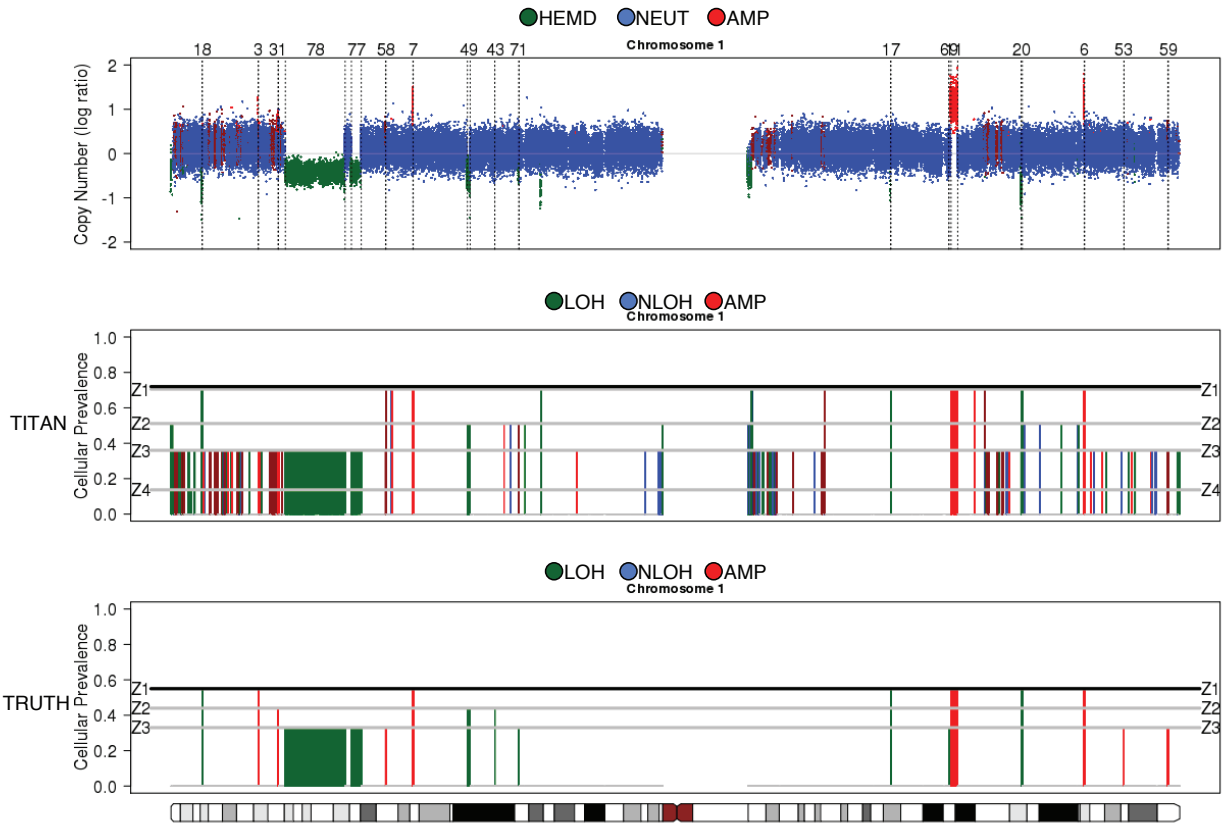
2 Supplementary Figures



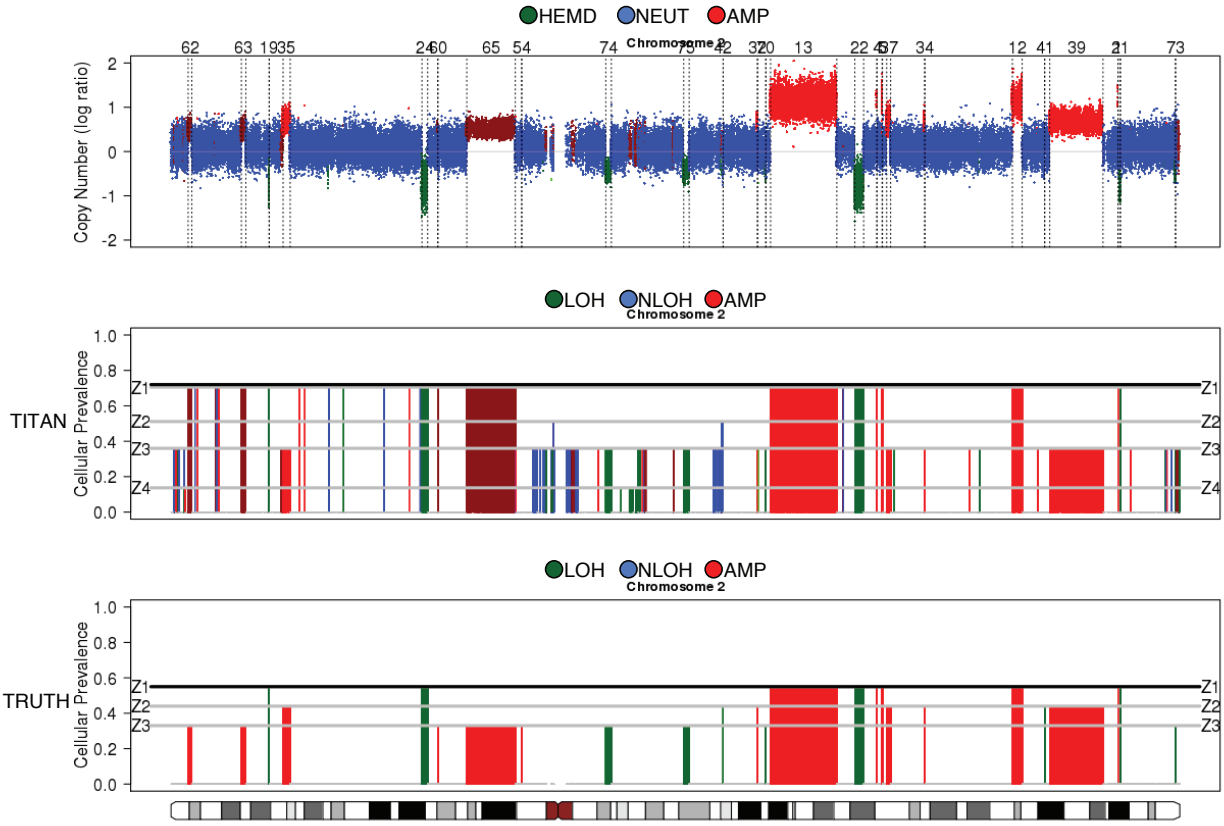
Supplementary Figure 1: Distribution of segment lengths (bp) for intra-patient samples of patient DG1136. CNA and LOH predictions made by HMMcopy and APOLLOH are shown.



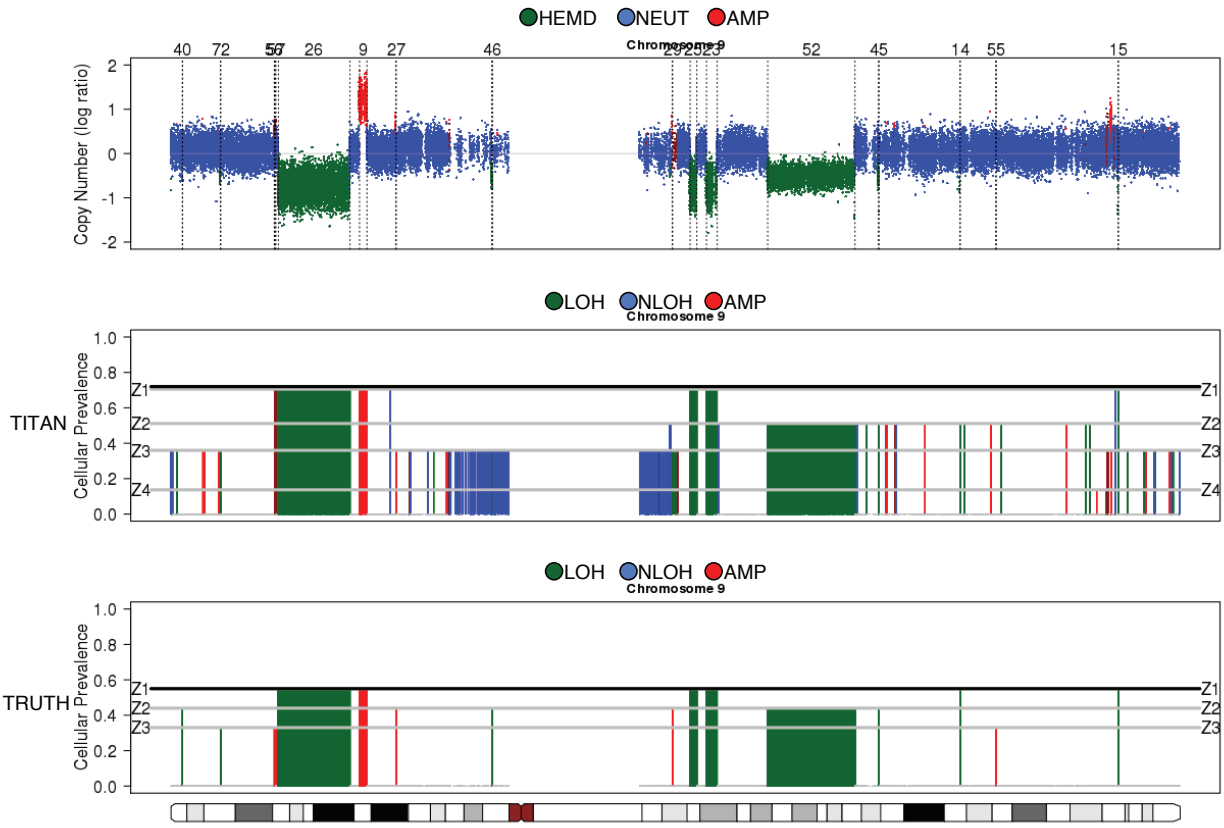
Supplementary Figure 2: HMMcopy (a) and APOLLOH (b) predictions of DG1136a used for the Spike-in simulation experiment. The log ratio and allelic ratio data for chromosomes 8 (chr8:97045605-144155272) and 16 (chr16:46464744-90173515) were randomly sampled and inserted into whole diploid heterozygous chromosomes of 1, 2, 9 and 18 as spike-in events of length 10, 100, 1000, and 10000 SNPs.



