Supplementary Note
Divergent Modes of Clonal Spread and Intraperitoneal Mixing in High-Grade Serous Ovarian Cancer

Andrew McPherson*[1,2,3,4] Andrew Roth*[1,3], Emma Laks[1], Tehmina Masud[1,4], Ali Bashashati[1], Allen W. Zhang[1,3,5], Gavin Ha[1,3,6,7], Justina Biele[1], Damian Yap[1], Adrian Wan[1], Leah M. Prentice[8], Jaswinder Khattra[1], Maia A. Smith[1,3], Cydney B. Nielsen[4], Sarah C. Mullaly[1], Steve Kalloger[1], Anthony Karnezis[8], Karey Shumansky[1], Celia Siu[1], Jamie Rosner[1], Hector Li Chan[8], Julie Ho[8], Nataliya Melnyk[8], Janine Senz[8], Winnie Yang[8], Richard Moore[9], Andrew J. Mungall[9], Marco A. Marra[9], Alexandre Bouchard-Côté[10], C. Blake Gilks[11], David G. Huntsman[1,8,4], Jessica N. McAlpine[12], Samuel Aparicio[1,4], Sohrab P. Shah[1,9,4]

1. Department of Molecular Oncology, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada
2. School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
3. Graduate Bioinformatics Training Program, University of British Columbia, Vancouver, BC, Canada
4. Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, V6T 2B5, Canada
5. Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Vancouver, BC, Canada
6. Dana Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215, USA
7. Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA
8. Centre for Translational and Applied Genomics, BC Cancer Agency, 600 West 10th Avenue, Vancouver, BC, V5Z 4E6, Canada
9. Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada
10. Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada
11. Vancouver General Hospital, 899 West 12th Ave, Vancouver, BC, V5Z 1M9, Canada
12. Department of Gynecology and Obstetrics, University of British Columbia, Vancouver, BC, Canada

* - equal contribution

Corresponding Authors:
Sohrab P Shah (sshah@bccrc.ca)
Samuel Aparicio (saparicio@bccrc.ca)

Keywords: tumour evolution, high grade serous ovarian cancer, phylogenetic reconstruction, whole genome sequencing, single nucleus sequencing, intraperitoneal spread.

Contents

Nature Genetics: doi:10.1038/ng.3573

List of Figures

Nature Genetics: doi:10.1038/ng.3573

List of Tables

1 Sample acquisition, patient consent and surgery

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. Cases of high-grade serous carcinoma where more than one sample were collected in different anatomic locations (e.g., different locations within the ovary, omentum) or where material was available over different time periods (e.g., at primary surgery and at recurrence) were chosen for this analysis. Clinicopathologic and outcome data was collected by chart review. Consistent with the practice at UBC and the British Columbia Cancer Agency all patients with high-grade serous cancer are referred to the hereditary cancer clinic and offered genetic testing for BRCA1 and BRCA2 mutations [1,2] (`http://www.bccancer.bc.ca/HPI/CancerManagementGuidelines/HereditaryCancerProgram/referralinformation/hboccriteria.htm`).

For consented patients, when multiple tumor sites were encountered intraoperatively, effort was made to bank multiple sites which were then flash frozen and stored according to conditions outlined below. For cases where multiple tumor sites were encountered intraoperatively but not all anatomic sites were banked (e.g., multiple ovarian samples banked but not other anatomic sites due to unavailability of trained staff) we looked to archival specimens stored within our pathology department to extract DNA and sequence. All samples were from removed structures during attempts at optimal debulking hence the majority of samples coming from omentum and ovarian samples. Within cases that were not optimally debulked, there were no additional samples studied from any sites not removed.

1.1 Specimen preservation and histologic evaluation In cases identified with high grade serous histology, multiple tissue samples were obtained from primary ovarian tumour and metastatic sites where adequate tumour volume permitted. When the ovary was pathologically enlarged, samplings were taken from up to five different areas with an effort made to equally space samples while staying within grossly apparent tumour tissue. Each sampling is cut into three pieces, yielding two end-pieces for cryovials and a middle portion placed in 10% buffered formalin. All paraffin-embedded blocks, including formalin-fixed tumour samples and molecular-fixed fallopian tubes, were sectioned and stained with hematoxylin and eosin prior to expert histopathological review (CBG) to confirm the presence of high grade serous carcinoma.

1.2 Library construction and sequencing A total of 10 patients' samples were submitted for library construction and sequencing. Sample size was determined by availability of resectable, cryo-preserved tissue, and also DNA quality. Samples from patients 5 and 6 were excluded due to low tumour cellularity as indicated by the absence of SNVs or copy number changes. Patient 8 was excluded as contaminated due to the high number of somatic variants present in dbSNP. For all tumour and normal samples, DNA extraction was followed by library construction and sequencing using Illumina HiSeq2000 whole genome shotgun v3 chemistry paired-end 100bp reads. Samples were sequenced to an average of 30x coverage (**Supplementary Table 2**).

2 WGSS analysis

2.1 Alignment Reads were aligned to the hg19 reference genome downloaded from `http://www.bcgsc.ca/downloads/genomes/9606/hg19/1000genomes/bwa_ind/genome/GRCh37-lite.fa`. Alignments were performed using bwa [3] using the `aln` and `sampe` commands. Duplicates were flagged with Picard `http://picard.sourceforge.net`.

2.2 SNV calling SNVs were called using both Strelka 1.0.14 and MutationSeq 4.2.0 with default parameters. Mappability scores were annotated for each position using precomputed values down-

Nature Genetics: doi:10.1038/ng.3573

loaded from UCSC (`http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/release3/wgEncodeCrgMapabilityAlign50mer.bigWig`). For downstream analysis we only considered mutations with mappability score of 1.0. We considered an SNV of high quality if it was predicted by both MutationSeq and Strelka to be present in any sample from a patient, not necessarily the same sample for each program. Gene name, predicted effect and impact of SNVs were annotated using SnpEff 4.0e [4].

2.3 Mutation Signatures Multiple mutational mechanisms may be active at one time in a single cancer, with the resulting signatures of each mechanism mixed to produce the observed pattern of mutations. To identify the relevant signatures for each sample, and the proportion of mutations attributed to each signature in each sample, we used a Latent Dirichlet Allocation (LDA) model. Within the context of an LDA model, mutations are generated by first selecting a signature with multinomial probability specific to each sample, and then selecting a mutation with multinomial probability specific to the signature [5]. We take the matrix of signature mutation probabilities from cosmic to be the signature probabilities of the LDA, resulting in a fitted LDA that we can use with our additional samples. We then learn the multinomial probability over signatures for each sample. Specifically, we use the `transform` function of the lda python package (`http://pythonhosted.org/lda/getting_started.html`), with $\alpha = 0.01$, and `max_iter` $= 1000$.

2.4 Rearrangement and breakpoint prediction Rearrangement breakpoints were predicted using deStruct v0.1.2 software derived from nFuse [6], available at `https://bitbucket.org/dranew/destruct.git`. In brief, discordant and non-mapping reads were extracted from bam files and realigned using a seed and extend strategy. Split alignment across a putative breakpoint was attempted for reads that did not fully align to a single loci. Discordant alignments were clustered according to the likelihood they were produced from the same breakpoint. Multiple mapped reads were assigned to a single mapping location using previously described methods [7]. Finally, heuristic filters removed predicted breakpoints with poor discordant read coverage of sequence flanking predicted breakpoints.

2.5 Copy number analysis We used a novel method named Demix to predict allele and clone specific copy number from WGS data. Demix is provided with deStruct 0.1.2 as the `demix.py` tool.

Based on an initial investigation, existing copy number prediction tools do not accurately model tetraploid genomes. Both Titan [8] and OncoSNP-SEQ [9] assume unaltered segments are diploid (2 copies). Segment copy number for a heterogeneous cell population is modelled as a mixture of 2 copies from normal cells, 2 copies from the unaltered tumour cells, and a tumour specific copy number of the altered tumour cells. In high grade serous ovarian cancer, genome doubling is an early event, and most cells are predominantly tetraploid. It is thus more accurate to model segment copy number as a mixture of diploid normal, and tumour specific copy numbers for a dominant and subdominant clones. An additional method, theta [10,11], accurately models multiple non-diploid tumour populations, but does not leverage allele specific read counts to infer allele specific copy number.

Demix provides several novel improvements over existing methods, and these improvements are critical to downstream analyses detailed in the main text. First, demix provides a joint segmentation of multiple samples from the same patient. Second, segment copy number is output in away that ensures alleles are identifiable across multiple samples. Distinguishing alleles as simply major (more copies) versus minor (less copies) is inadequate when comparing between samples, since 2 major 1 minor in sample A may be different from 2 major 1 minor in sample B if a different allele has been amplified to 2 copies. Third, copy number per segment is modelled as a mixture of 2 tumour specific copy numbers, allowing demix to more accurately model a tumour population for which the dominant clone is tetraploid and a minor clone has lost or gained copies relative to a base tetraploid state. For increased accuracy with respect to measurement of allele specific read counts, we use patient haplotype block information as predicted using shapeit [12] and 1000 Genomes data. The major steps of demix are detailed below.

Nature Genetics: doi:10.1038/ng.3573

Haplotype block prediction Normal reads covering 1000 Genomes SNP positions are classified as supportive of the reference or alternate allele. SNPs are classified as wild type, homozygous or heterozygous based on calculation of the posterior probability of each genotype g given the data. Specifically, alternate read count x of total count t is modelled as binomial distributed with binomial parameter $p = 0.5$ for heterozygous and $p = p_{\text{err}}$ for wild type and homozygous where $p_{\text{err}} = 0.01$ is the sequencing error (Equation 1). SNPs for which $P(g|x,t) > 0.9$ are taken to be heterozygous.

$$\begin{aligned} P(x,t|g) &= Bin(x|t,p) \\ P(g) &= \frac{1}{3} \\ P(g|x,t) &\propto \frac{P(x|g)P(g)}{P(x,t)} \end{aligned} \tag{1}$$

Heterozygous SNPs are used with shapeit to predict haplotype blocks. We use the preprocessed 1000 Genomes reference panel downloaded from `http://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1integrated_v3_impute.tgz` The `shapeit` executable was called once per chromosome with the following parameters `--no-mcmc`, and `--chrX` for chromosome X. We then used the `shapeit -convert` command to sample 100 haplotype block configurations from the haplotype graph output from the previous step. Adjacent heterozygous SNPs were considered confidently *phased* (ref/alt configuration on known for each parental allele) if the two SNPs co-occur in the same block for 95% of the samples.

Read count data preparation Segment boundaries are prepared from high to medium quality deStruct breakpoint predictions (`valid_prob` $*$ `chimeric_prob` > 0.5). Germline breakpoints and breakpoints seen in other patients are filtered as potential germline rearrangements or artefacts. The genome is segmented according to breakends of the remaining breakpoints. Haplotype blocks are also segmented by breakends, segmented blocks will be contained within a single genomic segment.

Concordantly aligning paired end reads are extracted from sample bam files and counted. Reads that fall entirely within segment boundaries are assigned to that segment, resulting in total read counts t_{si} per segment per sample. Additionally, reads covering heterozygous germline SNPs are classified as supportive of the reference or alternate allele. Reads are assigned to an allele of a haplotype blocks according to which heterozygous SNPs the read supports, and the allele and block to which those SNPs have been assigned. Reads that span multiple blocks or support conflicting alleles are not assigned to an allele of any block. Let $a_{sij\ell}$ denote the read count of sample s, segment i, block j, allele ℓ.

Multiple haplotype blocks within a segment are phased in a way that leverages multiple samples. For each segment in each sample, allelic imbalance b_{si} is calculated as the sum of absolute read count difference normalized by read count total (Equation 2). For each segment, the sample with highest allelic imbalance is selected as the *phasing* sample q_i (Equation 3). Major u_{ij} and minor v_{ij} phased alleles in the phasing sample are determined by selecting the alleles with higher or lower read counts (Equation 4-5). Finally, major m_{si} and minor n_{si} read counts for each sample for each segment are calculated by taking the maximum and minimum respectively of summed read counts for each

phased allele (Equation 6-7). Observed major x_{si} and minor y_{si} coverage is calculated as given by (Equation 8-9).

$$b_{si} = \frac{\sum_j |a_{sij1} - a_{sij2}|}{\sum_j \sum_\ell a_{sij\ell}} \tag{2}$$

$$q_i = \operatorname*{argmax}_s b_{si} \tag{3}$$

$$u_{ij} = \operatorname*{argmax}_\ell a_{sij\ell}|_{s=q_i} \tag{4}$$

$$v_{ij} = \operatorname*{argmin}_\ell a_{sij\ell}|_{s=q_i} \tag{5}$$

$$m_{si} = \max(\sum_j a_{sij\ell}|_{\ell=u_{ij}}, \sum_j a_{sij\ell}|_{\ell=v_{ij}}) \tag{6}$$

$$n_{si} = \min(\sum_j a_{sij\ell}|_{\ell=u_{ij}}, \sum_j a_{sij\ell}|_{\ell=v_{ij}}) \tag{7}$$

$$x_{si} = \frac{m_{si}}{m_{si} + n_{si}} t_{si} \tag{8}$$

$$y_{si} = \frac{n_{si}}{m_{si} + n_{si}} t_{si} \tag{9}$$

Probability Model We model bulk sequencing data as produced by a mixture of normal cells and two tumour cell populations. The larger tumour population is referred to as the *dominant* population, and the smaller as a *minor subclone*. The global parameters of the model include f, the frequency of the minor subclone, s, the number of reads per genomic length contributed by the normal cell population to the read count of each allele, and t, the number of reads per single copy per genomic length contributed by the dominant tumour population to the read count of each allele.

Model parameters

$f \in (0, 0.5)$	minor subclone frequency
$s \in \mathbb{R}^+$	normal haploid coverage
$t \in \mathbb{R}^+$	tumour haploid coverage
$b_i \in \mathbb{N}^2$	genotype base
$d_i \in \mathbb{N}^2$	genotype deviation
$g_i \in \mathbb{B}^2$	b_{ij} or $b_{ij} + d_{ij}$ is dominant
$u_i \in \mathbb{R}^2 \to \mathbb{R}$	maternal/paternal to major
$v_i \in \mathbb{R}^2 \to \mathbb{R}$	maternal/paternal to minor
$h_i \in (0, 1)$	base / deviation mixture
$a_i \in \mathbb{N}^2$	base / deviation mixture
$p_i \in \mathbb{R}$	major allele coverage
$q_i \in \mathbb{R}$	minor allele coverage

Per segment, copy number is modelled as a mixture of a base allele copy number pair b_i and a deviated allele copy number pair $b_i + d_i$ where d_i is the copy number deviation, always positive. The copy number of the dominant tumour population is either b_i or $b_i + d_i$ dependant on binary indicator pair g_i. The mixing coefficient h_i of the copy number of the two populations is calculated from g_i and f as given by Equation 10, and mixed copy number is calculated as given by Equation 11. Furthermore, u_i and v_i model the unknown mapping between allelic copy number a_i and observed major and minor alleles. Finally, haploid major and minor allele coverage of each segment p_i and q_i

Nature Genetics: doi:10.1038/ng.3573

are calculated as given by Equations 12 and 13.

$$h_i = g_i f + (1 - g_i)(1 - f) \tag{10}$$
$$a_i = (1 - h_i)b_i + h_i(b_i + d_i) \tag{11}$$
$$p_i = s + a_i u_i t \tag{12}$$
$$q_i = s + a_i v_i t \tag{13}$$

Observed major and minor read coverage x_i and y_i are modelled as normally distributed with mean p_i and q_i respectively and variance σ_x^2 and σ_y^2 (Equations 12, 13). Variance terms are estimated offline by assuming pairs of adjacent segments have identical copy number, and identifying the σ_x^2 and σ_y^2 that maximize the likelihood of seeing each pair of deviated allele coverages. A geometric prior is placed over the deviation d_i of each segment, reflecting the belief that the majority of the segments will have identical copy number in both dominant and subclonal populations. Uniform priors are used for the remaining unobserved variables including s, t, b_i, g_i, u_i, and v_i.

$$\begin{aligned} d_i &\sim \text{Geometric}(\phi) \\ x_i|s,t,f,g_i,b_i,d_i,u_i &\sim \mathcal{N}(p_i,\sigma_x^2) \\ y_i|s,t,f,g_i,b_i,d_i,v_i &\sim \mathcal{N}(q_i,\sigma_y^2) \end{aligned}$$

Inference Algorithm Parameters f, s, and t are inferred using a combination of Gibbs sampling and expectation maximization. Define auxiliary variable z_i such that $z_i = 1$ if and only if $d_i = (0, 0)$. Iteratively sample z_i from $P(z_i|x_i, y_i, f, s, t)$, then for regions with $z_i = 1$, calculate f, s, and t to maximize the expected value of the complete data likelihood $P(X, Y, B, D, G, U, V|f, s, t)$ under the conditional distribution of b_i, d_i, g_i, u_i, and v_i given the previous setting of f, s, and t. After sufficient iterations, take the f, s and t that maximized the marginal likelihood $P(X, Y|f, s, t)$ across all iterations. Copy number for each segment is taken to be the b_i, d_i, and g_i that maximizes the likelihood of the segment given inferred f, s and t.

Performance of the algorithm depends significantly on initial conditions. We generate a set of initial estimates for s and t based on an analysis of the multi-model distribution of the minor allele coverage y_{si}. For samples with a significant amount of ancestral loss of heterozygousity, a significant number of segments will have zero minor copies for all tumour cells. Coverage for the minor allele of these segments originates from normal cells only, and thus coverage for these segments will be approximately equal to s. Furthermore, we expect a significant number of segments will have a single minor copy for all tumour cells. Coverage of the minor allele for these segments will be approximately $s+t$. Thus we take the lowest value mode of y_{si} as a reasonable initialization of s. Higher value modes of y_{si} are used for multiple initializations of s_t. The algorithm will converge to a small number of locally optimal solutions. We select, by hand, the solution that minimizes the sizes of the dominant and subclonal genomes, while ensuring that the amount of homozygous deletion across dominant and subclonal genomes never exceeds more than 5% of the genome.

2.6 Phylogenetic inference of sample phylogenies We applied a novel statistical method (Supplemental Methods Section 6) which allows for a mutation to originate once along a tree, but to be lost at multiple points. SNV loss is governed by a global rate of loss parameter inferred during training. Measurement error is modelled by a binomial likelihood and incorporates 3 factors affecting the number of measured variant reads: allele specific copy number, tumour cellularity, and sequencing error. To infer the SNV trees we first performed an exhaustive search over all possible trees using only high confidence SNVs to compute tree likelihoods. For each tree, numerical methods were used to maximize the likelihood with respect to the loss parameter. We then used the maximum likelihood tree to compute origin and loss counts for all SNVs in the union of MutationSeq and Strelka calls.

Breakpoint trees were inferred using similar principles as described above for SNV trees. Breakpoints with less than total 4 supporting reads across samples were considered as low quality events and filtered. Rather than modelling a likelihood, for breakpoints we calculated presence/absence based on whether at least one read supported the breakpoint in the given sample. We calculated used a fixed error rate to account for measurement error: both the likelihood of observing at least one read given that the breakpoint is absent, and the likelihood of observing no reads given the breakpoint is present, was set to the value ϵ. We used $\epsilon = 0.01$ for the purposes of this study.

For both SNVs and breakpoints, we calculated the posterior probability of each mutation originating at, being present at, or being lost at each node in the sample tree. We also computed the maximum likelihood origin node, node specific presence, and node specific loss for each mutation. For SNVs we computed a binary indicator of whether a deletion of the encompassing segment was possible given a maximum parsimony reconstruction of ancestral copy number changes.

We used the likelihood ratio test to determine if allowing loss improved the fit of the phylogentic model to the data. This is possible because the null hypothesis that there is no loss ($\pi_\ell = 0$) is nested with respect to the alternate hypothesis ($\pi_\ell \geq 0$). It is a standard statistical [13] result that the likelihood ratio statistic, $D = -2 \times \ln \frac{L(\theta_0|x)}{L(\theta_1|x)}$ converges to a chi square distribution with one degree of freedom.

We remark that losses on some branches are unidentifiable within the context of our copy number naive model of SNV evolution. For instance, an ancestral mutation, lost in a clone directly descended from the ancestral clone, will exhibit evidence identical to a mutation originating in the sibling of the descendant clone. For a significant number of descendant SNVs identified in our cohort, an equally likely explanation of the data would be an origin at a more ancestral node, followed by a loss in a descendant node corroborated by an inferred copy number change (Figure **1**d). Within the context of our model, loss is always given lower probability than origin on a more descendant branch. Thus, our analysis is likely to underestimate the number of losses. Future work will involve extending the model, adding support for branch lengths and incorporating copy number information to identify SNVs that have been lost via deletion. Within the context of such a model, previously unidentifiable losses may become identifiable. For instance, an SNV may be more likely to have been deleted by an inferred copy number change as opposed to originating on very short descendant branch.

3 ERBB2/Her2 validation for patient 9

3.1 Fluorescence in situ hybridization for ERBB2 and CCNE1 Five-micron formalin-fixed paraffin embedded sections were hybridized with FISH probes to the appropriate patient sample. For patient 9, probes from the LSI Her-2/neu and CEP 17 with the Path-Vysion HER-2 DNA Probe Kit (Abbott Molecular) were used. For the CCNE1 locus in patient 2, in-house FISH probes were constructed using BACs labelled with a Nick Translation Kit (Roche Life Sciences) as previously described [14]. For the CCNE1 locus, BAC RP11-345J21 was labelled with a spectrum orange fluorophore and the reference probe, BAC RP11-81M8, was labeled with a spectrum green fluorophore. The reference probe was chosen as a region of neutral copy number based on bioinformatic CNA analysis. Each slide was counterstained with DAPI and visualized on a Zeiss Axioplan epifluorescent microscope. Probe signals were enumerated in 40-100 individual nuclei for each patient sample using either 60X or 100X magnification.

3.2 Immunohistochemistry for Her2 Formalin-fixed paraffin-embedded tissue blocks were cut at 4 micron thickness, dried, deparaffinized and stained using a Ventana Discovery XT stainer (Ventana, Tucson, AZ, USA) with antibodies against HER2 (cat # RM-9103, clone SP3, 1:100, Thermo Scientific, Ottawa, ON, Canada) and p16 (cat # 9518, clone E6H4, 1:2, Roche mtm laboratories, Heidelberg, Germany). Sections were thoroughly washed, dehydrated and coverslipped with Consul-Mount (9990440; Thermo Scientific, Ottawa, ON, Canada).

- Antibody: HER2
- Supplier: Thermo Scientific
- Catalogue: RM-9103
- Host: Rabbit
- Clone: SP3
- Stainer: Ventana Discovery XT
- Detection Kit: DAB Map
- Antigen Retrieval: Standard CC1 (Cell Conditioning 1)
- Primary Incubation: 60 mins with heat
- Primary Dilution: 1:100
- Secondary: Universal Secondary
- Secondary Incubation: 32 mins

4 Targeted bulk sequencing analysis

4.1 Illumina TruSeq/Nextera sequencing SNVs in protein coding regions of the genome were chosen based on distribution of prevalence in all or a subset of samples in each patient. The aggregate set of targets was then used to design a custom panel using the Illumina TruSeq Custom Amplicon design. Targets are listed in **Supplementary Table 15**. Amplicon-based libraries were constructed for all DNA templates (31 tumour samples and 7 normal samples) using the previously designed TruSeq Custom Amplicon Kit as per manufacturer's instructions. The constructed indexed libraries were pooled and loaded onto the Illumina MiSeq Personal Desktop Sequencer using the MiSeq Reagent 300 cycle kit V2 resulting in 151bp paired-end reads.

4.2 FFPE archival tissue Operative and pathology reports from the 7 cases were reviewed. Returning to FFPE archival specimens or FFPE specimens in our tumor bank where applicable we identified representative sections of multiple anatomic sites of tumor burden. Genomic DNA was extracted using the Gentra Purgene Tissue DNA extraction kit as per manufacturer's protocol (Qiagen[JM1]).

4.3 Primer design Primers flanking SNV positions or spanning breakpoints were designed using primer3 [15]; the list of primers used are appended as **Supplementary Table 14**. Primers were designed to produce products 140-200 nt in length for SNVs and 200-300 nt in length for breakpoints. For SNVs used as phylogenetic markers, primers were required to pass the following filters: maximum of 5 alignments to the genome as given by blat [16] for each primer sequence, maximum of 5 products produced throughout the genome as predicted by isPCR (`git://genome-source.cse.ucsc.edu/kent.git`, commit `21790480620a9bfea0e561427d17e17960ad8685`), and each primer sequence at least 30nt from the SNV position. Breakpoints used as phylogenetic markers were required to pass the following filters: maximum of 5 alignments to the genome as given by blat [16] for each primer sequence, maximum of 1 product produced throughout the genome as predicted by isPCR (`git://genome-source.cse.ucsc.edu/kent.git`, commit `21790480620a9bfea0e561427d17e17960ad8685`), and each primer sequence at least 50nt from the breakpoint position. For SNVs and breakpoints representing important biological events such as TP53 or ERBB2 breakpoints, the same filters were progressively relaxed until a primer pair could be designed to pass the relaxed set of filters.

