High clonal diversity and spatial genetic admixture in early prostate cancer and surrounding normal tissue

Nicola Crosetto (✉ nicola.crosetto@scilifelab.se)

Karolinska Institute <https://orcid.org/0000-0002-3019-6978>

Ning Zhang

Karolinska Institute

Luuk Harbers

Karolinska Institute <https://orcid.org/0000-0003-3910-6497>

Michele Simonetti

Karolinska Institute

Gabriel Longo

Institute of Molecular Biology (IMB) <https://orcid.org/0000-0003-2028-1068>

Enrico Berrino

Candiolo Cancer Institute

Niklas Schultz

Karolinska Institutet

Firas Tarish

Stockholm Southern Hospital

Peng Su

Qilu Hospital of Shandong University

Bo Han

Shandong University QiLu Medical College

Wanzhong Wang

Karolinska University Hospital

Sofia Onorato

Karolinska Institutet

Qifeng Yang

Qilu Hospital of Shandong University

Thomas Hellday

Karolinska Institutet

Vassilis Roukos

Institute of Molecular Biology <https://orcid.org/0000-0002-5065-3937>

Caterina Marchiò

Candiolo Cancer Center

Magda Bienko

Karolinska Institutet, SciLifeLab <https://orcid.org/0000-0002-6499-9082>

Biological Sciences - Article

Keywords: Copy number alterations (CNAs), human cancers, surrounding normal tissue, prostate cancer

Posted Date: November 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1093341/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Copy number alterations (CNAs) are pervasive in advanced human cancers, but their prevalence in early-stage, localized tumors and their surrounding normal tissues is poorly characterized. To investigate this phenomenon, here we developed a method for spatially resolved single-cell CNA profiling and applied it to characterize the CNA landscape in 10,007 nuclei extracted from 70 tumor and normal tissue regions (~125 mm³ tissue cubes) from prostatectomies performed in six patients with localized prostate cancer. We identified two distinct groups of cells with abnormal karyotype, one mainly consisting of sparse alterations ('pseudo-diploid' cells) and the other characterized by genome-wide karyotypic changes ('monster' cells). Pseudo-diploid cells displayed high clonal diversity and formed numerous small sized clones ranging from highly spatially localized to broadly spread clones, whereas monster cells were singular events detected throughout the prostate. We observed a remarkable correlation between the fraction of the genome affected by CNAs and the number of tissue regions in which pseudo-diploid cells were found. Highly localized pseudo-diploid clones were enriched in tumor regions and carried deletions of known or putative tumor suppressors, including APC, CDKN1B, FOXO1, FOXP1, and RB1. Spatially resolved targeted deep sequencing of 523 cancer genes detected non-synonymous mutations in both normal and tumor regions, including mutations in FOXA1, FOXP1, and SPOP genes previously implicated in prostate cancer. Strikingly, in two regions in which targeted deep sequencing detected a point mutation affecting the DNA-binding activity of the FOXA1 transcription factor, we also found a co-deletion of FOXO1 and FOXO3 genes in cells from two different pseudo-diploid clones, implicating combinatorial perturbations of Forkhead transcription factors as an early driver of prostate carcinogenesis. Our study reveals that CNAs and mutations are widespread across normal and tumor regions in the prostate glands of patients with localized prostate cancer and suggests that a subset of alterations—most likely small deletions causing the loss of key tumor suppressors—confer a fitness advantage and channel cells towards tumorigenesis.

Main

Chromosomal instability and copy number alterations (CNAs) are pervasive in human cancers^{1–3}. It has been proposed that CNAs form in the early stages of tumorigenesis in punctuated evolutionary bursts^{4–7}. Indeed, CNAs have been detected in pre-malignant lesions, such as in Barrett's esophagus years before the onset of invasive esophageal cancer⁸, and in oral leukoplakias undergoing progressive transformation to oral squamous cell carcinomas⁹. However, the exact timing and mechanisms of CNA formation during carcinogenesis remain poorly understood and a detailed map of CNA events in early-stage, localized tumors and their surrounding normal tissues is missing.

In prostate cancers, CNAs represent the main type of genomic alteration and are thought to be the major drivers of these tumors¹⁰. Early-stage localized prostate cancers are often characterized by the presence of multiple spatially confined tumor foci¹¹. Thus, they represent an ideal model to investigate the spatial distribution of CNAs in early-stage tumors and adjacent normal tissues. Multi-region sequencing of

medium-grade (Gleason 7) localized prostate cancers previously showed extensive intratumor heterogeneity between distinct tumor foci at the level of single-nucleotide variants (SNVs), CNAs and complex genomic rearrangements, suggesting an independent clonal origin of different foci¹². Polyclonal origin was also suggested by another study, in which SNVs were detected in morphologically normal prostate tissue distant from cancer lesions, consistent with so-called cancerization field effects¹³. However, these studies interrogated a small number of regions and lacked single-cell resolution, preventing an in-depth characterization of the diversity and spatial distribution of CNAs and mutations across the prostate gland.

In recent years, multiple single-cell DNA-sequencing (scDNA-seq) methods have been developed and applied to profile CNAs in single cells and to reconstruct the evolutionary history of different tumor types, including breast^{5,14–19}, colon²⁰, prostate^{21–23}, esophagus²⁴, and liver²⁵ cancers, as well as acute leukemias^{26,27}. Although greatly enabling, current scDNA-seq methods typically lack spatial resolution, with few exceptions^{18,19}. Here, we aimed at developing an agile and robust workflow for spatially resolved single-cell CNA profiling in tumor samples and applied it to obtain a comprehensive map of CNAs in prostatectomy samples from patients diagnosed with localized prostate cancer.

Development and validation of scCUTseq

To perform spatially resolved single-cell CNA profiling in clinically relevant tumor samples, we devised a method—scCUTseq—for preparing sequencing libraries from hundreds of nuclei extracted from dozens of frozen tissue cubes and then immediately fixed to stabilize genomic DNA and prevent batch effects (**Fig. 1a**). scCUTseq builds on the CUTseq method for highly multiplexed CNA profiling in tumor samples, which we previously developed²⁸, but features an extra step of whole genome amplification (WGA) by multiple annealing and looping based amplification cycles (MALBAC)²⁹ prior to cell indexing in order to increase library complexity (see step-by-step protocol in **Supplementary Methods** and at DOI: 10.21203/rs.3.pex-1602/v1). To minimize reagent costs, we scaled down the volume of MALBAC reagents by leveraging a nanodispensing device (I.DOT), which we previously used for high-throughput CUTseq²⁸ (**Methods**). Copy number profiles and genome coverage remained stable down to 1:200 MALBAC reagent volume scaling (sMALBAC) and were highly similar between libraries prepared from both fixed and non-fixed cells, either using a standard library preparation method or CUTseq following MALBAC (**Fig. 1b, Supplementary Fig. 1a-e and Supplementary Table 1**). Importantly, scCUTseq could clearly resolve the copy number profiles of different cell lines and was free of cross-contamination when different cell types were sorted in alternating columns of 96-well plates (**Fig. 1c and Supplementary Fig. 1f**), highlighting its specificity.