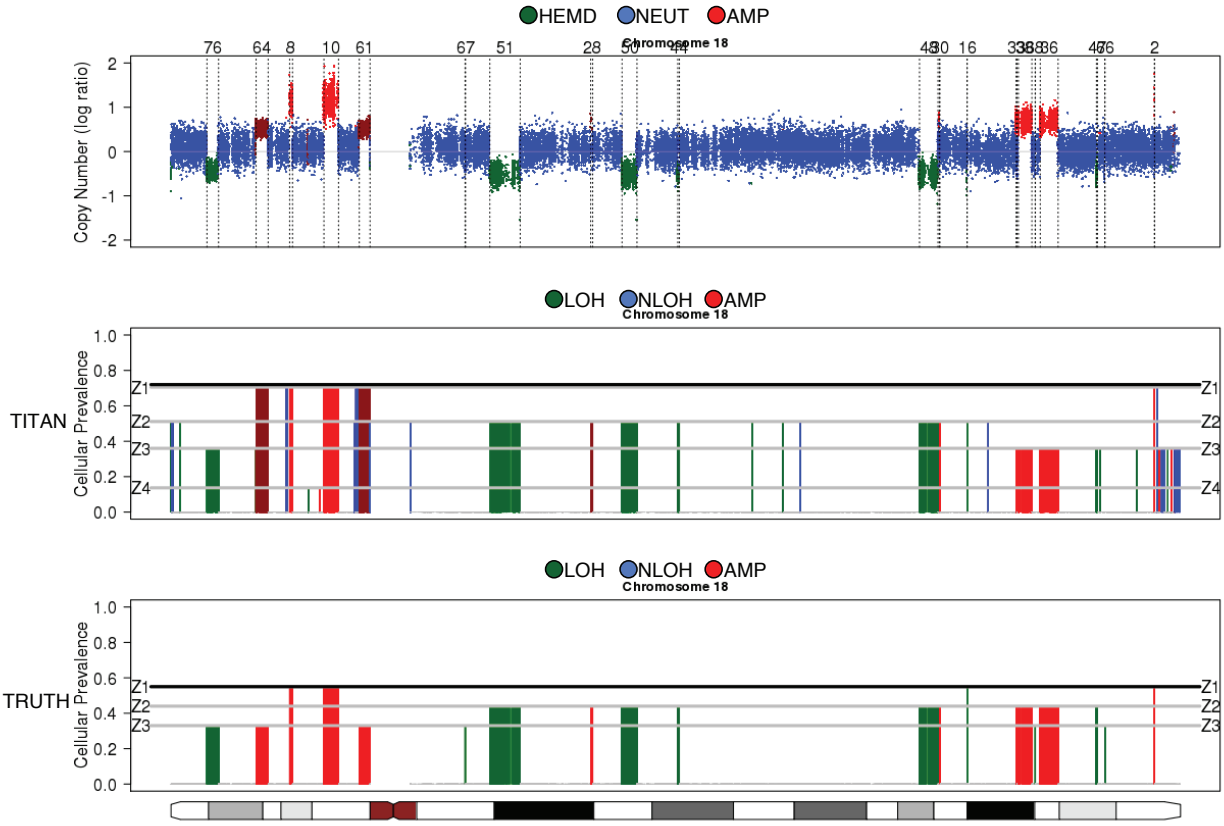
Supplementary Figure 3: TITAN CNA (top) and cellular prevalence (middle) results for chromosome 1 of the Spike-In simulation experiment using DG1136a. Spike-in events of length 10, 100, 1000, and 10000 SNPs were inserted. The vertical lines correspond to the known inserted (spiked-in) data; the number labels correspond to the list of events of the same ordering in Supplementary Table 2. The truth and TITAN-predicted cellular prevalence results for the spike-in events at chromosomes 1, 2, 9, and 18 are shown. TITAN cellular prevalence parameters were estimated on the entire genome including all original DG1136a events plus the spike-in events at the designated chromosomes. For log ratio plots, hemizygous deletion (HEMD), copy neutral (NEUT), and copy amplification (AMP) results are shown.The cellular prevalence value indicates the proportion of tumour cells in the whole sample. The plot follows the same colour legend as per the allelic ratio plot. Clonal clusters are shown in horizontal lines labeled with a 'Z'; tumour content is denoted with the black horizontal line. Deletion LOH (DLOH), copy neutral LOH (NLOH), diploid heterozygous (HET), and allele-specific amplification (ASCNA) are shown with green, blue, dark red, and red, respectively.



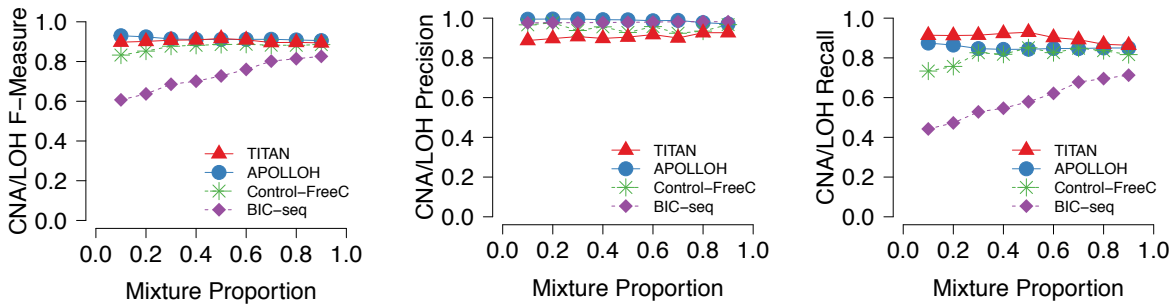
Supplementary Figure 4: TITAN CNA (top) and cellular prevalence (middle) results for chromosome 1 of the Spike-In simulation experiment using DG1136a. Spike-in events of length 10, 100, 1000, and 10000 SNPs were inserted. The vertical lines correspond to the known inserted (spiked-in) data; the number labels correspond to the list of events of the same ordering in Supplementary Table 2. The truth and TITAN-predicted cellular prevalence results for the spike-in events at chromosomes 1, 2, 9, and 18 are shown. TITAN cellular prevalence parameters were estimated on the entire genome including all original DG1136a events plus the spike-in events at the designated chromosomes. For log ratio plots, hemizygous deletion (HEMD), copy neutral (NEUT), and copy amplification (AMP) results are shown.The cellular prevalence value indicates the proportion of tumour cells in the whole sample. The plot follows the same colour legend as per the allelic ratio plot. Clonal clusters are shown in horizontal lines labeled with a 'Z'; tumour content is denoted with the black horizontal line. Deletion LOH (DLOH), copy neutral LOH (NLOH), diploid heterozygous (HET), and allele-specific amplification (ASCNA) are shown with green, blue, dark red, and red, respectively.



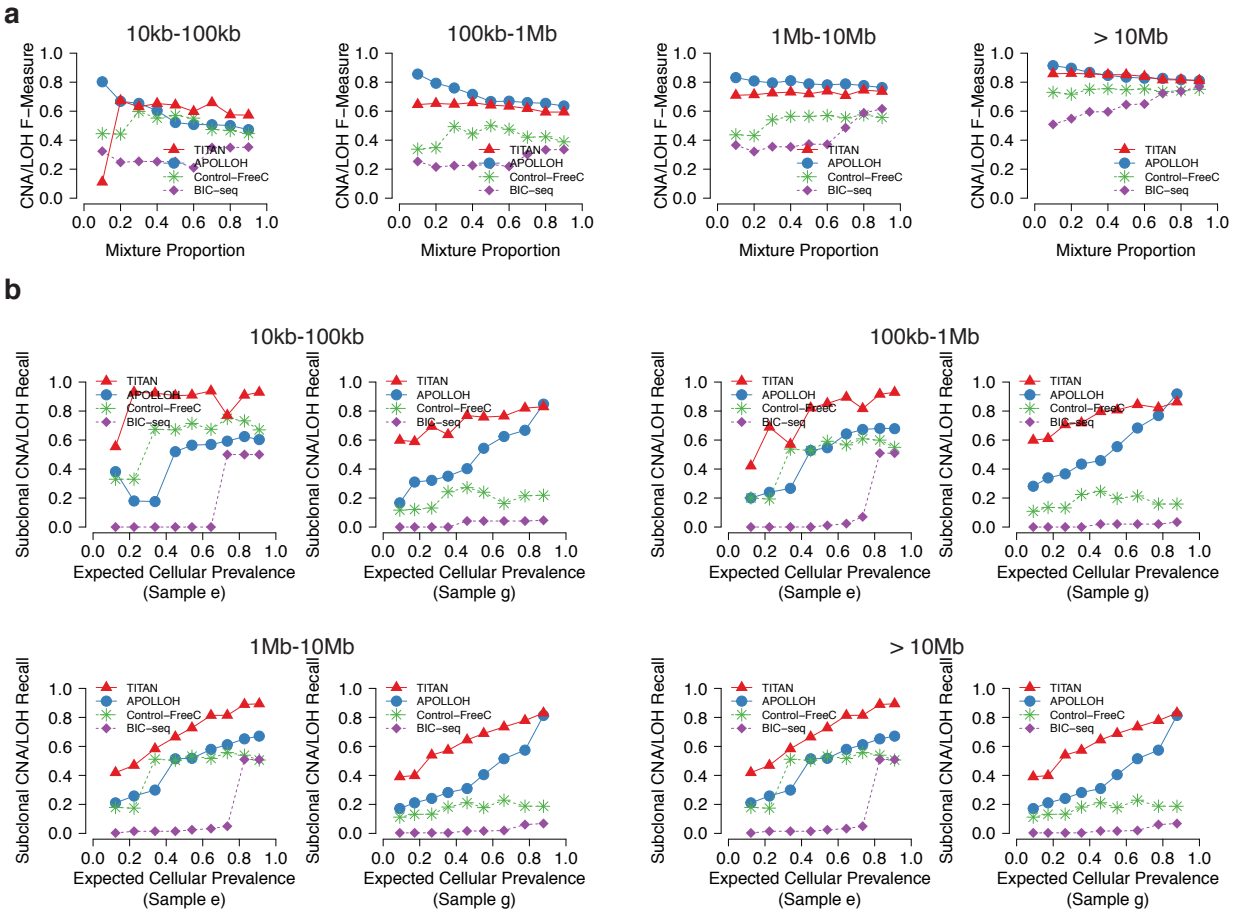
Supplementary Figure 5: TITAN CNA (top) and cellular prevalence (middle) results for chromosome 1 of the Spike-In simulation experiment using DG1136a. Spike-in events of length 10, 100, 1000, and 10000 SNPs were inserted. The vertical lines correspond to the known inserted (spiked-in) data; the number labels correspond to the list of events of the same ordering in Supplementary Table 2. The truth and TITAN-predicted cellular prevalence results for the spike-in events at chromosomes 1, 2, 9, and 18 are shown. TITAN cellular prevalence parameters were estimated on the entire genome including all original DG1136a events plus the spike-in events at the designated chromosomes. For log ratio plots, hemizygous deletion (HEMD), copy neutral (NEUT), and copy amplification (AMP) results are shown.The cellular prevalence value indicates the proportion of tumour cells in the whole sample. The plot follows the same colour legend as per the allelic ratio plot. Clonal clusters are shown in horizontal lines labeled with a 'Z'; tumour content is denoted with the black horizontal line. Deletion LOH (DLOH), copy neutral LOH (NLOH), diploid heterozygous (HET), and allele-specific amplification (ASCNA) are shown with green, blue, dark red, and red, respectively.