4.4 PCR and Illumina MiSeq sequencing Targeted deep-sequencing was performed according to internal lab standard operating procedures which was previously described [17] and the respective manufacturers' specifications. Amplicons for both SNVs and breakpoints were generated from genomic DNA (or whole-genome amplified DNA, QIAgen REPLI-g Screening Kit), cleaned up using QIAamp Mini Kit (QIAgen, Germany) and quantified using the Qubit dsDNA HS Assay kit (Life Technologies, UK). The first step amplifications were performed individually (ie singleplex in 384-well format) using SYBR Select Mastermix (Life Technologies), on an ABI 7900HT qPCR machine (Center for Translational and Applied Genomics, CTAG, Vancouver, Canada & Terry Fox Laboratories, BC Cancer Research Centre, Vancouver, Canada). The amplicons were pooled according to sample and cleaned up using the ExoSAP-IT kit (USB Corp, OH, USA). For the second step amplification, each sample was barcoded with Illumina indexes (Illumina, CA, USA) using the FastStart High Fidelity PCR System (Roche, Germany) on an Applied Biosystem Veriti or Bio-Rad T100. The products ($>$ 200bp, including adaptor-additions) were then selected and cleaned up using Left Side Size Selection with 1.2 x sample volume of SPRI Select beads (Beckman Coulter, CA, USA). DNA was quantified using Qubit Broad Range on a Qubit Fluorometer (Life Technologies) and the size analyzed with an Agilent DNA1000 Chip on a Agilent 2100 Bioanalyzer (Agilent Technologies, Germany). Individual samples were pooled to form a library and each library was denatured, diluted to 4 nM and 11 pM loaded on a MiSeq (running MiSeq Control Software v2.5) (Illumina, CA, USA) at CTAG (Vancouver, Canada) according to manufacturers protocol for 2 x 151bp pair-end runs using MiSeq Reagent Kit v2. Demultiplexing was done by the onboard MiSeq Reporter (v2.5.1, Illumina).

4.5 Rearrangement breakpoint validation We removed all reads less than 100bp in length before alignment. We constructed a custom genome by creating chromosomes from the predicted rearrangement sequences. Alignments were performed using bwa (v0.7.5a) [3] using the `aln` and `sampe` commands. We removed reads with more than 5 mismatching bases. We then counted the number of reads aligned to each rearrangement using a Python script.

4.6 SNV validation We removed all reads less than 100bp in length before alignment. Reads were aligned to the hg19 reference genome downloaded from `http://www.bcgsc.ca/downloads/genomes/9606/hg19/1000genomes/bwa_ind/genome/GRCh37-lite.fa`. Alignments were performed using bwa (v0.7.5a) [3] using the `aln` and `sampe` commands. We removed reads which aligned more than 10bp away from the start or end of a target amplicon or reads with more than 5 mismatching bases. We extracted counts of the reference allele and predicted variant allele at each site. We counted a read only if the mapping quality and base quality (at target loci) was ≥ 30.

To determine the presence or absence of each allele for each SNV loci we used a binomial exact test as described in [14]. Briefly, we computed the background error for each target loci by looking at 30 bases upstream and downstream of the target loci and computing the proportion of reads with the most frequently observed non-reference base at that position. We ignored germline and somatic position in this calculation, and used the mean value across positions as the predicted background error rate. We performed the binomial exact test for each allele at each loci, deeming an allele as present if the p-value was less than 0.000001.

4.7 Clonal analysis We used PyClone 0.12.7 to perform clonal analysis. Copy number and tumour content predictions from Demix were used with the parental_copy_number for `--var_prior` and normal_variant for the `--ref_prior` flags. Counts for ref and alt alleles were extracted from the deep sequencing data as described in Section 4.6. We ran the MCMC chain for 100,000 iterations, discarding the first 50,000 as burnin. Four independent MCMC chains where run and the posterior plots visually inspected to determined convergence.

To compute the posterior distribution for each SNV cluster we used the mpear method to perform hard clustering. We then computed the posterior density for cluster c conditional on cluster assignment $\mathbf{Z}$ using the following

equation

$$p(\boldsymbol{\phi}^c|\mathbf{X},\mathbf{Z},\boldsymbol{\pi}) \quad \propto \quad \prod_{m=1}^{M} \prod_{n:Z_n=c} \sum_{\psi_m^n \in \mathcal{G}^3} p(b_m^n|d_m^n,\phi_m^c,\psi_m^n,t_m)p(\psi_m^n|\boldsymbol{\pi}_m^n)$$

where $\mathbf{X}$ is the input data and $\boldsymbol{\phi}^c$ is the cellular prevalence for cluster c. The rest of the notation is the same as that used in the original PyClone paper [18].

4.8 Clonal phylogeny analysis We used two strategies to infer clonal genotypes and phylogenies. In what follows we define a clonal genotype as the presence or absence of a set of PyClone clusters.

For patients 1, 4, 7, and 10 we treated the sample trees as clonal phylogenies under the assumption that discovery samples were pure for clones from a single lineage. For these patients we computed the probability a PyClone cluster was dominant in a leaf node (sample) as the posterior probability that the prevalence of the cluster was ≥ 0.5. We input the probability of dominance into the stochastic Dollo model described above as the probability of presence. We fixed the predicted site tree as a clone phylogeny for these patients and reconstructed the clonal genotypes of interior nodes in the phylogeny using maximum likelihood. The rational for this approach is that clusters of SNVs above 0.5 prevalence necessarily co-occur in the genotypes of clones, since mutual exclusivity would imply the sum of prevalences exceeds 1.

For patients 2, 3 and 9 we had multiple lines of evidence that at least one sample was a mixture of divergent lineages. For these patient we performed targeted single cell sequencing of 48 loci for patients 2 and 96 loci for patient 9. We then inferred clonal phylogenies and genotypes as described in section 5.7.

Both approaches yielded a predicted clone phylogeny and set of clonal genotypes for all nodes in the phylogeny. Redundancy was removed from the phylogeny of each patient by contracting branches between clones with identical genotype.

Define G_j^c to be a binary variable indicating whether the genotype of clone j contains PyClone cluster c. Further, let Z_n be a variable indicating the PyClone cluster SNV n is assigned to. To infer the prevalence of clones in samples (discovery and archival) we use the predicted values of G_j^c from the analysis above, and the values Z_n predicted from PyClone analysis of the discovery samples. Using the notation from the original PyClone paper [18] we define the following generative model. We suppress the indices for samples as these can be treated independently.

$$\begin{aligned}
\boldsymbol{f} &\sim \text{Dirichlet}(\boldsymbol{f}|\boldsymbol{\kappa}) \\
\phi^c &= \sum_{j:G_j^c=1} f_j \\
b^n|\phi^c,d^n,Z_n=c,\boldsymbol{\pi}^n\sigma,t &\sim \sum_{\boldsymbol{\psi}} \pi_{\boldsymbol{\psi}}^n \text{BetaBinomial}(b_n|d_n,\xi(\boldsymbol{\psi},\phi^c,t),\sigma)
\end{aligned}$$

The generative model can be described informally as follows. First, we generate the prevalences of the clones in a sample. Next, we compute the cellular prevalence of a mutation by summing the prevalences of all clones which contain the associated PyClone cluster in their genotype. Finally, we apply the standard PyClone likelihood model to simulate allelic count data.

To apply this model we needed to supply several parameters to compute the PyClone likelihood, which we could not easily measure in the archival samples. These include the predicted set of genotypes for a mutation and the tumour content of a sample. To estimate the tumour content for archival samples we used the variant allelic frequency of the identified TP53 mutation, as these mutations should be clonaly dominant and homozygous. For discovery samples we use the tumour content predicted from copy number analysis of the WGS data. To compute the set of predicted mutational genotypes for all samples, we first enumerate the set of possible mutational genotypes based on the copy number data from discovery samples. We then take the union set of these genotypes as the possible set of genotypes for any sample. We set the precision of the BetaBinomial distribution, σ, to 200 for all analyses.

We designed a simple Markov Chain Monte Carlo inference procedure to estimate $\boldsymbol{f}$. Specifically, we used the Metropolis-Hastings algorithm with a symmetric Dirichlet distribution with parameter 1 as a proposal distribution. We ran the sampler for 100,000 iterations, discarding the first 50,000 as burnin. We then estimate the prevalence of a clone as the mean value of the post-burnin trace.

4.9 Minimum migration analysis We used a maximum parsimony approach to evaluate all possible migration scenarios that would produce the predicted clone phylogeny and the presence of observed clones at each anatomic site. A categorical variable with levels representing anatomic sites was associated with each clonal genotype in the clone phylogeny. For clone X with direct ancestor Y, the site variable of X represents the anatomic site in which Y gained additional mutations, evolving into clone X. The site variable of the root clone represents the site in which the tumour originated. A branch between two clones with different inferred site represents a migration events. To distinguish between clones with the same genotype observed at multiple sites, we augmented the tree, adding leaf nodes representing observed genotype/site combinations. Thus for clone X observed in site P and Q, we added a X/P clone and an X/Q clone as direct descendants of X. All leaf nodes in the augmented tree were initialized with appropriate site states (X/P clone was initialized to site P). We then used standard maximum parsimony ancestral reconstruction techniques (Sankoff's algorithm), calculating the site state of internal nodes so as to minimize the total number of migrations. Where multiple solutions existed, we selected the solution that assigned the root clone to the putative primary ovary site if such a solution was amongst the multiple optimal solutions.

Presence of each clone in each site was calculated as follows. First, samples were grouped into more broad anatomic locations, for instance ROv1, 2 etc. were grouped as the ROv site. A clone was said to be present in a sample if it was assigned a clonal prevalence greater than 0.01. Parent clones are often predicted to coexist with child clones at small proportions, though based on single cell data, such a scenario is likely an artifact. Parent and child clones are usually distinguished by an absence of mutations in the parent. If the mutations that identify the child clone are estimated to have slightly lower prevalence than more ancestral mutations, that slight deviation will results in prediction of a minor population of the parent clone in the sample. On the other hand, a sample composed of 95% parent and 5% child is more likely to represent a true mixture. To account for these issues, we use an additional rule for determining whether a parent clone co-exists with its child in a sample. A predicted cellular prevalence of non-leaf clone X must be at least 50% of the combined cellular prevalence of X and its descendants to be considered present in the sample. A clone is said to be present in a site if the clone, by the above rules, is present in any of the samples of that site.

5 Single nuclei analysis

5.1 Target Selection We selected SNVs and breakpoints that were biologically relevant, phylogenetically informative, or useful for discerning between tumour cells and normal cells. Phylogetically informative SNVs were selected at random from a specific set of high confidence informative PyClone clusters. Informative PyClone clusters were those with cellular prevalences that indicated the cluster of SNVs originated in a descendant tumour lineage, or were lost in a descendant tumour lineage. Ancestral SNVs were selected from the PyClone cluster with the highest cellular

prevalence across samples. Germline heterozygous SNPs putatively lost ancestrally in the tumour were selected by searching for heterozygous SNPs contained within chromosomal segments with ubiquitous LOH across samples, for which the alternate allele had significantly fewer reads than the reference allele. Ancestral SNVs were used as markers of tumour cells, while the presence of germline heterozygous SNPs lost ancestrally were used as markers of normal cells. Breakpoints were selected manually based on the results of demix.

5.2 Primer design Primers flanking SNV positions or spanning breakpoints were designed using primer3 [15]; the list of primers used are appended as **Supplementary Table 17**. Primers were designed to produce products 185-215 nt in length. For SNVs used as phylogenetic markers, primers were required to to pass the following filters: maximum of 5 alignments to the genome as given by blat [16] for each primer sequence, maximum of 5 products produced throughout the genome as predicted by isPCR (`git://genome-source.cse.ucsc.edu/kent.git`, commit `21790480620a9bfea0e561427d17e17960ad8685`), and each primer sequence at least 30nt from the SNV position. For SNVs and breakpoints representing important biological events such as TP53 or ERBB2 breakpoints, the same filters were progressively relaxed until a primer pair could be designed to pass the relaxed set of filters.

5.3 Nuclei preparation and sorting Single nuclei were released into suspension by using a rotor-stator homogenizer (Polytron PT1000) on solid tumour cryosections in Nuclei EZ lysis buffer (Sigma-Aldrich). The resulting tumour lysates were passed through a 70-micron filter twice 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) with Trypan blue stain. Nuclei were stained with propidium iodide and single nuclei were directly flow sorted into individual wells of microtiter plates using a FACSAria II or FACSAria III sorter (BD Biosciences).

5.4 Multiplex and singleplex PCRs Multiplex (48) PCRs were performed using a Biorad C1000 Touch thermal-cycler and SYBR GreenER qPCR Supermix reagent (Life Technologies). The 48-plex reaction products from each nucleus were treated with ExoSAP-IT (Affymetrix) and used as input template to perform 48 singleplex PCRs using 48x48 Access Array IFCs according to the manufacturer's protocol (Fluidigm). Empty plate wells and wells with flow sorted FITC labelled CaliBRITE beads (BD Biosciences) were used as negative controls and 10ng gDNA aliquots were used for positive control reactions.

5.5 Nuclei-specific amplicon barcoding and nucleotide sequencing. Pooled singleplex PCR products from each nucleus were assigned unique Nextera XT molecular barcodes (Illumina) and adapted for MiSeq flow-cell NGS sequencing chemistry using a PCR step. Barcoded amplicon libraries were pooled and purified by E-Gel SizeSelect gel electrophoresis (Life Technologies). Library quality and quantitation was performed using a 2100 Bioanalyzer (Agilent Technologies) and a Qubit 2.0 Fluorometer (Life Technologies). DNA sequencing was conducted using a MiSeq sequencer according to the manufacturer's protocols (Illumina).

5.6 Bioinformatic analysis Count data for breakpoints was generated as described in Section 4.5 and for SNVs as described in Section 4.6. We determined each allele of an SNV to be present independently using the Binomial exact test with a p-value of 10^{-6} as described in Section 4.6. SNV loci with fewer than 50 reads were treated as missing. We determined a breakpoint to be present if 5 or more reads aligned to the predicted breakpoint sequence.

We used the single cell genotyper (SCG) model version 0.3.0 [19] with position specific error rates, sample specific clone prevalences and doublet modelling, to cluster the nuclei and infer clonal genotypes. Input data was provided as three states (A, B, AB) for SNV data and two states for breakpoints (presence, absence). We performed 1,000 random restarts and selected the restart with best lower bound score. We determined the number of clusters by running the model with 40 clusters. We then assigned each nuclei to the most probable cluster and discarded empty clusters. Genotypes for each cluster were predicted by taking the most probable state of the posterior distribution. Hyper-parameters for SNV error rate prior distribution were

	A	AB	B
A	99	0.5	0.5
AB	1	1	1
B	0.5	0.5	99

and for the breakpoint error rate prior distribution

	Absent	Present
Absent	9	1
Present	1	9

We set the hyper-parameter for the Beta prior distribution on doublet probabilities to (99, 1).

5.7 Clone phylogeny analysis We combined our novel phylogenetic algorithm with the genotype clustering results from the SCG model to infer the clone phylogeny of the tumour. To ensure that we were working with accurately predicted clonal genotypes, we focused on clones with 10 or more nuclei assigned to them by the single cell genotyping model. We removed SNV events which were missing in $\geq 80\%$ of cells. Naive application of the stochastic Dollo model to the output of the SCG model was not possible. SNVs selected for single nucleus sequencing where not an unbiased representation of events across the genome. As a result, the maximum likelihood tree based on this data would infer loss of SNVs which were not predicted from the sample phylogeny analysis and not corroborated by copy number changes. To address the issue of bias we collapsed the somatic SNVs into the groups predicted from the PyClone analysis. For each PyClone cluster of SNVs, we determined that the cluster was present in the clonal genotype if any of the SNVs were predicted to be present. Specifically, for each SNV we computed the probability that the mutant allele was predicted to be present for the genotype by summing the posterior probabilities of the AB and B genotypes states from the SCG model. If this value was less than 0.1 we deemed the SNV absent in the genotype, and if it was greater than 0.9 we deemed the SNV present. All other values were treated as missing. We then deemed a PyClone cluster present in the genotype if any SNV was present in the genotype. This provided a binary representation genotypes in terms of presence/absence of PyClone clusters which could be used for phylogenetic inference. In order to correct for sampling bias we assigned all predicted SNVs in the genome (from WGS analysis) to one of the PyClone clusters inferred from the targeted deep sequencing data. To do this we computed the mean prevalence of each cluster at each site, $\bar{\phi}^c$. We then computed the posterior probability that a mutation belonged to cluster c according to the following equation

$$p(Z_n = c|\mathbf{X}, \bar{\phi}^c) \quad \propto \quad \prod_{m=1}^{M} \sum_{\psi_m^n \in \mathcal{G}^3} p(b_m^n|d_m^n, \bar{\phi}_m^c, \psi_m^n, t_m) p(\psi_m^n|\pi_m^n)$$

where Z_n is a categorical variable indicating the cluster membership of SVN n. We assigned each SNV to the most probable cluster and computed the number of SNVs in each cluster. The number of SNVs was then used to weight the likelihood computation in the stochastic Dollo for the SNV cluster. This had the effect of making trees which supported the loss of smaller SNVs clusters to become more likely than those which supported the loss of larger clusters. We then annotated the status of germline SNPs and breakpoints onto the inferred tree manually post-hoc.

6 Infinite sites with loss model

6.1 Introduction The number of mutations in a human cancer can reach tens of thousands [20]. Given that the size of the human genome itself is 3 billion nucleotides, the probability a specific nucleotide will be mutated is small, and

the probability that the same nucleotide will be independently mutated twice in the evolutionary history of a tumour is exceptionally small. Thus it is reasonable to assume that each nucleotides is mutated at most once throughout the evolutionary history of a tumour (the *infinite sites* assumption [21]). In genomically unstable cancers, deletion of large chromosomal segments is common due to errors in segregation during mitosis. Deletions have the side effect of deleting large numbers of mutations resident on deleted chromosomal segments. Furthermore, large deletions on several branches of a tree can span a shared locus, and thus a given mutation may be deleted independently multiple times. Thus we model somatic mutation as a process governed by single origins and multiple losses. Our model can be seen as a simplified version of the stochastic Dollo process [22,23].

Somatic mutation data is overwhelmingly binary (as a corollary of the infinite site assumption) in nature, with two states encoding presence and absence of a variant allele. Phylogenetic relationships between tumour clones are inferred by patterns of observed mutation presence/absence across related clones. An absence in a specific clone has two evolutionary explanations: 1) the mutation never arose is any of the clones ancestors, 2) the mutation arose in an ancestor and was deleted in a subsequent ancestor. Current sequencing technologies do not provide perfect observation of clonal genotypes. Observation of an absence in sequencing data has an additional explanation: the mutation is present in the sequenced clone(s) but was not detected due to under-sampling of the mutation sequence. Furthermore, observation of a presence in sequencing data could be the result of sequencing error. To overcome these challenges, we combine the assumptions of an infinite sites model of somatic evolution with an emission model to capture the measurement uncertainty present in tumour sequencing data.

We propose an emission model for bulk whole genome sequencing (WGS) of tumour samples. The WGS emission model incorporates three factors affecting mutation detection: tumour purity, allele specific copy number, and sequencing error. Additionally, we propose an algorithm for calculation of the maximum likelihood phylogenetic tree. Phylogenetic tree inference can be easily combined with the WGS emission model, or in fact any other model that provides a likelihood of presence absence per mutation per sequencing dataset.

6.2 Bulk whole genome sequencing emission model

6.2.1 Parameter definitions

WGS emission model parameters and variables

parameter	description
s	Haploid normal coverage
t	Haploid tumour coverage
c_m	Major tumour copy number at variant site
c_n	Minor tumour copy number at variant site
c_t	Total tumour copy number at variant site
c_x	Number of copies of the variant
n_x	Detected number of tumour reads with the variant
n_t	Total number of reads at the variant site
e_s	Sequencing error rate
z	Indicator for variant presence (z=1 $\iff$ c_x>0)

Nature Genetics: doi:10.1038/ng.3573

6.2.2 Likelihood of a single mutation

The expected ratio of mutant reads depends on the number of copies of the mutant, and the haploid coverage of the tumour and normal (Equation 14).

$$r = \begin{cases} \frac{c_x t}{2s + c_t t} & c_x > 0 \\ e_s & c_x = 0 \end{cases} \tag{14}$$

The likelihood of observing n_x mutant reads is distributed as a binomial given the expected mutant ratio (Equation 15).

$$P(n_x | n_t, s, t, c_x) = \text{Binomial}(n_x | n_t, r) \tag{15}$$

To obtain a likelihood given presence, we marginalize across positive copies of the variant (Equation 16). To obtain a likelihood given absence, we condition on zero copies of the variant (Equation 17). In both cases we assume a non-informative prior for the number of variant copies (Equation 18).

$$P(n_x | n_t, s, t, c_m, z{=}1) = \sum_{c=1}^{c_m} P(n_x | n_t, s, t, c_x{=}c) P(c_x{=}c) \tag{16}$$

$$P(n_x | n_t, s, t, z{=}0) = P(n_x | n_t, s, t, c_x{=}0) \tag{17}$$

$$P(c_x) = \frac{\mathbb{I}(c_x \in \{1, \ldots, c_m\})}{c_m} \tag{18}$$

6.3 Infinite sites model of somatic evolution

Nature Genetics: doi:10.1038/ng.3573

6.3.1 Parameter definitions

Evolutionary model parameters and variables

$V(T)$	Vertices of tree T
$L(T)$	Leaves of tree T
$D(i)$	Nodes descendent from node i
$L(i)$	Leaves descendent from node i
$C(i)$	Children of node i
$p(i)$	Parent of node i
$A(i)$	Ancestors of node i
w	Tree node at which the variant originated
z_i	Indicator for variant presence at node i
π_l	Probability of losing a variant
$\ell(z_i\|\cdot)$	Likelihood of variant presence

6.3.2 Likelihood of mutational profile given a fixed phylogeny

The single origin constraint adds dependencies between branches in the tree: if a mutation originated in one branch it cannot also originate independently in another branch. Conditioned on the originating branch ($p(w)$, w), the losses are Markovian on the sub-tree $D(w)$. Herein, we describe a mutation as originating at a specific node, which is equivalent to a mutation originating on the branch from that nodes direct ancestor. Furthermore, a mutation described as lost at a node is equivalent to that mutation being lost on the branch from that nodes direct ancestor.

Pick a node w as the node at which a mutation originated. Given w, we can efficiently calculate the likelihood of the mutation as the product of two terms: the likelihood of the mutation being absent at the leaves not descendent from w, and $Q(j{=}w, T)$, the likelihood of the presence/absences marginalizing all possible combinations of losses in the sub-tree rooted at w (Equation 19). The term $Q(j, T)$ can be efficiently calculated using dynamic programming [24] (Equation 21). The marginal likelihood of a mutation given T can be calculated by marginalizing w (Equation 22) with a non-informative prior over w (Equation 23). The likelihood of the full set of mutations can be then be calculated as a product of the likelihoods of individual mutations (Equation 24)

Nature Genetics: doi:10.1038/ng.3573

$$P(x|T,w) = Q(j{=}w,T) \prod_{i \in L(T) \setminus L(w)} \ell(z_i{=}0|\cdot) \tag{19}$$

$$Q(j,T) = \begin{cases} \pi_l \ell(z_j{=}0|\cdot) + (1-\pi_l)\,\ell(z_j{=}1|\cdot) & \text{if } j \in L(T) \\ \pi_l \prod_{i \in L(j)} \ell(z_i{=}0|\cdot) + (1-\pi_l) \prod_{i \in C(j)} Q(i,T) & \text{if } j \notin L(T) \end{cases} \tag{20}$$

$$\tag{21}$$

$$P(x|T) = \sum_{w \in V(T)} P(x|T,w)P(w) \tag{22}$$

$$P(w) = \frac{1}{|V(T)|} \tag{23}$$

$$P(X|T) = \prod_{x \in X} P(x|T) \tag{24}$$

6.3.3 Empirical Bayesian tree inference

We take an empirical Bayesian approach to tree inference, and infer a single best tree by maximizing the posterior probability of the data X over the space of all rooted full binary trees $\mathcal{T}$ (Equations 25-28).

$$P(X|T) = \prod_{x \in X} P(x|T) \tag{25}$$

$$P(T|X) = \frac{P(X|T)P(T)}{\sum_{T'} P(X|T')P(T')} \tag{26}$$

$$T^{\text{opt}} = \text{argmax}_T\, P(T|X) \tag{27}$$

$$P(T) = \frac{1}{|\mathcal{T}|} \tag{28}$$

Rooted trees are appropriate in the context of cancer, since the root has specific meaning as a normal genome free of all somatic mutations. Furthermore, full binary trees are sufficient, since the maximum posterior tree will be a full binary tree for almost all realistic datasets. A tree with higher branching factor will have fewer internal nodes, and fewer degrees of freedom, allowing for a poorer fit to the data for datasets of sufficiently realistic complexity. A tree for which some nodes have out degree 1 will be unidentifiable from the same tree with that node removed.

