Supplementary Methods

Sequencing and processing of bulk and single-nucleus libraries

We selected 50 and 60 best single-nucleus DNA libraries from BT340 and BT325, respectively, for deeper sequencing. SNS libraries were sequenced to a median coverage of 0.5x, and combined depth ~30x. Bulk tumor and germline reference DNA were sequenced to 15-30x. All sequencing were done on the Illumina HiSeq platform in the paired-end mode. Read pairs were aligned to the human genome reference (hg19/GRCh37) using BWA (http://bio-bwa.sourceforge.net/) with default parameters; duplicate reads were removed by the MarkDuplicate module from Picard (http://picard.sourceforge.net/) and processed by GATK (http://www.broadinstitute.org/gatk/) following best practices (1).

Two-step strategy for somatic variant detection by SNS

Whole-genome amplification by MDA is known to generate artifacts including single-base errors and random chimera. Artifacts generated early in the amplification are indistinguishable from true somatic mutations. Therefore, detection of *de novo* mutations within one individual nucleus is prone to error. Moreover, there is no way to independently validate the detected variant as all genomic DNA from that nucleus was used for MDA.

Given these considerations, we adopted a two-step strategy to analyze somatic mutations by SNS of many tumor nuclei. First, we detect all clonal and subclonal somatic variants, including single-nucleotide variants and chromosomal rearrangements, from the pool of SNS complemented with bulk whole-genome sequencing. Second, we search for the detected variants within each nucleus and in the bulk tumor to infer the clonality of these variants and construct the clonal hierarchy of these variants.

Detection of somatic chromosomal rearrangements and single-nucleotide variants

We detected somatic chromosomal rearrangements and single-nucleotide variants (sSNV) jointly from all single-nucleus sequencing data combined with the bulk tumor data. Rearrangement events were nominated by at least two discordant read pairs with aberrant insert sizes connecting two loci followed by a search for split read alignments to pinpoint the breakpoint; sSNVs were detected with the HaplotypeCaller program from GATK with a maximum number of alleles set to five.

By definition, any clonal or subclonal variant is present in a population of cells. Therefore true (sub)clonal somatic variants should be present either in the bulk tumor, or in at least two single tumor cells (to be defined as a “clone”). We

removed all variants that were only present in one individual SNS library even though they were supported by multiple variant reads from that library. On the other hand, we kept the variants if they had supporting reads coming from multiple libraries, even if there was only one variant read from each library. This population-based strategy ensured that low allelic frequency variants—either due to low clonal fraction in the bulk sample or due to low allelic coverage in single nucleus libraries—can be confidently emitted when they are found in two or more libraries ("recurrent"). By contrast, random base errors or chimera due to whole-genome amplification do not replicate between independent SNS libraries or get introduced in the bulk library, therefore they are excluded from the list of variants detected in two or more libraries.

From the pool of SN and the bulk tumor sequencing data there were more than 20,000 events in BT325 and more than 10,000 events in BT340 nominated as somatic rearrangements. We then applied the criterion that true somatic rearrangements should either appear in the bulk library or have supporting discordant read pairs from at least two SN libraries. Finally, we applied additional filters for chromosomal rearrangement detection, including mapping quality threshold (≥ 30) and enforced split read alignment. The final lists consisted of 100-200 true somatic events, all of which were also independently confirmed by read depth difference in the bulk tumor library.

We also detected somatic single-nucleotide variants (including indels) by HaplotypeCaller, which searches for genomic variants by local genome assembly (http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_haplotypecaller_HaplotypeCaller.html). Most of the detected somatic variants that were present in two or more libraries were also found in 120x whole-exome sequencing of the bulk tumor. We did not find sub-clonal mutations of cancer genes to be used as markers for inferring clonal segregation in the tumor nuclei population.

Interrogation of detected SCNA events within SNS libraries

Even with accurate segmentation, the copy-number profile for individual SNS is still subject to errors due to MDA amplification bias. We found that MDA amplification bias is predominant at length scales shorter than 1Mb and this bias is attenuated over long distances (manuscript in preparation). Thus, arm-level or long-range (>10Mb) SCNA events can be inferred from sequence read depths in MDA libraries with decent accuracy, as deviations caused by MDA bias are generally small relative to integer copy-number changes. However, distinguishing true focal copy-number alterations from MDA-generated bias is difficult and requires additional approaches.

We employed different strategies to detect focal amplifications and focal deletions. For focal amplifications, we relied on the detection of chromosomal translocations at the amplicon boundaries to verify the presence of the amplification. Once the presence of amplification is confirmed, we then calculated the copy number of the amplicon from the read depth. This

methodology was utilized to verify the presence of *EGFR* amplifications in BT325 and BT340 (Figure 3) as well as the presence of *EGFR* vII and *EGFR* vIII truncations in either tumor (as the truncated gene is also amplified). Detection of focal deletions required a different approach as allelic dropout can more frequently obliterate the signatures of chromosomal translocations at the deletion boundaries. We therefore developed a novel approach to detect both focal and arm-level deletions by loss-of-heterozygosity.