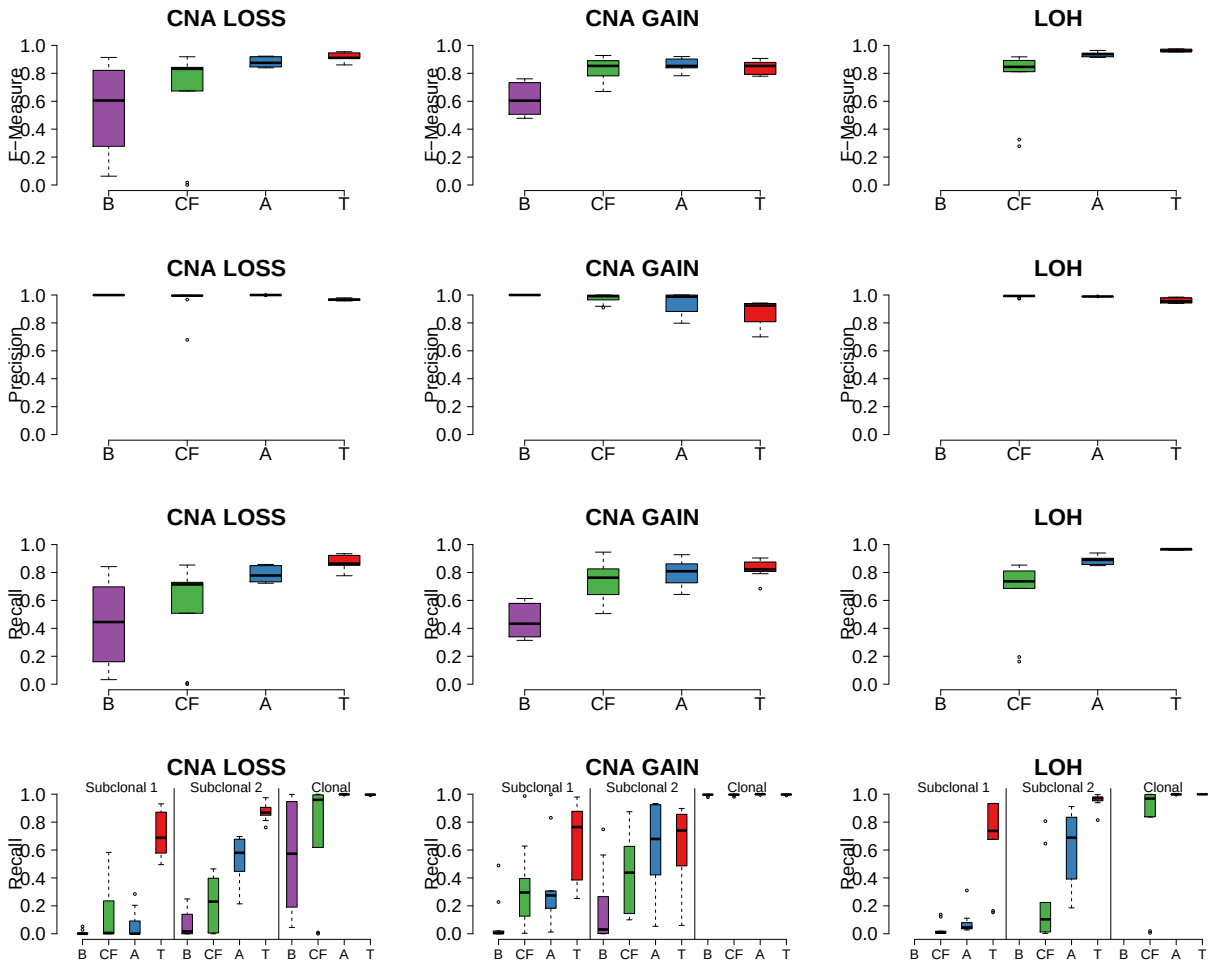
Supplementary Figure 6: TITAN CNA (top) and cellular prevalence (middle) results for chromosome 18 of the Spike-In simulation experiment using DG1136a. Spike-in events of length 10, 100, 1000, and 10000 SNPs were inserted. The vertical lines correspond to the known inserted (spiked-in) data; the number labels correspond to the list of events of the same ordering in Supplementary Table 2. The truth and TITAN-predicted cellular prevalence results for the spike-in events at chromosomes 1, 2, 9, and 18 are shown. TITAN cellular prevalence parameters were estimated on the entire genome including all original DG1136a events plus the spike-in events at the designated chromosomes. For log ratio plots, hemizygous deletion (HEMD), copy neutral (NEUT), and copy amplification (AMP) results are shown.The cellular prevalence value indicates the proportion of tumour cells in the whole sample. The plot follows the same colour legend as per the allelic ratio plot. Clonal clusters are shown in horizontal lines labeled with a 'Z'; tumour content is denoted with the black horizontal line. Deletion LOH (DLOH), copy neutral LOH (NLOH), diploid heterozygous (HET), and allele-specific amplification (ASCNA) are shown with green, blue, dark red, and red, respectively.



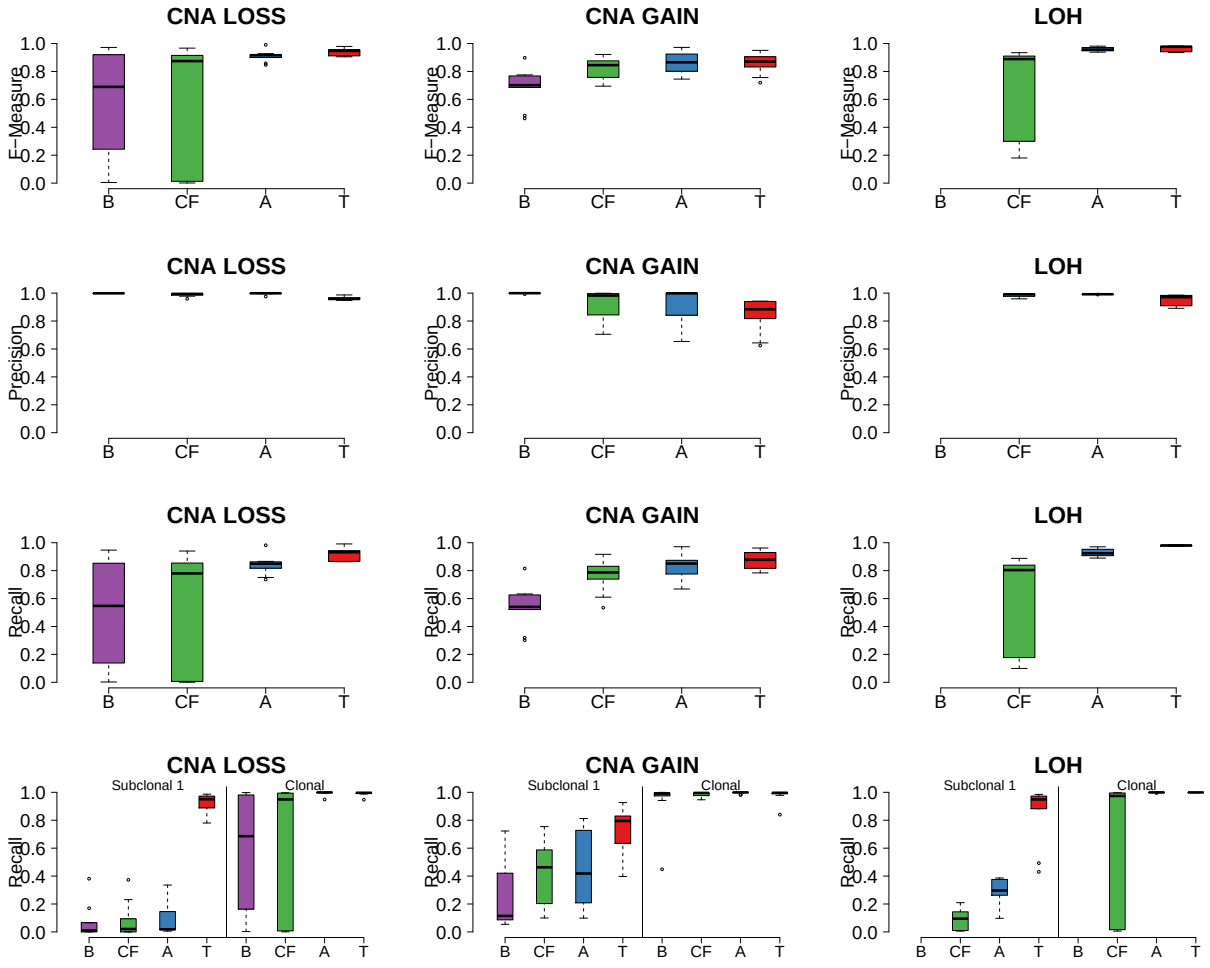
Supplementary Figure 7: Performance of TITAN for serial simulation of intratumour samples from an ovarian tumour. a) F-measure, precision, and recall performance across the mixture proportions comparing TITAN, APOLLOH (Ha et al., 2012) (including HMMcopy), Control-FREEC (Boeva et al., 2012), and BIC-seq (Xi et al., 2011). Performance for events for deletions, gains and LOH were averaged; see Supplementary Methods for how these metrics were computed. 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. 'Mixture Proportion' is defined as the ideal mixing fractions (e.g. 10%, 20%, etc.); expected 'cellular prevalence' is defined as the expected tumour contribution, at a given mixture proportion, from each individual sample making up the mixture. Performance was computed as described in Supplementary Methods.