The proposed application of the method involves a limited number of samples, making exhaustive iteration possible. Trees are enumerated using existing methods [25]. Furthermore, trees can be scored in parallel, and thus reasonable run times are possible for trees with 8 or fewer samples.The number of trees increases factorially with the number of leaf nodes. Thus, beyond 8 samples stochastic annealing or heuristic search methods become necessary.

Nature Genetics: doi:10.1038/ng.3573

Number of full binary trees for a given number of samples

number of samples	number of trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10395
8	135135

6.3.4 Maximum posteriori estimates of origin, presence and loss

Conditioned on the tree topology T^{opt}, we can calculate the maximum posteriori (MAP) estimate of a mutation's origin node, loss in descendant branches, and resulting presence/absence pattern in descendant nodes. Specifically, we would like to maximize the joint posterior probability of the origin w and presences z given by Equation 29. We can do this by maximizing instead of marginalizing during the recursion (Equation 30), and keeping track of which choice of z_j maximizes $Q(j,T)$ at each step. We then take the origin w, and corresponding z_j that maximizes Equation 29.

$$P(w,z|x,T^{\text{opt}}) \propto P(x|w,z,T^{\text{opt}})P(z|w)P(w) \tag{29}$$

$$Q^*(j,T) = \begin{cases} \max\left(\pi_l \ell(z_j{=}0|\cdot), (1-\pi_l)\,\ell(z_j{=}1|\cdot)\right) & \text{if } j \in L(T) \\ \max\left(\pi_l \prod_{i \in L(j)} \ell(z_i{=}0|\cdot), (1-\pi_l) \prod_{i \in C(j)} Q^*(i,T)\right) & \text{if } j \notin L(T) \end{cases} \tag{30}$$

6.4 Maximum parsimony copy number inference In genomically unstable cancers, mutations can be lost as a result of ancestral deletion of chromosomal segments. Some of these ancestral deletion events will be identifiable given the copy number profiles of extant tumour clones. Calculation of ancestral deletion events will be useful for corroborating evidence for mutations predicted as lost. Future algorithms may also benefit from using inferred ancestral deletions to inform phylogenetic inference.

We use maximum parsimony to infer ancestral deletions based on allele specific copy number profiles. We assume as given a joint segmentation of tumour samples, with allele specific copy number predictions for each segment. Furthermore, we assume that segment alleles are consistently labelled across samples. This is important, since 2 major copies and 1 minor copies in one sample may be different from 2 major copies and 1 minor copies in a related sample if a different allele has been amplified to 2 copies.

Nature Genetics: doi:10.1038/ng.3573

Parameters and variables for ancestral copy number inference

parameter	description
$a_{n\ell i}$	CN of segment n, allele $\ell \in \{1, 2\}$, node $i \in V$
$E(T)$	Edges of tree T
$C(i)$	Children of node i
r	Normal root node
l_n	Length of segment n

Our objective is to infer the unobserved ancestral copy number that minimizes the length weighted sum of copy number changes throughout the tree T (Equation 32). Unlike for mutations, we explicitly instantiate the normal genome with one copy per allele for all segments. The normal genome is placed at the root of the tree, with the root of the (full binary) somatic tree as its only child. Let $t(k, k', l)$ denote the cost of a segment of length l transitioning from copy number k to copy number k', defined as given in Equation 31. Note that $k = 0$ is an absorbing state since a segment-allele that reaches 0 is permanently removed from the genome and cannot be reacquired.

$$t(k, k', l) = \begin{cases} \infty & \text{if } k = 0 \\ l|k - k'| & \text{else} \end{cases} \tag{31}$$

$$a_{\text{opt}} = \underset{a}{\text{argmin}} \sum_n \sum_\ell \sum_{(i,j) \in E(T)} t(a_{n\ell i}, a_{n\ell j}, l_n) \tag{32}$$

The optimal score (length weighted CN changes) can be calculated efficiently using a using Dynamic programming as given by Equation 33. Unobserved copy number states can be inferred by recording child copy number states that maximize sub-tree scores, and backtracking from the root copy number that results in the optimal score.

$$S(i, k, l, a) = \begin{cases} a_i & \text{if } i \in L(T) \\ \sum_{j \in C(i)} \min_{k' \in \{0..K\}} [t(k, k', l) + S(j, k', l, a)] & \text{if } i \notin L(T) \end{cases}$$

$$S_{\text{opt}} = \sum_n \sum_\ell S(r, 1, l_n, a_{n\ell}) \tag{33}$$

It is often the case that multiple solutions for the unobserved copy number will be equivalent under the given scoring function. Potentially only one of these solutions will be corroborated by mutation loss. To increase the potential for corroborating mutation loss with ancestral deletions, we enumerate all potential deletions that result from equivalent solutions by backtracking all optimal paths and annotating any decrease in copy number between parent and child.

Nature Genetics: doi:10.1038/ng.3573

7 Divergent CCNE1 validation for patient 2

Patient 2 was characterized by a highly divergent tree topology between omentum and right ovary samples, including divergent sets of breakpoints surrounding the CCNE1 locus on chromosome 19. We sought additional confirmation of the predicted phylogeny and presence absence patterns of breakpoints. For both breakpoints and SNVs, the optimal tree (supporting divergence) is considerably more likely than any other tree, with log odds 297 and 3680 respectively (**Supplementary Fig. 5**, and **Supplementary Fig. 2**).

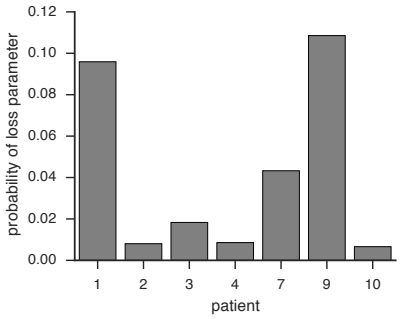
We then sought to confirm that counts of reads supporting breakpoints could be used to accurately determine presence or absence of those breakpoints in each sample. Determining the absence of breakpoints by deep sequencing was confounded by the observation that all samples, including the normal control, contained reads supporting putative somatic breakpoints. We used a 2 component Gaussian Mixture Model to fit the log read counts of deep sequenced breakpoints across all samples (**Supplementary Fig. 45**). Breakpoints with 0 read count were excluded from the GMM fit and were post-hoc added to the component with lower mean log read count. For all but 2 breakpoints, read counts in the normal were classified into the component with lower mean log read count, confirming this component as most likely the result of either background contamination by circulating tumour DNA or clones at very low prevalence (**Supplementary Fig. 46**). Assuming the component with higher mean represents a set of breakpoints present in the dominant clone, and the component with lower mean represents a set of breakpoints absent or present in a very minor clone, our type I and type II error rates for determining presence and absence of breakpoints by a threshold of > 1 supporting WGS reads is 3% and 7% respectively. Reanalysis of the WGS breakpoint data using these error rates had no effect on tree topology or placement of breakpoint origin and loss within the tree.

Next we defined the CCNE1 amplified region as the contiguous region of chromosome 19 containing CCNE1 for which all segments within that region were predicted to have copy number 8 or greater in at least one WGS sample. For each breakpoint with a break end within the CCNE1 amplified region, we calculated the posterior probability that the breakpoint originated on the ancestral branch given the observed WGS read counts. For 28 of the 28 breakpoints, the posterior probability of originating on the ancestral branch is less than 0.001, confirming site specific acquisition. For 19 of the 28 breakpoints, one or both break ends could be unambiguously assigned to the start or end of a segment (**Supplementary Tables 13** and **7**). Conditions for assignment were that the associated segment be at least 2000 nt in length, and no more than 100 nt from the break end. For the resulting set of 24 break ends, we calculated the difference in raw copy number between segments on either side of that break end in each sample. Break ends predicted as present in a sample were associated with significantly higher copy number differences in that sample when compared to break ends predicted as absent (p-value $= 2.8 \times 10^{-7}$, Mann-Whitney U test), confirming that site specific breakpoints induced site specific copy number changes.

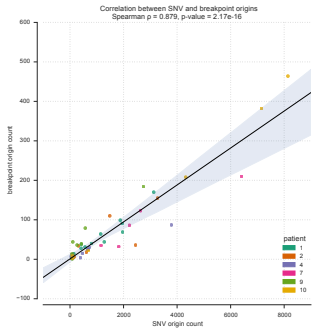
Nature Genetics: doi:10.1038/ng.3573

Supplementary Figures

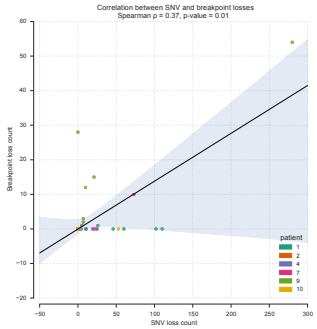
Supplementary Figure 1 Supplementary analysis of sample phylogenies



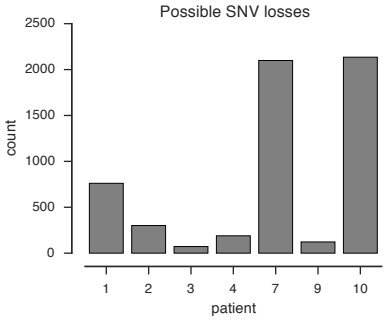
(a) Inferred probability of an SNV being lost on a given branch for each patient.



(b) Correlation between SNV and breakpoint origin count excluding patient 3.

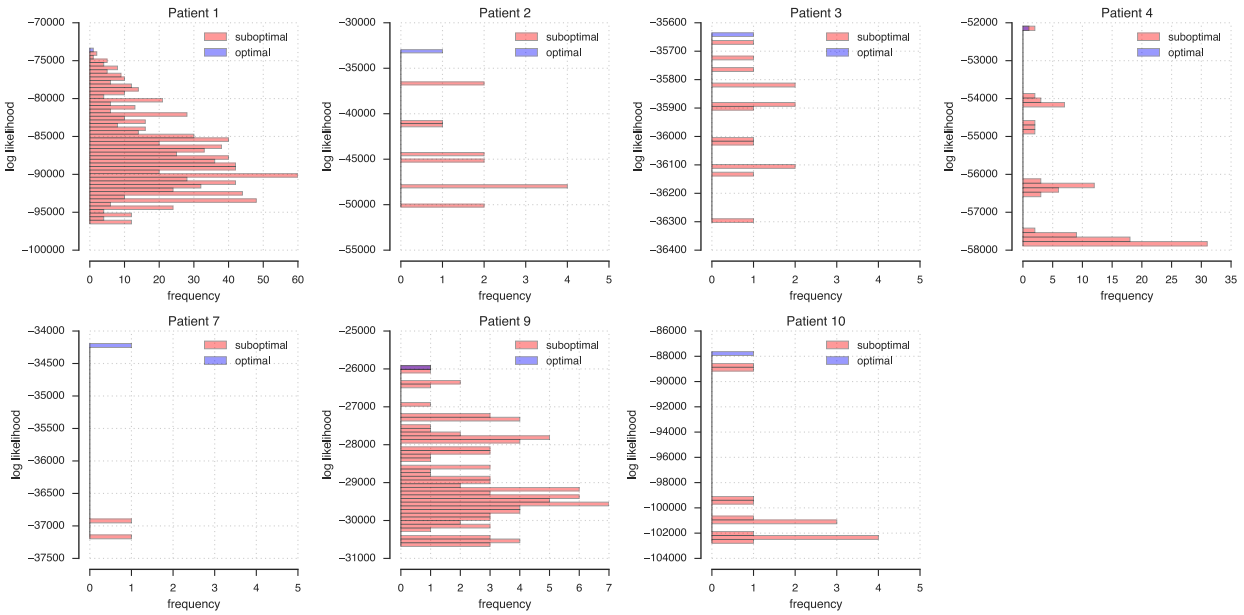


(c) Correlation between SNV and breakpoint loss count excluding patient 3.

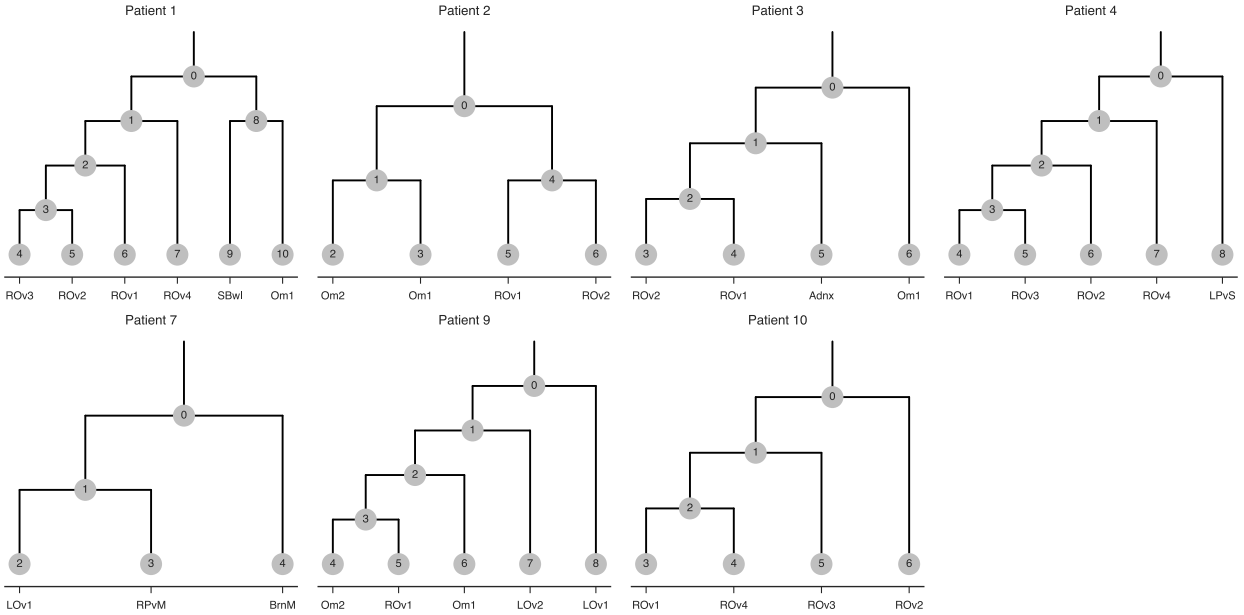


(d) Number of SNVs that are statistically unidentifiable as lost, per patient, within the context of our model of SNV evolution.

Nature Genetics: doi:10.1038/ng.3573

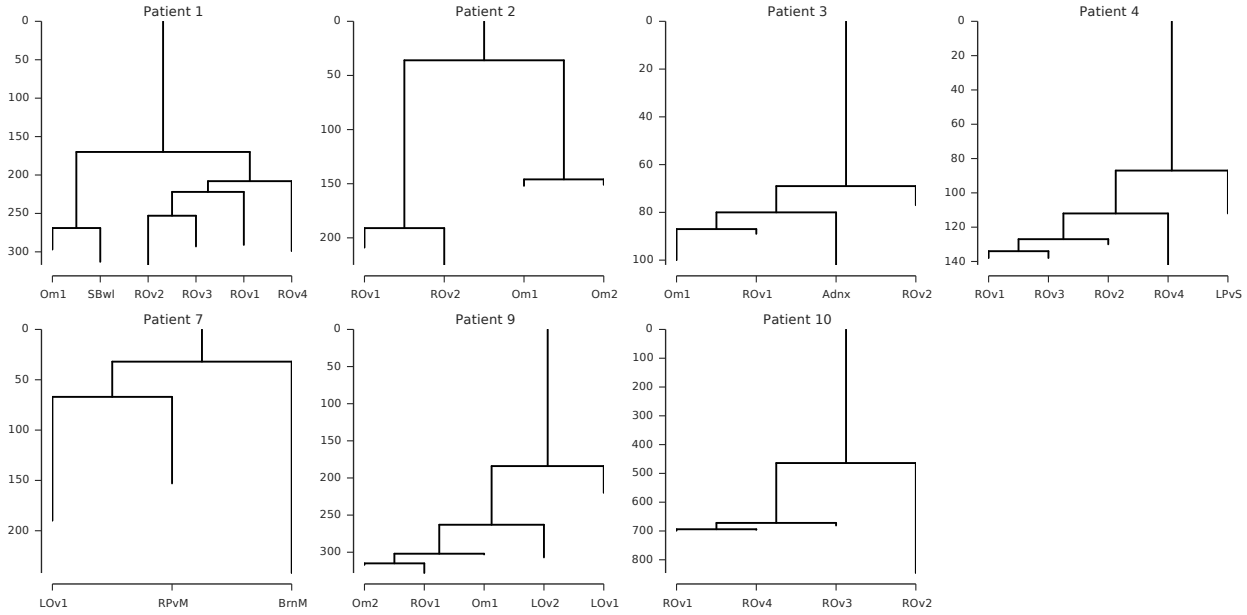


Supplementary Figure 2 Shown is the frequency (x-axis) of binned log likelihoods (y-axis) for all possible trees relating WGS samples by SNVs detected in those samples. The optimal tree log likelihood is shown in blue, suboptimal in red.



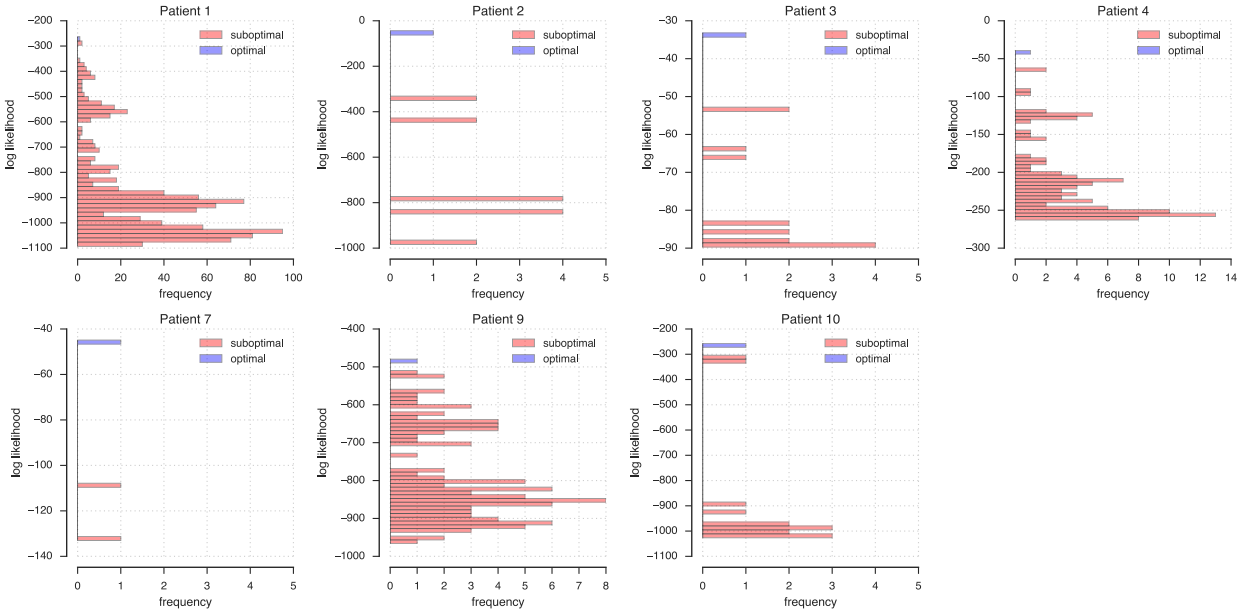
Supplementary Figure 3 SNV phylogenies of 7 HGSOvCa patients with nodes labeled.

Nature Genetics: doi:10.1038/ng.3573

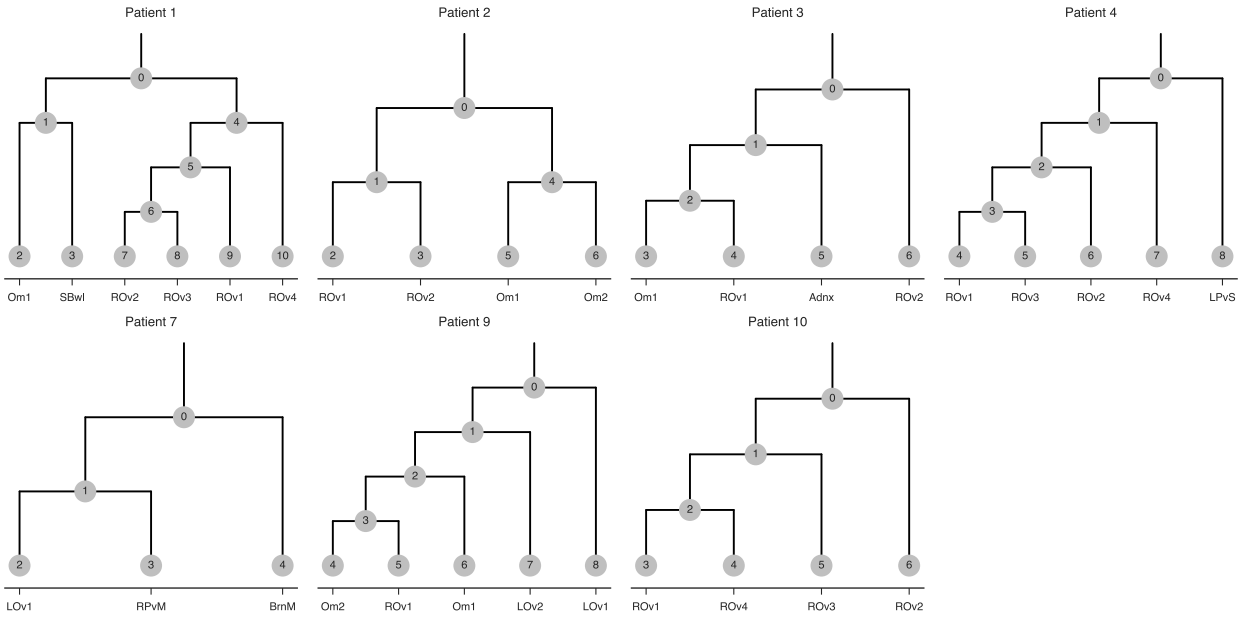


Supplementary Figure 4 Rearrangement breakpoint phylogeny of 7 HGSOvCa patients. Anatomic sites sampled for whole genome sequencing in 7 HGSOvCa patients. Phylogeny inferred from rearrangement breakpoints predicted with destruct. Branch lengths represent counts of the number of breakpoints originating on each branch. Branches are annotated with the number of breakpoints lost along the branch.

Nature Genetics: doi:10.1038/ng.3573

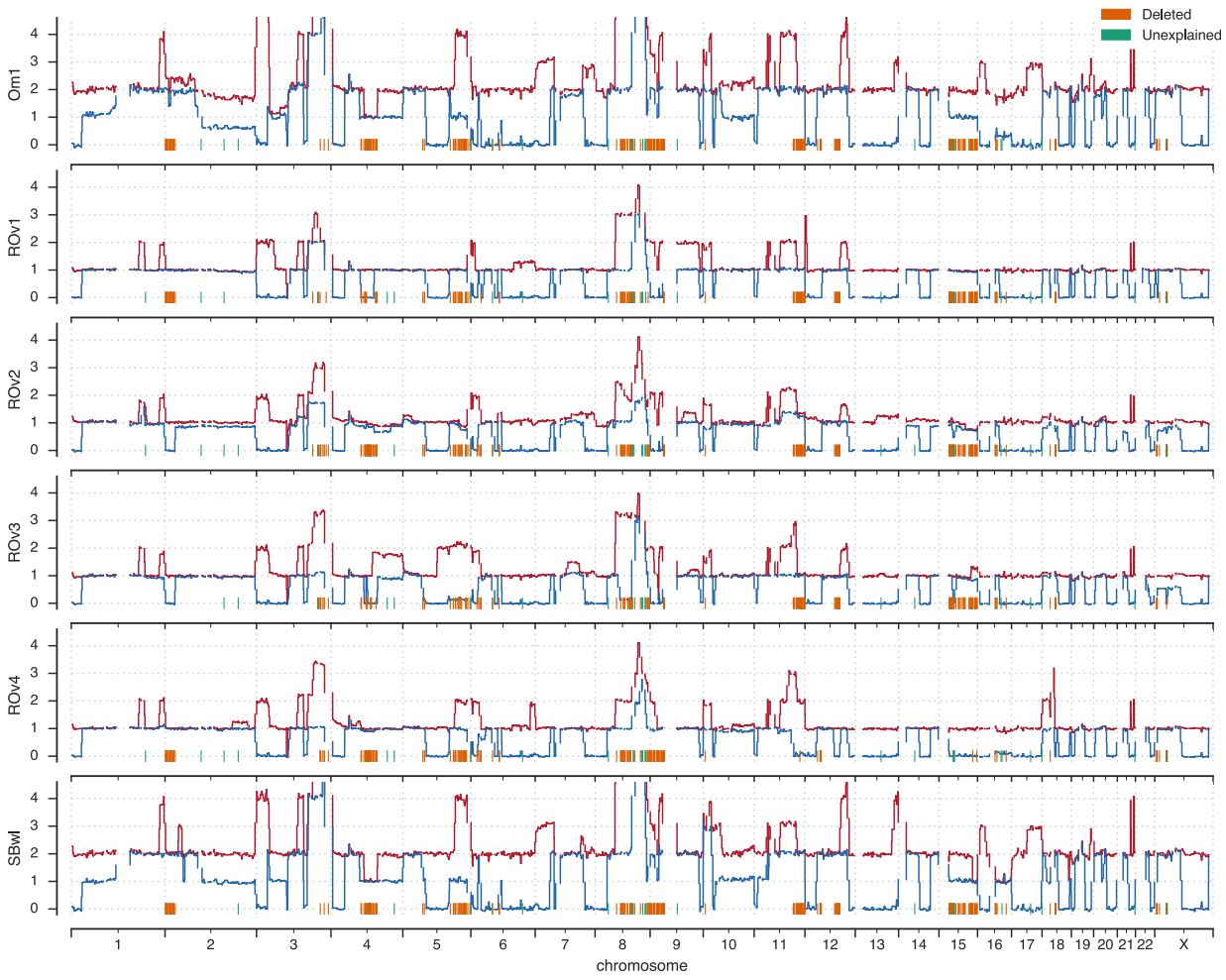


Supplementary Figure 5 Shown is the frequency (x-axis) of binned log likelihoods (y-axis) for all possible trees relating WGS samples by breakpoints detected in those samples. The optimal tree log likelihood is shown in blue, suboptimal in red.



