The spatial landscape of clonal somatic mutations in benign and malignant tissue

Joakim Lundeberg (joakim.lundeberg@scilifelab.se)
KTH Royal Institute of Technology https://orcid.org/0000-0003-4313-1601

Andrew Erickson
University of Oxford

Emelie Berglund
KTH Royal Institute of Technology

Mengxiao He
KTH Royal Institute of Technology https://orcid.org/0000-0001-5905-8467

Maja Marklund
KTH Royal Institute of Technology

Reza Mirzazadeh
KTH Royal Institute of Technology

Niklas Schultz
Karolinska Institutet

Linda Kvastad
KTH Royal Institute of Technology

Alma Andersson
KTH Royal Institute of Technology

Ludvig Bergenstråhle
KTH Royal Institute of Technology

Joseph Bergenstråhle
KTH Royal Institute of Technology

Ludvig Larsson
Karolinska Institute https://orcid.org/0000-0003-4209-2911

Alia Shamikh
Karolinska Institutet

Elisa Basmaci
Karolinska Institutet

Teresita Diaz de Ståhl
Karolinska Institutet

Timothy Rajakumar
University of Oxford

Kim Thrane

Royal Institute of Technology
Andrew Ji
Stanford University https://orcid.org/0000-0001-9688-5680
Paul Khavari
Stanford University https://orcid.org/0000-0003-0098-4989
Firaz Tarish
Karolinska Institutet
Anna Tanoglidi
University Uppsala Hospital
Jonas Maaskola
KTH Royal Institute of Technology
Richard Colling
University of Oxford
Tuomas Mirtti
University of Helsinki
Freddie Hamdy
Nuffield Department of Surgical Sciences
Dan Woodcock
University of Oxford https://orcid.org/0000-0003-0576-044X
Thomas Helleday
Karolinska Institutet
Ian Mills
University of Oxford
Alastair Lamb
University of Oxford

Biological Sciences - Article

Keywords: clonal somatic mutations, cancer diagnosis, focal therapy

Posted Date: July 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-745238/v1

Version of Record: A version of this preprint was published at Nature on August 10th, 2022. See the published version at https://doi.org/10.1038/s41586-022-05023-2.

The spatial landscape of clonal somatic mutations in benign and malignant tissue

Andrew Erickson[1,§], Emelie Berglund[2,§], Mengxiao He[2¤], Maja Marklund[2,¤], Reza Mirzazadeh[2¤], Niklas Schultz[3,¤], Linda Kvastad[2], Alma Andersson[2], Ludvig Bergenstråhle[2], Joseph Bergenstråhle[2], Ludvig Larsson[2], Alia Shamikh[4,5], Elisa Basmaci[4,5], Teresita Diaz De Ståhl[4,5], Timothy Rajakumar[1], Kim Thrane[2], Andrew L Ji[6], Paul A Khavari[6], Firaz Tarish[3], Anna Tanoglidi[7], Jonas Maaskola[2], Richard Colling[1,8], Tuomas Mirtti[9,10,11], Freddie C Hamdy[1,12], Dan J Woodcock[1,13], Thomas Helleday[3,14], Ian G Mills[1], Alastair D Lamb[1,12‡] and Joakim Lundeberg[2,‡*]

[1]Nuffield Department of Surgical Sciences, University of Oxford, UK
[2]Department of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.
[3]Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Solna, Sweden.
[4]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden
[5]Department of Clinical Pathology and Cytology, Karolinska University Hospital, Stockholm, Sweden
[6]Program in Epithelial Biology, Stanford University School of Medicine, Stanford, USA
[7]Department of Clinical Pathology, University Uppsala Hospital, Uppsala, Sweden.
[8]Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, UK
[9]Department of Pathology, University of Helsinki & Helsinki University Hospital, Finland
[10]Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Finland.
[11]iCAN-Digital Precision Cancer Medicine Flagship, Helsinki, Finland.
[12]Department of Urology, Oxford University Hospitals NHS Foundation Trust, UK
[14]Weston Park Cancer Centre, Department of Oncology and Metabolism, University of Sheffield, UK
[13]Big Data Institute, Old Road Campus, University of Oxford, UK

[§]These authors contributed equally to this work.
[¤]These authors contributed equally to this work.
[‡]Joint senior authors.
[*]Correspondence should be addressed to J.L. (joakim.lundeberg@scilifelab.se)

Abstract

Defining the transition from benign to malignant tissue is fundamental to improve early diagnosis of cancer. Here, we provide an unsupervised approach to study spatial genome integrity in situ to describe previously unidentified clonal relationships. We employed spatially resolved transcriptomics to infer spatial copy number variations in >120 000 regions across multiple organs, in benign and malignant tissues. We demonstrate that genome-wide copy number variation reveals distinct clonal patterns within tumours and in nearby benign tissue using an organ-wide approach focused on the prostate. Our results suggest a model for how genomic instability arises in histologically benign tissue that may represent early events in cancer evolution. We highlight the power of capturing the molecular and spatial continuums in a tissue context and challenge the rationale for treatment paradigms, including focal therapy.

Main

Mutations can either be inherited or acquired (somatic). Inherited genomic alterations are readily identifiable as these are present in all cells while somatic mutations are usually only present in a small fraction of cells. In order to obtain spatial information of these rarer non-heritable genetic events occurring in cancer, studies have commonly used laser capture microdissection to retrieve histologically (or biomarker) defined tissue regions or even single cells. These studies have an inherent bias as only a limited number of spatial regions or single cells per tissue section can be collected and examined. The possibility to perform spatial genome analysis without being confined by histological boundaries would therefore provide an important contribution to delineate the clonal architecture in tumours and co-existing benign tissue.

Spatially resolved transcriptomics has emerged as a genome-wide methodology to explore tissues in an unsupervised manner[1]. In this study we infer genome-wide copy-number variations (CNV) from spatially resolved mRNA profiles *in situ* (Fig. 1a). Gene expression has previously been used to inferCNVs in single cells, successfully identifying regions of chromosomal (chr) gain and loss[2]. Here we expand into a spatial modality generating CNV calls in each spatial region represented by barcoded spots. First, we sought corroboration that inferCNV (iCNV) data could accurately mirror DNA-based phylogenies, using simultaneously extracted single cell RNA, and DNA[3] (Extended Data Fig. 1a). Next, we successfully recapitulated published DNA-based phylogenies in prostate cancer using RNA from the same samples[4–6] (Extended Data Fig. 1b, c). To ensure that we robustly could capture sufficient and accurate CNV information from individual spots from a multifocal tumour model, and use this information to deduce clonal relationships between cells, we then designed an *in-silico* system to synthesise a tissue containing multiple clones determined by stochastic copy number (CN) mutations in a single

artificial chromosome. Using a probabilistic method to generate gene expression from such mutations we then interrogated the expression data using iCNV, while blind to the underlying 'ground-truth' CN status, and successfully recapitulated both the CN status and the clonal groupings (Extended Data Fig. 2a-c).

Next, we used a cross section of an entire prostate organ to explore the spatial iCNV landscape of a commonly multifocal malignancy[7]. The specimen was obtained by open radical prostatectomy and an axial section was taken from the mid-gland. The axial section was subdivided into cubes (Fig. 1a, b) and corresponding tissue sections were histologically graded using the Gleason grading system[8] identifying extensive intratumoural heterogeneity (ITH) in the context of surrounding benign tissue (Fig. 1b, e). We obtained organ-wide transcriptional information from 21 cubes (tissue sections) and > 21 000 barcoded regions (100 micron spots) with a mean average of 3500 expressed genes detected per barcoded spot[9]. We then analysed the transcriptional data using factorized negative binomial regression (Extended Data Fig. 3a). This provided an unsupervised view of gene expression factors (GEFs) over the cross section of the prostate (Fig. 1c). Twenty-five factors showed overlap between histology and GEFs representing tumour, hyperplasia and benign epithelia annotated by the factor marker genes, as previously reported[10] (Fig. 1f). Several GEFs provided distinct 'clonal' appearances and were associated with tumour regions (Fig. 1f, right panel). Next, we undertook a spatial iCNV analysis to provide an overall landscape of genome integrity (Fig. 1d) identifying certain regions with increased iCNV activity (V1_1, H2_1, H1_1, H1_5, H2_5; Fig. 1g) while the majority of the tissue area appears to be CN neutral. These initial results suggested that iCNVs could identify tissue regions, at organ scale, with inferred genomic variability, distinct from morphology or expression analysis.