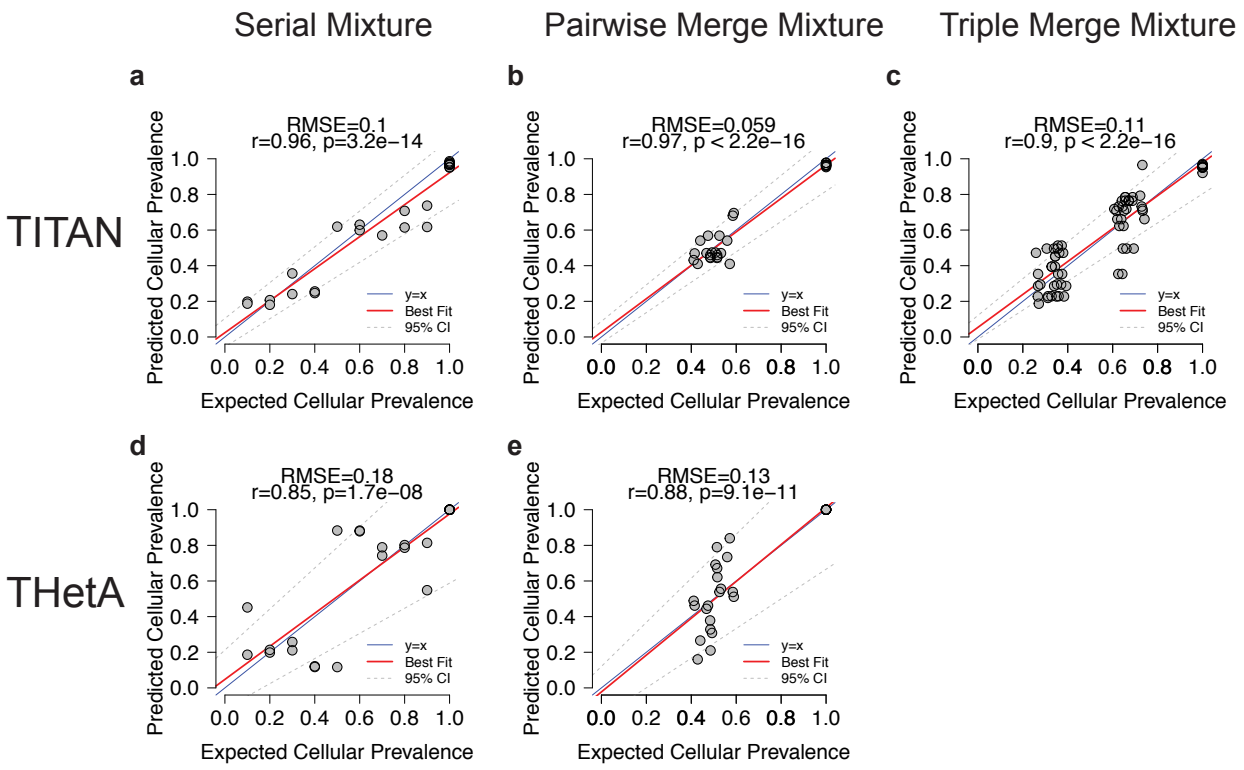
Supplementary Figure 8: Performance of TITAN for serial simulation of intratumour samples from an ovarian tumour evaluated at different event size groups. Sample DG1136e and DG1136g were mixed at known proportions (Supplementary Table 3). Events were grouped into ranges of lengths 10kb-100kb, 100kb-1Mb, 1Mb-10Mb, and greater than 10Mb as predicted in the ground truth on the samples, individually. a) F-measure performance across the mixture proportions comparing TITAN with Control-FREEC (Boeva et al., 2012), APOLLOH (Ha et al., 2012) (including HMMcopy), and BIC-seq (Xi et al., 2011). Events for deletions, gains and LOH are averaged. b) Recall performance for TITAN subclonal prediction results shown for the expected cellular prevalence computed from the original tumour contribution of each sample in the mixture (Supplementary Table 3). For each size range, performance is shown for subclonal events found only contributing from DG1136e and events only contributing from DG1136g. Cellular prevalence is defined as the proportion of tumour cells harbouring the events. Performance was computed as described in Supplementary Methods.



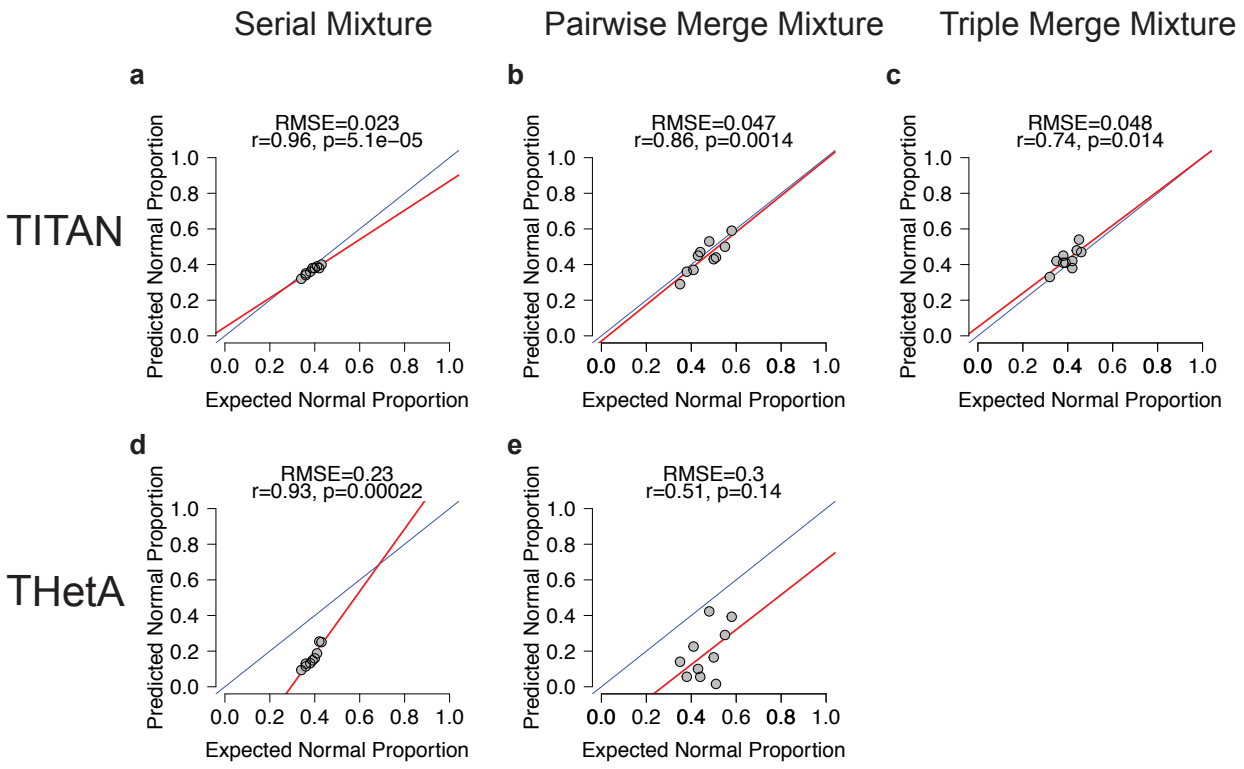
Supplementary Figure 9: Triplet merging simulation performance for TITAN (T), APOLLOH (Ha et al., 2012) (A, including HMMcopy), Control-FREEC (Boeva et al., 2012) (CF), and BIC-seq (Xi et al., 2011) (B). Combinations of three individual intratumour biopsy samples from an ovarian tumour were mixed at approximately equal proportions (see Supplementary Table 3). F-measure (first row), precision (second row), and recall (third row) for all events (both clonal and subclonal) are shown, separated into CNA loss, gains, and LOH. Recall for subclonal events (fourth row) are presented based on the number of individual samples within the mixture events are present. 'Subclonal 1' denotes events that are present in exactly one sample in the mixture and therefore considered subclonal in the simulation. Similarly, 'Subclonal 2' denotes events that are present in exactly two out of three samples in a triplet merge simulation. 'Clonal' denotes events present in exactly three samples and thus are clonally dominant in the simulation. Performance was computed as described in Supplementary Methods. Ground truth events were identified in the individual samples of the mixture using APOLLOH/HMMcopy and expected prevalence values are shown in Supplementary Table 3c.