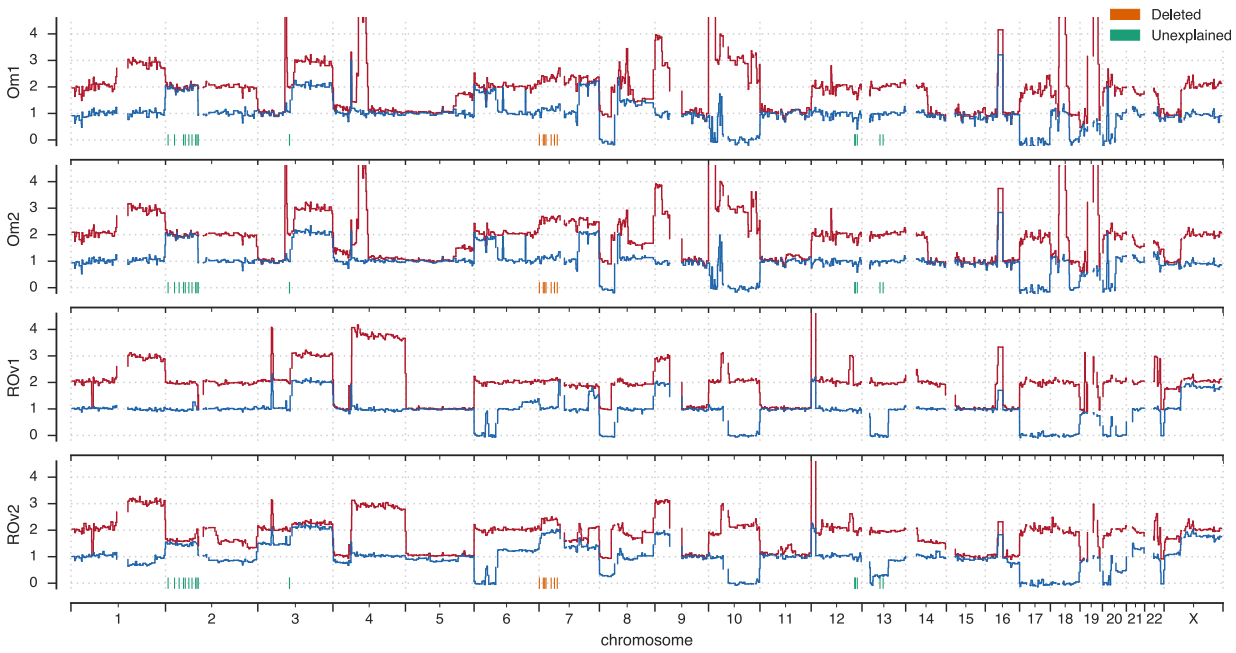
Supplementary Figure 6 Rearrangement breakpoint phylogenies of 7 HGSOvCa patients with nodes labeled.

Nature Genetics: doi:10.1038/ng.3573



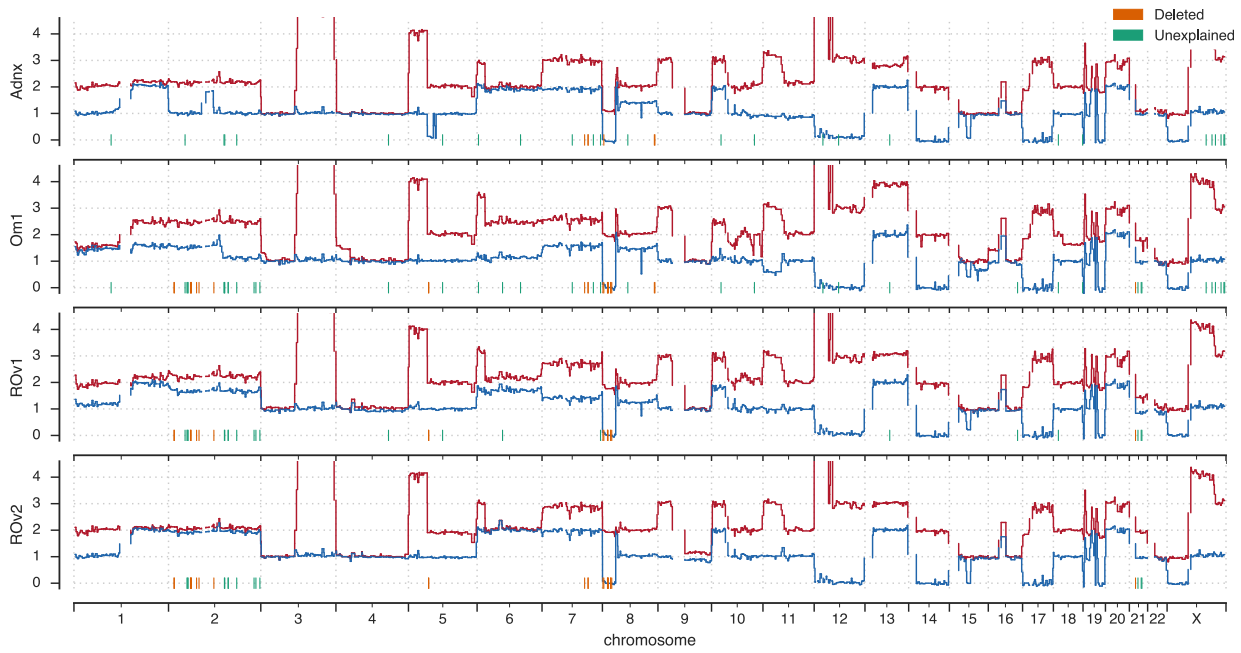
Supplementary Figure 7 Allele specific copy number profile, predicted by ReMixT, of the whole genome for patient 1. The minor allele (arbitrarily assigned) is represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Tick marks on the x-axis show positions of SNVs predicted as lost, coloured by whether the loss is corroborated by inferred ancestral deletion. In all samples chromosome 17 exhibits only a single allele, indicating a complete loss of heterozygosity.

Nature Genetics: doi:10.1038/ng.3573



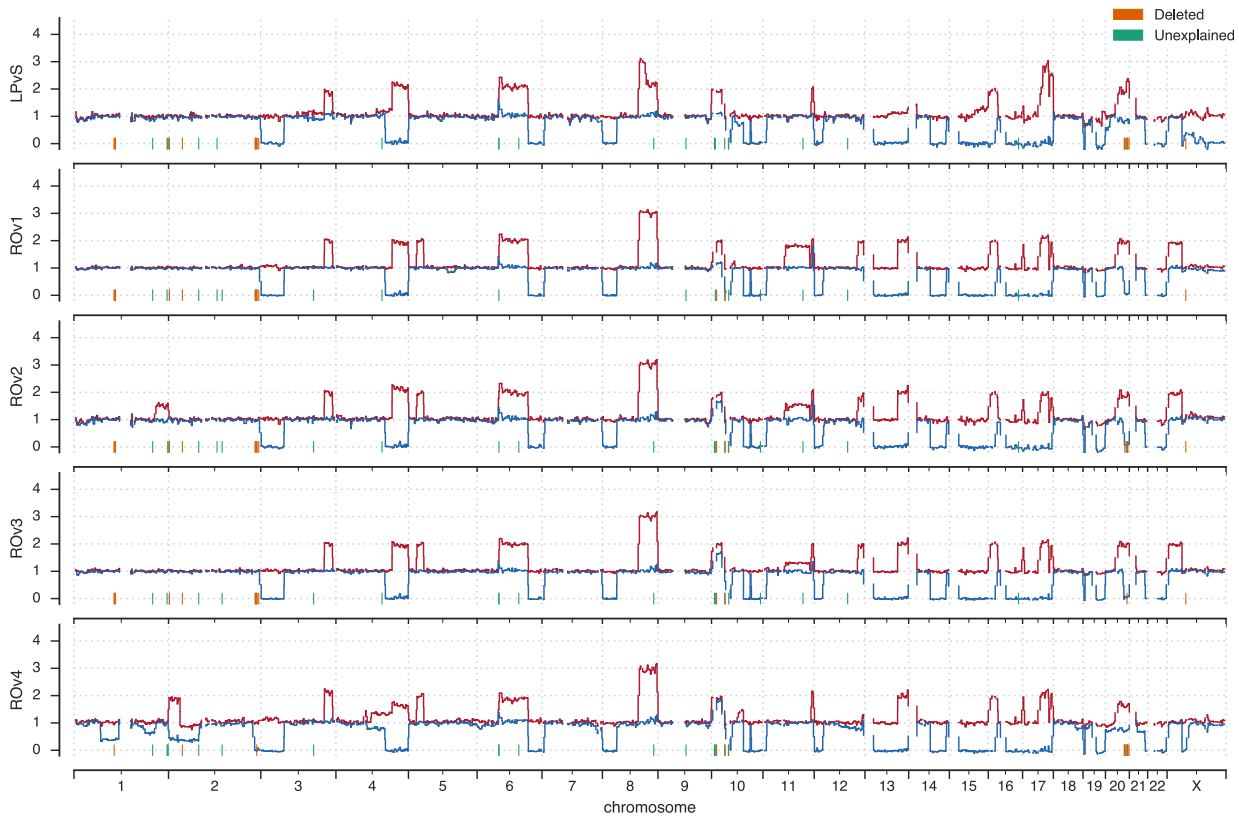
Supplementary Figure 8 Allele specific copy number profile, predicted by ReMixT, of the whole genome for patient 2. The minor allele (arbitrarily assigned) is represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Tick marks on the x-axis show positions of SNVs predicted as lost, coloured by whether the loss is corroborated by inferred ancestral deletion. In all samples chromosome 17 exhibits only a single allele, indicating a complete loss of heterozygosity.

Nature Genetics: doi:10.1038/ng.3573



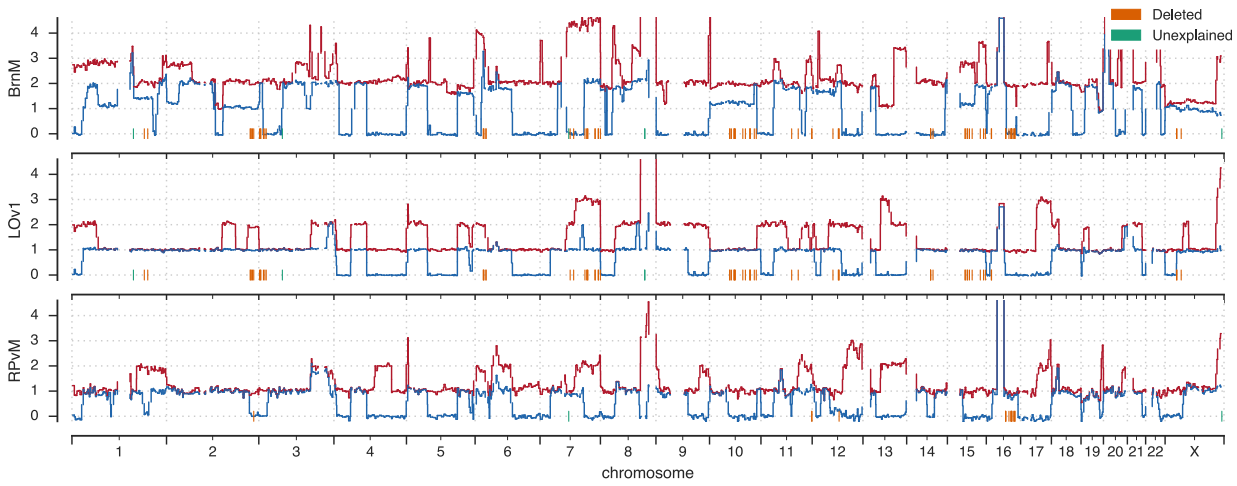
Supplementary Figure 9 Allele specific copy number profile, predicted by ReMixT, of the whole genome for patient 3. The minor allele (arbitrarily assigned) is represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Tick marks on the x-axis show positions of SNVs predicted as lost, coloured by whether the loss is corroborated by inferred ancestral deletion. In all samples chromosome 17 exhibits only a single allele, indicating a complete loss of heterozygosity.

Nature Genetics: doi:10.1038/ng.3573



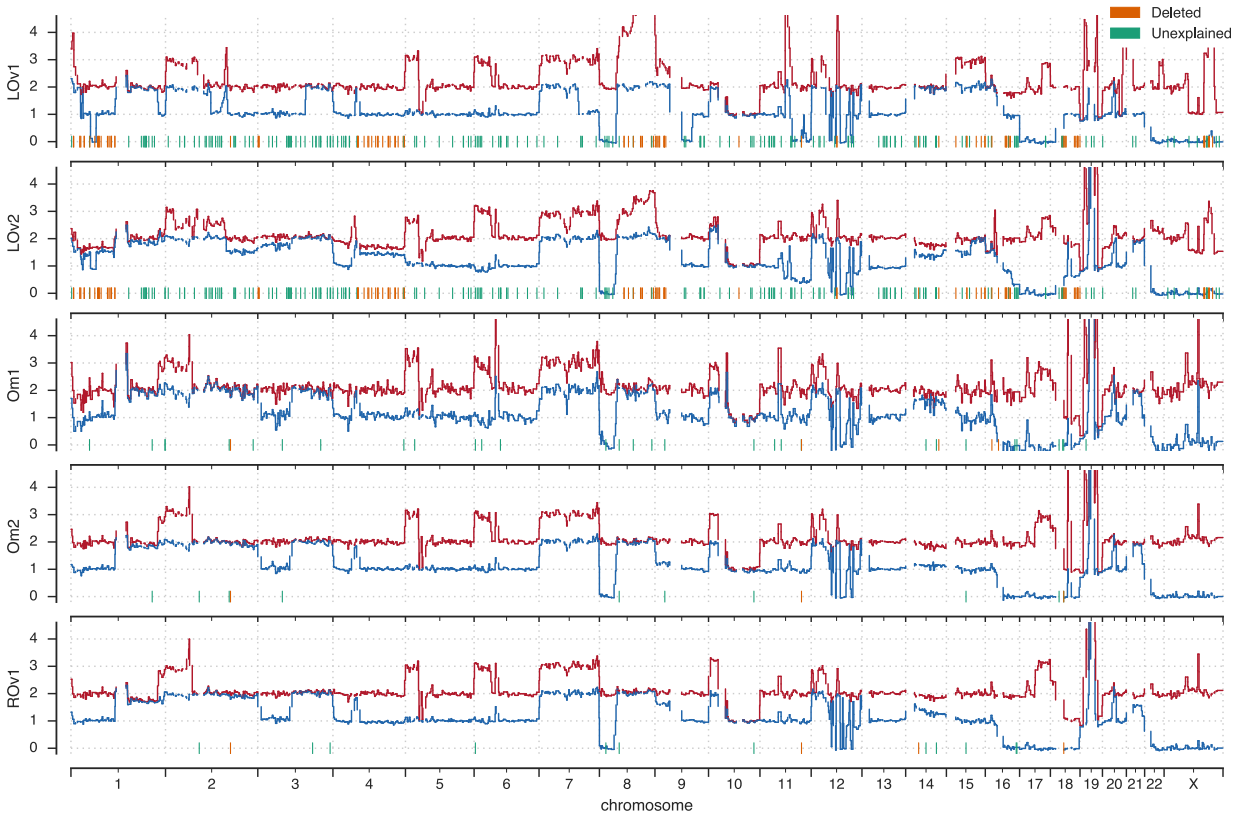
Supplementary Figure 10 Allele specific copy number profile, predicted by ReMixT, of the whole genome for patient 4. The minor allele (arbitrarily assigned) is represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Tick marks on the x-axis show positions of SNVs predicted as lost, coloured by whether the loss is corroborated by inferred ancestral deletion. In all samples chromosome 17 exhibits only a single allele, indicating a complete loss of heterozygosity.

Nature Genetics: doi:10.1038/ng.3573



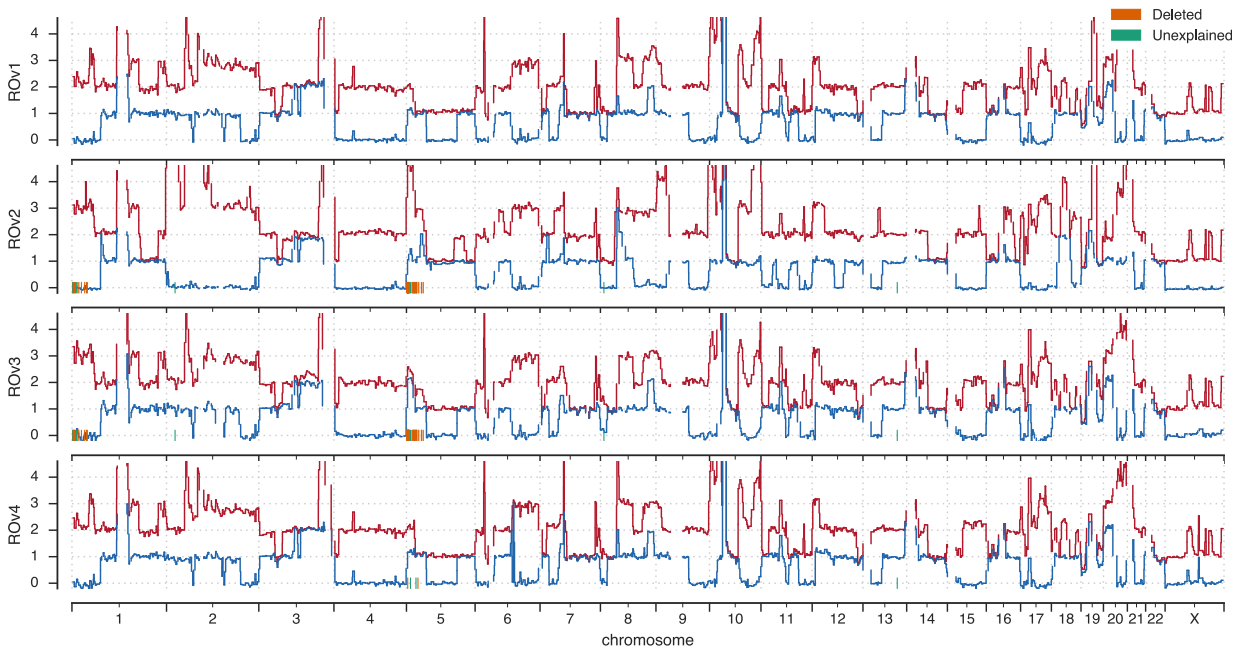
Supplementary Figure 11 Allele specific copy number profile, predicted by ReMixT, of the whole genome for patient 7. The minor allele (arbitrarily assigned) is represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Tick marks on the x-axis show positions of SNVs predicted as lost, coloured by whether the loss is corroborated by inferred ancestral deletion. In all samples chromosome 17 exhibits only a single allele, indicating a complete loss of heterozygosity.

Nature Genetics: doi:10.1038/ng.3573



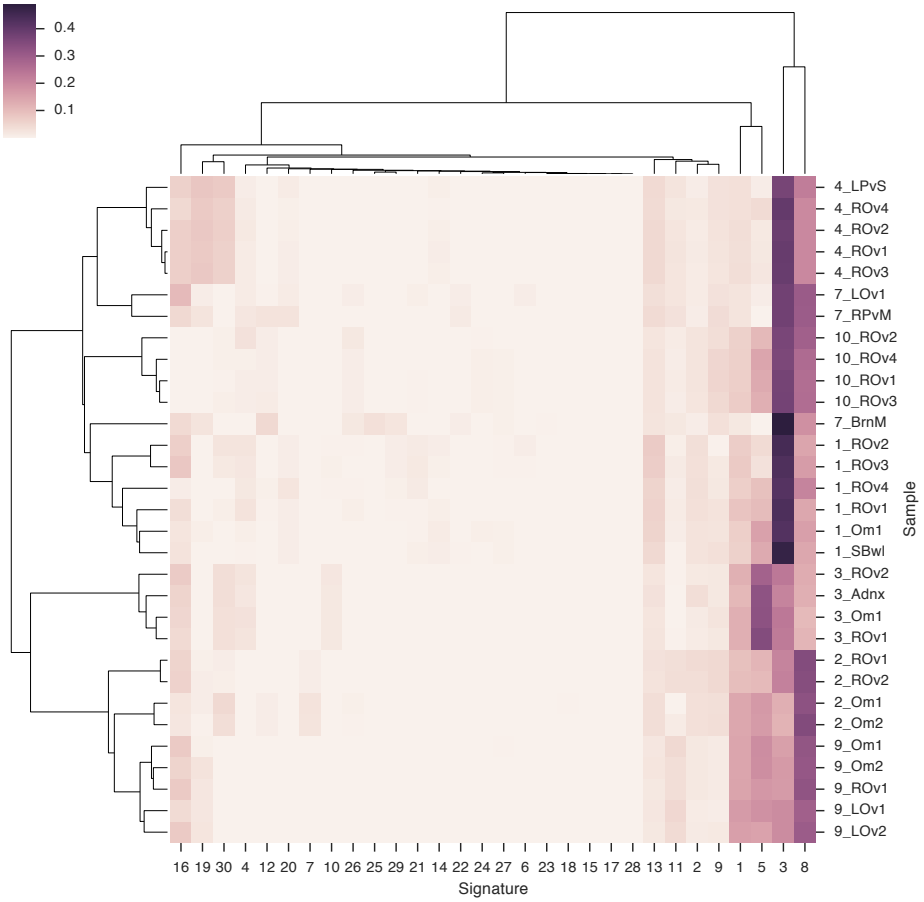
Supplementary Figure 12 Allele specific copy number profile, predicted by ReMixT, of the whole genome for patient 9. The minor allele (arbitrarily assigned) is represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Tick marks on the x-axis show positions of SNVs predicted as lost, coloured by whether the loss is corroborated by inferred ancestral deletion. In all samples chromosome 17 exhibits only a single allele, indicating a complete loss of heterozygosity.

Nature Genetics: doi:10.1038/ng.3573



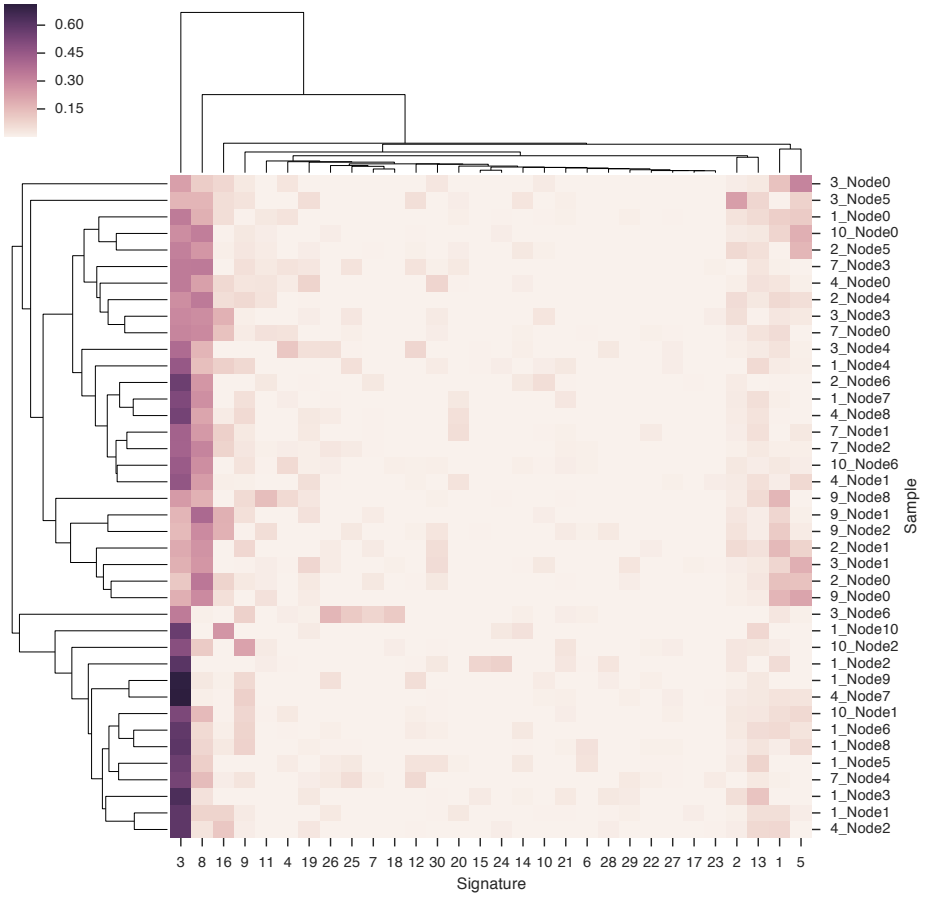
Supplementary Figure 13 Allele specific copy number profile, predicted by ReMixT, of the whole genome for patient 10. The minor allele (arbitrarily assigned) is represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Tick marks on the x-axis show positions of SNVs predicted as lost, coloured by whether the loss is corroborated by inferred ancestral deletion. In all samples chromosome 17 exhibits only a single allele, indicating a complete loss of heterozygosity.