We next assessed the sensitivity of scCUTseq. To this end, we used CRISPR-Cas9 with two small guide RNAs (sgRNA) targeting, respectively, the *KMT2A* and *HYLS1* locus on chromosome (chr) 11 to engineer a 7 megabase (Mb) deletion between them in human TK6 lymphoblastoid cells (**Fig. 1d and Methods**). ~3% of the cells transfected with both sgRNAs acquired the expected 7 Mb deletion on one copy of chr11, while in ~3% of the cells the q-arm of chr11 downstream of the 7 Mb deletion was additionally either lost

or amplified, as assessed by quantitative high-throughput fluorescence in situ hybridization (FISH) (**Supplementary Fig. 2a-e and Methods**). We then fixed and performed scCUTseq on 746 TK6 cells, including 666 sgRNA-treated and 80 control cells, followed by sequencing on the Illumina NovaSeq 6000 platform. We detected a single copy of the exact 7 Mb deletion in 22 (3.3%) of the cells transfected with both sgRNAs, but in none of the control cells (**Fig. 1e, Supplementary Fig. 2f and Supplementary Table 1**). In 21 (3.2%) of the sgRNA transfected cells, the deletion extended until the end of the q-arm of chr11, while in 15 (2.3%) of them the chr11 q-arm downstream of the deletion was amplified, in line with the quantification by FISH. Clustering of individual copy number profiles using uniform manifold approximation and projection (UMAP)³⁰ revealed a major cluster corresponding to non-edited TK6 cells and a few small clusters mostly populated by cells that had acquired the expected 7 Mb deletion (**Fig. 1f, g**). In 523 (70.1%) of all the cells sequenced, scCUTseq also detected multiple CNAs unrelated to CRISPR editing, including a fully clonal chr13 trisomy and a subclonal amplification of the q-arm of chr3 (**Supplementary Fig. 2f**). Together, these results demonstrate that scCUTseq is a sensitive method for single-cell CNA profiling in fixed cells, which can also be used to study off-target genomic alterations in genome editing experiments.

scCUTseq detects the clonal architecture of breast cancer samples

Single-cell CNA profiling of breast cancer samples has previously revealed branched clonal structures^{14,15} consistent with a punctuated evolution model^{5,17}. To assess whether scCUTseq would detect the same type of clonal architecture in breast cancers, we applied it to fixed nuclei extracted from two frozen breast cancer surgical biopsies (B1 and B2) (**Supplementary Table 1 and Methods**). To automatically filter out low-quality copy number profiles, we trained a random forest classifier on 2,304 manually annotated single-cell profiles, reaching a very high classification accuracy (area under the curve, AUC: 99.3%) (**Supplementary Fig. 3a-d, Supplementary Table 2, and Methods**). Using the fraction of cells in B1 and B2 that contained CNAs as a proxy of the percentage of cancer cells in these tumors, we estimated the tumor purity to be 62% and 71% in B1 and B2, respectively (**Fig. 2a-c and Supplementary Fig. 4a-c**). Both tumors displayed a well-defined clonal organization consisting of clones and super-clones and a branched phylogenetic tree structure, in line with the findings of a recent study performed on eight triple-negative breast cancers using a different scDNA-seq method¹⁷ (**Fig. 2d-g and Supplementary Fig. 4d-h**). Importantly, the median copy number profiles of the identified clones mirrored the profile obtained by applying bulk CUTseq to the same samples, even though individual single-cell profiles captured subtle differences that were washed out in the bulk sample (**Fig. 2g and Supplementary Fig. 4g**). These results indicate that scCUTseq can be applied to assess the tumor purity and reconstruct the clonal architecture of clinical tumor samples.

Spatially resolved single-cell CNA profiling across prostate specimens harboring tumor foci

Having implemented and validated scCUTseq, we then sought to apply it to portray the prevalence and spatial distribution of CNAs in prostate tissues from patients diagnosed with localized prostate cancer. To this end, we prospectively collected whole prostate samples from six patients undergoing endoscopic

prostatectomy following a diagnosis of localized prostate cancer (**Supplementary Table 3 and Methods**). For each sample, we obtained a transversal section of the entire prostate and sliced it into multiple ~125 mm³ tissue cubes (herein named 'regions') (**Fig. 3a**). We classified each region as tumor (>50% of tumor cells), focal (10–50% of tumor cells) or normal based on morphological examination of one H&E-stained tissue section from the top and one from the bottom of each cube by two board-certified pathologists independently (**Supplementary Fig. 5 and 6**). In two cases (samples P2 and P5), we isolated nuclei from 34 and 28 regions, respectively, whereas for each of the remaining four samples (P1, P3, P4, P6) we isolated nuclei from two regions. In total, we obtained high-quality profiles for 10,007 cells from 70 regions and generated spatial heatmaps showing the distribution of cells with at least one CNA in P2 and P5 (**Fig. 3b-g, Supplementary Fig. 7a, b, and Supplementary Table 1**). To our surprise, in both samples, scCUTseq detected the presence of CNA-harboring cells in all the regions examined—not only in tumor or focal regions—although, as expected, the CNA frequency was higher in tumor regions while, conversely, diploid cells were more frequently detected in normal regions (**Fig. 3b-g and Supplementary Fig. 8a-d**).

To explore the diversity of CNAs in these samples, we visually inspected all the copy number profiles from P2 and P5, which revealed the existence of three major groups of cells (**Fig. 3h**): (i) diploid cells; (ii) 'pseudo-diploid' cells with an altered copy number profile harboring multiple sparse alterations mainly consisting of deletions; and (iii) 'monster' cells marked by genome-wide copy number alterations, mainly consisting of whole-chromosome amplifications at varying levels, often alternating with deletions, reminiscent of the genomic landscape of so-called 'hopeful monsters' recently described in colorectal cancer⁷. Pseudo-diploid cells accounted for 35.2% and 42.3% of all the cells sequenced in P2 and P5, respectively, and were more frequently detected in tumor regions (**Fig. 3h and Supplementary Fig. 8e, f**). In contrast, monster cells accounted for a lower percentage of all the cells (5.8% and 7.5%, respectively, in P2 and P5) and were detected at similar frequencies in all the regions examined (**Fig. 3h and Supplementary Fig. 8g, h**). Notably, we detected similar proportions of diploid, pseudo-diploid, and monster cells in the other four prostate samples profiled by scCUTseq (**Supplementary Fig. 9a, b**). These results indicate that CNAs are ubiquitous in prostates from patients with localized prostate cancer and that pseudo-diploid and monster cells co-exist in these tissues, most likely reflecting different mechanisms of CNA formation.

High clonal and spatial heterogeneity of pseudo-diploid cells

Although the CNA profiles in the prostate samples did not display any obvious clonal structure, we hypothesized that some degree of clonal organization might exist among pseudo-diploid cells. To this end, we applied UMAP to the CNA profiles of the pseudo-diploid cells identified in the P2 and P5 samples and identified numerous (51 and 48, respectively) small sized clones (range: 14–156 and 7–69 cells per clone in P2 and P5, respectively) forming together 3–4 larger superclones (**Fig. 4a and Supplementary Fig. 10a**). The number of clones identified in prostate samples is considerably higher than the number of clones in the breast cancer samples B1 and B2 profiled by scCUTseq (13 in both cases). Accordingly, the phylogenetic trees of the P2 and P5 samples consisted of many parallel branches of similar length (**Fig. 4b, c and Supplementary Fig. 10b, c**). In contrast, the B1 and B2 trees displayed 2–3 major branches,

each with multiple smaller branches and sprouts (**Fig. 2f and Supplementary Fig. 4f**). These results are consistent with different biological processes and evolutionary strategies underpinning the evolution of prostate and breast cancers, as reflected in their profoundly different genomic landscapes^{31,32}.

We then wondered how the cells belonging to different pseudo-diploid cell clones are spatially distributed. Most of the clones mapped to more than one region and the clonal composition markedly differed between regions, in both the P2 and P5 sample (**Supplementary Fig. 11a, b**). Some clones (C), such as C6 in P2, were the dominant clone found in a single region, whereas other clones, such as C28 in P5, ranked among the top-3 most frequent clones in multiple regions (**Supplementary Fig. 11a, b**). To gain further insights into the spatial distribution of different clones, we defined a locality index by computing the fraction of regions in which the cells belonging to a given clone were detected (**Methods**). For both P2 and P5, the locality index distribution formed a continuum ranging from values close to 0 for broadly diffused clones to values close to 1 for highly localized clones (**Fig. 4d, e, Supplementary Fig. 10d, e, Supplementary Fig. 12 and 13, and Supplementary Table 4**). The mean locality index was higher for tumor compared to normal regions and, accordingly, all the five most localized clones in P2 and 4 out of the 5 most localized clones in P5 had most of their cells present in regions classified as tumor or focal, where they also represented the dominant clones (**Fig. 4f, Supplementary Fig. 10f, Supplementary Fig. 11a, b, and Supplementary Table 4**). These results suggest that highly localized pseudo-diploid cell clones might contribute to driving the tumorigenic process inside tumor regions.