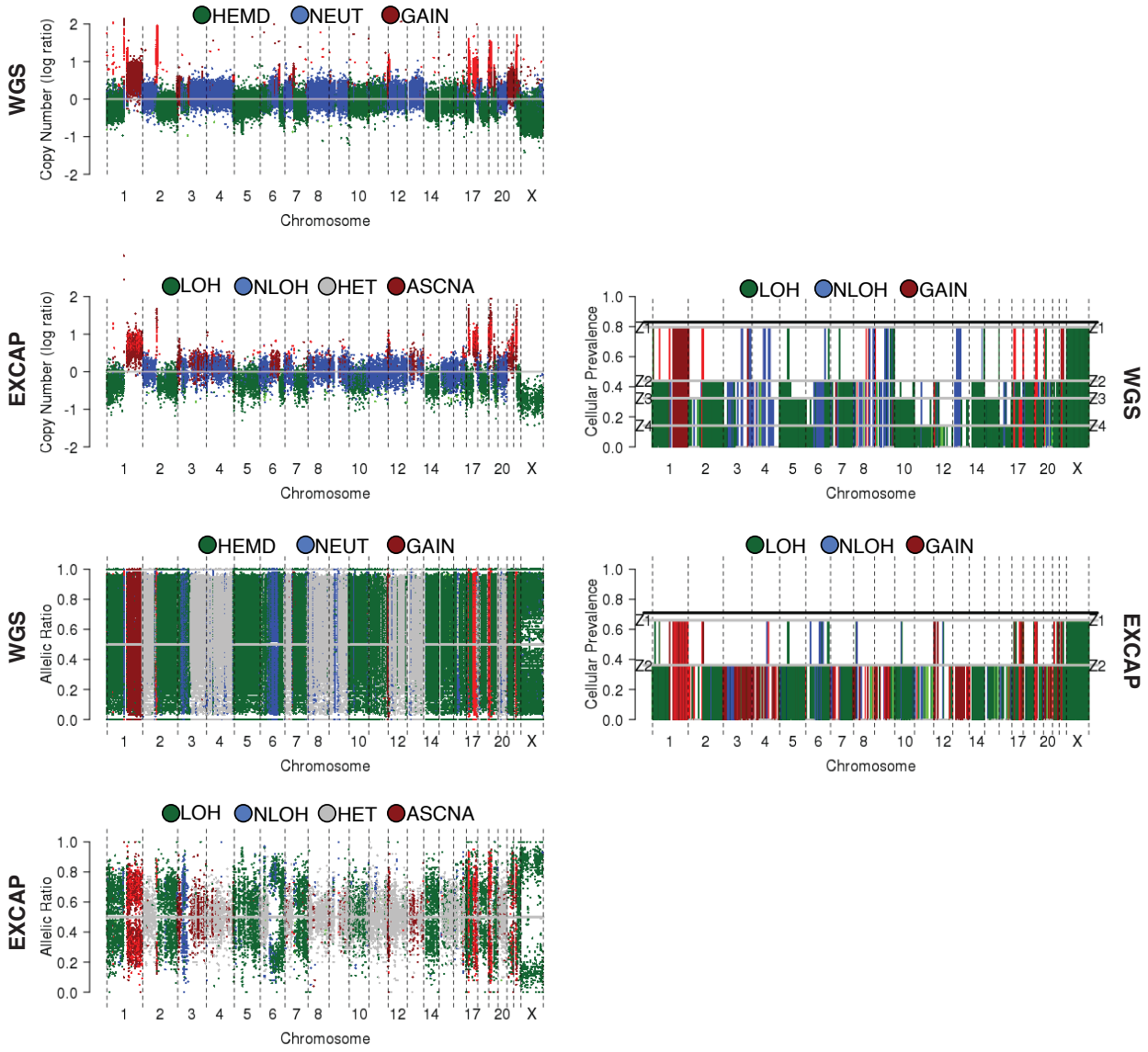
Supplementary Figure 10: Pairwise merging simulation performance for TITAN (T), APOLLOH (Ha et al., 2012) (A, including HMMcopy), Control-FREEC (Boeva et al., 2012) (CF), and BIC-seq (Xi et al., 2011) (B). Combinations of three individual intratumour biopsy samples from an ovarian tumour were mixed at approximately equal proportions (see Supplementary Table 3). F-measure (first row), precision (second row), and recall (third row) for all events (both clonal and subclonal) are shown, separated into CNA loss, gains, and LOH. Recall for subclonal events (fourth row) are presented based on the number of individual samples within the mixture events are present. 'Subclonal 1' denotes events that are present in exactly one sample in the mixture and therefore considered subclonal in the simulation. 'Clonal' denotes events present in exactly two samples and thus are clonally dominant in the simulation. Performance was computed as described in Supplementary Methods. Ground truth events were identified in the individual samples of the mixture using APOLLOH/HMMcopy and expected prevalence values are shown in Supplementary Table 3d.



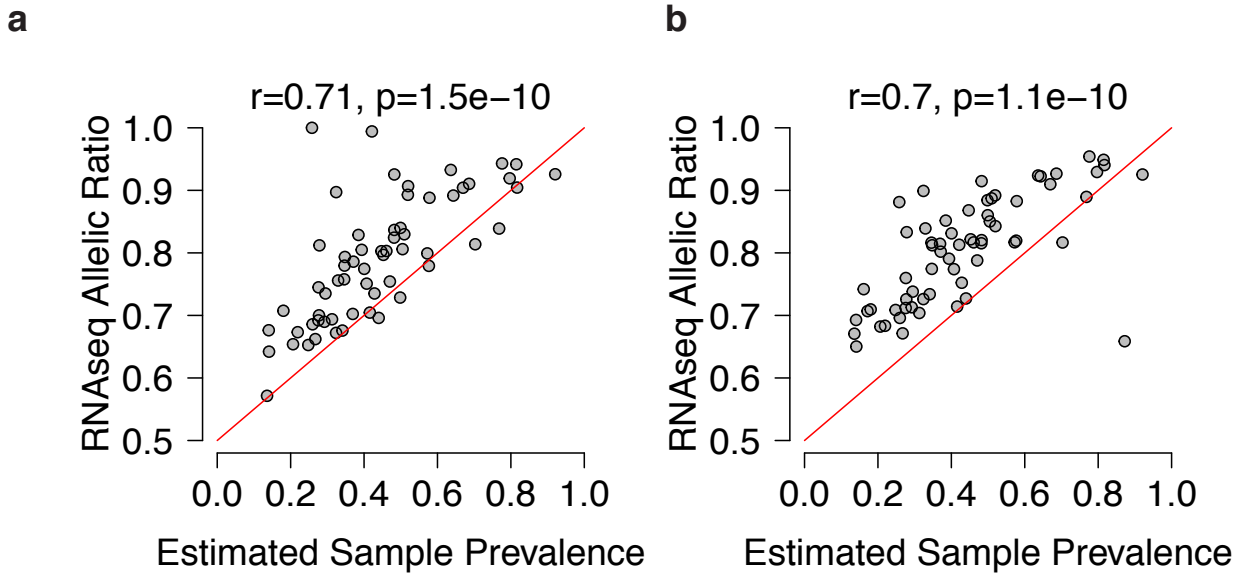
Supplementary Figure 11: Performance of TITAN cellular prevalence and normal proportion estimates for serial and pairwise/triplet merging simulations of intratumour samples from an ovarian tumour. Pearson correlation coefficients are shown and all correlations were significant. The root mean squared error (RMSE) is also presented. Expected normal proportion was determined as the consensus of the pathologist and Control-FREEC (Boeva et al., 2012) estimates.



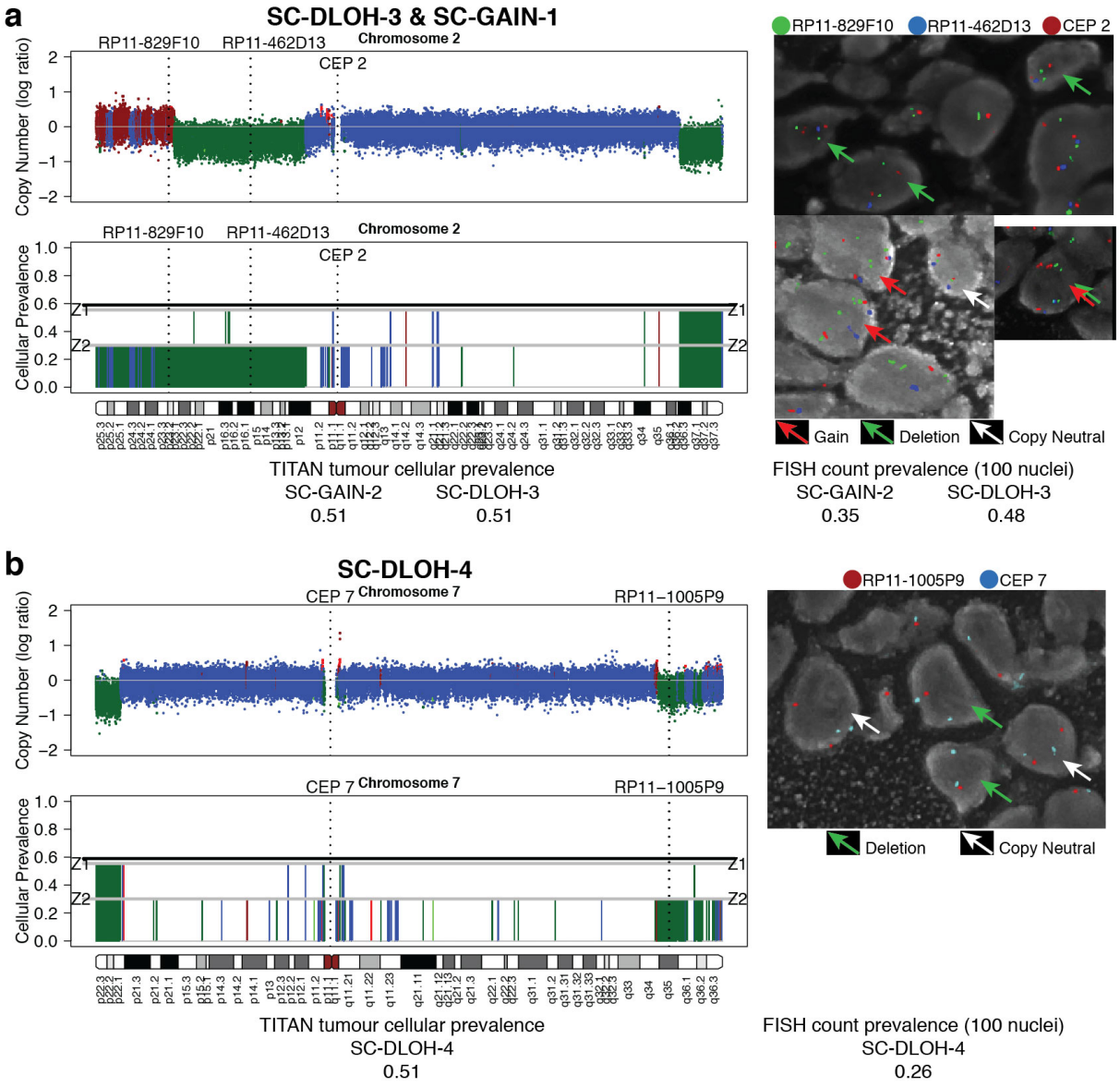
Supplementary Figure 12: Performance of TITAN cellular prevalence and normal proportion estimates for serial (30X) and pairwise (60X)/triplet (90X) merging simulations of intra-tumour samples from an ovarian tumour. Pearson correlation coefficients are shown for TITAN (a-c) and THetA (Oesper et al., 2013) (d-e) estimates where each data point represents a sample in the mixture. The root mean squared error (RMSE) is also presented. Ground truth events were identified in the individual samples of the mixture using APOLLOH (Ha et al., 2012) and expected normal proportion was determined as the consensus of the pathologist and APOLLOH estimates (Supplementary Table 3b-d).