Haplotype-based detection of loss-of-heterozygosity

A unique feature of many tumor genomes is the loss of heterozygosity of large chromosomal regions as well as smaller homozygous deletions, leading to the elimination of one or both germline haplotypes (2-4) (Supplementary Fig. 5A). For cells with loss-of-heterozygosity (LOH) in a region of the genome, the retained haplotype, or allelotype (5), can be directly determined from the sequencing data (Supplementary Fig. 5B). If whole-genome amplification from single-cells were complete and represented all alleles, then the haploid genotype of cells with LOH could be readily distinguished from the heterozygous genotype of diploid cells (Supplementary Fig. 5B, left panel). However, single-cell sequencing is inevitably incomplete and not all alleles are represented in the resulting data (Supplementary Fig. 5B, right panel). Single-base errors introduced during MDA can further confound the inference of true genotype.

We devised a clustering-based approach to dissect the cell populations harboring heterozygous and/or homozygous genotypes in each genomic region. First, the pairwise distance between single-cell genotypes is calculated from the fraction of germline heterozygous sites covered in both libraries that show different genotypes (Supplementary Fig. 5C); this distance is large if cells are discordant for heterozygosity/homozygosity, and small if they are both derived from a common ancestor cell where the initial LOH occurred (concordant). We then perform clustering analysis of cellular populations to identify cells with loss of heterozygosity (LOH) in a given region, which form a very tight cluster (Supplementary Fig. 5C, right panel). Supplemental Figure 5D showed the two clusters formed by 60 SNS libraries from BT325 identified by LOH in Chromosome 10, one cluster displaying a unique haploid genotype (LOH), and a second cluster that were heterozygous.

By pooling the genotype data of LOH cells in Cluster 1, we can determine both the retained haplotype “R” and the deleted haplotype “D” and assign individual or combined “RD” haplotypes to each nucleus at all covered germline heterozygous sites (Supplementary Fig. 5E). This analysis showed that Cluster 1 contains BT325 nuclei with only allele “R” while Cluster 2 contains an admixture of alleles “R” and “D” on chromosome 10 (Supplementary Fig. 5F). The ratio of the deleted to retained haplotypes is approximately 0 in LOH nuclei, and close to 1 in diploid nuclei. Interestingly, the D:R ratio is near 0.5 in another population of nuclei (Supplementary Fig. 5G).

We inferred that the population with D:R ratio of 0.5 represents libraries derived from a 1:1 mixture of tumor and normal nuclei. The normal/tumor mixture was later confirmed by the presence of clonal drivers at subclonal quantities in these samples (Supplementary Fig. 7). Therefore, the haplotype analysis also provides an accurate method to identify and characterize cellular mixtures in presumably single-nuclear samples.

Estimation of MDA single-base error

Haplotype analysis also allowed us to estimate the single-base error rate due to whole-genome MDA amplification. As the tumor clone expands from a common ancestor, all tumor cells should inherit the same preserved haplotype at sites of LOH with deviations representing either inaccurate alignments, MDA-generated errors, or *de novo* subclonal mutations. Out of these 3 categories, *de novo* subclonal mutations are presumed to be much more rare, and alignment artifacts can be replicated among independent samples. To estimate the error rate, we randomly separated the 56 LOH samples into two groups each containing 28 samples, and used the inferred haplotype from one group to evaluate the frequency of errors of each sample in the other group (Supplementary Fig. 5H). The average percentage of reads containing at least 1 error base is 1.2% (std. dev. 0.3%); on average, 1.9% of all covered bases contain at least 1 error read (std. dev. 0.5%). Due to the low sequencing depth, MDA-introduced errors (multiple error reads in one sample) cannot be distinguished from sequencing errors; however, the above estimate suggests that MDA error rate is at most comparable to sequencing errors (0.1-1%).

1. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012;150:1107-20.
2. Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM, et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science. 1989;244:217-21.
3. Kern SE, Fearon ER, Tersmette KW, Enterline JP, Leppert M, Nakamura Y, et al. Clinical and pathological associations with allelic loss in colorectal carcinoma [corrected]. Jama. 1989;261:3099-103.
4. Schutte M, Rozenblum E, Moskaluk CA, Guan X, Hoque AT, Hahn SA, et al. An integrated high-resolution physical map of the DPC/BRCA2 region at chromosome 13q12. Cancer Res. 1995;55:4570-4.
5. Vogelstein B, Fearon ER, Kern SE, Hamilton SR, Preisinger AC, Nakamura Y, et al. Allelotype of colorectal carcinomas. Science. 1989;244:207-11.