Monster cells are singular events detected throughout the prostate

Next, we turned our attention to monster cells. In both P2 and P5, the CNA landscape of these cells was dominated by large amplifications, often affecting the entirety or a large portion of every chromosome, including chrX, where the androgen receptor (AR) gene locus is located (**Supplementary Fig. 14a, b**). We applied UMAP to the segmented copy number profiles and identified several clusters, including clusters almost exclusively composed of cells with 4–5 copies of each chromosome (*e.g.*, cluster 1 in P2) as well as clusters featuring a mix of large amplifications alternating with smaller deletions (*e.g.*, cluster 6 and 3 in P2 and P5, respectively) (**Fig. 4g-k and Supplementary Fig. 14c-h**). However, unlike in the case of pseudo-diploid cells, we did not manage to identify clones of monster cells sharing highly similar CNA profiles, indicating that these cells might represent singular events that most likely fail to propagate due to incompatibility of their highly rearranged karyotype with cell proliferation.

We then checked the spatial distribution of monster cells across all the regions in P2 and P5. Independently of the UMAP cluster to which they were assigned, we detected the presence of monster cells in both tumor and normal regions, although with seemingly different patterns between different clusters (**Fig. 4l-r and Supplementary Fig. 14i-l**). For example, monster cells assigned to cluster 6 in P2—the largest cluster in this sample—were homogeneously distributed across all the regions (**Fig. 4r**), while the cells in P2 cluster 4 and 5 and in P5 cluster 1 displayed a more heterogeneous spatial pattern (**Fig. 4p, q and Supplementary Fig. 14j**). These results indicate that monster cells with highly aberrant karyotypes are ubiquitous in the prostate glands of patients diagnosed with localized prostate cancer and can be

found in normal prostate tissue regions distant from tumor foci, most likely reflecting the presence of ongoing genomic instability in these tissues.

Tumor suppressors are deleted in localized pseudo-diploid clones

The finding that highly localized pseudo-diploid cell clones were enriched in tumor regions, while the other pseudo-diploid clones were spread throughout the prostate gland, made us wonder whether the former might somehow contribute to tumorigenesis. To tackle this question, we examined whether the CNA profiles differed between highly localized and widespread pseudo-diploid clones. Intriguingly, we observed that highly localized pseudo-diploid clones seemed to harbor smaller alterations (usually deletions) compared to more spread clones, which in most cases carried one or more chromosome-wide alterations (**Supplementary Fig. 12 and 13**). Indeed, we found a striking inverse correlation between the locality index and the mean size of genomic deletions, in both P2 and P5 (**Fig. 5a, b**). Cells belonging to the five most localized pseudo-diploid clones in P2 and P5 carried 1–10 deletions ranging from 1.7 to 58.2 megabases (Mb) (43.0 ± 41.6 Mb, mean $\pm$ s.d.), whereas cells belonging to the five most spread pseudo-diploid clones in the same samples contained a lower number (1–5) of larger deletions (123.8 ± 71.0 Mb, mean $\pm$ s.d.) (**Fig. 5a, b and Supplementary Fig. 12 and 13**). We therefore hypothesized that the small deletions in the highly localized pseudo-diploid clones might cause the loss of key tumor suppressor genes, thus contributing to the tumorigenic process in the tumor regions where they were detected. To test this hypothesis, we checked which genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC)³³ were amplified or deleted in the five most localized pseudo-diploid clones identified in the P2 and P5 samples (**Methods**). We found that several known or putative tumor suppressor genes were deleted in one or more pseudo-diploid clones, including classical tumor suppressors such as *RB1*, *APC* and *CDKN1B* (**Fig. 5c, d and Supplementary Table 5**). Among the putative tumor suppressor genes deleted in sample P2, we identified several members of the Forkhead family of transcription factors—which are involved in numerous physiological processes and many of which have been shown to act as tumor suppressors in different tissue contexts, including *FOXO1*, *FOXO3*, and *FOXP1*³⁴. The latter encodes for a transcription factor that is part of the AR regulatory network together with FOXA1—another Forkhead transcription factor frequently mutated in prostate cancers³⁵—and was shown to exert tumor-suppressive effects in the prostate³⁶. Notably, *FOXO1* and *FOXO3* were co-deleted in 52 (96.3%) and 23 (100%) of the cells belonging to the C6 and C17 localized pseudo-diploid clones in P2, respectively, and these cells were distributed across five adjacent regions (3 tumor and 2 focal) at the prostate mid-section left-anterior pole, and in one tumor region on the opposite side (**Fig. 5e-i**). *FOXP1* was deleted in 14 (93.3%) of the cells belonging to another pseudo-diploid clone (C8) localized in one focal region adjacent to a tumor region containing cells with *FOXO1* and *FOXO3* co-deleted (**Fig. 5j**). These findings suggest that loss of these genes in the same or spatially close cells from the same region in the prostate gland might be an early event during prostate carcinogenesis.

Mutations in cancer genes are detected in tumor and normal regions

Finally, we wondered whether the observed spatial patterns of CNAs are reproduced at the level of somatic single-nucleotide variants (SNVs). To this end, we performed targeted deep sequencing (656 average depth) of 523 cancer genes (Illumina TSO500 panel) using genomic DNA (gDNA) extracted from each tissue region in the P2 sample and using peripheral blood gDNA from the same patient as reference (**Supplementary Methods**). Consistent with the notion that prostate cancer is CNA- rather than mutation-driven, we detected a relatively small number (55) of non-synonymous SNVs in 35 (6.7%) of the targeted genes, with an average variant allelic frequency of 3.7% (range: 1.1%–11.8%) (**Fig. 5k and Supplementary Table 6**). Each SNV was detected in 1–3 regions, except for *NOTCH3* mutations, which were detected in 6 regions, including 5 (83%) regions classified as normal (**Supplementary Fig. 15**). *SPOP*—a gene frequently mutated in prostate cancers^{32,37}—was mutated in three tumor regions and one focal region at the left-anterior corner of the prostate midsection examined, while *FOXP1* was mutated in a normal region (**Fig. 5l, m**). In one tumor and in one focal region of P2, where we found cells harboring a co-deletion of *FOXO1* and *FOXO3*, as well as in two normal regions on the opposite side of the prostate mid-section analysed, we also detected a mutation (p.R219A) in the *FOXA1* gene, which was previously shown to affect the DNA-binding and pioneering activity of the FOXA1 transcription factor and perturb normal luminal epithelial differentiation programs in the prostate³⁵ (**Fig. 5n**). Consistent with these observations, we found that 5.9% of prostate cancers in TCGA harbor a dual deletion of *FOXO1* and *FOXO3*, and that 3.9% of the tumors carry any combination of *FOXO1*, *FOXO3* or *FOXP1* deletions together with *FOXA1* mutations (**Fig. 5o**), suggesting that combinatorial perturbations of Forkhead transcription factor regulatory networks by CNAs and SNVs might drive tumorigenesis in a subset of patients with prostate cancer.

Conclusions

Leveraging scCUTseq to perform spatially resolved single-cell profiling of CNAs across normal and tumor regions in six prostates from patients with localized early-stage prostate cancer, we have discovered the existence of two major types of genomically altered cells—pseudo-diploid and monster cells—characterized by highly diverse CNA profiles and forming spatially admixed, heterogeneous spatial patterns. Monster cells are reminiscent of ‘hopeful monsters’, which were recently shown to emerge continuously in propagating colorectal cancer organoid cultures as a result of catastrophic mitotic errors⁷. We detected monster cells and pseudo-diploid cells carrying chromosome-wide alterations (typically deletions) in all the prostate tissue regions examined, including normal and tumor regions, suggesting that they originate from different, yet widespread genome instability processes that are active throughout the prostate gland. Nonetheless, the fact that pseudo-diploid cells—but not monster cells—form different clones with similar CNA profiles indicates that CNAs in these cells are still compatible with cell proliferation and thus can be propagated, unlike the highly aberrant karyotypes of monster cells, representing singular events that cannot be propagated.