Nature Genetics: doi:10.1038/ng.3573



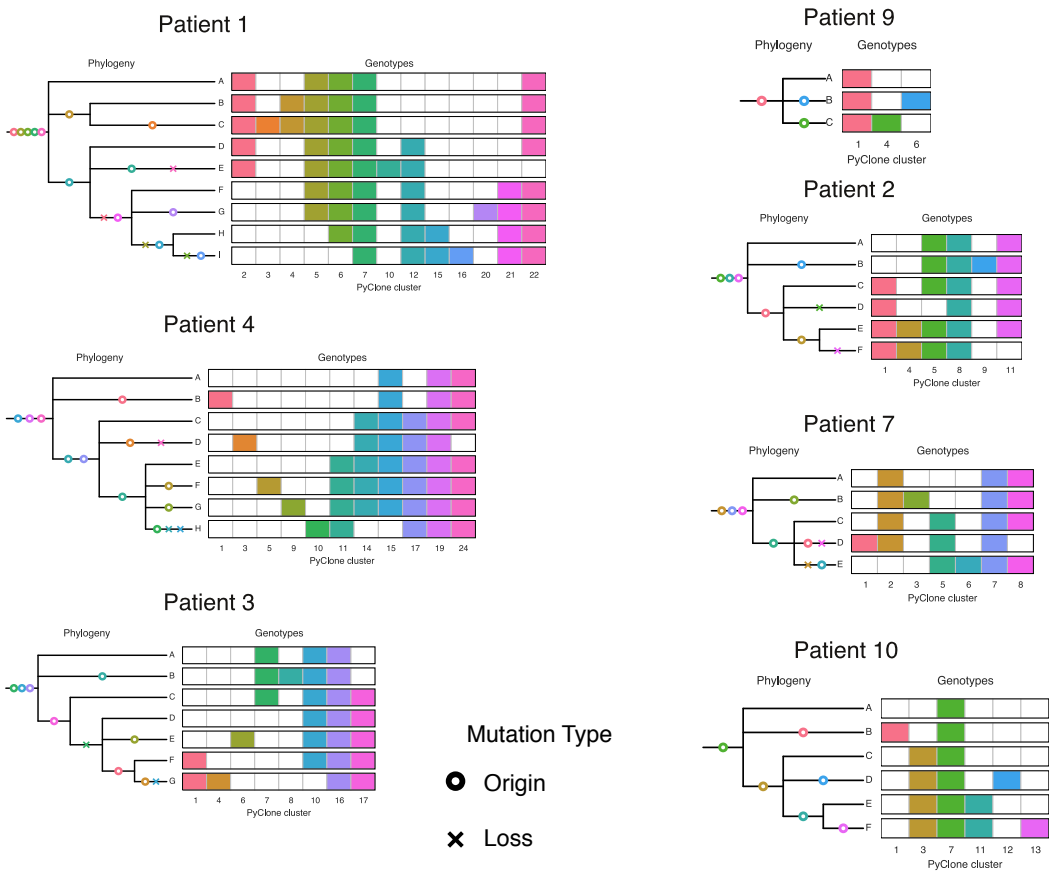
Supplementary Figure 14 Heatmap of proportions of each signature in each sample. Rows are labeled with the patient id and sample id separated by an underscore. Columns are labeled with the number of the curated cosmic signature. Intensity represents the proportion of each signature that generated the SNVs in each sample of each patient.

Nature Genetics: doi:10.1038/ng.3573



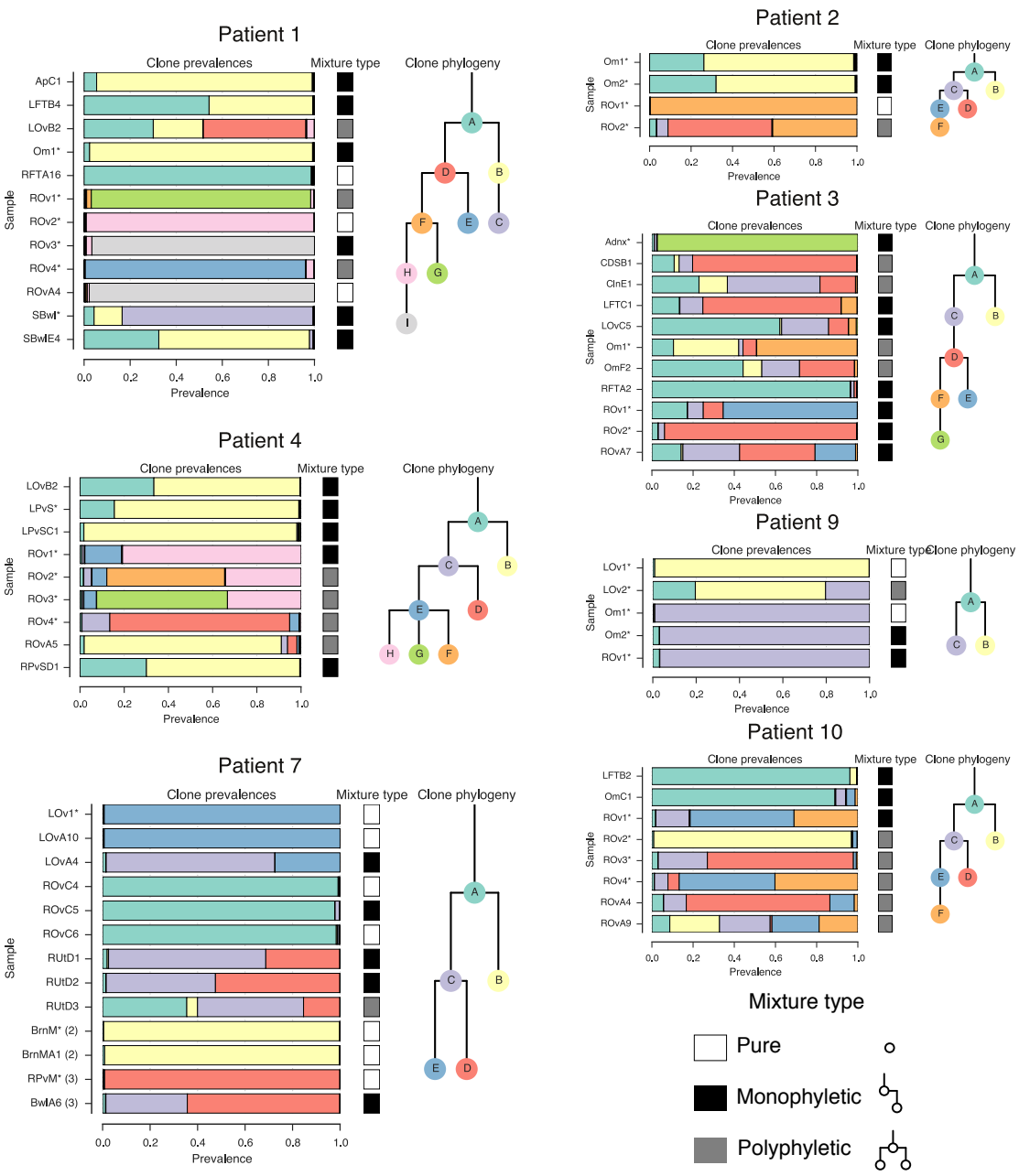
Supplementary Figure 15 Heatmap of proportions of each signature for each branch of the patient's sample tree. Rows are labeled with the patient id and tree node id separated by an underscore. Tree nodes are used synonymously with the branch entering the node from the ancestral node. Columns are labeled with the number of the curated cosmic signature. Intensity represents the proportion of each signature that generated the SNVs on each branch of each patient.

Nature Genetics: doi:10.1038/ng.3573



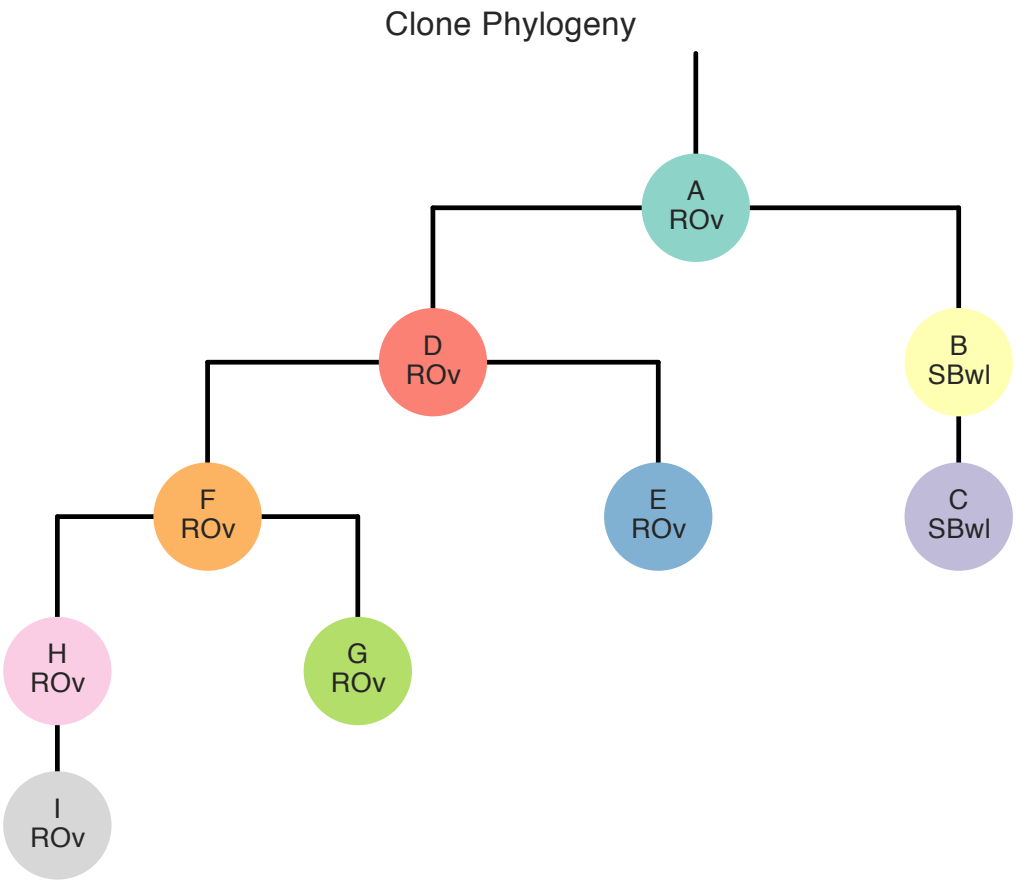
Supplementary Figure 16 Clonal phylogeny with inferred origin and loss of PyClone clusters by the stochastic Dollo process. The left part of each figure shows the clone phylogeny, and the right the clonal genotype matrix. Coloured circles / crosses represent origin / losses of PyClone clusters (**Supplementary Table 16**) respectively.

Nature Genetics: doi:10.1038/ng.3573



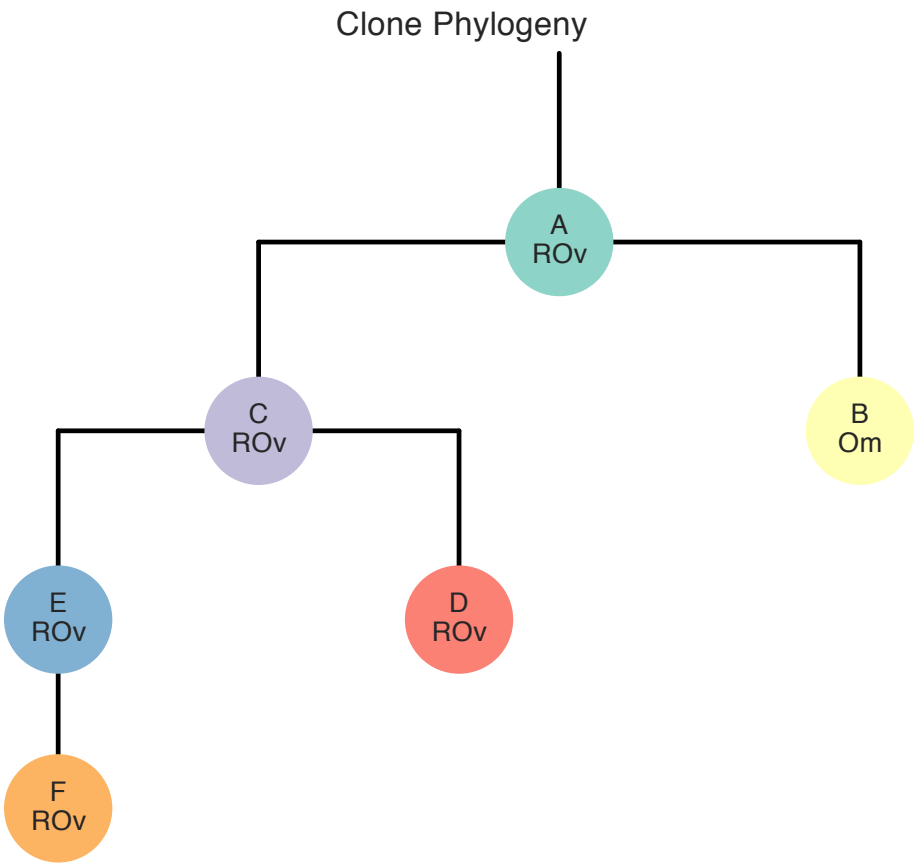
Supplementary Figure 17 Clone phylogenies and sample clone mixtures. A clone phylogeny is shown for each patient. The left stacked bar plot shows prevalence (x axis) of clones (coloured as for the clone phylogeny), across samples (y axis). Discovery samples are denoted by a '*' after the sample name. For patient 7, second and third timepoint samples are denoted by (2) and (3) respectively. The sample mixture type is shown to the right of the stacked bar plot, with white as clonally pure, black as mixed monophyletic, and grey as mixed polyphyletic.

Nature Genetics: doi:10.1038/ng.3573



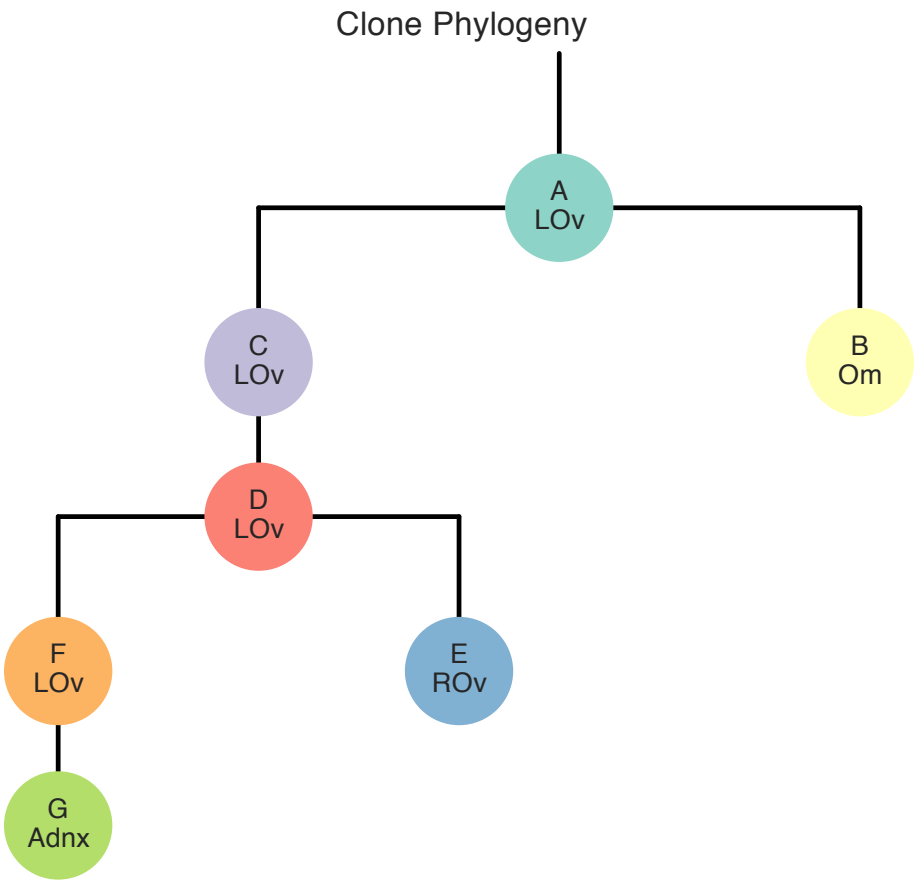
Supplementary Figure 18 Clonal migration patient 1. For each patient, we calculated the minimum number of migrations required to produce the observed distribution of clones to sites, given the clone phylogeny. Shown is the maximum parsimony assignment of clones to the general anatomic site in which the clone originated. Parent child nodes with different anatomic sites represent a migration, followed by the addition of mutations defining the clonal genotype of the child. Where multiple solutions exist at the root, the putative primary is selected if it is among the maximum parsimony set of solutions.

Nature Genetics: doi:10.1038/ng.3573



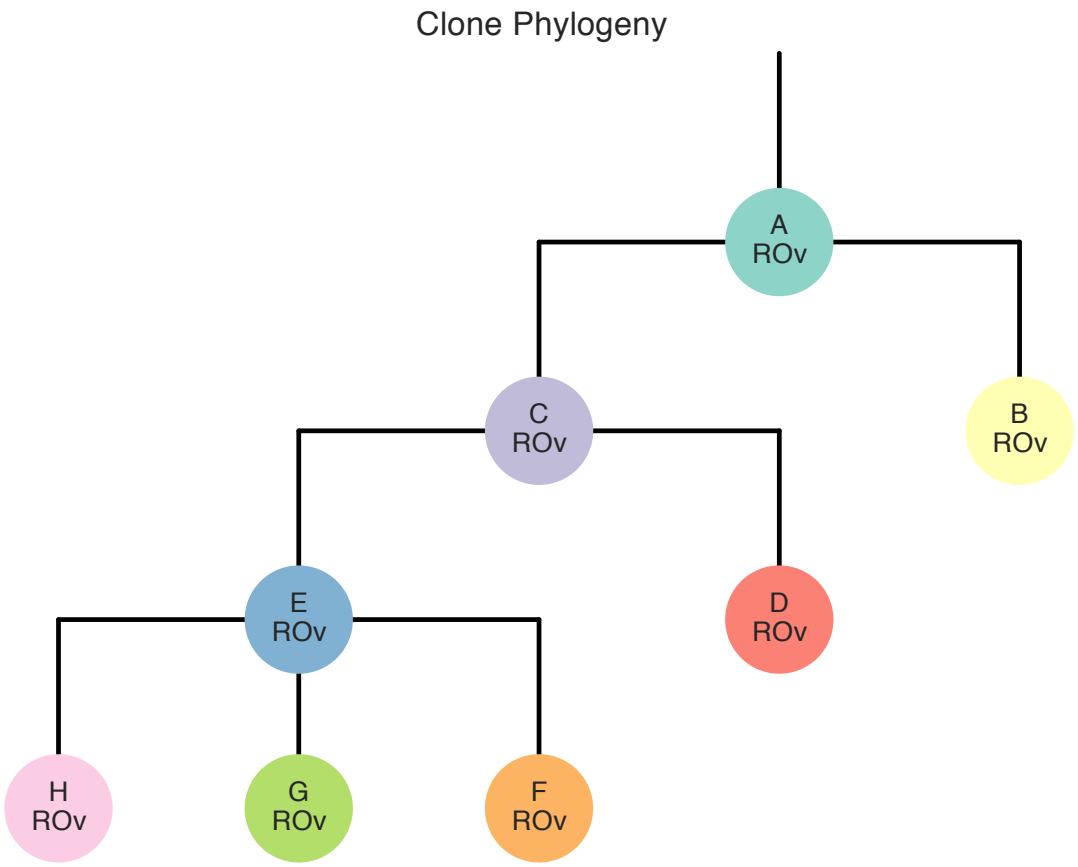
Supplementary Figure 19 Clonal migration patient 2. For each patient, we calculated the minimum number of migrations required to produce the observed distribution of clones to sites, given the clone phylogeny. Shown is the maximum parsimony assignment of clones to the general anatomic site in which the clone originated. Parent child nodes with different anatomic sites represent a migration, followed by the addition of mutations defining the clonal genotype of the child. Where multiple solutions exist at the root, the putative primary is selected if it is among the maximum parsimony set of solutions.

Nature Genetics: doi:10.1038/ng.3573



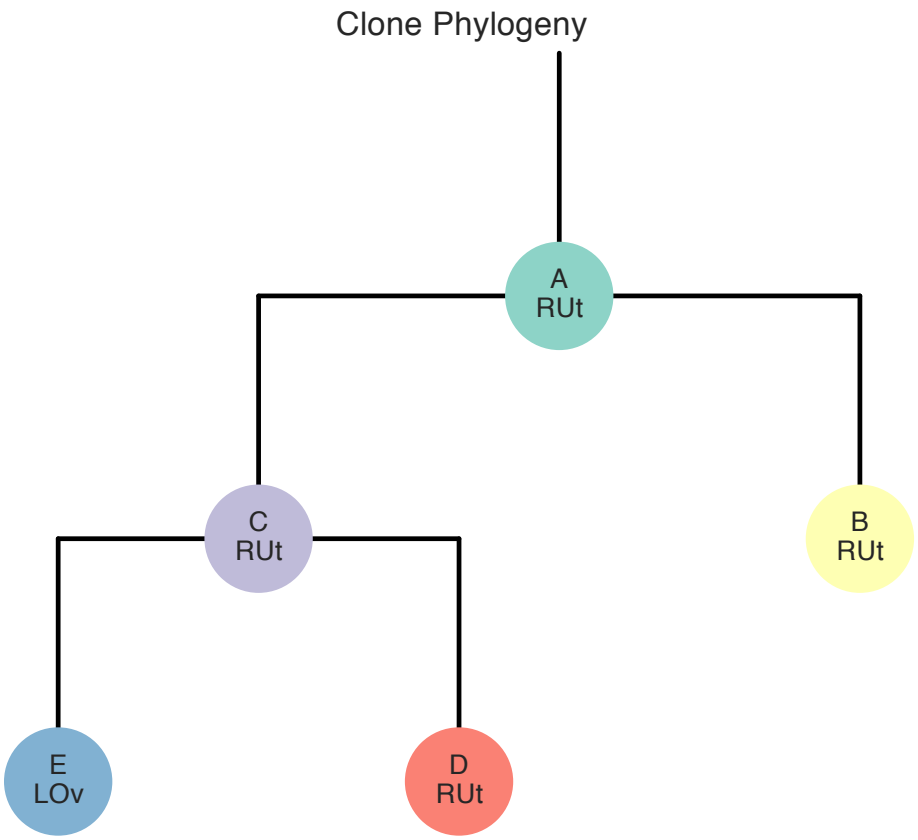
Supplementary Figure 20 Clonal migration patient 3. For each patient, we calculated the minimum number of migrations required to produce the observed distribution of clones to sites, given the clone phylogeny. Shown is the maximum parsimony assignment of clones to the general anatomic site in which the clone originated. Parent child nodes with different anatomic sites represent a migration, followed by the addition of mutations defining the clonal genotype of the child. Where multiple solutions exist at the root, the putative primary is selected if it is among the maximum parsimony set of solutions.