To increase the fidelity of our analysis of variable iCNV regions we took advantage of smaller 55 micron diameter barcoded spots (Visium, 10x Genomics), reducing the number of cells to approximately 5-10 per spot, to perform a more detailed interrogation of seven key sections. We first validated the increased precision of this higher resolution platform using the synthetic tissue method (Extended Data Fig. 2d, e). We next obtained data from approximately 30 000 spots using factorized negative binomial regression resulting in 24 spatially distinct GEFs (Extended data Fig. 3b). Two pathologists independently annotated each spot to provide consensus pathology and histology scoring (Fig. 1e). We then investigated clonal relationships across the investigated tissue using iCNVs. Having established the association between gene expression factors and certain regions of interest (Fig. 1c, f) we wanted to determine the degree of clonal CN heterogeneity in these regions. After designating all histologically benign spots as a reference set (Extended Data Fig. 3c) it was immediately apparent that while certain GEFs displayed a fairly homogenous inferred genotype (e.g. GEF 7, 14 and 22, Extended Data Fig. 3d), others were strikingly heterogeneous (e.g. GEF 10, Extended Data Fig. 3e).

Prompted by the realization that certain regions annotated as histologically benign displayed CN heterogeneity (Fig. 1d), we refined the reference set to those spots which were both histologically benign (outside the regions of interest) and also lacking any iCNV (Extended Data Fig. 4). This constituted a ‘pure benign’ reference set for all subsequent iCNV analyses, unique to each patient. It was apparent from cancer-wide inferred genotype (Fig. 2a-e) that there was evidence of clonally distributed CN heterogeneity within areas of spatially homogeneous Gleason pattern (Fig. 2a, d, e). We constructed a clone-tree to describe sequential clonal events versus independently arising cancer-clones (Fig. 2b). It was apparent that two cancer clones lacked key truncal events including a loss of a region of chr 16q and 8p which were otherwise

ubiquitous across all cancer clones (clones A and B, Fig. 2a, b). These clones were spatially restricted to section H1_2 containing a region of low-grade Gleason Grade Group 1, discussed later. The majority of clonally related spots were located around the largest focus of Gleason Grade Group 4 disease with a striking pattern of truncal and branching events (clones H, I, J and K). We therefore focused on this dominant region of cancer (spanning sections H1_4, H1_5 and H2_5), to establish a first view of the interplay between spatial architecture and clonal dynamics (remaining sections in Extended data Fig. 5a, b).

To construct clone-trees, we assumed that: (i) groups of cells containing identical CN profiles were more likely to be related, than to have arisen by chance; and (ii) somatic CN events must be acquired sequentially over time, the more numerous the events, the more distinct the clone. Using this approach, we observed a common ancestral clone (clone H, Fig. 2b) containing truncal events including CN loss on chr 6q and 16q, and CN gain on 12q and 16q. These were clearly located in two tissue regions: an area of Gleason Grade Group 2 on the medial side of the main tumour focus (section H1_4) and a region described as 'transition state' by consensus pathology at the upper mid edge (section H2_5). These conserved iCNV features in distinct spatial locations are noteworthy. A possible explanation is that clone H represents a linear sequence of branching morphology in the prostatic glandular system[11], and that further somatic events took place giving rise to clones I, J and K forming a high-grade tumour focus (Fig. 2b), which pushed apart the branching histology due to an aggressive expansile phenotype. For the first time we have a spatial imprint of these events in prostate tissue. We also propose that some CNVs may be of particular pathological significance (Extended data Fig. 4d) based on spatial molecular phylogeny. Our analysis therefore provides insight into processes of tumour clonal evolution, identifying discriminating events by spot-level CNV calling in a spatial context.

Given this discovery of a discordance between cellular phenotype and inferred genotype, we then undertook a detailed interrogation of section H2_1 in the left peripheral zone of the prostate (Fig. 1c, 2c) containing roughly equal proportions of cancer and benign tissue. We profiled CN status of every spot in this section and ordered these spots by hierarchical clustering into 'clones' A to G based on defined levels of cluster separation (Fig. 3a, b). Spatially, we observed that these data-driven 'clone' clusters were located in groups, broadly correlating with histological subtype, but with some important distinctions (Fig. 3c, d). We observed that many CNVs already occurred in benign tissue, clone C (Fig. 3a-d), most notably in chr 8, which has been well-described in aggressive prostate cancer[12–14], but also several other CN gains and losses. Spatially, this clone constituted a region of exclusively benign acinar cells branching off a duct lined by largely copy neutral cells in nearby clone A and B (Fig. 3d). The unobserved ancestor to clone C then gave rise to a further unobserved clone, and then cancer-containing clones E, F and G. While clone G was made up exclusively of Gleason Grade Group 2 cancer cells, clones E and F were mixed with up to 25% benign cells (Fig. 3d). The presence of somatic events in histological benign cells highlights, for the first time in prostate biology, that these clone groups traverse histological boundaries.

In order to validate that this inferred CN status was truly representative of underlying genotype we used fluorescence in situ hybridization (FISH) probes to target two specific genes of discriminatory interest, *MYC* and *PTEN*, encompassed in the notable chromosomal changes in benign tissue clone C as well as high grade tumour clones, but absent in low grade disease. This confirmed that while the status of both genes was diploid in normal benign tissue (Clone A), *MYC* amplification and *PTEN* loss were evident in altered benign (Clone C) as well as in tumour clones (Clone F; Fig. 3e, Extended data Fig. 6). Going forward, we hypothesise that other

homogenous iCNV calls are accurate, based on the evidence provided by these two selected loci. This evidence suggests that somatic events, creating a mosaic of branching clones during ductal morphogenesis, are present even in histologically benign disease. It therefore follows that an understanding of this somatic mosaicism could distinguish which regions of benign glandular tissue may give rise to lethal cancer, and which will not.

We considered the place of branching morphogenesis in the sequential acquisition of transformative events in a predominantly benign section of the prostate (section H2_1, also in section H2_2) (Extended data Fig. 7). Here we noted that such events seem to occur during the development of prostatic ducts and acinar branches, with changes occurring at key branching points, and the altered genotype passed on to daughter cells lining the ducts and glands of associated branches. Interestingly, not all cells in such branches displayed the same cellular structure, raising important questions as to why epithelial glands with seemingly identical inferred genotypes might display divergent histological phenotypes.

In view of the above findings, we hypothesized that analysis of the inferred genotype of low grade cancer might reveal important differences to that of high grade cancer. Section H1_2 contains a region of Gleason Grade Group 1 prostate cancer (Extended data Fig. 5d). As noted previously there were two clones (Fig. 2a, clones A and B) which lacked key changes in both chr 8 and 16, with little in common with other cancer-bearing clones (Fig. 2c). A spot-wise re-analysis of section H1_2 (including benign spots) revealed that these two clones, now labelled F and G were spatially grouped as two approximately equal halves of this region of Gleason Grade Group 1 cancer (Extended Data Fig. 5c, d). This is evidence that low grade prostate cancer is indeed fundamentally distinct from high grade, and raises the hypothesis that such cancer cannot become higher grade because it lacks essential somatic events.