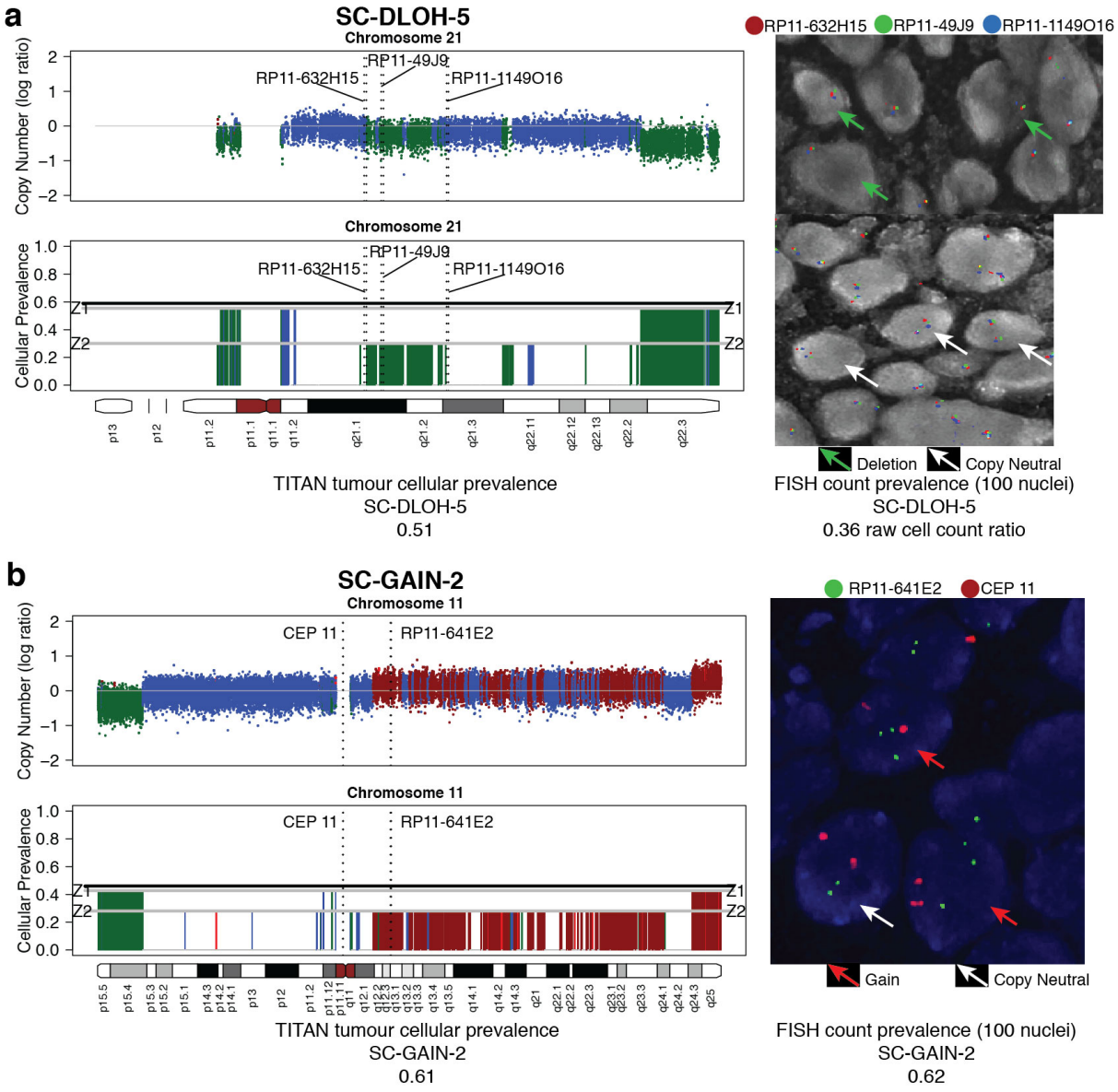
Supplementary Figure 13: Comparison of TITAN results for whole exome capture (EXCAP) sequencing and whole genome sequencing (WGS) of triple negative breast cancer sample SA052. For copy number plots, copy neutral, deletion, amplification are represented by blue, green, red, respectively. For log ratio plots, hemizygous deletion (HEMD), copy neutral (NEUT), and copy gain (GAIN) results are shown. For allelic ratio plots, LOH, copy neutral LOH (NLOH), diploid heterozygous (HET), and allele-specific amplification (ASCNA) are shown. The cellular prevalence value indicates the proportion of tumour cells in the whole sample. Clonal clusters are shown in horizontal lines labeled with a 'Z'; tumour content is denoted with the black horizontal line.



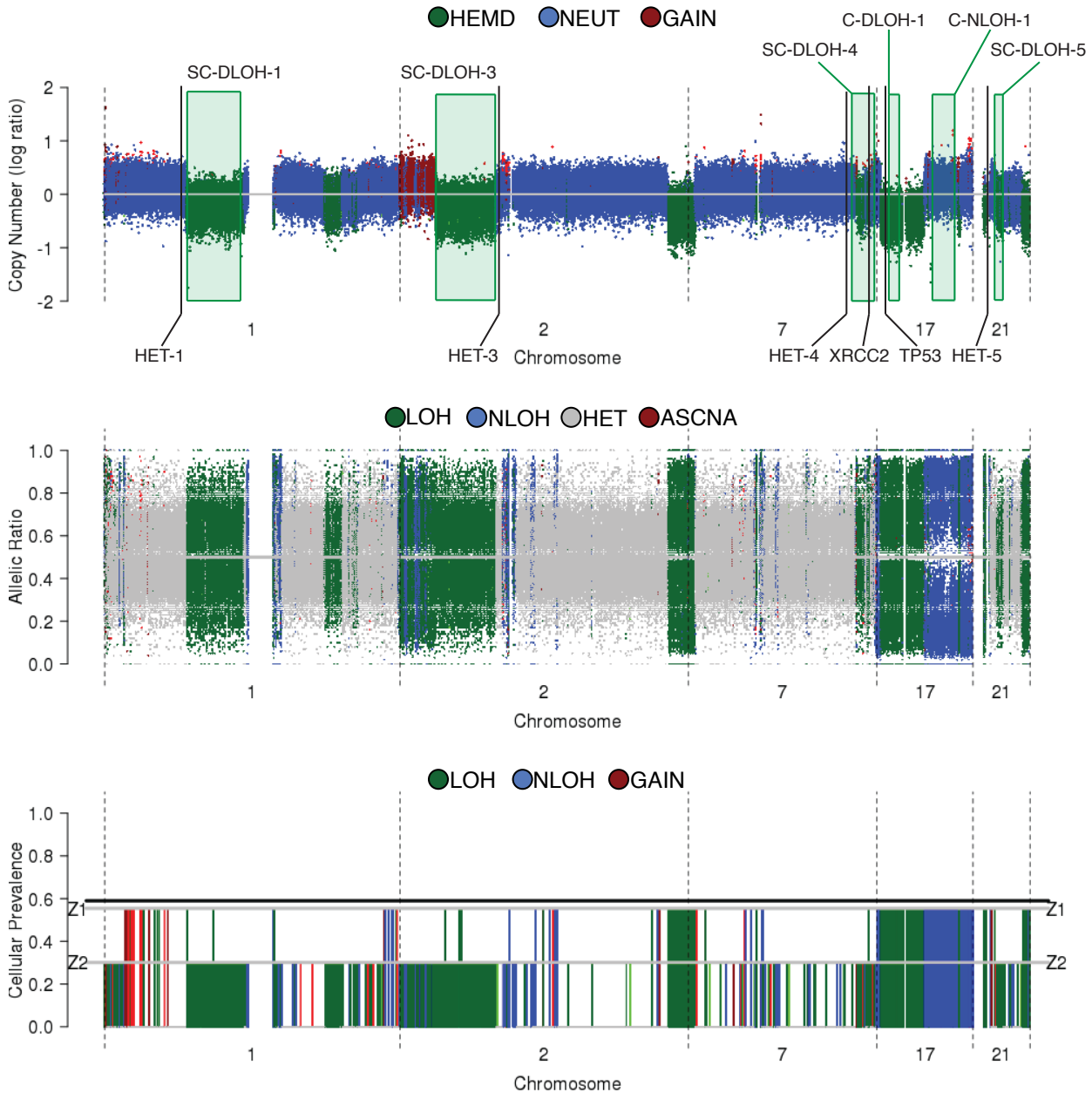
Supplementary Figure 14: Comparison of TITAN cellular prevalence and RNA-seq transcriptome allelic ratios (TAR). Sample prevalence (proportion within sample including normal contamination) for all LOH segments (**a**, deletion LOH, copy neutral LOH, and amplified LOH) and only deletion LOH (**b**) in all clonal clusters of all samples are shown (x-axis). The mean RNA-seq allelic ratio ($max(\frac{ref}{depth}, 1 - \frac{ref}{depth})$), for transcriptomic positions overlapping LOH regions for each clonal cluster across all samples are shown (y-axis). The Pearson correlation coefficient in this comparison was 0.71. The red line indicates the expected allelic ratio for the given sample prevalence assuming cells (tumour and normal) without the event are diploid heterozygous and both alleles are expressed equally, thus is a function of the cellular prevalence $s_z + (1 - sz)/2$. Allelic ratios may be more imbalanced due to epigenetic factors and higher copy numbers in cells with LOH. RNA-seq data was filtered based on depth threshold > 10, mapping quality > 30, and base quality > 5.