Nature Genetics: doi:10.1038/ng.3573



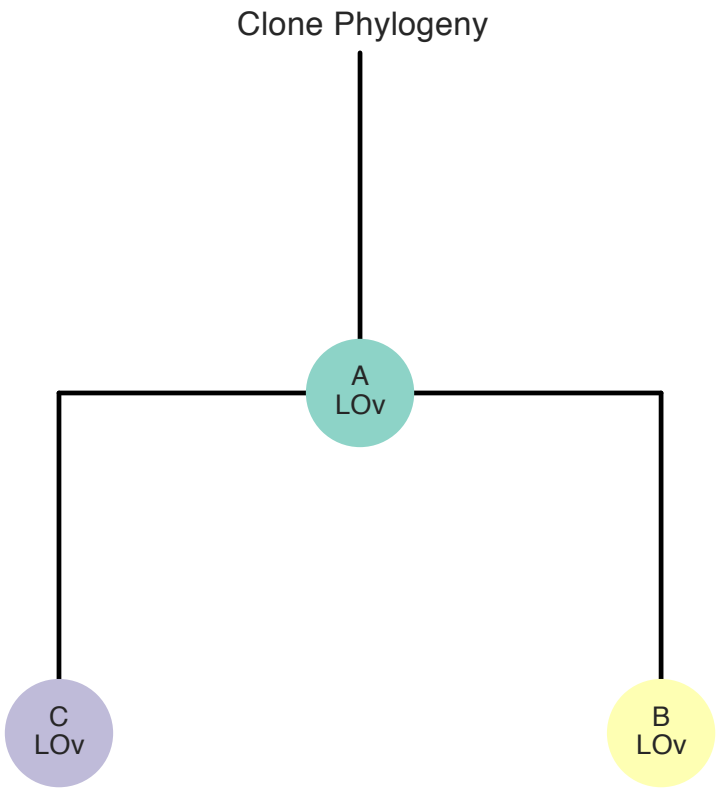
Supplementary Figure 21 Clonal migration patient 4. For each patient, we calculated the minimum number of migrations required to produce the observed distribution of clones to sites, given the clone phylogeny. Shown is the maximum parsimony assignment of clones to the general anatomic site in which the clone originated. Parent child nodes with different anatomic sites represent a migration, followed by the addition of mutations defining the clonal genotype of the child. Where multiple solutions exist at the root, the putative primary is selected if it is among the maximum parsimony set of solutions.

Nature Genetics: doi:10.1038/ng.3573



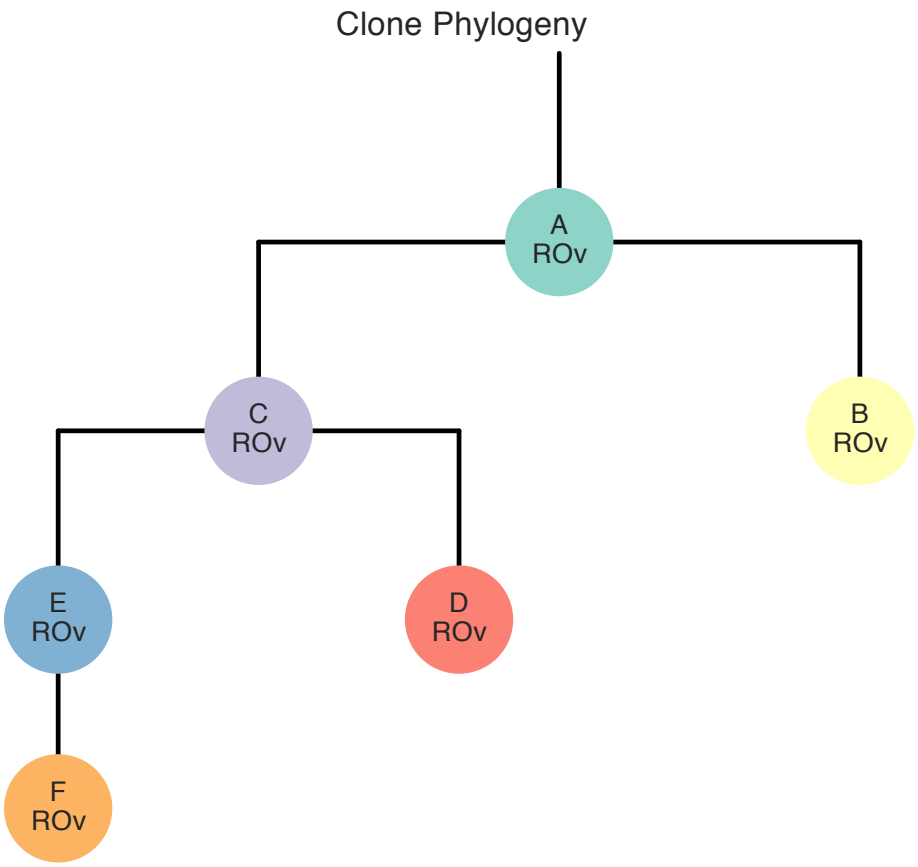
Supplementary Figure 22 Clonal migration patient 7. For each patient, we calculated the minimum number of migrations required to produce the observed distribution of clones to sites, given the clone phylogeny. Shown is the maximum parsimony assignment of clones to the general anatomic site in which the clone originated. Parent child nodes with different anatomic sites represent a migration, followed by the addition of mutations defining the clonal genotype of the child. Where multiple solutions exist at the root, the putative primary is selected if it is among the maximum parsimony set of solutions.

Nature Genetics: doi:10.1038/ng.3573



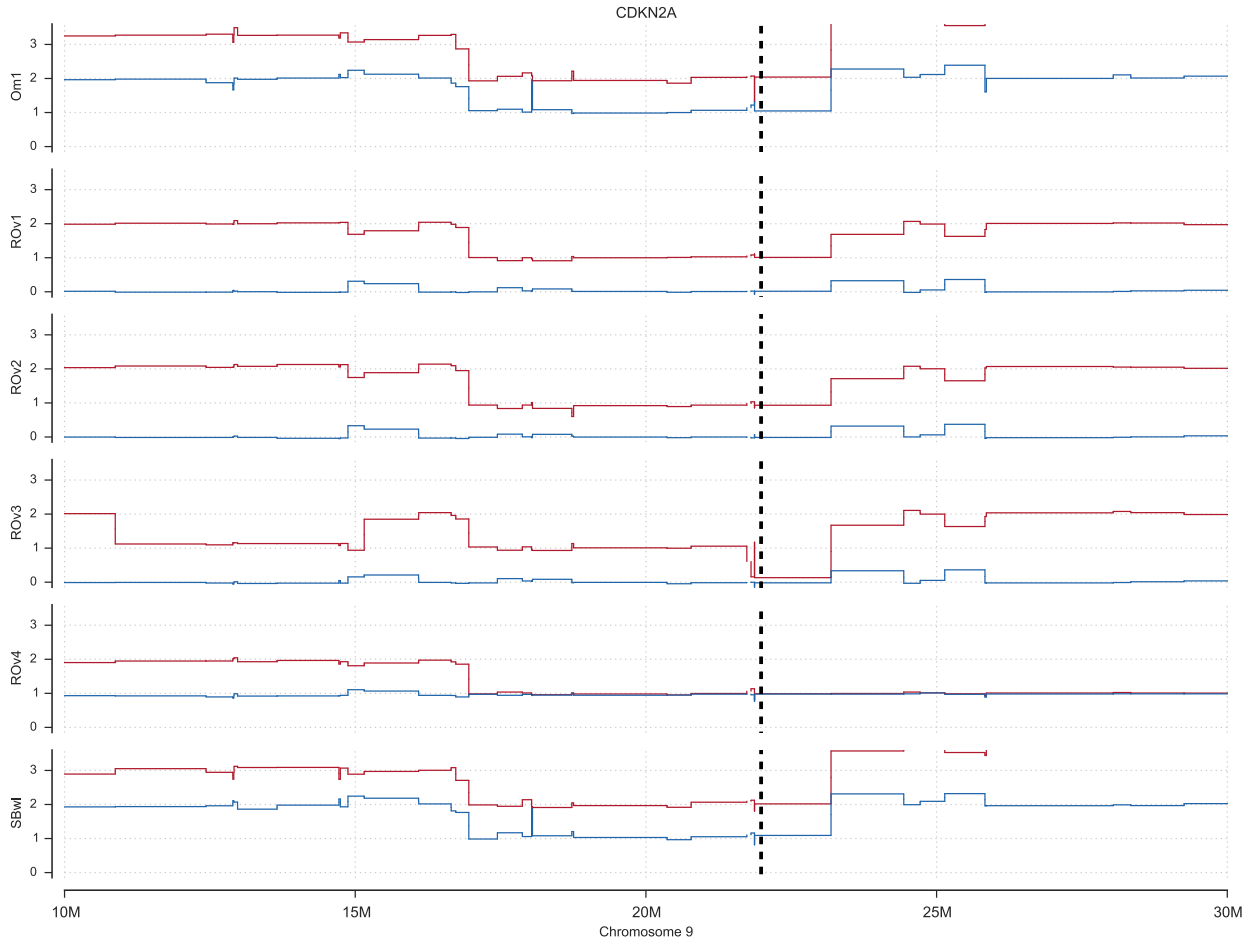
Supplementary Figure 23 Clonal migration patient 9. For each patient, we calculated the minimum number of migrations required to produce the observed distribution of clones to sites, given the clone phylogeny. Shown is the maximum parsimony assignment of clones to the general anatomic site in which the clone originated. Parent child nodes with different anatomic sites represent a migration, followed by the addition of mutations defining the clonal genotype of the child. Where multiple solutions exist at the root, the putative primary is selected if it is among the maximum parsimony set of solutions.

Nature Genetics: doi:10.1038/ng.3573



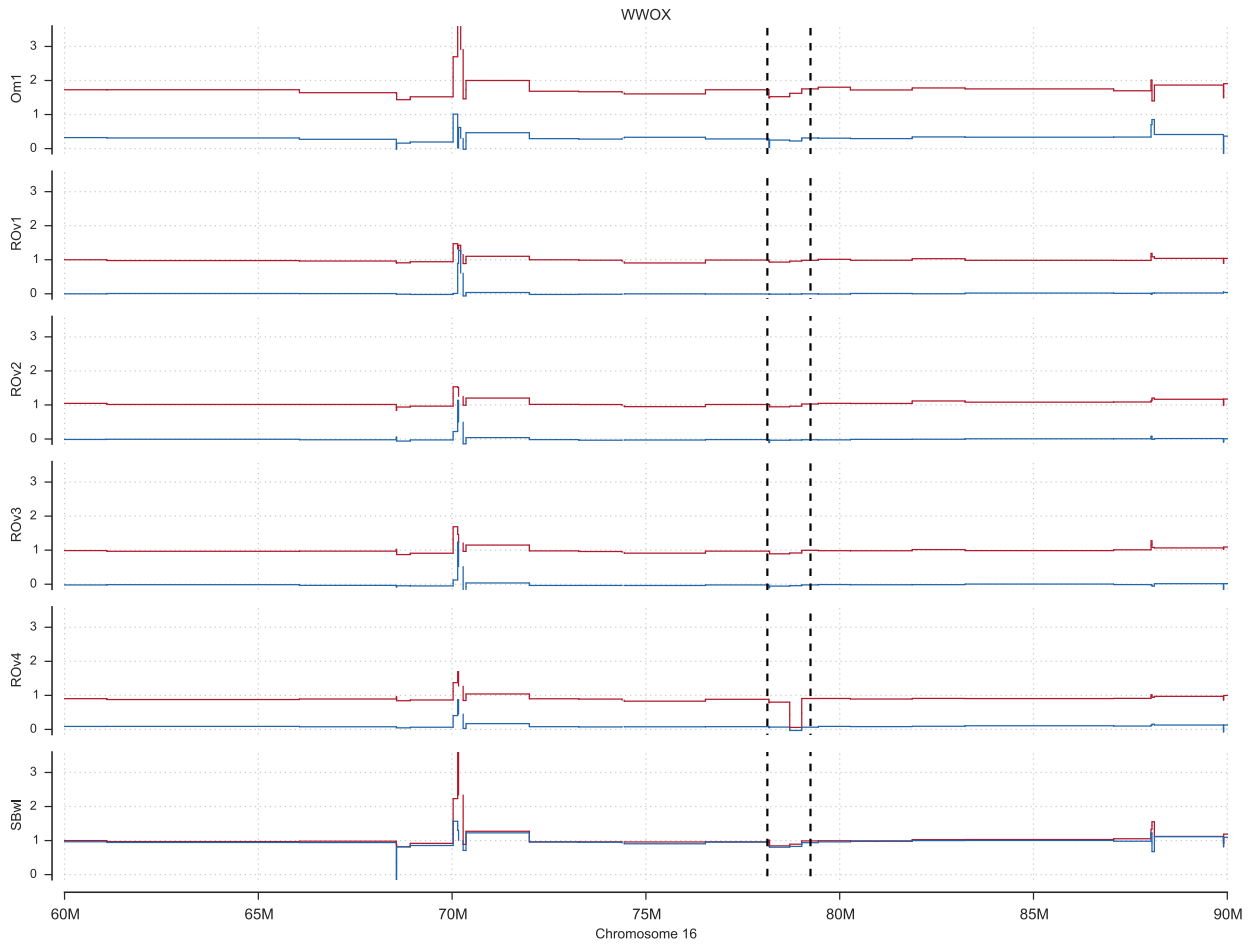
Supplementary Figure 24 Clonal migration patient 10. For each patient, we calculated the minimum number of migrations required to produce the observed distribution of clones to sites, given the clone phylogeny. Shown is the maximum parsimony assignment of clones to the general anatomic site in which the clone originated. Parent child nodes with different anatomic sites represent a migration, followed by the addition of mutations defining the clonal genotype of the child. Where multiple solutions exist at the root, the putative primary is selected if it is among the maximum parsimony set of solutions.

Nature Genetics: doi:10.1038/ng.3573



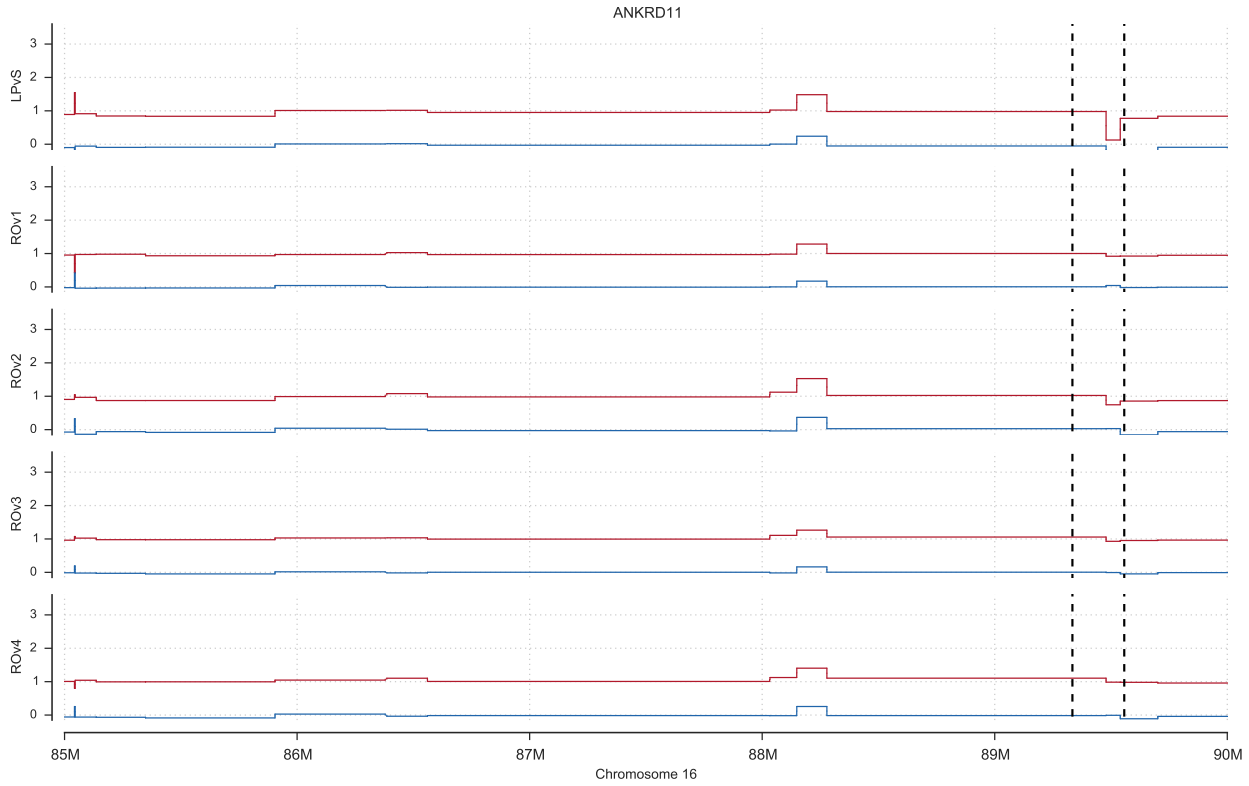
Supplementary Figure 25 Sample specific homozygous deletion of the CDKN2A locus in right ovary site 3 of patient 1. Allele specific copy number profile, predicted by ReMixT is shown with the minor allele (arbitrarily assigned) represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Dashed vertical lines indicate gene boundaries.

Nature Genetics: doi:10.1038/ng.3573



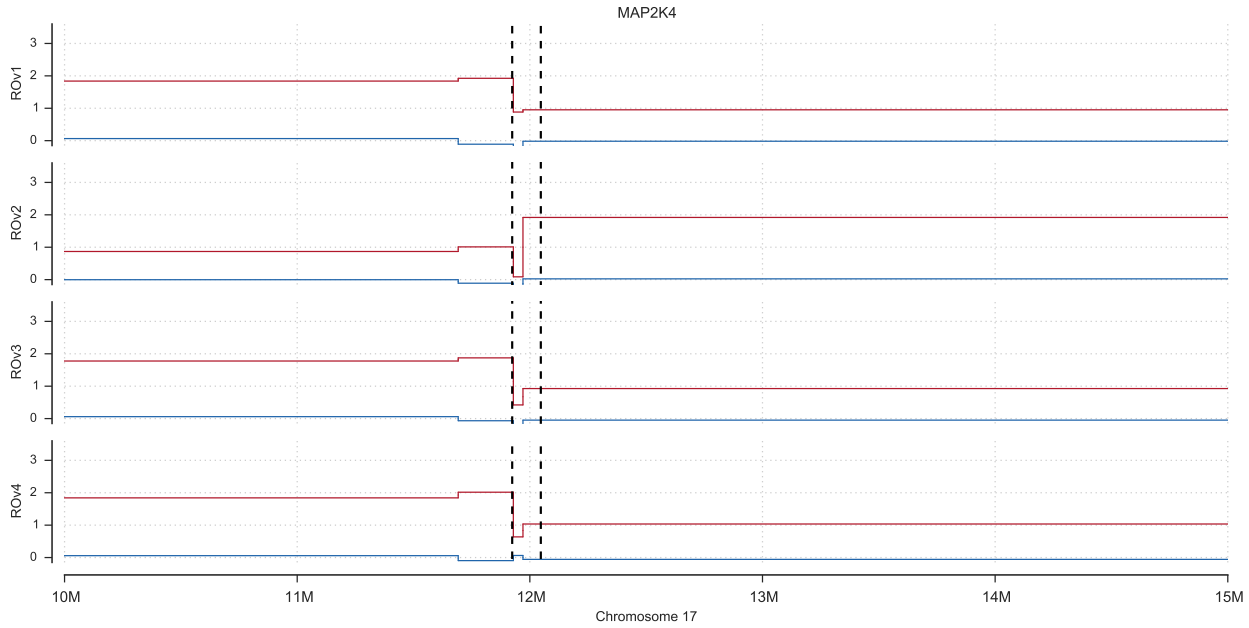
Supplementary Figure 26 Sample specific homozygous deletion of the WWOX locus in right ovary site 4 of patient 1. Allele specific copy number profile, predicted by ReMixT is shown with the minor allele (arbitrarily assigned) represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Dashed vertical lines indicate gene boundaries.

Nature Genetics: doi:10.1038/ng.3573



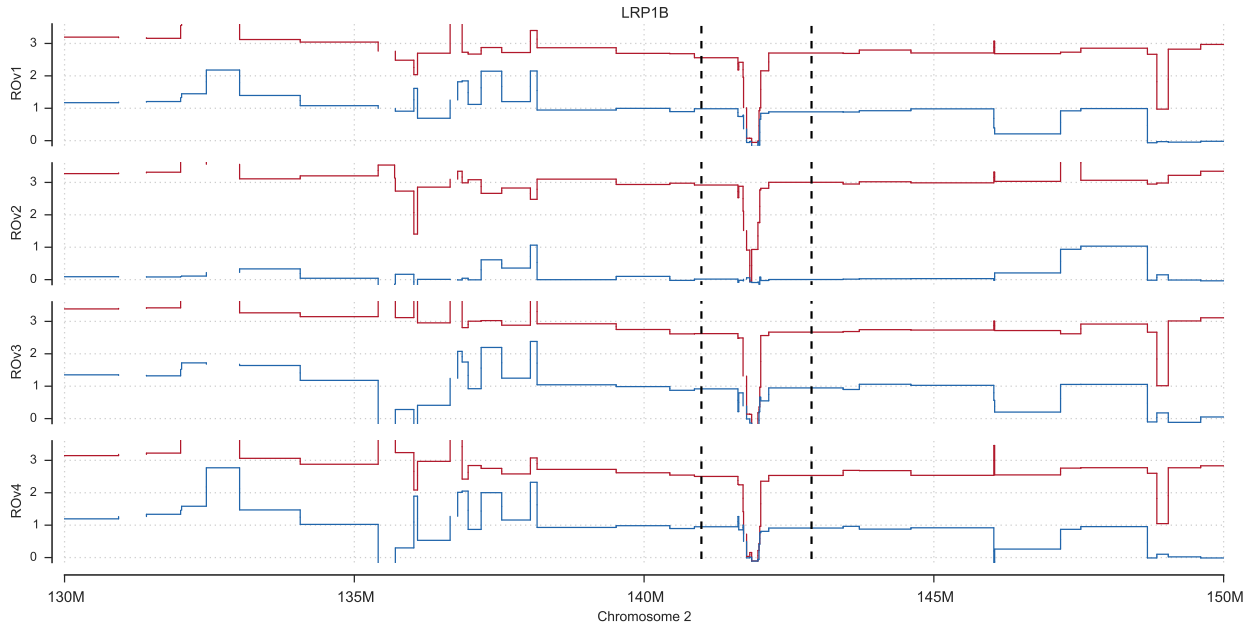
Supplementary Figure 27 Sample specific homozygous deletion of the ANKRD11 locus in left pelvic site of patient 4. Allele specific copy number profile, predicted by ReMixT is shown with the minor allele (arbitrarily assigned) represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Dashed vertical lines indicate gene boundaries.

Nature Genetics: doi:10.1038/ng.3573

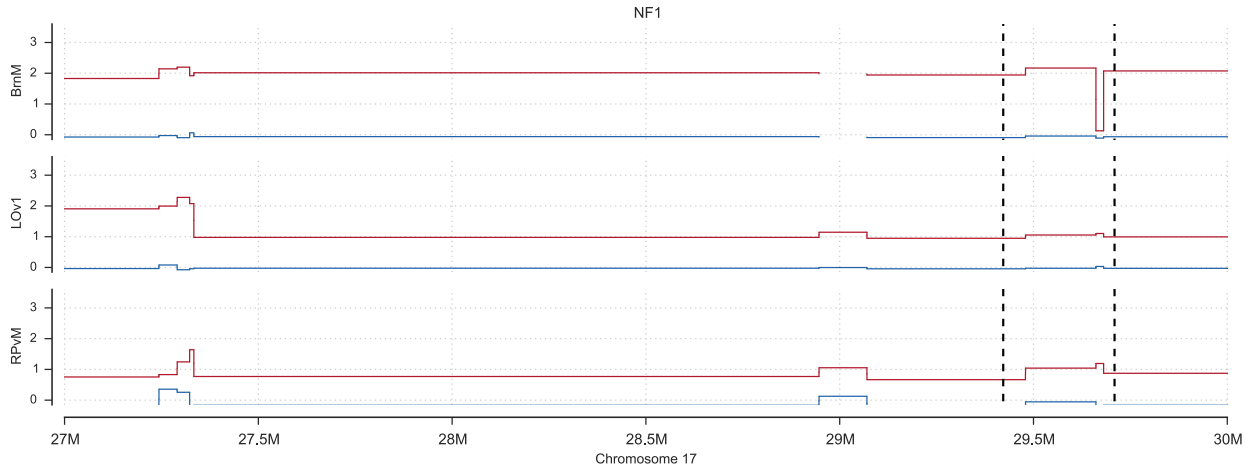


Supplementary Figure 28 Sample specific homozygous deletion of the MAP2K4 locus in right ovary sites 2 and 3 of patient 10. Allele specific copy number profile, predicted by ReMixT is shown with the minor allele (arbitrarily assigned) represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Dashed vertical lines indicate gene boundaries.