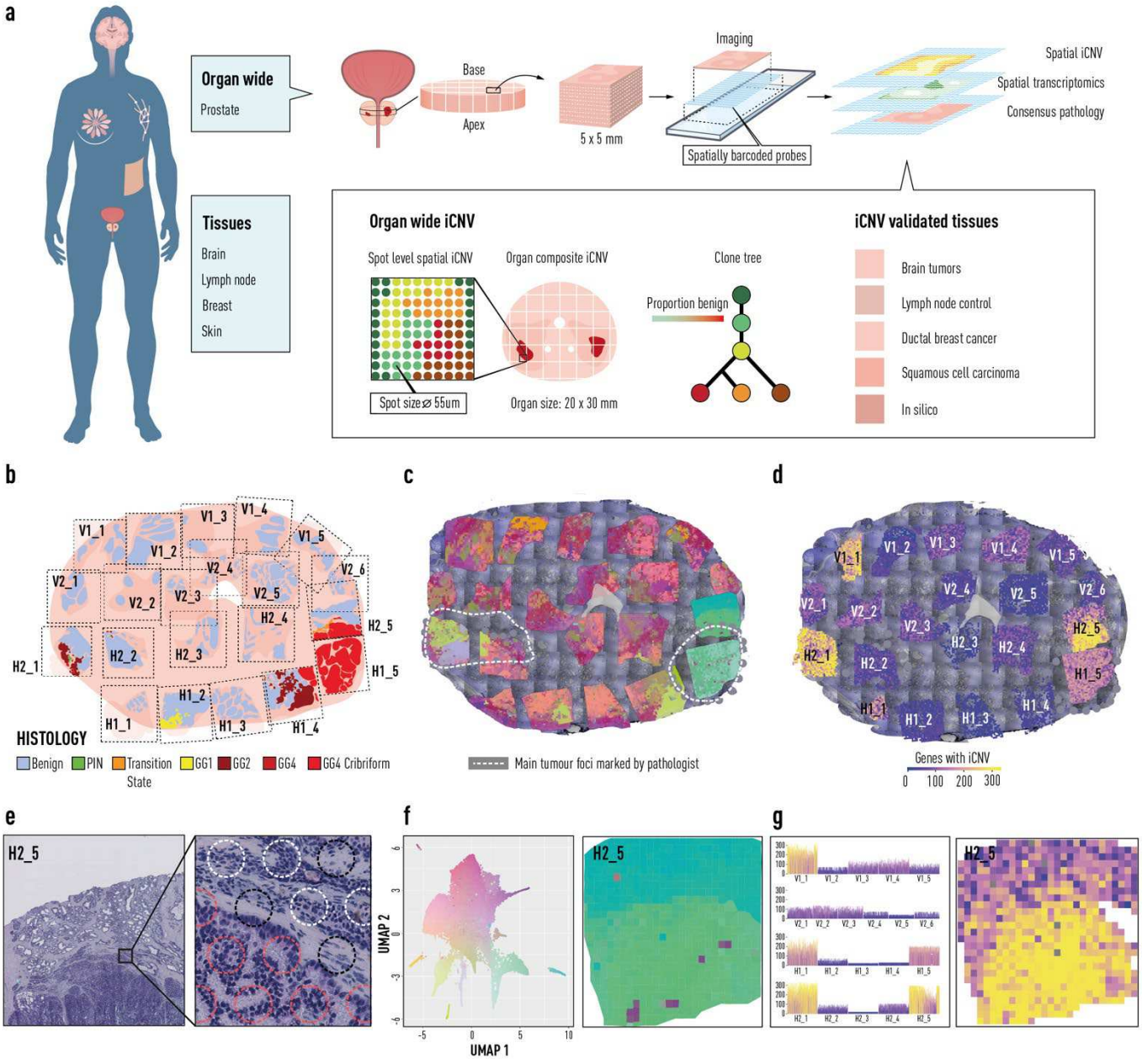
To corroborate our findings, we first performed validation through an additional 37,000 spots from a cross-section of a further prostatectomy that confirmed the spatial continuum of benign clones in proximity to cancer with shared truncal events. We also confirmed the high degree of ITH of iCNV clones within prostate tumour loci (Extended Data Fig. 8), and the presence of key somatic events in benign prostate glands. We then generalized our findings in multiple organs (Fig. 4, Extended data Fig. 9). First, we analysed a benign lymph node displaying distinct gene expression clusters for different histological entities (such as germinal centres) and the iCNV analysis provided, as expected, a copy neutral profile for the entire tissue section (Fig. 4a, b). This provided further validation that iCNV clones are distinct from gene expression. We next analysed skin tissue containing both benign squamous epithelia and squamous cell carcinoma (SCC). For this, we obtained a patient-matched, benign reference set of RNA-sequenced single skin cells with confirmation from adjacent sections of benign histology[15]. Spatial iCNV identified four clones within the tissue, one of which corresponded to SCC containing several CN events. Importantly, two key events (partial chr 1 and 12 gain) were shared with another nearby clone composed entirely of histologically benign tissue (Fig. 4c, d). This substantiates our finding of iCNV clones traversing histological boundaries for an additional tumour type. To contrast these observations, we performed analysis of a sonic hedgehog (SHH) paediatric medulloblastomas (Fig. 4e, f) with sex and age matched samples. The results show a uniformly homogeneous spatial iCNV clone type throughout the tumour with key expected genetic alterations such as 3q gain (encoding *PIK3CA*) and a 9q deletion (encoding the *PTCH1* gene) as well as short gain in 9p. These homogenous findings were validated by whole genome sequencing (WGS) of the tumour displaying distinct CNV calls for the three altered chromosomal regions identified by our iCNV analysis (Extended Data fig. 10). We further

analysed two additional tumour types without reference sets: ductal breast cancer and an adult glioblastoma (Extended Data Fig. 9). Here we confirmed a multifaceted spatial iCNV tumour landscape with multiple co-existing clone types in histological similar-appearing tumour tissue. For example, in ductal breast cancer (Extended Data Fig. 9k, l) we observed two distinct clone types (C and F), separated by stroma, with little or no CNV overlap). In the glioblastoma tissue we similarly identified five clone types that had sharp spatial demarcations separating the iCNV clones, despite being histologically similar (Extended Data Fig. 9m, n). Overall, the clonal appearances of ITH were clear as well as the overlap with tumour morphology.

The tissue clone diversity over the five investigated tissue types was strikingly variable from homogenous to highly heterogeneous genomes in both tumours and benign tissue (Fig. 4k). We therefore believe that combining the iCNV information with the spatial gene expression patterns, which provides some functional understanding, and cell type mapping (using scRNAseq) could enable targeted treatment options for individual clones, 'benign' or tumour, that would not be easily attainable by any other means.