Supplementary Figure 15: Fluorescence in-situ hybridization (FISH) validation of TITAN predictions for chromosomes 2 and 7 in DG1136g. (**a**) A subclonal gain, SC-GAIN-1, and a subclonal hemizgyous deletion, SC-DLOH-3, in chromosome 2 was validated BAC probes RP11-829F10 (green, chr2:28,154,550-28,364,468) and RP11-462D13 (blue, chr2:59,904,520-60,114,863), respectively. The centromeric probe, CEP 2, was used as a control (orange). Nuclei harbouring only the gain, only the deletion, and both as co-occurring events were observed. (**b**) Subclonal hemizygous deletion, SC-DLOH-4, in chromosome 7 was validated using BAC probe RP11-1005P9 (orange, chr7:145530552-145724648). The centromeric probe, CEP 7, was used as the control (blue). The prevalence observed in the FISH was lower than that predicted by TITAN. 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.



Supplementary Figure 16: Fluorescence in-situ hybridization (FISH) validation of TITAN predictions for chromosome 21 in DG1136g and chromosome 11 in DG1136c. (**a**) Subclonal hemizygous deletion, SC-DLOH-5, in chromosome 21 of DG1136g was validated using BAC probe RP11-49J9 (green, chr21:22060503-22231762). The BAC probes RP11-632H15 (orange, chr21:20742595-20912882) and RP11-1149O16 (blue, chr21:27104756-27246972) were used as the controls. The FISH results indicates that the control probes were also deleted as part of the same event as SC-DLOH-5; therefore, there was no appropriate control for this deletion, and the raw cell count ratio of 0.36 was used. (**a**) A subclonal gain, SC-GAIN-2, in chromosome 11 of DG1136c was validated using BAC probe RP11-641E2 (green, chr17: 3294803-3452243). The centromeric probe, CEP 11, was used as the control (orange). The FISH prevalence (0.62) validates the TITAN-predicted cellular prevalence (0.61). 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.



Supplementary Figure 17: TITAN predictions selected for validation by single-cell sequencing of DNA from individual nuclei. Two clonally dominant LOH regions (C-DLOH-1 and C-NLOH-1) were selected from chr17. Four subclonal regions were selected from chr1 (SC-DLOH-1), chr2 (SC-DLOH-3), chr7 (SC-DLOH-4), and chr21 (SC-DLOH-5). For each region, 10-11 germline SNP loci were selected for deep amplicon sequencing in individual nuclei of single-cells (Supplementary Methods). Control sets of 2-3 SNP loci were selected from diploid heterozygous regions (HET-1, HET-3, HET-4, HET-5) nearby the subclonal regions. A set of somatic mutations (SNVs) were also selected as controls to distinguish cell types of normal and tumour nuclei (Supplementary Table 11a, 12a for full list of positions). For the log ratio plot (top), hemizygous deletion (HEMD), copy neutral (NEUT), and copy gain (GAIN) results are shown. For the allelic ratio plot (middle), LOH, copy neutral LOH (NLOH), diploid heterozygous (HET), and allele-specific amplification (ASCNA) are shown. The sample cellular prevalence plot (bottom) indicates the proportion of tumour cells in the whole sample. The plot follows the same colour legend as per the allelic ratio plot. Clonal clusters are shown in horizontal lines labeled with a 'Z'; tumour content is denoted with the black horizontal line.

3 Supplementary Tables

Supplementary Table 1: Copy number alteration (CNA) predictions for five individual biopsy samples of ovarian carcinoma DG1136. The sample IDs are DG1136a, c, e, g, i. a) HMMcopy segments are presented. The copy number ('state.name') are categorized as homozygous deletion (HOMD), hemizygous deletion (HETD), copy neutral (NEUT), gain (GAIN), amplification (AMP) and high-level amplicon (HLAMP). 'num.mark' is the number of 1kb bins within and included in a segment. 'state.num' is the integer state assigned based on HMMcopy output. b) The SNP loci from the APOLLOH analysis, which integrates HMMcopy results, formed the ground truth data used in the spike-in, serial and merging mixture simulation experiments. The number of SNP positions for deletions, amplifications, and LOH are given for each sample.

Supplementary Table 2: Spike-In simulation experiment. a) Randomly sampled deletion (from chr16) and amplification (from chr8) data was inserted into chr1, 2, 9 and 18. The 'Event ID' indicates which admixture sample the data originated from: clonally dominant (tum100), 80% tumour-normal mixture (tum80-norm20), and 60% tumour-normal mixture (tum60-norm20). The length, median allelic ratio and log ratio for each segment is given. b) Segment-based true positive rate (TPR) for each inserted spike-in event. The TPR is computed as the proportion of correctly predicted SNPs in the event. An event is true positive if TPR≥ 0.9. Cellular prevalence TPR is computed as the proportion of SNPs that is within ± 0.05 of the expected cellular prevalence of 0.65 (clonally dominant), 0.52 (80% admixture) and 0.36 (60% admixture; see Supplementary Methods). c) Size-based performance summarized across all spike-in events with 10, 100, 1000 and 10000 SNPs. A global false positive rate was computed on all negative (diploid heterozygous) positions in chr1, 2, 9, 18, which was where the spike-in events were inserted.

Supplementary Table 3: Simulation experiments using serial and merging mixtures of spatially related ovarian intra-tumoural samples. a) Patient DG1136 sample information including primary and metastatic tumour site information, sequencing coverage, tumour (content) cellularity estimates by the pathologist and predicted by APOLLOH. The consensus mean tumour content between the pathologist and APOLLOH was used to compute the expected cellular prevalence in the mixture simulations. b) Serial mixture experiment showing the tumour and normal cell contributions from DG1136e and DG1136g to each mixture. Proportion of each sample in the mixture was pre-defined at 10%-90%, 20%-80%, etc. '% tumour' columns are the sample cellular prevalence values for the mixture. The tumour cellular prevalence for Sample e is computed as '% tumour e'/ ('% tumour e' + '% tumour g'). TITAN results for number of clusters and normal and cellular prevalence estimates for each cluster are also presented. c) Pairwise merging mixture experiment tumour and normal cell contributions from pairwise combinations of DG1136a,c,e,g,i. Two samples were mixed at approximately equal proportions with differences attributed to difference in individual sample read coverage ('% of 1' and % of 2'). The sample cellular prevalence is given by '% tumour' columns. TITAN results are also shown. d) Triplet merging mixture experiment tumour and normal cell contributions from triplet combinations. Three samples were mixed at approximately equal proportions with differences attributed to difference in individual sample read coverage. The sample cellular prevalence is given by '% tumour' columns. TITAN results are also shown. e) TITAN results for the individual samples of DG1136. Parameter estimates for normal proportion, ploidy, and cellular prevalence for one and two clonal clusters are presented.

Supplementary Table 4: Performance of TITAN, APOLLOH/HMMcopy, Control-FREEC, and BIC-seq for serial (**a**) and pairwise (**b**) and triplet (**c**) merging simulation experiments. Ground truth data was determined from APOLLOH/HMMcopy predictions on the individual DG1136 samples. Performance metrics (precision, recall, F-measure) was computed for clonal and sub clonal events using ground truth status at germline heterozygous SNP positions. See Supplementary Methods for details.

Supplementary Table 5: Simulation experiments using serial and merging mixtures of spatially related ovarian intra-tumoural samples. a) Patient DG1136 sample information including primary and metastatic tumour site information, sequencing coverage, tumour (content) cellularity estimates by the pathologist and predicted by Control-FREEC. The consensus mean tumour content between the pathologist and Control-FREEC was used to compute the expected cellular prevalence in the mixture simulations. b) Serial mixture experiment showing the tumour and normal cell contributions from DG1136e and DG1136g to each mixture. Proportion of each sample in the mixture was pre-defined at 10%-90%, 20%-80%, etc. '% tumour' columns are the sample cellular prevalence values for the mixture. The tumour cellular prevalence for Sample e is computed as '% tumour e'/ ('% tumour e' + '% tumour g'). TITAN results for number of clusters and normal and cellular prevalence estimates for each cluster are also presented. c) Pairwise merging mixture experiment tumour and normal cell contributions from pairwise combinations of DG1136a,c,e,g,i. Two samples were mixed at approximately equal proportions with differences attributed to difference in individual sample read coverage ('% of 1' and % of 2'). The sample cellular prevalence is given by '% tumour' columns. TITAN results are also shown. d) Triplet merging mixture experiment tumour and normal cell contributions from triplet combinations. Three samples were mixed at approximately equal proportions with differences attributed to difference in individual sample read coverage. The sample cellular prevalence is given by '% tumour' columns. TITAN results are also shown.

Supplementary Table 6: Predicted CNA/LOH segments for 23 TNBCs using TITAN. 'Median_Ratio' is computed as the median symmetric ($max(\frac{ref}{depth}, 1 - \frac{ref}{depth})$) allelic ratio for positions overlapping the segment. 'Median_logR' is computed as the median logR for positions overlapping the segment. 'TITAN_state' and 'TITAN_call' are assigned from one of the states listed in Table S14. 'Copy_Number' represents the discrete number copies of the segment. 'MinorCN' is the number of copies from the allele having fewer copies. 'MajorCN' is the number of copies from the allele having more copies. 'Clonal_Cluster' is the clonal cluster state predicted by TITAN. 'Cellular_Prevalence' is the assigned prevalence estimate to the event. Coordinates are from NCBI build 36 (hg18).