Nature Genetics: doi:10.1038/ng.3573

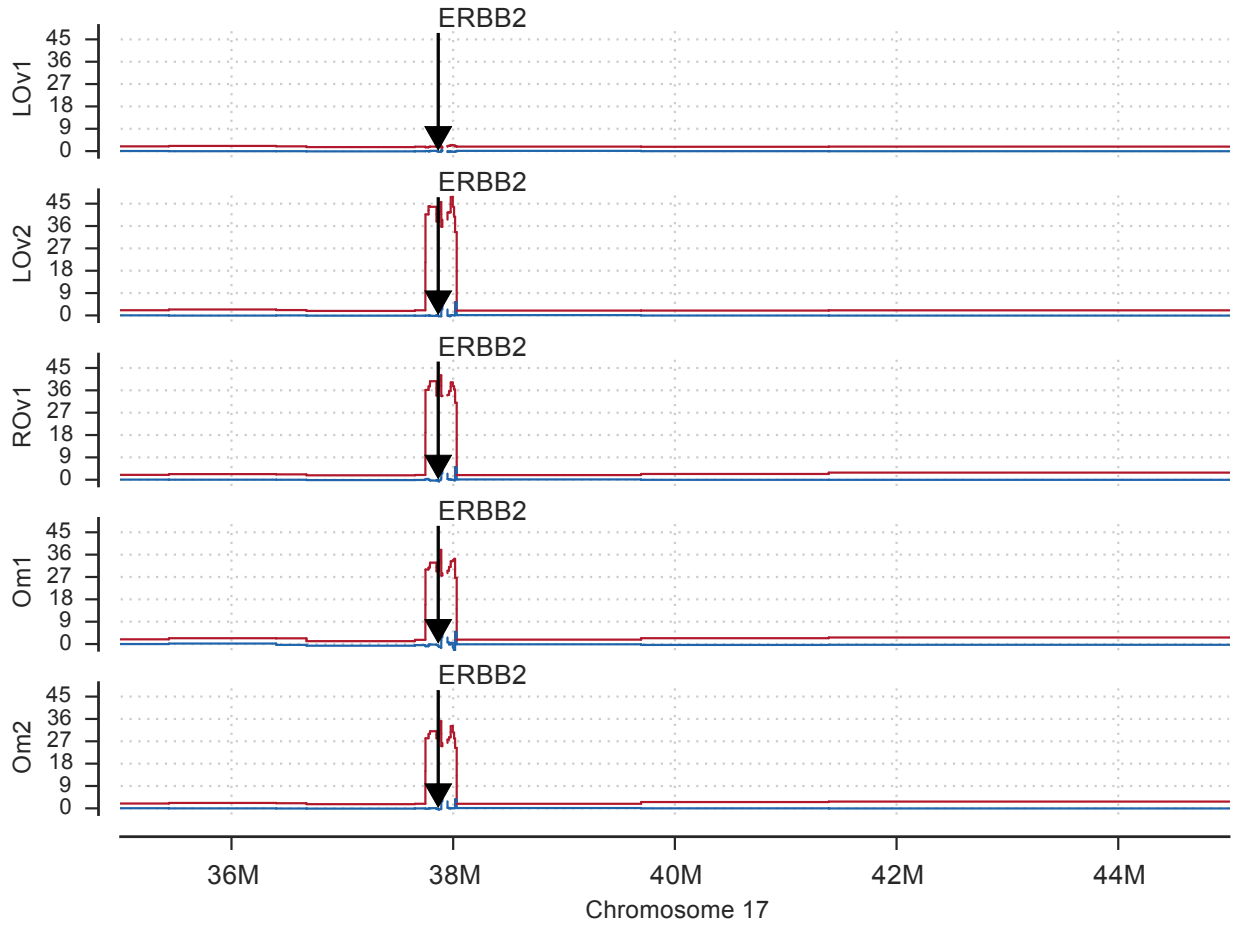


Supplementary Figure 29 Homozygous deletion of the LRP1B locus in patient 10. Allele specific copy number profile, predicted by ReMixT is shown with the minor allele (arbitrarily assigned) represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Dashed vertical lines indicate gene boundaries.



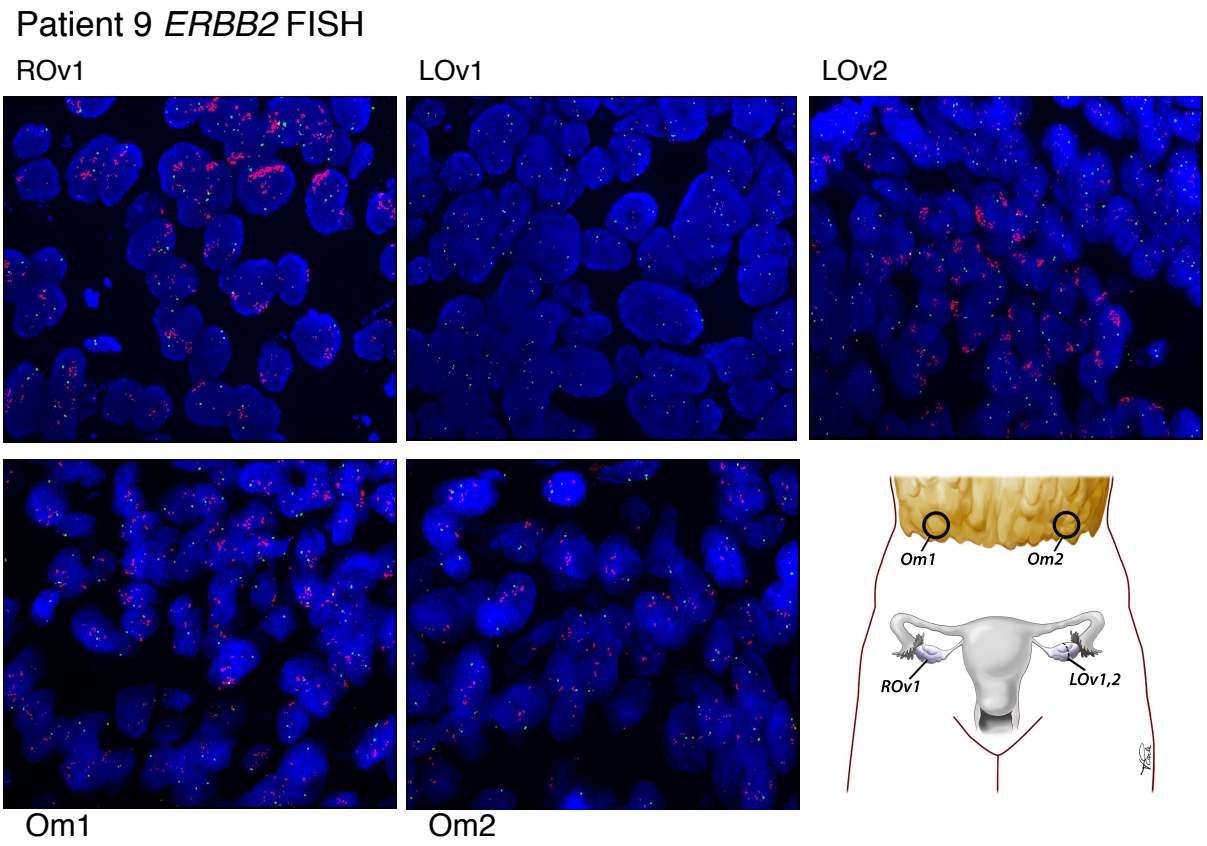
Supplementary Figure 30 Sample specific homozygous deletion of the NF1 locus in brain metastasis site of patient 7. Allele specific copy number profile, predicted by ReMixT is shown with the minor allele (arbitrarily assigned) represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37. Dashed vertical lines indicate gene boundaries.

Nature Genetics: doi:10.1038/ng.3573



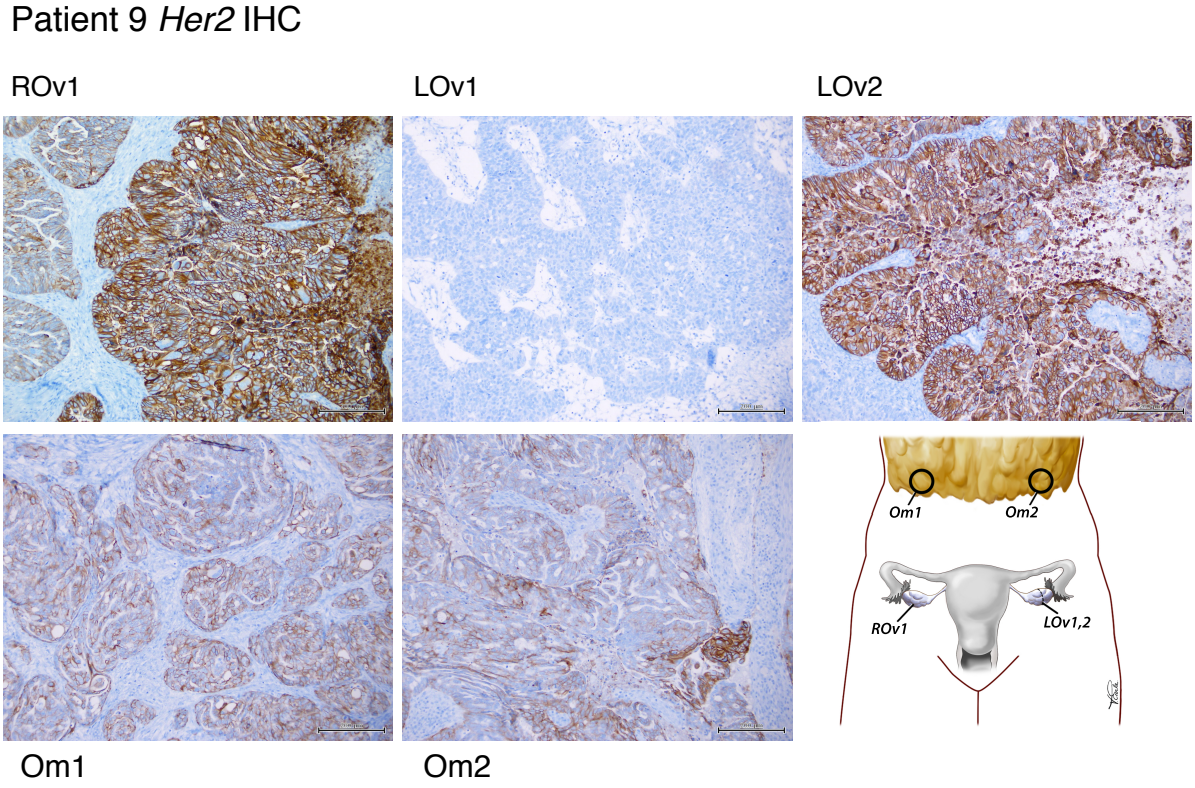
Supplementary Figure 31 High level amplification of ERBB2 locus in patient 9. Allele specific copy number profile, predicted by ReMixT is shown with the minor allele (arbitrarily assigned) represented in blue and the major allele in red. Y-axis indicates the number of copies, x-axis indicates genomic coordinates on the human genome reference build GRCh37.

Nature Genetics: doi:10.1038/ng.3573



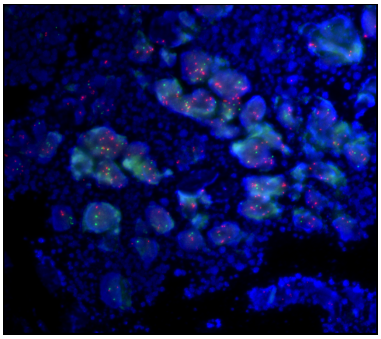
Supplementary Figure 32 Fluorescence in situ hybridization for the ERBB2 locus. The clinical assay (red probe) used for breast cancer was applied to FFPE sections relative to control (green probe) corresponding to LOv1, LOv2, Om1, Om2 and ROv1. Clear amplification is visible in all but LOv1, consistent with predictions in whole genome sequencing.

Nature Genetics: doi:10.1038/ng.3573

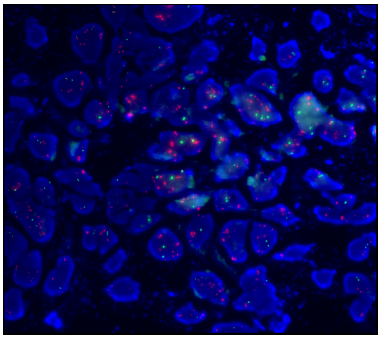


Supplementary Figure 33 Immnohistochemistry. Antibodies against HER2 (cat RM-9103, clone SP3, 1:100, Thermo Scientific, Ottawa, ON, Canada) were applied to LOv1, LOv2, Om1, Om2 and ROv1. Her2 protein was expressed clearly in all samples but LOv1, consistent with FISH and whole genome sequencing results.

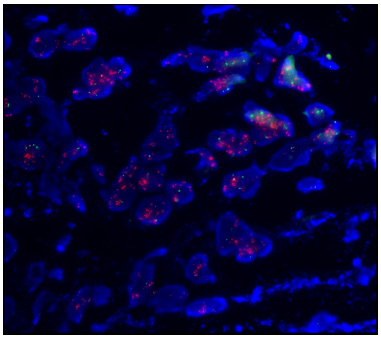
Nature Genetics: doi:10.1038/ng.3573



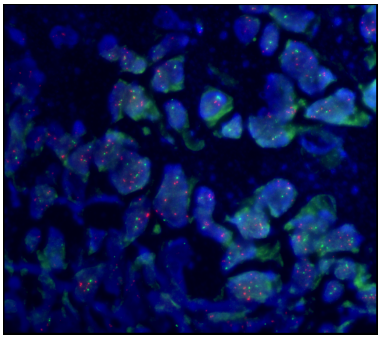
(a) Example picture from omentum site 1 (Om1).



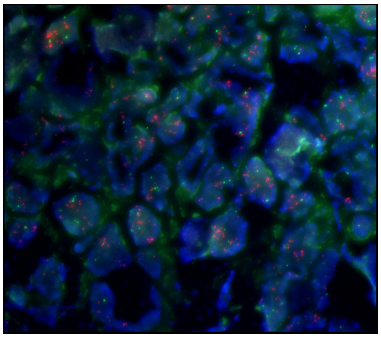
(b) Example picture from omentum site 2 (Om2) showing a lower amplification set of nuclei.



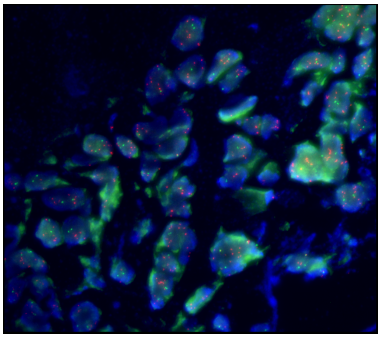
(c) Example picture from omentum site 2 (Om2) showing a higher amplification set of nuclei.



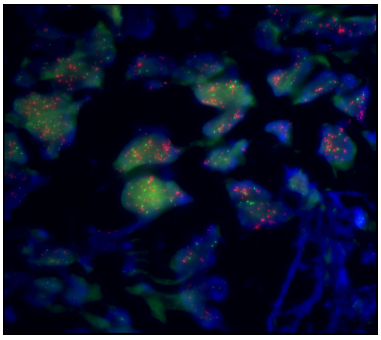
(d) Example picture from right ovary site 1 (ROv1) showing a lower amplification set of nuclei.



(e) Example picture from right ovary site 1 (ROv1) showing a higher amplification set of nuclei.



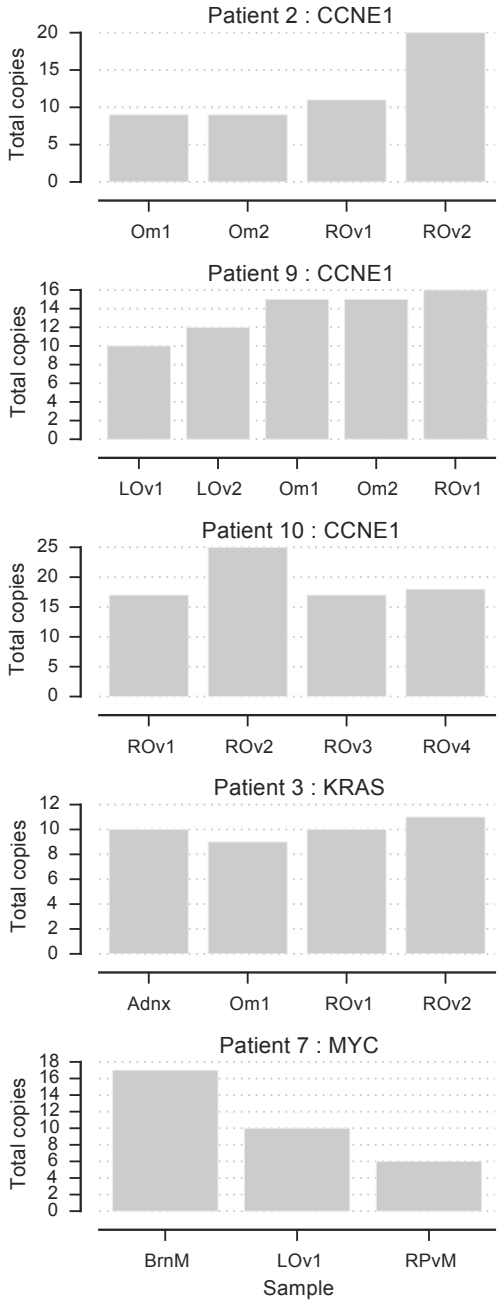
(f) Example picture from right ovary site 2 (ROv2) showing a lower amplification set of nuclei.



(g) Example picture from right ovary site 2 (ROv2) showing a higher amplification set of nuclei.

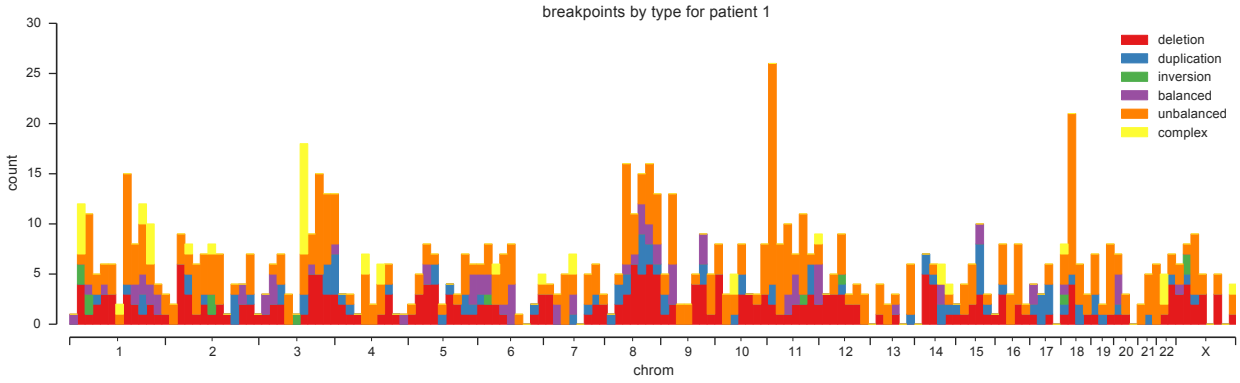
Supplementary Figure 34 Example pictures from FISH analysis of patient 2 CCNE amplification.

Nature Genetics: doi:10.1038/ng.3573

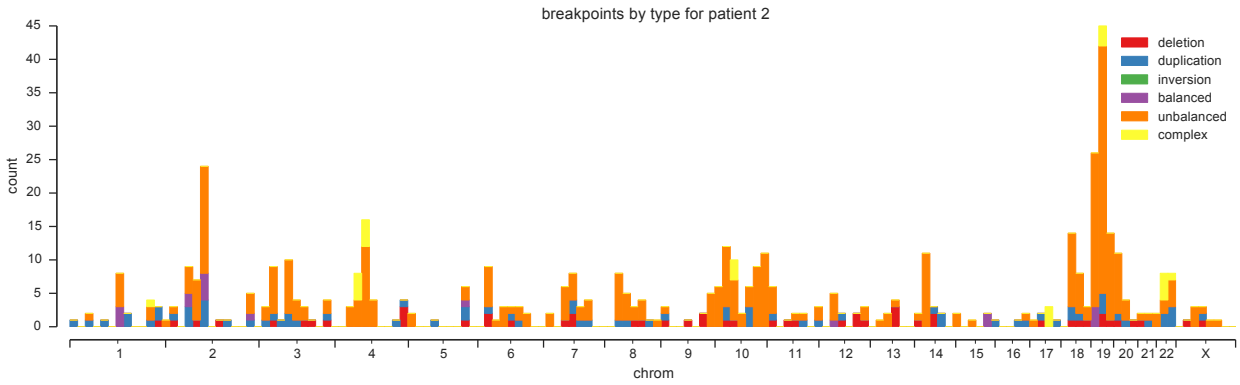


Supplementary Figure 35 Number of copies predicted by ReMixT per gene per site for patients with KRAS, CCNE1 and MYC amplifications.

Nature Genetics: doi:10.1038/ng.3573

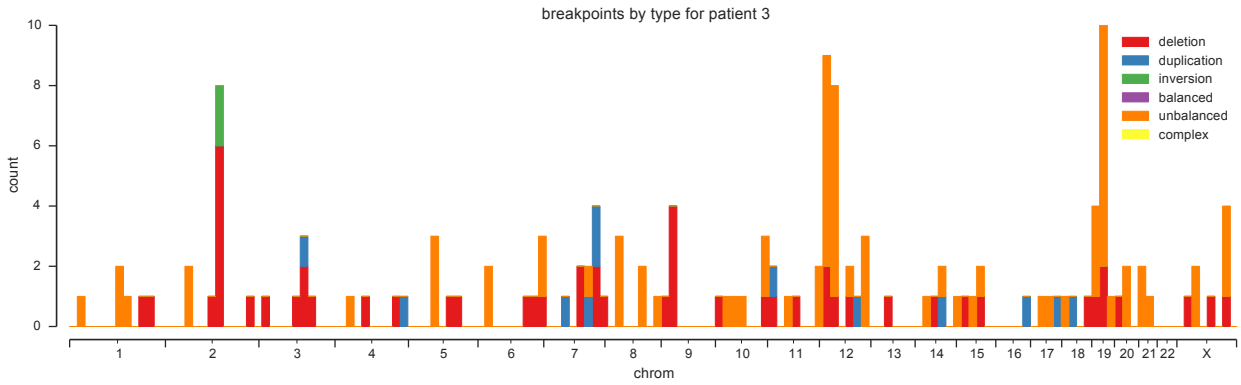


Supplementary Figure 36 Rearrangement type distribution across the genome. Rearrangement breakends are binned in 20Mb intervals across the genome. For each bin, the counts of rearrangement type is shown for breakends within that bin.

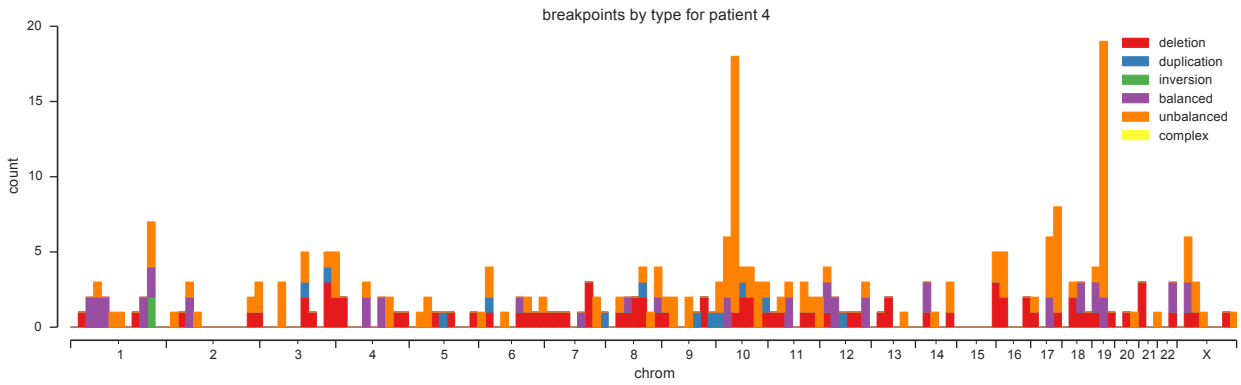


Supplementary Figure 37 Rearrangement type distribution across the genome. Rearrangement breakends are binned in 20Mb intervals across the genome. For each bin, the counts of rearrangement type is shown for breakends within that bin.

Nature Genetics: doi:10.1038/ng.3573

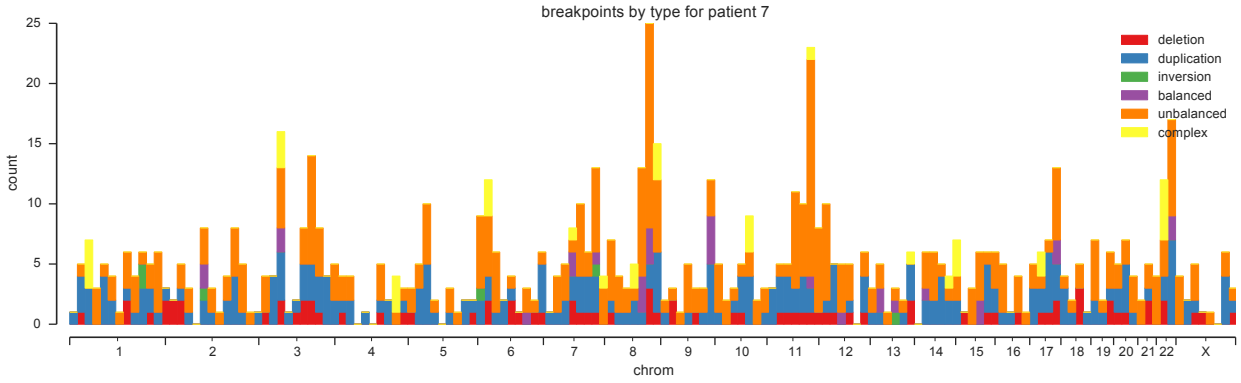


Supplementary Figure 38 Rearrangement type distribution across the genome. Rearrangement breakends are binned in 20Mb intervals across the genome. For each bin, the counts of rearrangement type is shown for breakends within that bin.