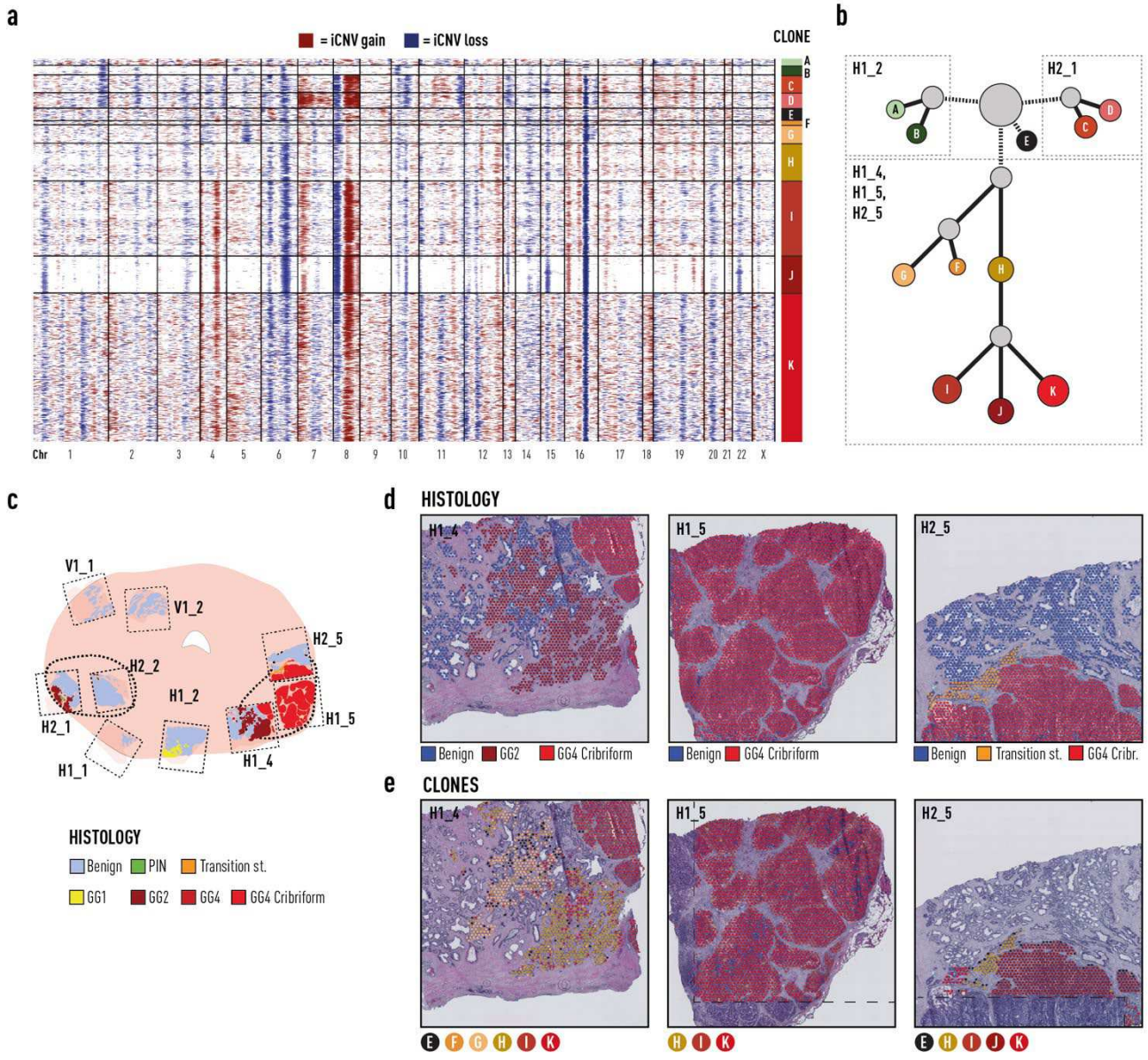
In summary, we show that spatial transcriptomic data across multiple cancer types can robustly be used to infer copy number variation, as validated by FISH and WGS. Specifically, we performed an in-depth spatial analysis of the prostate organ that generated an unprecedented atlas of up to 50 000 tissue domains in a single patient, and 120 000 tissue domains across 10 patients. For these domains we inferred genome-wide information in each spot, which facilitated data driven clone-generation in a tissue-wide fashion at high resolution. Importantly, the spatial information allowed us to identify small clonal units not evident from morphology which would therefore be overlooked by histologically-guided laser microdissection or even random sampling of single cells. We continue to show that in some tumour types, particularly prostate, glioma and

breast cancers, CNV analysis reveals distinct clonal patterns within tumours. Focusing on prostate cancer, those patterns, as defined by the conservation of CNVs across morphological entities, indicate hitherto unappreciated molecular relationships between histologically benign and cancerous regions. It is known that CNVs occur early in tumorigenesis[14]. We propose that CNVs can precede tumorigenesis and are a feature of glandular morphogenesis, with propagation of particular variants traversing disease pathology. This study shows that CNVs in regions of the genome that encode cancer drivers are truly early events, occurring in tissue regions currently unknown to and therefore ignored by pathologists (Extended Data Fig. 4d). Currently the risk stratification delivered by pathologists dictates to a significant degree treatment decisions and subsequent clinical outcome. Our study therefore provides a new and unbiased avenue to interrogate genomic integrity, adding to the armamentarium of cancer molecular pathology. Our findings provide a basis for improved early detection of clinically important cancers and improved patient outcomes for ubiquitous malignancies such as prostate cancer. Overall, our study raises important questions about cancer evolution, somatic mosaicism and tissue development.



1 **Fig. 1 | Organ-wide spatial determination of transcript and CNV status. a,** For organ-wide
2 assessment axial segments of the prostate were divided into 5x5mm blocks for spatial
3 transcriptomic analysis with spatially barcoded probes. The resulting spatial gene expression
4 profile was accompanied by inferred copy number profile supported by spot-by-spot consensus
5 pathology calls. Copy number features were used to detect clonal groups and instruct
6 phylogenetic tree construction. Tissue specific analyses of multiple phenotypes were performed.

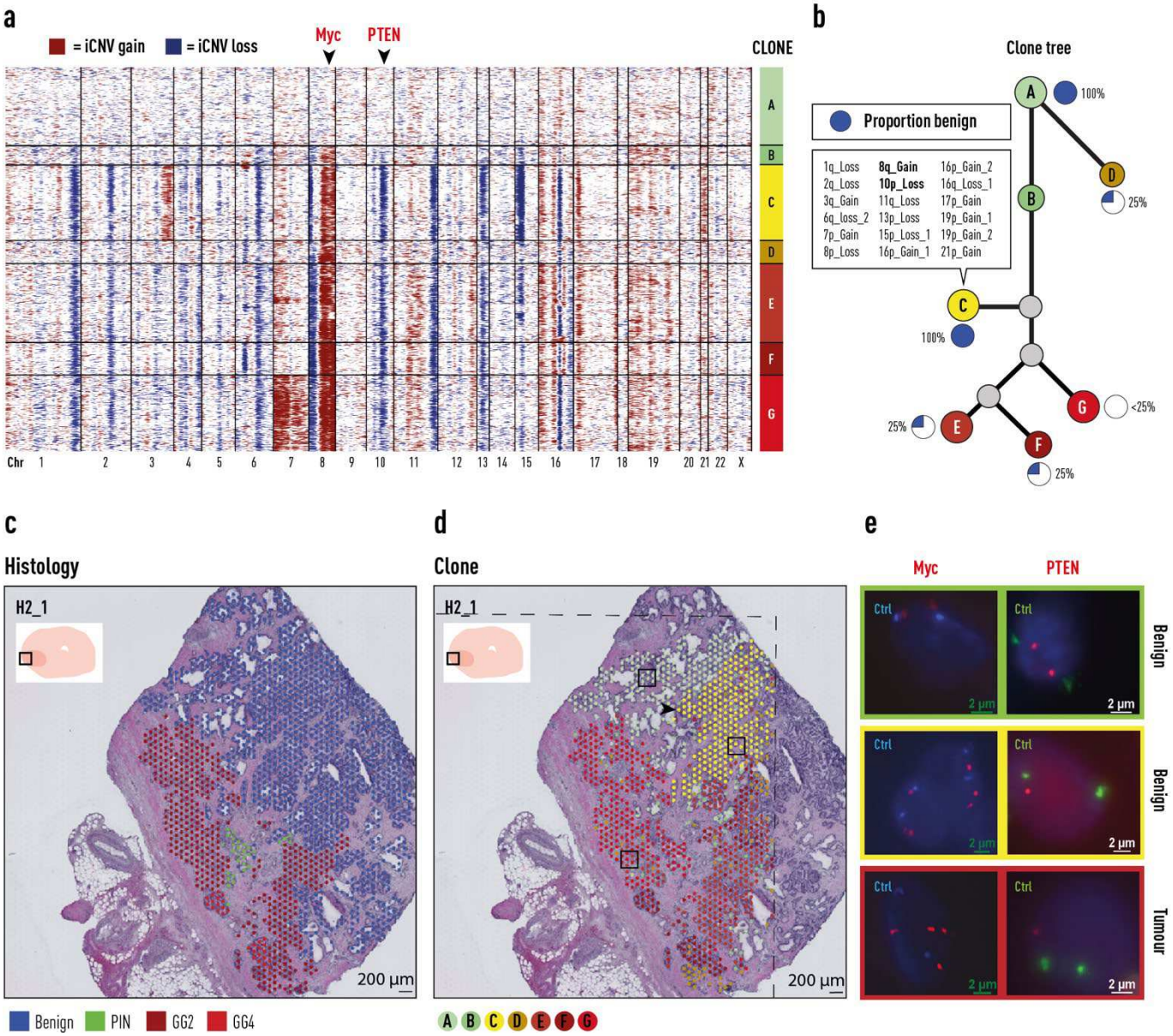
b, Histology status for each organ-wide section. Black dotted lines represent the area covered by spatial transcriptomics array surface. **c,** Spatial distribution of gene expression (see panel f). **d,** Spatial distribution of summed copy number events (see panel g). **e,** Representative spot-level consensus pathology for Section H2_5. Red circles = >50% cancer. White circles = >50% benign epithelium. Black circles = <50% of a single cell-type. **f,** UMAP principal component analysis of gene expression factors with representative close-up for Section H2_5. **g,** Total copy number events for each section with representative close-up for Section H2_5.