Supplementary Table 7: TITAN results for 23 triple negative breast cancer (TNBC) WGS samples. a) TITAN parameter summary that includes the number of clonal clusters and their cellular prevalences, normal proportion and tumour ploidy estimates. b) Proportion of the length (bp) of the TNBC genome that is altered by clonal and subclonal events.

Supplementary Table 8: Comparison of TITAN results for whole exome (EXCAP) and genome (WGS) sequencing data. Concordance was computed based on overlapping germline heterozygous SNP positions between the EXCAP and WGS sample for the same patient sample. A match for a deletion ('DEL.Match'), amplification ('AMP.match'), or copy neutral ('HET.match') at an overlapping position if both were less than 2, both greater than 2, or both equal to 2, respectively. 'Concordance' was computed as the proportion of overlapping positions that matched.

Supplementary Table 9: Validation of TITAN predictions using fluorescence in-situ hybridization (FISH). a) BAC and centromeric probes used for event Groups 1-5 (for DG1136g) and Group 6 (for DG1136c). Subclonal deletions (SC-DLOH-X) and gains (SC-GAIN-X) and clonal deletion (C-DLOH-X) and clonal copy neutral LOH (C-NLOH-X) are labelled. Coordinates are from genome build GRCh37 (hg19). b-g) FISH cell counts for 100-200 nuclei for each event group. h) Summary of the FISH cell counts and the event ratios (event:control). For 'Raw cell counts', 'Loss', 'Neutral', and 'Gain' are counts of nuclei that contain < 2, 2, and > 2 copies, respectively. For 'Event Ratios', 'Loss', 'Neutral', and 'Gain' are counts of nuclei that contain event:control ratio < 1, 1, and > 1, respectively. The final FISH count prevalence used are 'Cell Prev' values highlighted in yellow.

Event ID	**Sample**	**Location**	**HMMcopy**	**Control-FreeC**	**THetA**	**TITAN**	**FISH**
C-DLOH-1	DG1136g	CEP 17	✔	✔	✔	0.94	0.77
SC-GAIN-2	DG1136c	11q13.1	✔	✔	✘	0.61	0.62
SC-DLOH-1	DG1136g	1p31.1	✔	✘	✔	0.51	0.48
SC-DLOH-3	DG1136g	2p16.1	✔	✘	✔	0.51	0.48
SC-GAIN-1	DG1136g	2p23.2	✔	✔	✘	0.51	0.35
SC-DLOH-4	DG1136g	7q35	✘	✘	✘	0.51	0.26
SC-DLOH-5	DG1136g	21q21.1	✘	✘	✘	0.51	0.36*

Supplementary Table 10: Summary of CNA predictions compared with fluorescence in-situ hybridization (FISH) results. Subclonal deletions (SC-DLOH-X), subclonal gains (SC-GAIN-X) and clonal deletion (C-DLOH-1) were assayed for DG1136g and DG1136c. 'TITAN' predicted tumour cellular prevalence and the 'FISH' prevalence, which is the proportion of nuclei that contain ratio (event:control) < 1, 1, and > 1 for deletion, neutral and gain, respectively (Supplementary Table 9h) are presented. Presence (check mark) and absence (x mark) of the CNA events are indicated for HMMcopy (Ha et al., 2012), Control-FreeC (Boeva et al., 2012), and THetA (Oesper et al., 2013). (*) indicates that raw cell count proportion was used.

Supplementary Table 11: Single-cell analysis for Set1 events in DG1136g. a) List of amplicon regions in Set1; position of interest (mutations and SNPs) are indicated in column 'Name' with format "[event type]-[number or gene]_[chr]_[position]". 'C-DLOH' stands for clonal deletion; 'SC-DLOH' stands for subclonal deletion. b) List of nuclei in Set1 labeled with cell type: Control, Tumour, Normal, low coverage. Tumour and normal nuclei were predicted from presence and absence of mutations. Sequencing data for the normal (c) and tumour (d) nuclei. Binomial exact tests for presence/absence of alleles are shown. The status of the reference ('ref_status_NTCbg') and variant ('var_status_NTCbf') alleles for all positions are indicated as 'present', 'absent', or 'low_coverage'. Low coverage positions were determined as having depth of less than 50 reads. Event-based analysis for normal (e) and tumour (f) nuclei. For each event and each nuclei, the number of heterozygous ('BOTH') and homozygous ('XOR') positions, median allelic ratio ('Median_AR'), binomial test for drop-out and Wilcoxon rank sum test for allelic ratios. 'Combined_qvalue' was used to determine LOH status of an event if < 0.05.

Supplementary Table 12: Single-cell analysis for Set2 events in DG1136g. a) List of amplicon regions in Set2; position of interest (mutations and SNPs) are indicated in column 'Name' with format "[event type]-[number or gene]_[chr]_[position]". 'C-DLOH' stands for clonal deletion; 'SC-DLOH' stands for subclonal deletion. b) List of nuclei in Set2 labeled with cell type: Control, Tumour, Normal, low coverage. Tumour and normal nuclei were predicted from presence and absence of mutations. Sequencing data for the normal (c) and tumour (d) nuclei. Binomial exact tests for presence/absence of alleles are shown. The status of the reference ('ref_status_NTCbg') and variant ('var_status_NTCbf') alleles for all positions are indicated as 'present', 'absent', or 'low_coverage'. Low coverage positions were determined as having depth of less than 50 reads. Event-based analysis for normal (e) and tumour (f) nuclei. For each event and each nuclei, the number of heterozygous ('BOTH') and homozygous ('XOR') positions, median allelic ratio ('Median_AR'), binomial test for drop-out and Wilcoxon rank sum test for allelic ratios. 'Combined_qvalue' was used to determine LOH status of an event if < 0.05.

Variable	Description	Value
$\boldsymbol{\pi_Z}$	Initial state distribution for clonal clusters	Estimated by EM in M-step
$\boldsymbol{\delta_Z}$	Prior counts; parameter of Dirichlet for π_Z	User-defined
$\boldsymbol{\pi_G}$	Initial state distribution for genotypes	Estimated by EM in M-step
$\boldsymbol{\delta_G}$	Prior counts; parameter of Dirichlet for π_G	User-defined
Z_t	Latent variable for clonal cluster at position t	Estimated by EM in E-step
G_t	Latent variable for genotype at position t	Estimated by EM in E-step
a_t	Reference count at position t	Observed
N_t	Total read depth at position t	Observed
l_t	Log ratio of tumour-normal depths at position t	Observed
s_z	Clonal parameter of cluster z	Estimated by EM in M-step
n	Global normal proportion parameter	Estimate by EM in M-step
$(\sigma^2)_g$	Variance parameter of Gaussian for genotype g	Estimated by EM in M-step
ϕ	Tumour ploidy parameter	Estimated by EM in M-step
α_z	Hyperparameter of Beta prior (shape) on s_z	Uniform setting
β_z	Hyperparameter of Beta prior (scale) on s_z	Uniform setting
α_g	Hyperparameter of Inverse Gamma prior (shape) on σ_g^2	User-defined
β_g	Hyperparameter of Inverse Gamma prior (scale) on σ_g^2	User-defined
α_ϕ	Hyperparameter of Inverse Gamma prior (shape) on ϕ	User-defined
β_ϕ	Hyperparameter of Inverse Gamma prior (scale) on ϕ	User-defined
$\boldsymbol{T_t}$	$Z \times Z$ clonal cluster transition matrix at position t	Fixed using ρ_Z
$\boldsymbol{A_t}$	$K \times K$ genotype transition matrix at position t	Fixed using ρ_G

Supplementary Table 13: Description of random variables and fixed quantities in the TITAN framework depicted in Figure 2b) and described in Methods. $a_{1:T}$, $N_{1:T}$ and $l_{1:T}$ are observed input quantities. All hyperparameters are user-defined. The position-specific HMM transition probabilities for genotypes $\boldsymbol{A_t}$ and clonal clusters $\boldsymbol{T_t}$ are fixed quantities. s_z, n, $(\sigma^2)_{1:21}$, π_G, π_Z and are unknown variables estimated during expectation maximization (EM).

State	Genotype (G)	Total copy number (c)	Call
-1	NA	NA	OUT
0	NA	0	HOMD
1	A	1	DLOH
2	B		DLOH
3	AA	2	NLOH
4	AB		HET
5	BB		NLOH
6	AAA	3	ALOH
7	AAB		GAIN
8	ABB		GAIN
9	BBB		ALOH
10	AAAA	4	ALOH
11	AAAB		ASCNA
12	AABB		BCNA
13	ABBB		ASCNA
14	BBBB		ALOH
15	AAAAA	5	ALOH
16	AAAAB		ASCNA
17	AAABB		UBCNA
18	AABBB		UBCNA
19	ABBBB		ASCNA
20	BBBBB		ALOH

Supplementary Table 14: Tumour genotype states used by TITAN. Descriptions of states: homozygous deletion (HOMD), hemizygous deletion LOH (DLOH), copy neutral LOH (NLOH), diploid heterozygous (HET), amplified LOH (ALOH), gain/duplication of 1 allele (GAIN), allele-specific copy number amplification (ASCNA), balanced copy number amplification (BCNA), unbalanced copy number amplification (UBCNA). State -1 represents the outlier state (OUT).

References

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.

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.

Colella S, Yau C, Taylor J. M, Mirza G, Butler H, Clouston P, Bassett A. S, Seller A, Holmes C. C, and Ragoussis J, *et al.*, 2007. Quantisnp: an objective bayes hidden-markov model to detect and accurately map copy number variation using snp genotyping data. *Nucleic Acids Res*, **35**(6):2013–2025.

Ding J, Bashashati A, Roth A, Oloumi A, Tse K, Zeng T, Haffari G, Hirst M, Marra M. A, Condon A, *et al.*, 2012. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, **28**(2):167–175.

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.

Li H and Durbin R, 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**(14):1754–1760.

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.

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.

Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, *et al.*, 2012. Jointsnvmix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, **28**(7):907–913.

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.

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.

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.

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).