In addition to spatially diffused monster and pseudo-diploid cells, we have revealed the existence of a small population of pseudo-diploid cells harboring one or more focal alterations (typically deletions)

localized in one or a few tumor regions. Notably, in one (P2) of the two prostate samples that we extensively profiled, we uncovered a combination of monozygous deletions of the *FOXO1* and *FOXO3* genes co-existing in the same cells inside one tumor and one focal region in which we also detected a specific mutation (p.R219A) in the *FOXA1* gene, which was previously shown to alter gene expression programs in the prostate epithelium³⁵. *FOXO1*, *FOXO3*, and *FOXA1* (as well as *FOXP1*, which was also deleted in a different region in the same P2 sample) are members of the large family of Forkhead transcription factors, involved in multiple physiological and pathological processes³⁴, and deregulation of *FOXA1* and *FOXP1* by mutations and deletions, respectively, was shown to drive prostate carcinogenesis^{38,39}. We speculate that combinatorial perturbations in Forkhead transcription factors might be an early event occurring at multiple, spatially distinct locations in the prostate gland, in a subset of patients with prostate cancer, explaining our finding of the co-existence of *FOXA1* mutations and cells with co-deletion of *FOXO1* and *FOXO3* at multiple distant sites in the prostate, including normal regions.

The Gleason grading system represents the main tool used for predicting the prognosis and guiding the therapy of patients affected by prostate cancer⁴⁰. This system, however, is blind to the genomic makeup of the cells analyzed and to spatial genetic heterogeneity across different regions of the prostate, which likely encode important prognostic information. Instead, the combined spatially resolved single-cell CNA profiling and targeted deep sequencing approach that we have implemented here can capture genetic diversity and spatial heterogeneity throughout the prostate in an unbiased manner, providing a quantitative portray of the type, relative proportions, and spatial distribution of cells with different karyotypes as well as of potential cancer driver mutations. We propose using this approach to derive a 'Genomic Diversity Index', which could then be integrated with the Gleason score to improve prognostication for patients affected by prostate cancer.

The scCUTseq method that we have developed and applied here is highly versatile and particularly suited for profiling CNAs across many regions of frozen tumor biopsies, since extracted cells or nuclei can be first stabilized by mild fixation and stored up to several weeks before being sorted and used to prepare sequencing libraries. This approach has not been demonstrated with other scDNA-seq methods, which instead require that sorting is performed immediately after the isolation of single cells or nuclei to avoid DNA degradation. One inherent limitation of scCUTseq is the low (~1%) breadth of genome coverage that is achieved due to the use of restriction enzymes to fragment the genome in CUTseq²⁸. However, we anticipate that a different cell barcoding and library preparation method, such as tagmentation, could be used following WGA to achieve higher genome coverage. Importantly, the scaled down version of MALBAC, which we have implemented here, is highly cost-effective (see Cost Analysis in **Supplementary Notes**) and could be translated to other WGA methods, such as multiple displacement amplification (MDA), and combined with tagmentation to enable cost-effective single-cell SNV profiling.

Methods

Samples

Cell lines. We purchased IMR90, SKBR3 and MCF10A cell lines from the American Tissue Culture Collection (ATCC, cat. no. CCL-186, HTB-30, and CRL-10317, respectively) and Drosophila S2 cells from Gibco (cat. no. R69007). TK6 cells were previously described⁴¹. None of these cell lines is registered in the International Cell Line Authentication Committee (ICLAC) database of misidentified cell lines. We periodically tested all the cell lines for mycoplasma contamination using the MycoAlert Mycoplasma Detection Kit (Lonza, cat. no. LT07-118). Culturing conditions were as follows: (i) IMR90 cells: Eagle's Minimum Essential Medium (EMEM) (Sigma, cat. no. M5650) supplemented with 10% heat-inactivated fetal bovine serum (FBS; Sigma, cat. no. F9665), 2 mM L-glutamine (Sigma, cat. no. 59202C), and 1% non-essential amino acids (Gibco, cat. no. 11140035); (ii) SKBR3 cells: McCoy's 5A Medium (Sigma, cat. no. M9309) supplemented with 10% heat inactivated FBS (Sigma, cat. no. F9665); (iii) MCF10A cells: Mammary Epithelial cell Growth Medium (MEGM) (Lonza, cat. no. CC-3150) supplemented with 100 ng/ml cholera toxin (Sigma, cat. no. C8052) according to the ATCC guidelines; (iv) S2 cells: Schneider's medium (Gibco, cat. no. 21720024) supplemented with 10% heat inactivated FBS (Sigma, cat. no. F9665); (v) TK6 cells: Roswell Park Memorial Institute Medium (GIBCO, cat. no. 11875093) supplemented with 5% horse serum (GIBCO, cat. no. 11510516), 1 mM sodium pyruvate (GIBCO, cat. no. 11360070), 100 U/mL penicillin-streptomycin (GIBCO, cat. no. 15140122) and 2 mM L-glutamine (GIBCO, cat. no. 25030081). We grew all the cell lines at 37 °C in 5% CO₂ air, except for S2 cells, which were grown at 28 °C in a non-humidified, ambient air-regulated incubator.

Breast cancer samples. We retrospectively collected $n=2$ frozen breast cancer specimens at the Pathology Unit of the Candiolo Cancer Institute (Italy). The samples had been previously collected and stored in the frame of a prospective study for the identification of molecular profiles conferring resistance to selected target therapies in oncological patients approved by the local Ethical Committee ("Profiling", 001-IRCC-00IIS-10). Following excision in the surgical theater, the samples were immediately placed in vacuum and stored them at +4 °C until further processing. ~ 110.5 cm³ tissue blocks were embedded in Tissue-Tek OCT compound (VWR, cat. no. 00411243), snap-frozen in isopentane, and stored at -80 °C.

Prostate cancer samples. We prospectively collected six prostatectomy samples from patients operated for localized prostate cancer by endoscopic surgery at the Södersjukhuset hospital in Stockholm, Sweden. The study was approved by the regional Ethical Committee (ethical permit #2018/1003-31). The main pathological characteristics of each tumor are summarized in **Supplementary Table 3**. For each tumor, we obtained a ~ 0.5 cm thick transversal mid-section of the prostate and then immediately cut it into multiple $\sim 0.50.50.5$ cm³ tissue blocks. We embedded each block in Tissue-Tek OCT compound (VWR, cat. no. 00411243), snap-froze it in isopentane, and stored it at -80 °C. We stained with hematoxylin-eosin (H&E) one section from the top and one from the bottom of each block, and then scanned each section using a Hamamatsu Nano Zoomer-XR Digital slide scanner. Two board-certified pathologists at Qilu Hospital of Shandon University, China independently evaluated each tissue section and performed Gleason grading following the International Society of Urological Pathology (ISUP) 2014/World Health Organization 2016 modified system⁴². Accordingly, regions containing more than 50% of tumor cells were classified as 'Tumor', whereas regions containing less than 50% of tumor cells

(typically, 10–20%) were labeled as ‘Focal’. All the prostate and breast cancer samples described above are unique biological samples the use of which is restricted to this study and that cannot be distributed to other researchers.

scCUTseq

A detailed step-by-step protocol is available at DOI: 10.21203/rs.3.pex-1602/v1.