1 **Fig. 2 | Specific somatic alterations in all cancer organ-wide analysis. a,** Genome-wide
2 derived analysis (inferCNV) for all Visium spots harbouring tumour from prostate patient 1.
3 Clonal groupings of spots (approx. 10-15 cells each) determined by hierarchical clustering. **b,**
4 Phylogenetic clone tree of tumour clones (from panel a), with grey clones representing
5 unobserved, inferred common ancestors. Clone circle area is proportional to number of spots and
6 branch length determined by weighted quantity of CNVs, both on a logarithmic scale. iCNV
7 changes for each clone is available in Supplementary Table 1. **c,** Representation of all tissue

sections from prostate patient 1. Thicker black lines denote original boundaries annotated by initial clinical pathology. **d,** Consensus epithelial histological annotations for sections H1_4, H1_5 and H2_5, corresponding to the right tumour focus. **e,** Spatial visualization of tumour clones (Panel a). Dashed line marks areas where no spatial transcriptomics data was obtained due to being outside of barcoded array surfaces.

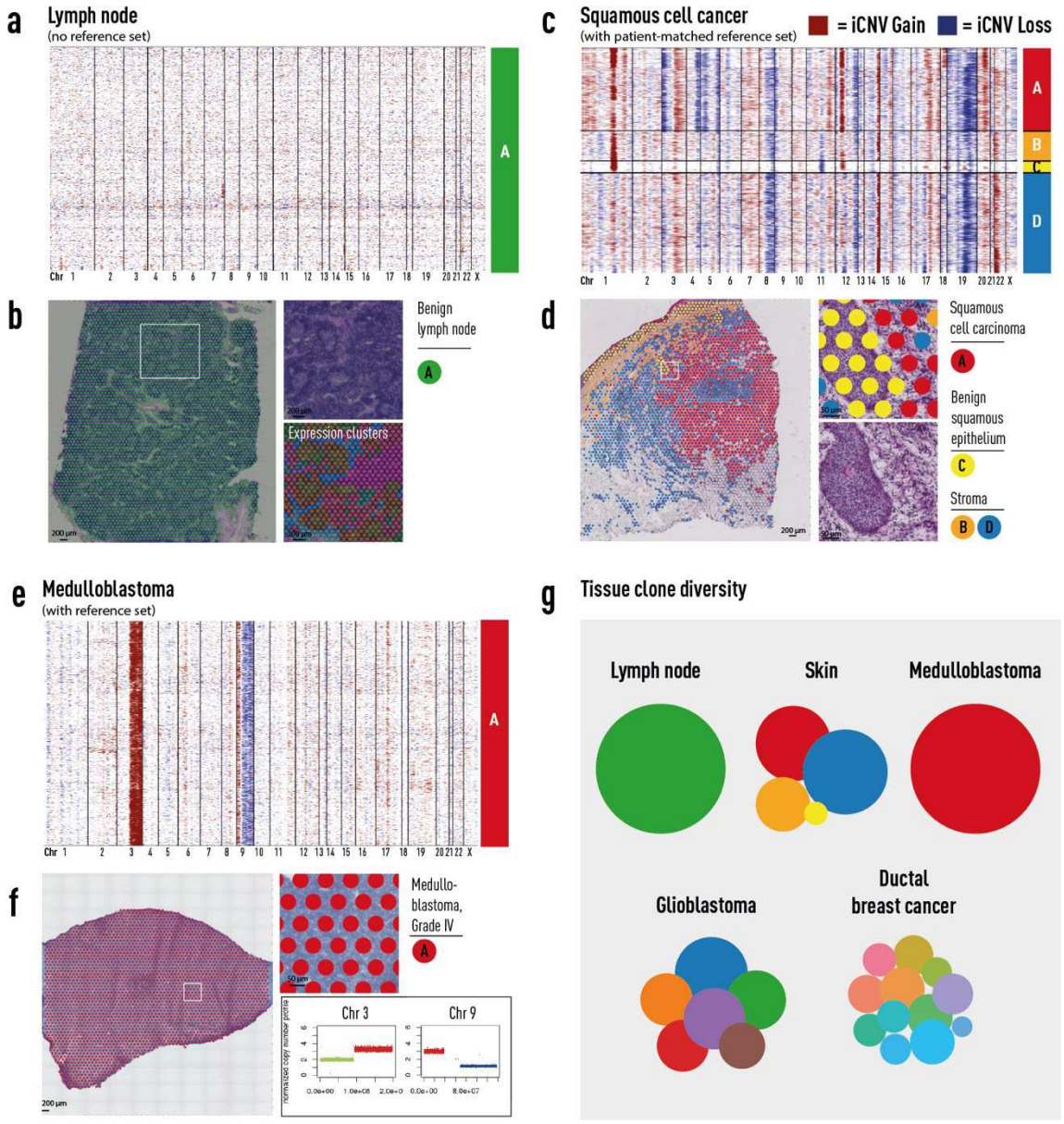
Chr = chromosome. iCNV = inferred copy number variant. PIN = prostatic intra-epithelial neoplasia. GG = ISUP Gleason 'Grade Group'



1 **Fig. 3 | Somatic events in both cancer and benign prostate epithelium. a,** Genome-wide
2 derived CNV analysis (iCNV) for each barcoded high-resolution spatial transcriptomic (ST) spot
3 from Section H2_1, which contained a mixture of tumour and benign epithelia (red = gain; blue
4 = loss). Clonal groupings of spots (approx. 10-15 cells each) determined by hierarchical
5 clustering. **b,** Phylogenetic clone tree of all clones (from panel a). Proportion of benign epithelial
6 cells in each clone as indicated. Specific CNV locations unique to Clone C are listed

(summarized by chr number where event is located, p/q arm and gain/loss; remainder of iCNV changes in Supplementary Table 2). **c,** Spatial visualization of histopathological status of each spot. Each spot assessed by two pathologists for consensus annotation with only spots >50% cellularity included. **d,** Spatial visualization of clone status of each spot. Clonal groupings cross histological boundaries. Branching point of prostatic duct (black arrow), indicates possible site of somatic events arising in Clone C (also Extended Data fig. 6). Dashed line marks areas where no spatial transcriptomics data was obtained due to being outside of barcoded array surfaces. **e,** FISH validation of two iCNVs: MYC, from chr8q; and PTEN, from chr10p (black arrow heads on panel a). Control probes (Ctrl) target centromeres for chr8 and chr10 respectively.

Chr = chromosome. CNV = copy number variant. PIN = prostatic intra-epithelial neoplasia. GG = ISUP Gleason 'Grade Group'



1 **Fig. 4 | Somatic copy-number alterations in cancer and benign histologies. a, b,** Benign
2 lymph node with distinct histological features and gene-expression heterogeneity (Panel b)
3 harbouring no detected copy-number alterations (Panel a). Gene expression clusters determined
4 by UMAP (Panel b). **c, d,** Skin containing squamous cell carcinoma (Clone A, red) as well as
5 benign squamous epithelium (Clone C, yellow). Subset of somatic events visualized in cancer
6 Clone A are also detected in adjacent benign epithelial Clone C. Panel d demonstrates

representative spot placement for adjacent benign and cancer regions. **e, f,** Monoclonal childhood medulloblastoma. iCNV in Chr 3 and 9 (Panel e), were corroborated by CN calls from WGS (Panel f, lower right). **g,** Clone distribution per tissue type. Circle area corresponds to numbers of spots per clone. Results from ductal breast cancer and glioblastoma are available in Extended Data Fig. 9.