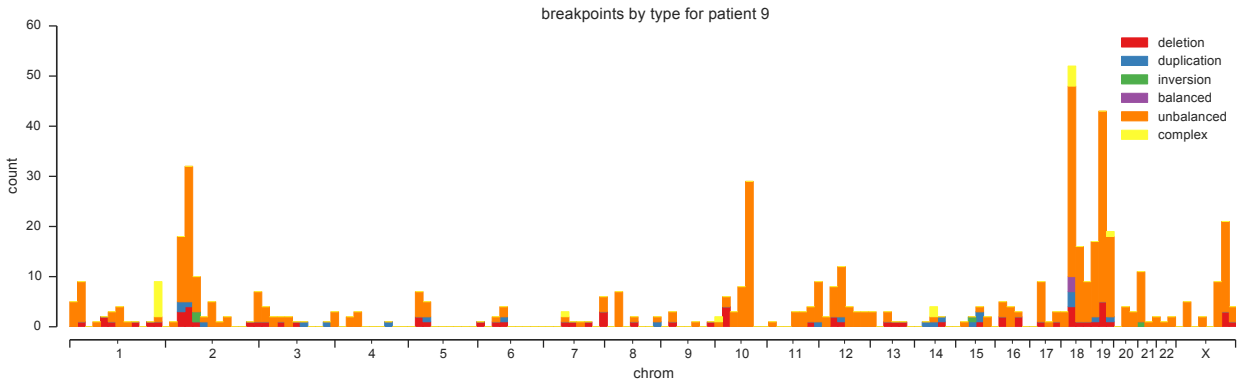


Supplementary Figure 39 Rearrangement type distribution across the genome. Rearrangement breakends are binned in 20Mb intervals across the genome. For each bin, the counts of rearrangement type is shown for breakends within that bin.

Nature Genetics: doi:10.1038/ng.3573

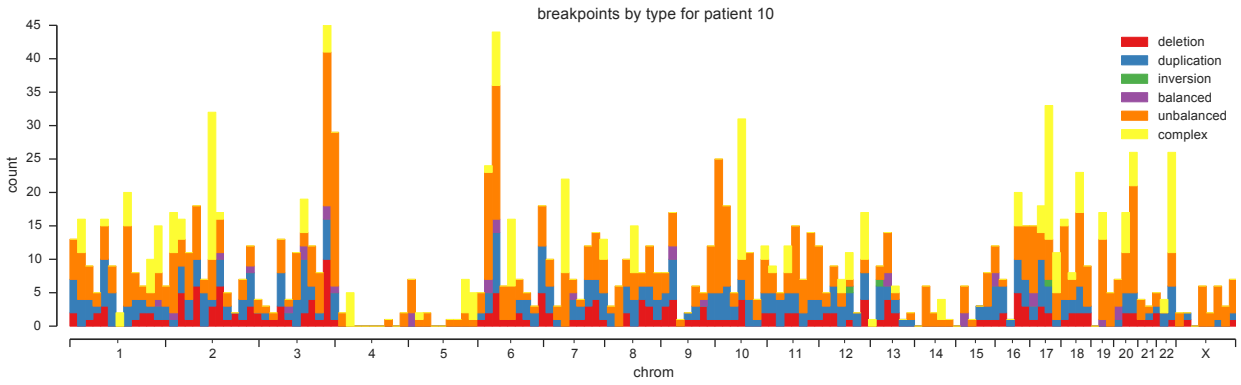


Supplementary Figure 40 Rearrangement type distribution across the genome. Rearrangement breakends are binned in 20Mb intervals across the genome. For each bin, the counts of rearrangement type is shown for breakends within that bin.

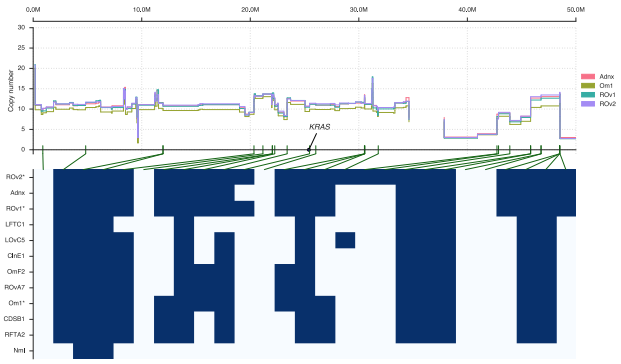


Supplementary Figure 41 Rearrangement type distribution across the genome. Rearrangement breakends are binned in 20Mb intervals across the genome. For each bin, the counts of rearrangement type is shown for breakends within that bin.

Nature Genetics: doi:10.1038/ng.3573

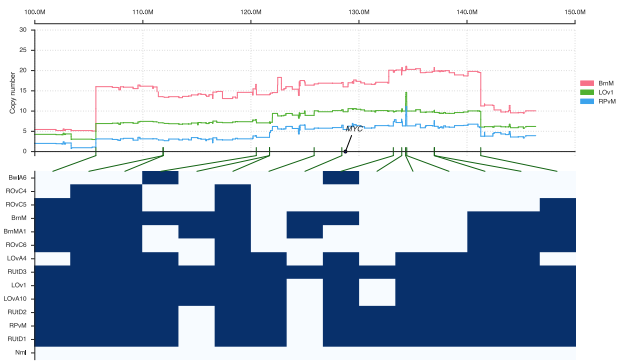


Supplementary Figure 42 Rearrangement type distribution across the genome. Rearrangement breakends are binned in 20Mb intervals across the genome. For each bin, the counts of rearrangement type is shown for breakends within that bin.

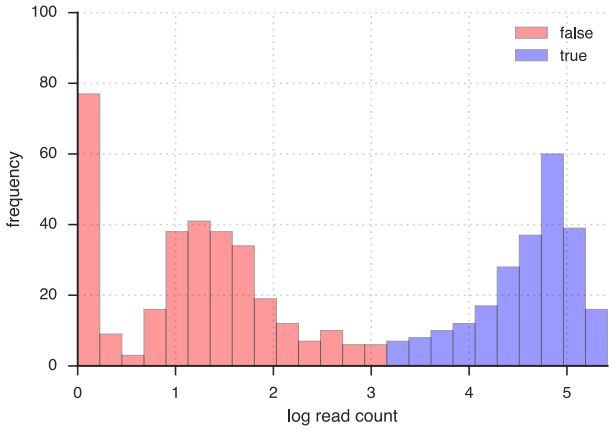


Supplementary Figure 43 The top plot shows total copy number for Adnx, Om1, ROv1, and ROv2 discovery samples in patient 3. The bottom plot shows log read count for deep sequenced breakpoints in the *KRAS* region. Green lines show correspondence between breakpoints in the heatmap and positions of breakends.

Nature Genetics: doi:10.1038/ng.3573

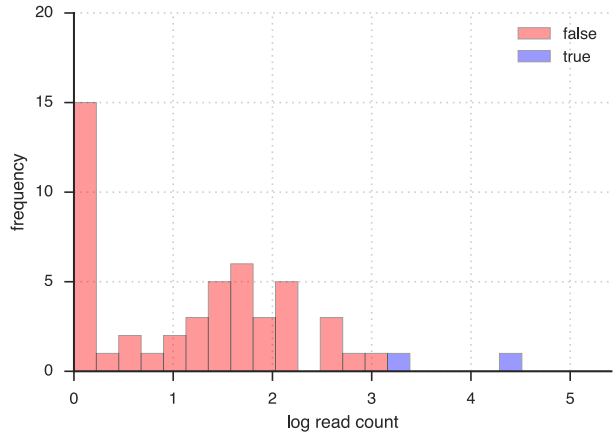


Supplementary Figure 44 The top plot shows total copy number for BrnM, LOv1 and RPvM discovery samples in patient 7. The bottom plot shows log read count for deep sequenced breakpoints in the *MYC* region. Green lines show correspondence between breakpoints in the heatmap and positions of breakends.

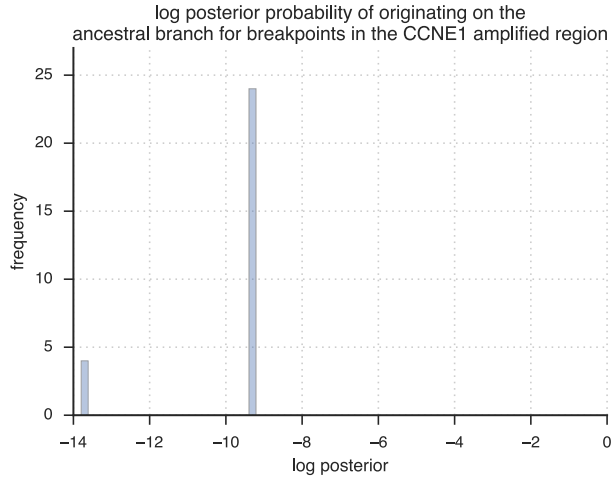


Supplementary Figure 45 Shown is the frequency (y-axis) of binned log read counts (x-axis) for deep sequenced breakpoints across all samples of patient 2. Breakpoints are classified by the membership in one of two components of a Gaussian mixture model, with the component with higher mean assumed to represent breakpoints present in the sample (blue) vs a component representing breakpoints absent, or only present due to contamination.

Nature Genetics: doi:10.1038/ng.3573



Supplementary Figure 46 Shown is the frequency (y-axis) of binned log read counts (x-axis) for deep sequenced breakpoints in the normal blood sample of patient 2. Breakpoints are classified by the membership in one of two components of a Gaussian mixture model, with the component with higher mean assumed to represent breakpoints present in the sample (blue) vs a component representing breakpoints absent, or only present due to contamination.



Supplementary Figure 47 Shown is the frequency (y-axis) of binned log posterior probabilities (x-axis) that breakpoints detected in the whole genome sequencing data originating in the ancestral branch of the sample tree.

Nature Genetics: doi:10.1038/ng.3573

Supplementary Tables

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 1 Description of all samples in study

- patient_id - Study identifier for patient
- sample_id - Anatomic location of tissue sample
- paper_id - Short identifier for a patient sample used in paper text and figures
- malignant - Indicates if samples was from malignant tissue
- discovery_sample - Indicates if samples was included in discovery cohort
- tissue_source - Indicates how tissue was preserved. fresh_frozen (cryo preserved) or ffpe (formalin-fixed, paraffin-embedded)
- anatomy - Anatomic location samples was from

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 2 Sample description and sequencing statistics for discovery cohort

- patient_id - Study identifier for patient
- sample_id - Anatomic location of tissue sample
- paper_id - Short identifier for a patient sample used in paper text and figures
- total_reads - Total number of reads sequenced
- aligned_reads - Number of reads aligned to reference genome
- coverage - Average haploid coverage
- mutation_seq_snvs - Number of SNVs predicted by mutationSeq
- strelka_snvs - Number of SNVs predicted by Strelka
- all_snvs - Number of SNVs in the union set of predictions from mutationSeq and Strelka
- high_quality_snvs - Number of highly mappable (mappability=1.0) SNVs in the intersection set of predictions from mutationSeq and Strelka
- validated_snvs - Number of SNVs validated as somatic using targeted deep sequencing
- ploidy - Ploidy of sample predicted by Demix
- tumour_cell_proportion - Proportion of cells in the sample which are cancerous predicted by Demix
- subclone_frequency - Prevalence of dominant copy number subclone predicted by Demix

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 3 Patient statistics.

- patient_id - Study identifier for patient
- num_samples - Number of samples analysed, discovery and archival
- num_discovery_samples - Number of samples in discovery set
- num_archival_samples - Number of samples in archival set
- snvs_all - Number of SNVs in the union set of predictions from mutationSeq and Strelka
- snvs_all_mappable - Number of SNVs with mappability = 1.0
- snvs_high_quality - Number of SNVs in the intersection set of predictions from mutationSeq and Strelka with mappability = 1.0
- snvs_proportion_sample_specific - Proportion of high quality SNVs that are not ubiquitous
- snvs_ml_rate_of_loss - Inferred rate of loss parameter from the SNV sample tree analysis
- snvs_loss_model_p_value - Likelihood ratio test p-value comparing SNV sample tree null model with no loss to model with loss
- snvs_lost_total - Number of SNVs predicted to be lost
- snvs_lost_deleted - Number of lost SNVs with corroborating copy number changes
- snvs_proportion_lost_deleted - Proportion of lost SNVs with corroborating copy number changes
- breakpoints - Number of rearrangement breakpoints predicted by destruct
- breakpoints_proportion_sample_specific - Proportion of breakpoints that are not ubiquitous
- breakpoints_ml_rate_of_loss - Inferred rate of loss parameter from the breakpoint sample tree analysis
- breakpoints_loss_model_p_value - Likelihood ratio test p-value comparing SNV sample tree null model with no loss to model with loss
- number_of_nuclei_sequenced - Number of nuclei sequenced per patient
- number_of_nuclei_used_for_analysis - Number of nuclei that passed pre-processing for clonal genotyping and phylogeny analysis
- snvs_single_nucleus_targets - Number of SNV targets for single nucleus sequencing which had a sufficient number of non-missing values for clonal genotyping and phylogeny analysis
- breakpoints_single_nucleus_targets - Number breakpoint targets for single nucleus sequencing

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 4 Statistics of sample trees inferred with SNVs and breakpoints

- patient_id - Study identifier for patient
- snv_node - Node in SNV sample tree
- breakpoint_node - Node in breakpoint sample tree
- count_origin_snv - Number of SNVs originating at node
- count_loss_snv - Number of SNVs lost at node
- count_origin_breakpoint - Number of breakpoints originating at node
- count_loss_breakpoint - Number of breakpoints lost at node

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 5 Table of proportions of each signature in each sample. The first column is the patient id and sample id separated by an underscore. Additional columns are named as per the 30 curated cosmic mutation signatures, with values in each column representing the proportion of that signature in the given patient sample.

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 6 Table of proportions of each signature in each branch. The first column is the patient id and tree node id separated by an underscore. Tree nodes are used synonymously with the branch entering the node from the ancestral node. Additional columns are named as per the 30 curated cosmic mutation signatures, with values in each column representing the proportion of that signature attributed to the given patient branch.

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 7 Predicted copy number segments using Demix

- patient_id - Study identifier for patient
- sample_id - Anatomic location of tissue sample
- chrom - Chromosome of the target SNV (hg19)
- start - Chromosome coordinate of start of segment (hg19)
- end - Chromosome coordinate of end of segment (hg19)
- major - Major copy number of segment in dominant clone
- minor - Minor copy number of segment in dominant clone
- major_sub - Major copy number of segment in subclone
- minor_sub - Minor copy number of segment in subclone
- subclonal - Probability the segment is present in subclone population

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 8 Predicted genotypes from clonal analysis

- patient_id - Study identifier for patient
- clone_id - Unique identifier of clone for a patient
- pyclone_cluster_id - Identifier of predicted PyClone cluster in genotype
- present - Indicates if genotype is predicted to contain PyClone cluster

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 9 Predicted clone prevalences

- patient_id - Study identifier for patient
- paper_id - Short identifier for a patient sample used in paper text and figures
- clone_id - Unique identifier of clone for a patient
- prevalence - Predicted prevalence of clone in sample

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 10 Single nucleotide variant targets sequenced using deep amplicon sequencing

- sample_id - Anatomic location of tissue sample
- patient_id - Study identifier for patient
- malignant - Indicates if samples was from malignant tissue
- primer_id - Identifier for PCR primer set for sequencing experiment
- chrom - Chromosome of the target SNV (hg19)
- coord - Chromosome coordinate of target SNV (hg19)
- ref - Nucleotide at target position in reference genome
- alt - Variant nucleotide observed
- ref_counts - Number of reads with ref nucleotide
- alt_counts - Number of reads with alt nucleotide
- depth - Number of reads covering loci
- alt_freq - Proportion of reads with alt nucleotide
- background_average_alt_freq - Average proportion of non-reference nucleotides at positions 30 bases upstream and downstream of target
- ref_p_value - P-value reference allele is present computed from binomial exact test
- alt_p_value - P-value alternate allele is present computed from binomial exact test
- status - Categorical variable indicating if SNV is somatic, germline, wildtype or unkown if corresponding normal is low depth. Does not apply to normal sample.
- gene_name - Name of gene containing SNV if applicable
- snpeff_impact - Impact of SNV predicted by SnpEff

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 11 Results from parsimony analysis of LOH events. The table list segments predicted to be LOH in one or more samples in a patient.

- patient_id - Study identifier for patient
- chrom - Chromosome of the target SNV (hg19)
- start - Chromosome coordinate of start of segment (hg19)
- end - Chromosome coordinate of end of segment (hg19)
- length - Length of the segment
- origin_node - Node(s) of SNV sample tree where the LOH event is predicted to occur
- is_concordant - Binary variable indicating whether the observed pattern of LOH events agrees with the SNV sample tree and a single origin

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 12 Results of fluorescence in situ hybridization analysis of patient 2 CCNE1 amplification across samples.

- sample_id - Anatomic location of tissue sample
- nucleus_id - Identifier of nucleus in sample
- number_of_green_probes_RP11-81M8_19p13.3 - Number of green control probes counted for nucleus
- number_of_orange_probes_RP11-345J21_19q12 - Number of orange event probes counted for nucleus
- picture_number - Identifier of picture containing nucleus
- x_coord - X-coordinate of picture
- y_coord - Y-coordinate of picture

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 13 Break ends and associated segments for breakpoints within the CCNE1 amplified region for patient 2.

- prediction_id - Prediction identifier from destruct
- chrom_1 - Chromosome of first side of breakpoint (hg19)
- coord_1 - Coordinate of first side of breakpoint (hg19)
- strand_1 - Strand of first side of breakpoint (hg19)
- chrom_2 - Chromosome of second side of breakpoint (hg19)
- coord_2 - Coordinate of second side of breakpoint (hg19)
- strand_2 - Strand of second side of breakpoint (hg19)
- prediction_side - Side of the breakpoint corresponding to this break end (1 or 2)
- sample_id - Anatomic location of tissue sample
- total_difference - Difference in total copy number between segments before and after the break end
- num_reads - Number of WGS reads supporting the breakpoint
- is_present - Breakpoint predicted as present in this sample
- segment_chrom - Chromosome of the segment to which the break end is associated
- segment_extremity - Start or end position of the associated segment

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 14 Information about primers used for the targeted bulk sequencing analysis.

- patient_id - Study identifier for patient
- primer_set - Identifier for PCR primer set for sequencing experiment
- chrom - Chromosome of the target SNV (hg19)
- coord - Chromosome coordinate of target SNV (hg19)
- gene_name - Name of gene containing SNV if applicable
- effect - Effect of SNV predicted by SnpEff
- category - Type of event
- ref - Nucleotide at target position in reference genome
- alt - Variant nucleotide observed
- left_primer - Sequence of first primer
- right_primer - Sequence of second primer
- product_start - Genomic coordinate were product starts for SNVs. Position in rearrangement sequence for breakpoints.
- product_end - Genomic coordinate were product ends for SNVs. Position in rearrangement sequence for breakpoints.
- strand - Strand of breakpoint ends. Not applicable for SNVs

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 15 Information about amplicons for the Illumina TruSeq/Nextera sequencing.

- chrom - Chromosome of the target SNV (hg19)
- amplicon_beg - Chromosome coordinate of where amplicon begins (hg19)
- amplicon_end - Chromosome coordinate of where amplicon ends (hg19)

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 16 Results from PyClone analysis of deep sequenced SNVs

- patient_id - Study identifier for patient
- sample_id - Anatomic location of tissue sample
- chrom - Chromosome of the target SNV (hg19)
- coord - Chromosome coordinate of target SNV (hg19)
- ref - Nucleotide at target position in reference genome
- alt - Variant nucleotide observed
- primer_set - Identifier for PCR primer set for sequencing experiment
- cluster_id - PyClone cluster of SNV
- mean - Mean cellular prevalence of SNV in sample
- std - Standard deviation of cellular prevalence estimates from MCMC chain
- ci_length - Length of 95% credible interval of cellular prevalence in sample
- ml_loss - Binary variable if the SNV was predicted to be lost in the sample based on sample tree analysis
- ml_origin - Binary variable indicating if SNV was predicted to originate in the sample based on sample tree analysis
- ml_presence - Binary variable indicating if the SNV was predicted to be present at sample based on sample tree analysis
- deletion - Binary variable indicating if the loss was corroborated by a deletion (not meaningful if ml_loss is 0)

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 17 Information about primers used for the targeted single cell (nucleus) sequencing analysis.

- patient_id - Study identifier for patient
- primer_set - Identifier for PCR primer set for sequencing experiment
- chrom - Chromosome of the target SNV (hg19)
- coord - Chromosome coordinate of target SNV (hg19)
- gene_name - Name of gene containing SNV if applicable
- effect - Effect of SNV predicted by SnpEff
- category - Type of event
- ref - Nucleotide at target position in reference genome
- alt - Variant nucleotide observed
- left_primer - Sequence of first primer
- right_primer - Sequence of second primer
- product_start - Genomic coordinate were product starts for SNVs. Position in rearrangement sequence for breakpoints.
- product_end - Genomic coordinate were product ends for SNVs. Position in rearrangement sequence for breakpoints.
- strand - Strand of breakpoint ends. Not applicable for SNVs

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 18 Results from single nucleus sequencing of SNVs

- patient_id - Study identifier for patient
- sample_id - Anatomic location of tissue sample
- primer_set - Identifier for PCR primer set for sequencing experiment
- well_id - Identifier of nuclei or control in well
- well_type - Variable indicating whether well contains a nucleus, positive control or negative control
- chrom - Chromosome of the target SNV (hg19)
- coord - Chromosome coordinate of target SNV (hg19)
- ref - Nucleotide at target position in reference genome
- alt - Variant nucleotide observed
- ref_counts - Number of reads with ref nucleotide
- alt_counts - Number of reads with alt nucleotide
- ref_p_value - Binomial exact test p-value testing for presence of ref allele
- alt_p_value - Binomial exact test p-value testing for presence of alt allele

Nature Genetics: doi:10.1038/ng.3573

Supplementary Table 19 Results from single nucleus sequencing of breakpoints

- patient_id - Study identifier for patient
- sample_id - Anatomic location of tissue sample
- primer_set - Identifier for PCR primer set for sequencing experiment
- well_id - Identifier of nuclei or control in well
- well_type - Variable indicating whether well contains a nucleus, positive control or negative control
- seq_id - Prediction identifier from destruct
- chrom_1 - Chromosome of first side of breakpoint (hg19)
- coord_1 - Coordinate of first side of breakpoint (hg19)
- strand_1 - Strand of first side of breakpoint (hg19)
- chrom_2 - Chromosome of second side of breakpoint (hg19)
- coord_2 - Coordinate of second side of breakpoint (hg19)
- strand_2 - Strand of second side of breakpoint (hg19)
- count - Number of reads aligned to rearrangement sequence

Nature Genetics: doi:10.1038/ng.3573

1. Schrader, K. A. *et al.* Germline brca1 and brca2 mutations in ovarian cancer: utility of a histology-based referral strategy. *Obstet Gynecol* **120**, 235–40 (2012).

2. McAlpine, J. N. *et al.* Brca1 and brca2 mutations correlate with tp53 abnormalities and presence of immune cell infiltrates in ovarian high-grade serous carcinoma. *Mod Pathol* **25**, 740–50 (2012).

3. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

4. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

5. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research* **3**, 993–1022 (2003).

6. McPherson, A. *et al.* nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res* **22**, 2250–61 (2012).

7. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**, 1270–8 (2009).

8. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**, 1881–93 (2014).

9. Yau, C. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29**, 2482–4 (2013).

10. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* **14**, R80 (2013).

11. Oesper, L., Satas, G. & Raphael, B. J. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* **30**, 3532–40 (2014).

12. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–81 (2012).

13. Casella, G. & Berger, R. *Statistical Inference*. Duxbury advanced series (Duxbury Thomson Learning, 2002).

14. Shah, S. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).

15. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic acids research* **40**, e115–e115 (2012).

16. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome research* **12**, 656–664 (2002).

17. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* (2014).

18. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature methods* (2014).

19. Roth, A. *et al.* Simultaneous inference of clonal genotypes and population structure from single cell tumour sequencing. *Nature methods* (Under review).

20. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).

21. Kimura, M. Theoretical foundation of population genetics at the molecular level. *Theoretical population biology* **2**, 174–208 (1971).

22. Alekseyenko, A. V., Lee, C. J. & Suchard, M. A. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Systematic biology* **57**, 772–784 (2008).

23. Ryder, R. J. & Nicholls, G. K. Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**, 71–92 (2011).

24. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* **17**, 368–376 (1981).

25. Felsenstein, J. & Felenstein, J. *Inferring phylogenies*, vol. 2 (Sinauer Associates Sunderland, 2004).