Cells and nuclei preparation. We prepared cultured cells as described for standard MALBAC in the **Supplementary Methods**. For tumor samples, we prepared single nuclei suspensions following the Tapestry frozen tissue nuclei extraction protocol (<https://missionbio.com>). Briefly, we first prepared a spermine solution composed of 3.4 mM sodium citrate tribasic dihydrate (Sigma, cat. no. C8532)/1.5 mM spermine tetrahydrochloride (Sigma, cat. no. S1141)/0.5 mM tris (hydroxymethyl) aminomethane (Sigma, cat. no. 252859)/0.1% v/v IGEPAL CA-630 (Sigma, cat. no. I8896) and stored it at +4 °C. We retrieved the frozen breast cancer or prostate cancer tissue blocks and kept them on dry ice until further processing. We then transferred one sample at a time into a pre-chilled Petri dish placed onto a dry-ice pad, added 200 µL of tissue lysis solution containing spermine solution (pH 7.6)/0.03 mg/mL trypsin-EDTA (0.25%), phenol red (Thermo Fisher scientific, cat. no. 25200072)/0.1 mg/mL collagenase (Worthington, cat. no. CLS-7 LS005332)/1 mg/mL Dispase II (Gibco, cat. no. 17105-041) on top of the tissue block and incubated it until the tissue lysis solution had frozen (~3 min). We minced the sample thoroughly with two pre-chilled sterile scalpels and then transferred it to room temperature, while continuing mincing until all the tissue fragments could flow through a 1 mL pipette tip without clogging it. We added additional 1.8 mL of tissue lysis solution to rinse the dish, transferred all the tissue fragments from the Petri dish into a clean 5 mL low-binding tube (Eppendorf, cat. no. 0030108.310), and rotated the tube at 20 rpm at room temperature for 15 min. We stopped the lysis by adding 2 mL of stop solution containing spermine solution (pH 7.6)/0.5 mg/mL trypsin inhibitor from chicken egg white, type II-O (Sigma, cat. no. T9253)/ 0.1 mg/mL ribonuclease A from bovine pancreas, type I-A (Sigma, cat. no. R4875-100mg) and mixed the solution by gently inverting the sample. Next, we filtered the solution through a 50 mm cell strainer and collected the flowthrough in a clean 5 mL low-binding tube. We centrifuged the flowthrough at 300g for 5 min at room temperature and discarded the supernatant. To fix the nuclei, we resuspended them in 400 µL of a nuclei fixation solution containing 66% vol./vol. methanol/33% vol./vol. acetic acid pre-chilled at +4 °C and thoroughly pipetted the nuclei suspension up and down. After 15 min incubation on ice, we centrifuged the sample at 300g for 5 min at room temperature to collect the fixed nuclei. We discarded the supernatant and resuspended the nuclei in 1X PBS/5 mM EDTA/0.05% NaN₃ and filtered the nuclei through a 20 mm cell strainer. We have successfully used fixed nuclei prepared in this manner, that were kept at 4 °C several weeks up to two months.

Preparation of CUTseq adapters. We used either 96 CUTseq adapters previously described²⁸ or 384 adapters that we newly designed. The sequences of all the oligonucleotides (oligos) used are available in **Supplementary Table 7**. We purchased the oligos from Integrated DNA Technologies (IDT) as standard desalted oligos at 100 µM in Nuclease-Free Water in 96-well plates. In each well of a 96- (for 96 adapters)

or 384-well plate (for 384 adapters), we dispensed 5 μ L of a sense oligo and then added 40 μ L of phosphorylation reaction mix containing 1 μ L of T4 Polynucleotide Kinase (PNK; NEB, cat. no. M0201S), 5 μ L of T4 PNK buffer (NEB, cat. no. M0201S) and 5 μ L of 10 mM ATP (Thermo Fisher Scientific, cat. no. PV3227) in Nuclease-Free Water and incubated at 37 °C for 30 min followed by inactivation at 65 °C for 20 min. Then, we added 5 μ L of the corresponding antisense oligo to each well and annealed the complementary sense and antisense oligos by incubating the plates at 95 °C for 5 min, followed by cooling down to 25 °C over a period of 45 min in a PCR thermocycler. Afterwards, we diluted the annealed adapters to 33 nM in Nuclease-Free Water and stored them at –20 °C. Adapter dilutions stored in this way are stable for 6–8 months.

Single nucleus sorting. Before sorting, we manually dispensed 5 μ L of Vapor-Lock (Qiagen, cat. no. 981611) into each well in the targeted region of a 384-well plate. We then transferred the samples into FACS-compatible tubes, added 2.46 ng/mL Hoechst 33342 (Thermo Fisher Scientific, cat. no. 62249) to them and incubated for 2 min on ice in darkness, while rotating. We sorted single nuclei in 96- or 384-well plates (one nucleus per well in 10 nL sorting volume) using the BD FACSJazz Cell Sorter (BD Biosciences) based on forward and side scatter properties. After sorting, we immediately sealed the plates, spun them at 3,220g for 5 min and stored them at –20 °C until further processing.

Scaled down MALBAC. We used the MALBAC kit (Yikon Genomics, cat. no. Y001A) for whole genome amplification and the I.DOT nanodispensing system (CELLINK) for dispensing all reagents in small volumes, to increase the efficiency of reactions while limiting the reagent costs per cell (see Cost Analysis in **Supplementary Notes**). For the cell lysis step, we prepared a lysis mix containing 30 nL of cell lysis buffer and 0.6 nL of cell lysis enzyme per cell. We then dispensed 30 nL of the lysis mix into each well and incubated the plates at 50 °C for 1 h and 80 °C for 10 min in a PCR thermocycler. For the pre-amplification step, we prepared a pre-amplification mix containing 150 nL of pre-amplification buffer and 5 nL of pre-amplification enzyme mix per cell. We then dispensed 150 nL of the pre-amplification mix into each well, and incubated the plates at 94 °C for 3 min to denature DNA, followed by 8 cycles of quasi-linear amplification (20 °C for 40 sec; 30 °C for 40 sec; 40 °C for 30 sec; 50 °C for 30 sec; 60 °C for 30 sec; 70 °C for 4 min; 95 °C for 20 sec; and 58 °C for 10 sec) in a PCR thermocycler. For the amplification step, we prepared an amplification mix containing 150 nL of amplification buffer, 4 nL of amplification enzyme mix and 11 nL of 4x SYBR Green (Thermo Fisher Scientific, cat. no. S7563) per cell, and kept the mix protected from light until we dispensed it. We dispensed 160 nL of the amplification mix into each well to reach a final volume of 350 nL per well and incubated the plates at 94 °C for 30 sec to denature DNA followed by 14 cycles of exponential amplification (94 °C for 20 sec; 58 °C for 30 sec; 72 °C for 3 min) in a PCR thermocycler. After each dispensing step, we shook the plates on a ThermoMixer at 1,000 rpm and then centrifuged the plates at 3,220g for 10 min and placed them on ice until the next step.

High-throughput CUTseq. We performed CUTseq on the MALBAC products largely following the high-throughput CUTseq protocol based on the I.DOT system (CELLINK), which we described before²⁸. Briefly, for the digestion step, we prepared a digestion mix containing 100 nL of NlaIII restriction enzyme (NEB, cat. no. R0125L) and 50 nL of CutSmart buffer (NEB, cat. no. R3104) per cell. We then dispensed 150 nL

of the digestion mix into each well and incubated the plates at 37 °C for 30 min followed by 65 °C for 20 min to inactivate the enzyme. After digestion, we dispensed 300 nL of 33 nM scCUTseq adapters (prepared as described above) into each well (one barcode per cell), followed by 700 nL of a ligation mix containing 200 nL of T4 rapid DNA ligase (Thermo Fisher Scientific, cat. no. K1423), 300 nL of T4 ligase buffer (Thermo Fisher Scientific, cat. no. K1423), 120 nL of 10 mM ATP (Thermo Fisher Scientific, cat. no. PV3227), 30 nL of 50 mg/mL bovine serum albumin (Thermo Fisher Scientific, cat. no. AM2616), and 50 nL of Nuclease-Free Water (Thermo Fisher Scientific, cat. no. 4387936). We then incubated the plates at 22 °C for 30 min, after which we dispensed 5 µL of Nuclease-Free Water/33 mM EDTA (for a final concentration of 25 mM) into each well and pooled the contents of all the wells of each plate by placing the plates upside down onto a collection plate covered with parafilm (NUNC, cat. no. 267060) and centrifuged them at 117g for 1 min. We carefully transferred the collected solution into a 5 mL low-binding tube and then removed as much as possible of the Vapor-Lock. We then purified the pooled samples with a 1.2 vol./vol. ratio of the sample and Agencourt Ampure XP bead suspension (Beckman Coulter, cat. no. A63881). After purification, we prepared 250 ng of DNA and sheared it using Covaris ME220 Focused-ultrasonicator with a target peak set at 200 base pairs (bp). To prepare sequencing libraries, we followed the CUTseq protocol²⁸, with the following minor modifications: (1) We increased the PCR volume to 200 µL and performed a split PCR reaction in 4-tubes strips; (2) We purified the final library with a 0.8 vol./vol. ratio of sample and Agencourt Ampure XP bead suspension and eluted the purified library in 50 µL of Nuclease-Free Water.