Chr=chromosome. InfCN = inferred Copy-Number Variant. UMAP = Uniform Manifold Approximation and Projection.

Methods and material

Tissue specimens

Whole prostates were obtained by open radical prostatectomy at Västerås Hospital. Each prostate was cut in two halves by a horizontal cut, the upper part (closest to the patient's head) was used and cut on a stepped 5 mm mould to obtain a 5 mm high cylinder. Next, stripes were cut out from the cylinder, and each stripe was cut into smaller cubes (total 21 for patient 1 and 28 for patient 2). All tissue cubes were fresh-frozen in liquid nitrogen and stored at −80 °C until embedding for cryosectioning. The childhood brain tumours were collected and provided by The Swedish Childhood Tumour Biobank and stored at −80 °C until embedding for cryosectioning.

Datasets

Human squamous cell carcinoma and case-matched dissociated normal skin cells (reference set) were provided from a published dataset[15]. The human lymph node, human adult glioblastoma multiforme (tumour grade IV) and human breast cancer (ductal carcinoma in situ, lobular carcinoma in situ, invasive carcinoma) datasets were provided from 10x Genomics (https://support.10xgenomics.com/spatial-gene-expression/datasets).

Spatial Transcriptomics (1k arrays)

For prostate (patient 1), all 21 tissue cubes were cryosectioned into 10 µm sections from the bottom (two sections per cube) for ST analysis. The sections were mounted onto spatially barcoded microarray slides. The protocol described in Ståhl et al. and Salmén et al was used to prepare all mounted sections with few modifications[9,16]. Fixation was performed for 10 min at room temperature and permeabilized using exonuclease I buffer for 30 min at 37 °C and 0.1X

pepsin (pH 1) for 10 min at 37 °C. The material was processed into libraries as described in Jemt et al. and sequenced on an Illumina Novaseq using paired-end 300-bp reads[17].

Spatial Transcriptomics (10X Genomics Visium)

Visium Spatial Tissue Optimization Slide & Reagent kit (10x Genomics, Pleasanton, CA, USA) was used to optimize permeabilization conditions for the tissue sections. One 10µm section from each patient was processed according to the manufacturer's instructions. Spatially barcoded cDNA of every tissue section was generated using Visium Spatial Gene Expression Slide & Reagent kit (10x Genomics). Tissue sections from prostate patient 1 were fixed according to manufacturer's instructions and permeabilization was performed for 8 min. Sections from prostate patient 2 were fixed for 10 min using acetone at -20 °C and permeabilized for 15 min. Childhood brain tumour sections of 12µm were permeabilized for 30 min. Libraries for all tissue sections were generated following 10x Genomics Visium library preparation protocol and sequenced on Illumina sequencing instruments.

Data processing

1k arrays: FASTQ files were processed using the ST Pipeline v.1.5.1 software[18]. Transcripts were mapped with STAR[19] to the GRCh38.79 human reference genomes. Mapped reads were counted using the HTseq count tool[20]. Spatial barcodes were demultiplexed using an implementation of TagGD UMI filtering[21] is carried out to remove duplicated reads. A mean of 3582 unique genes and 10734 unique transcripts was obtained per spot after removing spots with less than 100 genes or transcripts.

10x Visium arrays: Specifics regarding data processing prior to data analysis after demultiplexing of fastq files has been described elsewhere for the human squamous cell carcinoma[15] and datasets provided from 10x Genomics

(https://support.10xgenomics.com/spatial-gene-expression/datasets). For the childhood brain tumour, read 2 was trimmed to remove both the TSO adapter sequence and polyA homopolymers using Cutadapt[22]. Trimmed fastq files were then run through Space Ranger (version 1.0.0, 10x Genomics) where reads were mapped to the human reference genome (GRCh38, release 93). The raw sequencing reads of the prostate samples were directly processed using Space Ranger (version 1.0.0 prostate 1, version 1.2.1 prostate 2, 10x Genomics) and mapped using the same human reference genome as above. A mean of 2334/2104 unique genes and 10221/5711 unique transcripts was obtained per spot after removing spots with less than 100 genes or transcripts for Patient 1/Patient 2.

Factorized negative binomial regression of prostate samples

GEF analysis was performed as previously described[10]. In all analyses, we factorized the data into $T = 25$, 24 & 20 GEFs (1k, Visium patient 1 and Visium patient 2) and ran the optimization for 5000 iterations. Spots were annotated based on their section to control for sample-wise batch effects.

Processing and visualization of non-prostate samples

Data processing and visualization was carried out using Seurat (version 3.2.2)[23] and STUtility (version 0.1.0)[24] R packages. UMI counts were filtered (Supplementary Information) and samples were normalized using SCTransform. Dimensionality reduction was performed using principal component analysis and the top principal components were used (Supplementary Information). The expression-based clustering was performed with resolution parameter set to 0.8 for all samples except childhood brain tumour where 0.2 was used (Supplementary Information). Finally, a two-dimensional UMAP embedding was constructed from the previously established top principal components for each tissue type.

Paediatric Tumour DNA sequencing and data analysis

Libraries for WGS were prepared using Illumina TruSeq PCR-free. WGS samples were sequenced 2x150bp paired-end, on HiSeqX v2.5 (sample 810) or NovaSeq6000 (samples 539 and 659) instruments (Illumina, SanDiego, California). DNA sequence data was processed with Sarek, following the GATK best-practice recommendations[25], on UPPMAX Clusters at Uppsala University (https://www.uppmax.uu.se/resources/systems/the-bianca-cluster/). Briefly, the steps run were: quality control of FASTQ files using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), alignment of short reads to the human reference genome sequence (GRCh38/hg38) using bwa-mem, with the ALT-aware option turned on[26], sorting of reads and marking of PCR duplicates with GATK MarkDuplicates and base quality scores recalibration and joint realignment of reads around insertions and deletions (indels), using GATK tools (https://github.com/broadinstitute). Tumour CNV profiles were generated using Control-FREEC[27]. The matched normal sample was used to call somatic CNVs.

DNA fluorescence in situ hybridization (FISH)

Optimal cutting temperature (OCT)-embedded block of fresh frozen prostate sample was sectioned at 5µm thickness and several consecutive sections were mounted on positively charged microscope slides (VWR) and placed at -80 °C until processing. Sections were fixed with methanol and acetic acid (3:1 ratio) for 15 minutes at room temperature, washed in 1x PBS, briefly air-dried followed by H&E staining and imaging. DNA FISH probes containing *MYC*/8cen (Cytocell, catalogue number MPD28000) or *PTEN*/10cent (Cytocell, catalogue number MPD15000) were added (10-15µl) on top of tissue sections, sandwiched with 18x18 coverslips and sealed with rubber glue (BioNordika AB, catalogue number PCN009). Slides were placed on a hot plate for exactly 6 minutes at 76 °C for DNA molecules to denature and

immediately placed inside an incubator with 100% humidity for overnight incubation at 37 °C. Coverslips were gently removed and slides were washed in a ceramic jar containing prewarmed 0.4x SSC for 3 minutes at 72 °C, transferred to 2x SSC/0.05 TWEEN® 20 for 2 minutes at room temperature, then quickly washed in 2x SCC and nuclease free water. To reduce the autofluorescence backgrounds, we applied quenching probes (Thermo Fisher Scientific, catalogue number R37630) on top of sections, incubated for 5 minutes at room temperature, washed in 1x PBS, nuclei were counterstained with DAPI and slides were mounted using mounting medium (Thermo Fisher Scientific, catalogue number S36936). Microscopy images were acquired using a ×100 1.45 NA objective mounted on an Eclipse inverted microscope system (Nikon) controlled by the NIS Elements. We collected multiple image stacks per sample, each consisting of 30-40 focal planes spaced 0.3µm apart.