Sequencing and data pre-processing

We sequenced all the libraries on the NextSeq 500 platform (Illumina) using the high output 75 cycles single-end kit (Illumina, cat. no. FC-404-2005). For large-scale experiments (sequencing of TK6 cells treated with CRISPR/Cas9 and of breast and prostate tumor samples), we sequenced high-quality libraries (pre-checked on Bioanalyzer) on the NovaSeq 6000 platform (Illumina) at the National Genomics Infrastructure in Stockholm, Sweden. After each sequencing run, we demultiplexed raw sequence reads to fastq files using the BaseSpace Sequence Hub cloud service of Illumina. In the case of scCUTseq, each fastq file typically contains 384 single cells. We further demultiplexed each fastq file using a custom Python script. In short, we extracted cellular barcodes (or sample specific barcodes in the case of bulk CUTseq) and UMIs from each read at the specified nucleotide locations. Following this, we matched the barcodes to a list of predefined barcodes allowing for two mismatches if using a set of 384 adapters, or one mismatch if using a set of 96 adapters. We then re-wrote the reads to cell-specific fastq files with the barcode and UMI appended to the read name. For sequence reads that extended through adapters, we trimmed the adapters from cell (or sample) specific fastq files using *fastp* (version 0.20.1)⁴³. We aligned the fastq files to the hg19 human reference genome using *bwa-mem* (version 0.7.17-r1188)⁴⁴ or to the dm6 *Drosophila melanogaster* reference in the case of S2 cells, and subsequently sorted and indexed using *samtools* (version 1.10)⁴⁵. We moved barcodes and UMI tags from the bam file read header to the tags using a custom Python script. Finally, we deduplicated the reads based on the UMI tag and read

position using *umi-tools* (version 1.1.1)⁴⁶. All the procedure described above is fully automated and streamlined using *snakemake* (version 5.30.1)⁴⁷.

Copy number calling

We counted reads in genomic bins of variable length (average length: 100 kb for bulk CUTseq; 250 kb for TK-6 scCUTseq; and 500 kb for all other scCUTseq datasets), taking into account mappability and GC-content as described before⁴⁸. Next, we filtered out read count blacklisted regions, normalized the read counts by GC content and segmented them using *DNAcopy* (version 1.66.0)⁴⁹. We calculated integer copy numbers and assigned them using a sum of least squares approach. In bulk CUTseq integer copy numbers cannot be assigned, therefore we computed the log2 ratio of the read counts instead.

Copy number quality control using a Random Forest

To include only high-quality copy number profiles of single cells in our analyses, we trained a Random Forest⁵⁰ based on 16 different copy number profile features (see **Supplementary Table 2**), using the profiles of 2,304 single cells. In short, we manually annotated 2,304 scCUTseq profiles as high or low quality. We trained a random forest on 80% of these cells using the R package *randomForest* (version 4.6-14) with `ntree = 500` and `importance = TRUE`. To assess the performance of the random forest we used the remaining 20% of the cells that were not used in the initial training as a validation set. Based on a receiver operating characteristic curve (ROC), we selected the most optimal threshold to classify the single-cell copy number profiles.

Cell classification in tumor samples

For phylogenetic analyses of B samples, we filtered diploid and highly aberrant cells from both B samples based on setting two thresholds of the mean integer copy number (we retained B1 cells if the mean integer copy number was higher than 2.15 and lower than 2.5; for B2 cells, we retained them if the mean integer copy number was higher than 2.05 and lower than 1.95). For prostate cancer samples, we classified the cells as diploid if less than 30 genomic bins were altered and as ‘hopeful monsters’ if the mean integer copy number was larger than 2.1. We classified cells that were neither assigned as diploid or hopeful monsters as ‘pseudo-diploid since they carried few focal or chromosome arm-level alterations.

Clone identification in tumor samples

We embedded single-cell integer copy number profiles of filtered tumor cells from breast cancer (B) samples and pseudo-diploid cells from prostate (P2 and P5) samples using UMAP with the R package *uwot* (version 0.1.10)³⁰. We performed UMAP embedding with the following sample-specific settings: B1: `distance = "manhattan"`, `seed = 59`, `min_dist = 0`, `n_neighbors = 40`; B2: `distance = "manhattan"`, `seed = 26`, `n_neighbors = 4`; P2: `distance = "manhattan"`, `min_dist = 0`, `seed = 678`, `n_neighbors = 20`; P5: `distance = "manhattan"`, `min_dist = 0`, `seed = 678`, `n_neighbors = 12`. We identified superclones from UMAP using a shared nearest neighbor graph with the R package *scrn* (version 1.20.1)⁵¹. We used the following k

values: B1: 25; B2: 25; P2: 50; P5: 40. We assigned superclones using the R package *igraph* (version 1.2.6)⁵². Following this, we identified subclones using *hdbscan* from the R package *dbscan* (version 1.1-8)⁵³ with the following sample-specific parameters: B1: minPts = 11; B2: minPts = 5; P2: minPts = 13; P5: minPts = 7. We discarded from the clonal analysis cells that were unable to form a distinct cluster with a non-diploid median copy number profile.

UMAP on monster cell CNA profiles

We clustered cells that were classified as monsters using UMAP as described above. We first embedded the cells using UMAP with the following parameters: P2: distance = "manhattan", seed = 678, min_dist = 0, n_neighbors = 3; P5: distance = "manhattan", seed = 678, min_dist = 0, n_neighbors = 3. Subsequently, we clustered the cells using *hdbscan* from the R package *dbscan* (version 1.1-8)⁵³ with the following parameters: P2: minPts = 14; P5: minPts = 14.

Phylogenetic analysis

We constructed phylogenetic trees using the R packages *ape* (version 5.5)⁵⁴ and visualized them using *ggtree* (version 3.04)⁵⁵. We then calculated pairwise Manhattan distances of integer copy numbers of single cells and computed a balanced minimum evolution tree. We simulated the diploid tree root by setting an integer copy number of two for all chromosomes (except chromosome X in prostate cancer samples). Following this, we calculated the consensus copy number profile of the subclones by taking the median copy number of all the cells assigned to the same subclone. We computed phylogenetic trees for consensus profiles and visualized them in the same way as described above for single cells.

Locality index computation

To quantify the spatial distribution of pseudo-diploid clones and hopeful monster clusters in prostate samples, we defined a locality index as $1 - (\text{\# of regions in which the clone/cluster was present} / \text{total \# of regions profiled by scCUTseq})$. A locality index close to 1 indicates a highly localized clone while a locality index close to 0 indicates widespread spatial distribution.

TCGA data analysis

We downloaded segmented copy number profiles of Prostate Adenocarcinoma samples ($n=499$, TCGA, Firehose Legacy) from the cBioPortal⁵⁶ (<https://www.cbioportal.org/>). We classified patients that had a Gleason score equal to or lower than 7 as "Low Gleason" and patients that had a Gleason higher than 7 as "High Gleason".

Code availability

All the custom code used for processing sequencing data generated by scCUTseq is available at <https://github.com/ljwharbers/scCUTseq>.

Declarations

Data availability

Due to privacy reasons, all the sequencing data obtained from clinical tumor samples have been deposited at the European Genome-Phenome Archive with restricted access. The link to the data will be made publically available before publication.