Pathologist Workflow – Spot-level annotation for prostate patient 1

All Visium spots were annotated on a spot-by-spot basis using the Loupe Browser Version 5.0 (10x Genomics) for the Visium sections by two uro-pathologists (R.C. and T.M.). Using a >50% cellular coverage threshold, the pathologists annotated spots by histological class, or "Exclude" (eg. mixed coverage, array regions not covering tissue such as lumens, or if scanning/sectioning artefact rendered it impossible to determine a histological class). The annotations were cleaned, unified and visualized in Loupe Browser for review. Next, a consensus workflow where-in R.C. and T.M. were asked to determine a final annotation class if there were discrepancies between benign or cancerous luminal epithelial cells. If there were discrepancies between luminal classes and stroma, A.E. performed review and re-classification, such that if over 50% of cells of one class could be identified, it was marked as the corresponding class. If there was uncertainty, the spot was marked as "mixed" and excluded from downstream analysis. The final consensus

annotation dataset consisted of a total n = 23 282 spots. We defined low-grade prostate cancers as Gleason Grade Group 1 and high-grade cancer as containing Gleason pattern 4.

Pathologist Workflow – Annotation for prostate patient 2

Prostatic luminal epithelial cells were annotated for 15 Visium sections from prostate 2 for presence of tumour histology. Luminal epithelial spots from benign tissue sections were analysed for selection of a benign reference set. Tumour histology was confirmed in sections H3_1, H2_1, H2_2, and H3_6 using Loupe Brower.

InferCNV – Data Pre-processing

In order to systematically interrogate the data, we developed an R package called SpatialInferCNV: *https://github.com/aerickso/SpatialInferCNV*. Additional analyses were performed using a series of R packages (tidyverse, Seurat, infercnv, hdf5r), python, and BASH scripts as follows. Histological annotations were imported from the final annotation consensus files for all sections, and the barcodes were appended with their section identifier. Next, the annotations were filtered for a given feature of interest. Files output from the cell ranger pipeline (filtered_feature_bc_matrix.h5) were imported, and barcodes were appended with information of their corresponding section name. The count files were then filtered only for those within the analysis of interest. The count files further underwent a quality control (QC) filter[3] wherein spots containing 500 counts or less were removed. The annotations file and counts file were joined for each section, and these were then all combined into a final matrix that was then output (.tsv file) for downstream analysis with inferCNV. The barcodes for only those that passed the annotation and QC filters were merged again with the annotations, and these were separately exported (.tsv) files for further InferCNV analysis. Lastly, a genomic positions file was created following the

instructions here: https://github.com/broadinstitute/inferCNV/wiki/instructions-create-genome-position-file.

Selection of Benign References

Inputs to InferCNV can include a reference set of UMI-barcoded objects, in order to improve precise inference of genomic CN events in the observed population. We first performed an unsupervised analysis of only the benign reference cells: (Parameters for inferCNV object: ref_group_names=NULL, Parameters for run: cutoff=0.1, cluster_by_groups=FALSE, denoise=TRUE). Using the denoised outputs, we identified by visual inspection a subgroup of all benign that harboured little-to-no inferredCNVs (Extended Data Fig. 4). The associated dendrogram file (harbouring the cluster structure and each barcode therein) was then further analysed for node selection.

InferCNV Parameters

For unsupervised SpatialInferCNV analysis, we included the following parameter for the function CreateInfercnvObject(): chr_exclude = c("chrM"). For the run() function, we used the following parameter values: the following inferCNV run() parameters were used: “cutoff=0.1, num_threads = 10, cluster_by_groups=FALSE, denoise=TRUE, HMM=FALSE”. A reference set was used for all analyses, with the exceptions of defining the reference set, or for if a suitable reference set was not available (Fig. 4, Extended Data Fig.9).

In supervised SpatialInferCNV analysis (to call inferCNV’s Hidden-Markov Model functions), InferCNV was run as follows. The node identity file was used in place of the annotation file. The following “InferCNV run” parameters were used: “cutoff=0.1, num_threads = 10, cluster_by_groups=TRUE, denoise=TRUE, HMM=TRUE).

For the global visualization of iCNV events in Fig. 1, we analysed spatial transcriptomics (1k arrays) data with inferCNV for all 21 sections in a global analysis without a reference set. We performed the analysis such that each individual spatial transcriptomics spot was ran with the following inferCNV run() parameters were used: cutoff=0.1, num_threads = 10, cluster_by_groups=FALSE, denoise=TRUE, HMM=TRUE, analysis_mode = "cells", HMM_report_by = "cell"". To spatially visualize global iCNV profiles across an entire prostate, we then determined the number of individual genes detected to harbour an inferred CN gain or loss. To reduce background noise in the visualization, the resultant HMM calls were thresholded for the number of gene-level iCNV events present in at least 35% of all spots across the entire dataset, and in at least 45% of the spots of a given section. These thresholds were selected after detailed interrogation of thresholds ranging from 10-90% in 5% increments with positive, neutral, and negative control sections for visual consistency.

Clone Selection

The dendrogram tree with numerical node identities was visualized, nodes were extracted, and the specific barcodes (Visium Spots) were digitally selected and assigned a clone identity. All members of a given analysis were merged, and a CSV file containing the clone identity and the barcode was output for each Visium section.

Clone Visualization

Loupe Browser Version 5.0 (10x Genomics) was used to spatially visualize resultant clones from clone selection. For the manuscript, if a clone in a given section had ≤ 10 1k or Visium spots, it was not visualized.

Clone Tree consensus iCNV event calling

Both HMM iCNVs, and manual interpretation of denoised outputs were used to identify putative subclonal CNVs. These were then merged in a final consensus set for building clone trees (Supplementary Table 1, 2). Briefly, trees were constructed by identifying where CNVs were shared across clusters identified above as, under the assumption that a CNV cannot be reversed once it occurs, this indicates the cells in those clusters share a common ancestry. We therefore used this logic to identify ancestral relationships between clusters and build the clone tree.

Clone Trees – Branch Lengths

To semi-quantitatively depict the 'evolutionary distance' between subclones, we determined the branch lengths by taking the logarithm (base 2) of the number of additional CNVs in the descendant clone, and adding an arbitrary value to ensure that branches were always visible even with few CNV differences. The formula is given as $b_k = 100\ log_2(|Z_{desc}| - |Z_{par}|) + 300$, where b_k is the length of branch k in pixels.

Clone Trees – Clone Diameters

We scaled the size of each circle denoting a clone by the proportion of spots in the sample that was assigned to a clone using the formula $d_l = 10\ log_2(s_l)$, where d_l denotes the circle diameter in pixels, s_l is the number of spots that correspond to a clone.

SpatialInferCNV Parameters (Fig. 4)

Patient-matched, scRNAseq data from dissociated normal skin cells were analysed for selection (previously described) of a benign reference set. This reference set was then used as a reference control for all ST spots in section T28. Node selection was performed (previously described). One pathologist (R.C.) annotated the resultant clones with the percentage of spots for each clone

that harboured either “Stroma”, “Tumour Epithelia”, or “Non-Invasive Epithelia” (table S9). For inferCNV analysis of the childhood brain tumour, Patient 2 and 3 were selected as reference samples to patient 1. The selected reference samples appeared to demonstrate little to no inferred CNV gains and losses as shown by Extended data fig. 10.

RNA vs DNA phylogenies analysis of previously published single-cell data

DNA and RNA data, co-extracted from single tumour cells, were obtained from publicly available data repositories[3]. Genomic and transcriptomic libraries were aligned to GRCh38.79. DNA-based CNV profiles were analysed and clustered by GINKGO: https://github.com/robertaboukhalil/ginkgo[30]. RNA-profiles were analysed by inferCNV, without a reference set, using default parameters. Tanglegrams of hierarchical clustering of both DNA-based Copy-Number Profiles and RNA-based inferred Copy-Number Profiles were then analysed by the R package *Dendextend*[31].

RNA vs DNA phylogenies from published prostate data

RNA data were obtained for patient A21 [4,28], patient 499 [5], and Cases 6, 7 and 8 [6]. For patients A21 and 499, only a subset of all specimens had transcript data available. For cases 6, 7, and 8, only RNA microarray data were available, precluding their analyses by inferCNV. The transcriptomes were aligned to GRCh38.79, and RNA counts were obtained. These were then processed into inferCNV objects, and ran with standard inferCNV settings, without a reference set**.** Dendrograms from the inferCNV outputs were then visualized using the R. For Cases 6, 7, and 8, we used R scripts to implement a framework to generate phylogenetic trees directly from transcriptomic data[21].