Acknowledgements

We thank Britta A.M. Bouwman (Bienko-Crosetto Lab) for critically reading the manuscript and Grazyna Lietzau (Stockholm Southern Hospital) for help with sectioning and H&E staining of prostate tissues. We acknowledge the Biomedicum Flow cytometry Core facility (Karolinska Institutet), supported by KI/SLL, for providing cell sorting services; the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council sequencing on NovaSeq 6000; the SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure for processing NovaSeq data; and the IMB Microscopy Core Facility for providing access to the Opera Phenix microscope (funded by the Deutsche Forschungsgemeinschaft, grant. no. INST 247/845-1 FUGG) for detecting and assessing the frequency of the CRISPR-Cas9 induced deletion on chr11 in TK6 cells. We thank Cecilia Söderberg-Nauclér (Karolinska Institutet) for providing access to a microscope slide scanner to digitalize H&E-stained tissue sections, and Vera Minneker (Roukos lab) for generating the TK6-Cas9 expressing cell line. This work was supported by grants from the National Natural Science Foundation of China (grant no. 81972475) and the Chinese Postdoctoral Science Foundation (grants no. 2019T120593, 2018M630787) to N.Z.; by a scholarship from the Marie Skłodowska-Curie Innovative Training Networks (ITN) (H2020-MSCA-ITN-2018, grant no: 812829 “aDDress”) to L.H.; by a scholarship from MIUR - Dipartimenti di Eccellenza 2018–2022 (Project No. D15D18000410001) to E.B.; by grants from the Swedish Research Council (grant no. 2015-00162) and the Swedish Cancer Society (grant no. CAN2018/0658) to T.H.; by grants from the Deutsche Forschungsgemeinschaft (DFG, grant no. 393547839–SFB1361, 402733153–SPP2202 and 455784893) to V.R.; by a grant from Fondazione AIRC per la Ricerca sul Cancro (grant no. 22850 under IG 2019) to C.M.; by a grant from the Swedish Cancer Research Foundation (grant no. 19 0130 Pj 03 H) to M.B.; by grants from the Swedish Cancer Research Foundation (grant no. CAN 2018/728) and the Strategic Research Programme in Cancer (StratCan) at Karolinska Institutet (grant. no. 2201) to N.C.; and by a joint grant from the Swedish Foundation for Strategic Research (SSF, grant no. BD15_0095) to N.C. and T.H.

Author contributions

Project conceptualization: N.C. scCUTseq implementation and application: N.Z., M.S. CRISPR experiments: G.L., V.R. Breast cancer samples and annotations: C.M. Prostate cancer samples: F.T., N.S. Prostate samples pathological examination and annotation: P.S., B.H., W.W. Data analysis and visualization: L.H. Random Forest classifier: L.H., S.O. Targeted sequencing: E.B. Funding acquisition:

N.C., M.B., T.H. Clinical knowledge support: Q.Y., N.Z. Supervision: N.C. Figures: N.C., M.B. Writing: N.C. with input from all the authors.

Competing interests

The authors declare no competing interests.

References

1. Watkins, T. B. K. *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).
2. Knouse, K. A., Davoli, T., Elledge, S. J. & Amon, A. Aneuploidy in Cancer: Seq-ing Answers to Old Questions. *Annu. Rev. Cancer Biol.* **1**, 335–354 (2017).
3. Harbers, L. *et al.* Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. *Front. Oncol.* **11**, 700568 (2021).
4. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
5. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
6. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
7. Bollen, Y. *et al.* Reconstructing single-cell karyotype alterations in colorectal cancer identifies punctuated and gradual diversification patterns. *Nat. Genet.* **53**, 1187–1195 (2021).
8. Killcoyne, S. *et al.* Genomic copy number predicts esophageal cancer years before transformation. *Nat. Med.* **26**, 1726–1732 (2020).
9. Cervigne, N. K. *et al.* Recurrent genomic alterations in sequential progressive leukoplakia and oral cancer: drivers of oral tumorigenesis? *Hum. Mol. Genet.* **23**, 2618–2628 (2014).
10. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* **173**, 1003-1013.e15 (2018).
11. Rebello, R. J. *et al.* Prostate cancer. *Nat. Rev. Dis. Primer* **7**, 9 (2021).
12. Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
13. Cooper, C. S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat.*

Genet. **47**, 367–372 (2015).

14. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
15. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
16. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893.e13 (2018).
17. Minussi, D. C. *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* **592**, 302–308 (2021).
18. Casasent, A. K. *et al.* Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172**, 205-217.e12 (2018).
19. Martelotto, L. G. *et al.* Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nat. Med.* **23**, 376–385 (2017).
20. Bolhaqueiro, A. C. F. *et al.* Ongoing chromosomal instability and karyotype evolution in human colorectal cancer organoids. *Nat. Genet.* **51**, 824–834 (2019).
21. Laks, E. *et al.* Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179**, 1207-1221.e22 (2019).
22. Su, F. *et al.* Spatial Intratumor Genomic Heterogeneity within Localized Prostate Cancer Revealed by Single-nucleus Sequencing. *Eur. Urol.* **74**, 551–559 (2018).
23. Alexander, J. *et al.* Utility of Single-Cell Genomics in Diagnostic Evaluation of Prostate Cancer. *Cancer Res.* **78**, 348–358 (2018).
24. Reuter, J. A., Spacek, D. V., Pai, R. K. & Snyder, M. P. Simul-seq: combined DNA and RNA sequencing for whole-genome and transcriptome profiling. *Nat. Methods* **13**, 953–958 (2016).
25. Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304–319 (2016).
26. Zachariadis, V., Cheng, H., Andrews, N. & Enge, M. A Highly Scalable Method for Joint Whole-Genome Sequencing and Gene-Expression Profiling of Single Cells. *Mol. Cell* (2020) doi:10.1016/j.molcel.2020.09.025.
27. Morita, K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.* **11**, 5327 (2020).

28. Zhang, X. *et al.* CUTseq is a versatile method for preparing multiplexed DNA sequencing libraries from low-input samples. *Nat. Commun.* **10**, 4732 (2019).
29. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
30. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).
31. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
32. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).
33. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
34. Calissi, G., Lam, E. W.-F. & Link, W. Therapeutic strategies targeting FOXO transcription factors. *Nat. Rev. Drug Discov.* **20**, 21–38 (2021).
35. Adams, E. J. *et al.* FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. *Nature* **571**, 408–412 (2019).
36. Takayama, K.-I. *et al.* Integrative analysis of FOXP1 function reveals a tumor-suppressive effect in prostate cancer. *Mol. Endocrinol. Baltim. Md* **28**, 2012–2024 (2014).
37. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
38. Gao, S. *et al.* Forkhead domain mutations in FOXA1 drive prostate cancer progression. *Cell Res.* **29**, 770–772 (2019).
39. Hieronymus, H. *et al.* Deletion of 3p13-14 locus spanning FOXP1 to SHQ1 cooperates with PTEN loss in prostate oncogenesis. *Nat. Commun.* **8**, 1081 (2017).
40. Epstein, J. I. Prostate cancer grading: a decade after the 2005 modified system. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **31**, S47-63 (2018).
41. Gothe, H. J. *et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Mol. Cell* **75**, 267-283.e12 (2019).
42. Epstein, J. I., Amin, M. B., Reuter, V. E. & Humphrey, P. A. Contemporary Gleason Grading of Prostatic Carcinoma: An Update With Discussion on Practical Issues to Implement the 2014 International Society

- of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *Am. J. Surg. Pathol.* **41**, e1–e7 (2017).
43. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinforma. Oxf. Engl.* **34**, i884–i890 (2018).
44. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
46. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
47. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* **34**, 3600 (2018).
48. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).
49. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat. Oxf. Engl.* **5**, 557–572 (2004).
50. Rigatti, S. J. Random Forest. *J. Insur. Med. N. Y.* **N47**, 31–39 (2017).
51. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).
52. Csardi G, Nepusz T. “The igraph software package for complex network research.” *InterJournal, Complex Systems*, 1695. (2006).
53. Hahsler, M., Piekenbrock, M. & Doran, D. dbscan: Fast Density-Based Clustering with R. *J. Stat. Softw.* **91**, 1–30 (2019).
54. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinforma. Oxf. Engl.* **35**, 526–528 (2019).
55. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* **69**, e96 (2020).
56. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

Figures

Figure 1

Figure 1

Development and validation of scCUTseq. (a) scCUTseq workflow. IVT, in vitro transcription. (b) Lorenz curves for genomic coverage by bulk CUTseq, standard or 1:200 scaled MALBAC, and scCUTseq on SKBR3 cells. (c) Hierarchically clustered copy number profiles obtained by applying scCUTseq to four

different cell lines. n, number of single cells. (d) Genomic location of two single-guide RNAs (sgRNAs) used to induce a ~7 Mb deletion on the q-arm of chromosome (chr) 11. The forward sequence of each sgRNA is shown. (e) Zoom-in view of the chromosomal region on chr11 targeted by the two sgRNAs, in cells in which the region had a non-diploid copy number. The color bars on the left refer to the colors in (f) and (g). (f) Dimensionality reduction by Uniform Manifold Approximation and Projection (UMAP) of the copy number profiles of edited (sgRNA) and non-edited (Cntr) TK6 cells. Each dot represents a single cell. n, number of single cells. (g) Same as in (f), but with different clusters identified shown in different colors.