Synthetic Data – Generative Process

To evaluate our application of the computational method InferCNV to spatial transcriptomics data, we designed a generative process that results in an in silico spatial transcriptomics experiment of a tissue with a known - and spatially structured - clonotype population. In short, we construct a spatial domain (representing a tissue region) in which we place a set of virtual cells with a common genome structure, and then let these cells populate the tissue region by simulating growth. In the process, at every time point, cells can move, generate offspring, die or stay stagnant. The generative process above is implemented in Python code and available as a CLI application that can be accessed at GitHub https://github.com/almaan/growmeatissue. The GitHub repository also contains more extensive documentation and examples of how to use the code, the exact parameters - defined in a TOML design file - used to produce the presented data are included in Supplementary Data 1.

Synthetic Data - Evaluation

The process described above was used to generate a set of synthetic data incorporating a single chromosome, from which the obtained spatial gene expression data together with associated annotations were entered as input to spatial InferCNV, this data can be found in the mendeley repository. The synthetic data was analysed according to the same procedure as previously outlined for the real data, providing, as output, information regarding the clonal population as determined from the inferred genomic state. To compare the results with the ground truth we focused exclusively on the set of cells not being used as a reference (non-benign). Spatial InferCNV assigns a state (either three or six depending on which hidden markov model approach is used) to every gene in each clone; we converted these states into a categorization that was more suitable for comparison according to the following scheme, given as “spatial inferCNV

state" - "new category": 1 - deletion, 2 - deletion, 3 - neutral, 4 - amplification, 5 - amplification, 6 - amplification. For the ground truth data, we computed the average copy number of all cells assigned to each spot and rounded this value to the nearest integer. We considered a gene (within a clone) as deleted if the rounded average copy number within the given clone was less than one, amplified if it was higher than one, and neutral if it was equal to one. Having cast the two data sets (real and synthetic) into comparable formats, we then computed the accuracy (within each clone) as the number of equal gene annotations (deletion, neutral, amplifications) between the ground truth and the (from spatial InferCNV) inferred results.

REFERENCES

1. Larsson, L., Frisén, J. & Lundeberg, J. Spatially resolved transcriptomics adds a new dimension to genomics. *Nat. Methods* **18**, 15–18 (2021).
2. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
3. Han, K. Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res.* **28**, 75–87 (2018).
4. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
5. Hong, M. K. H. *et al.* Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat. Commun.* **6**, 6605 (2015).
6. 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).
7. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
8. Egevad, L. *et al.* Standardization of Gleason grading among 337 European pathologists. *Histopathology* **62**, 247–256 (2013).
9. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
10. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9**, 2419 (2018).

11. Grossmann, S. *et al.* Development, maturation, and maintenance of human prostate inferred from somatic mutations. *Cell Stem Cell* (2021) doi:10.1016/j.stem.2021.02.005.
12. Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
13. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
14. Ross-Adams, H. *et al.* Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine* **2**, 1133–1144 (2015).
15. Ji, A. L. *et al.* Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell* **182**, 497-514.e22 (2020).
16. Salmén, F. *et al.* Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. *Nat. Protoc.* **13**, 2501–2534 (2018).
17. Jemt, A. *et al.* An automated approach to prepare tissue-derived spatially barcoded RNA-sequencing libraries. *Sci. Rep.* **6**, 37137 (2016).
18. Navarro, J. F., Sjöstrand, J., Salmén, F., Lundeberg, J. & Ståhl, P. L. ST Pipeline: an automated pipeline for spatial mapping of unique transcripts. *Bioinformatics* **33**, 2591–2593 (2017).
19. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
20. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
21. Costea, P. I., Lundeberg, J. & Akan, P. TagGD: Fast and Accurate Software for DNA Tag

Generation and Demultiplexing. *PLoS One* **8**, e57521 (2013).

22. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).

23. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

24. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, 482 (2020).

25. Garcia, M. *et al.* Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res.* **9**, 63 (2020).

26. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

27. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).

28. Bova, G. S. *et al.* Integrated clinical, whole-genome, and transcriptome analysis of multisampled lethal metastatic prostate cancer. *Cold Spring Harb Mol Case Stud* **2**, a000752 (2016).

29. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).

30. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).

31. Kin, K., Nnamani, M. C., Lynch, V. J., Michaelides, E. & Wagner, G. P. Cell-type Phylogenetics and the Origin of Endometrial Stromal Cells. *Cell Rep.* **10**, 1398–1409 (2015).

ACKNOWLEDGMENTS

This study was supported by The Swedish Cancer Society, Swedish Foundation for Strategic Research, AstraZeneca, and Science for Life Laboratory. The authors also acknowledge The Swedish Childhood Tumour Biobank, supported by The Swedish Childhood Cancer Fund, for access and handling of patient biobank material/sequencing data and The Swedish Childhood Cancer Fund. We would like to thank the National Genomics Infrastructure (NGI), Sweden for providing infrastructure support. We thank Annelie Mollbrink, Monica Nistér and Anders Ståhl, for helpful assistance and discussions. Alastair D. Lamb was supported by a Cancer Research UK Clinician Scientist Fellowship award (C57899/A25812) that also funded Andrew Erickson. Freddie C. Hamdy, Richard Colling and Alastair D. Lamb have received support from the Oxford National Institute for Health Research (NIHR) Biomedical Research Centre Surgical Innovation and Evaluation Theme. Ian G. Mills is grateful to the John Black Charitable Foundation for support and Dan J. Woodcock to the Cancer Research UK Oxford Centre. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

DATA AND CODE AVAILABILITY

Sequence data have been deposited at the European Genome-Phenome Archive (EGA, www.ebi.ac.uk/ega/), which is hosted by the European Bioinformatics Institute (EBI), under accession numbers EGAXXX. Raw fastq files for the childhood brain tumour samples are available through a Materials Transfer Agreement with Monica Nister (monica.nister@ki.se), in

line with GDPR regulations. Count matrices, high-resolution histological images and additional material, are available on Mendeley: https://data.mendeley.com/datasets/svw96g68dv/draft?a=3f263217-2bd3-4a3c-8125-8c517c3a9e29. Details of the ST analysis pipeline can be found at https://github.com/SpatialTranscriptomicsResearch/st_pipeline. The factor analysis software (STD) is available under the GNU General Public License v3 at https://github.com/SpatialTranscriptomicsResearch/std-nb.

AUTHOR CONTRIBUTION

EB, MH, MM, RM, and NS performed the experiments; AE, LB, LK, JB, LL, TR, KT, ALJ, MM, TDDS analysed the data; LB and JM developed the factorization approach; AA, AE designed and implemented the method for synthetic data generation; FT, TH provided tissue; RC, TM, AT, AS undertook pathological analysis; EB, ALJ, PAK, FT, IGM, FCH, DJW provided biological insight; EB, AE, MH, ADL, J.L, drafted the manuscript; all authors read and approved the final manuscript. ADL, JL supervised and managed the study.

Correspondence should be addressed to Alastair Lamb or Joakim Lundeberg.

ETHICS DECLARATION

The study was performed according to the Declaration of Helsinki, Basel Declaration and Good Clinical Practice. The study was approved by the Regional Ethical Review Board (REPN) Uppsala, Sweden before study initiation (Dnr 2011/066/2, Landstinget Västmanland, Sari Stenius), Regional Ethical Review Board (EPN), Stockholm, Sweden (DNR 2018/3-31, Monica Nister). All human subjects were provided with full and adequate verbal and written information

about the study before their participation. Written informed consent was obtained from all participating subjects before enrolment in the study.

COMPETING INTEREST

MH, MJ, RM, LK, AA, LL and JL are scientific consultants to 10x Genomics Inc.

Supplementary Files

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

- SupplementaryInformation.docx
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx
- ExtendedData.pdf