Figure 2

Figure 2

scCUTseq detects the clonal architecture of breast cancers. (a) Hierarchically clustered copy number profiles obtained by applying scCUTseq to nuclei extracted from a frozen breast cancer (B) surgical biopsy. n, number of single nuclei. (b) Distribution of the fraction of the genome amplified (> 2 copies) in the cells shown in (a). (c) Same as in (b), but for deletions. (d) Balanced minimal evolution tree of the non-diploid cells in (a) based on their copy number profiles. n, number of single cells. (e) UMAP dimensionality reduction of the copy number profiles of the cells clustered in (d). n, number of cells. Colored dot groups represent clones, whereas colored outlines indicate so-called superclones. Each dot represents a single cell. (f) Balanced minimal evolution tree of the clones shown in (e). (g) Median copy number profile of each clone shown in (f). The heatmap on the bottom shows the copy number profile obtained by applying bulk CUTseq to gDNA extracted from leftover nuclei that were not used for sorting.

Figure 3

Figure 3

Spatially resolved single-cell CNA profiling across the prostate. (a) Scheme of sample preparation and assays performed. (b) Schematic map displaying the pathological classification of each tissue cube (region) from which single nuclei were extracted, in one of the two prostate samples (P2) in which scCUTseq was performed on all the regions. Each cell in the grid represents a tissue cube from which nuclei were isolated and profiled by scCUTseq as shown in (a). White cells indicate absence of prostate

tissue. (c) Percentage of cells carrying at least one CNA in each region of P2. (d) Same as in (c), but for diploid cells. (e-g) Same as in (b-d), but for the other sample (P5) in which scCUTseq was performed on all the regions. The anatomical orientation of the maps in (b-g) is shown by the four arrows at the bottom. A, anterior. P, posterior. L, left. R, right. (h) Fraction of diploid, 'pseudo-diploid' and 'monster' cells in all the sampled regions in P2 and P5 samples. Representative copy number profiles for each group are shown on the right. n, number of single cells.

Figure 4

High clonal and spatial heterogeneity of pseudo-diploid and monster cells. (a) UMAP dimensionality reduction of the copy number profiles of pseudo-diploid cells identified in sample P2. n, number of cells. Dots of the same color represent clones, whereas colored outlines indicate so-called superclones. Each dot represents a single cell. (b) Balanced minimal evolution tree of the clones shown in (a). (c) Median copy number profile of each clone shown in (b). n, number of clones. (d) Distribution of the locality index for each pseudo-diploid cell clone shown in (a). The color of each clone is the same as in (a). (e) Schematic maps of the spatial distribution of the most localized (top) and most spread (bottom) pseudo-diploid clones in P2. Each grid represents the tissue map, with each cell in the grid representing a tissue cube from which nuclei were isolated and profiled by scCUTseq (see scheme in Fig. 3a). Grey cells indicate tissue regions in which the corresponding clone was not detected. White cells indicate absence of prostate tissue. The bar on top of each grid shows the genome-wide median copy number profile of the corresponding clone. Blue bars indicate deletions, orange bars amplifications. Vertical black bars mark the boundaries between consecutive chromosomes. The anatomical orientation of the grids is shown by the four arrows on the bottom right. A, anterior. P, posterior. L, left. R, right. (f) Distributions of the mean locality index of pseudo-diploid clones in different regions classified based on the corresponding H&E pattern in sample P2 (see Methods for the classification criteria). n, number of regions. P, Wilcoxon's test, two-tailed. In each boxplot, the box spans from the 25th to the 75th percentile and whiskers extend from the minimum to the maximum value. The horizontal line represents the median. (g) UMAP dimensionality reduction of the CNA profiles of monster cells in P2. n, number of cells. Dots of the same color belong to the same cluster. Each dot represents a single cell. (h) Hierarchically clustered single-cell copy number profiles of monster cells assigned to the UMAP cluster 1 in (g). (i-k) Same as in (h), but for UMAP cluster 2, 3, and 4, respectively. (l) Schematic map displaying the pathological classification of each region profiled by scCUTseq in P2 (same as in Fig. 3b). (m) Map showing the relative frequency and number of monster cells assigned to UMAP cluster 1 in each region of P2. (n-r) Same as in (m), but for cells assigned to UMAP cluster 2 to 6. In all the maps in (l-r), each cell in the grid represents a tissue cube from which nuclei were isolated and profiled by scCUTseq (see scheme in Fig. 3a). Grey cells indicate tissue regions in which the corresponding clone was not detected. White cells indicate absence of prostate tissue. The anatomical orientation of all the maps is shown by the four arrows on the bottom right. A, anterior. P, posterior. L, left. R, right.

Figure 5

Figure 5

Localized pseudo-diploid clones harbor deletions in key tumor suppressor genes. (a) Anti-correlation between the mean genomic size of deletions and the locality index, for each pseudo-diploid cell clone found in P2. Each dot represents one clone. Dashed red line, linear regression fit. PCC, Pearson' correlation coefficient. P, Wilcoxon's test, two-tailed. (b) Same as in (a), but for the second sample (P5) in which we performed scCUTseq for every tissue region (cube). (c) Genes in the Catalogue of Somatic

Mutations in Cancer (COSMIC) found to be amplified (AMP) or deleted (DEL) in the five most localized pseudo-diploid cell clones in P2 (shown on the top). Bold blue, genes encoding for members of the Forkhead of transcription factors. Black bold, classical tumor suppressor genes. (d) Same as in (c), but for sample P5. (e) Schematic map displaying the pathological classification of each region profiled by scCUTseq in P2 (same as in Fig. 3b). (f) Schematic map showing the spatial location of the C6 pseudo-diploid cell clone harboring a monozygous deletion of the FOXO1 gene shown in (c). (g) Same as in (f), but for the C17 clone. (h) Same as in (f), but for the FOXO3 gene. (i) Same as in (h), but for the C17 clone. (j) Same as in (f), but for the FOXP1 gene in clone C8. (k) Schematic map showing the number of unique non-synonymous somatic SNVs in cancer associated genes identified by targeted deep sequencing (see Methods) in all the regions of P2. (l) Schematic map showing the spatial location of SPOP gene mutations. (m) Same as in (l), but for mutations in the FOXP1 gene. (n) Same as in (l), but for mutations in the FOXA1 gene. In all the maps in (e-n), each grid represents the tissue map, with each cell in the grid representing a tissue cube from which nuclei were isolated and profiled by scCUTseq (see scheme in Fig. 3a). Grey cells indicate tissue regions in which the indicated gene was not deleted or mutated. White cells indicate absence of prostate tissue. The anatomical orientation of all the maps is shown by the four arrows at the bottom. A, anterior. P, posterior. L, left. R, right. (o) Prevalence of FOXO1, FOXO3, FOXP1 and FOXA1 genomic alterations in prostate cancers sequenced by The Cancer Genome Atlas (TCGA). AMP, amplifications. DEL, deletions. SNV, single-nucleotide variants. n, number of TCGA samples with one or more alterations affecting the corresponding gene. The percentage refers to all the prostate cancers samples available for download from TCGA (see Methods).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)
- [SupplementaryTable1.xlsx](#)
- [SupplementaryTable2.xlsx](#)
- [SupplementaryTable3.xlsx](#)
- [SupplementaryTable4.xlsx](#)
- [SupplementaryTable5.xlsx](#)
- [SupplementaryTable6.xlsx](#)
- [SupplementaryTable7.xlsx](#)
- [SupplementaryTable8.xlsx](#)