Title: Extensive heterogeneity in somatic mutation and selection in the human bladder

Authors: Andrew R. J. Lawson[1], Federico Abascal[1], Tim H. H. Coorens[1], Yvette Hooks[1], Laura O'Neill[1], Calli Latimer[1], Keiran Raine[1], Mathijs A. Sanders[1,2], Anne Y. Warren[3], Krishnaa T. A. Mahbubani[4,5], Bethany Bareham[4,5], Timothy M. Butler[1], Luke M. R. Harvey[1], Alex Cagan[1], Andrew Menzies[1], Luiza Moore[1], Alexandra J. Colquhoun[6], William Turner[6], Benjamin Thomas[7,8], Vincent Gnanapragasam[9,10], Nicholas Williams[1], Doris M. Rassl[11], Harald Vöhringer[12], Sonia Zumalave[13], Jyoti Nangalia[1], José M. C. Tubío[13,14,15], Moritz Gerstung[12], Kourosh Saeb-Parsy[4,5], Michael R. Stratton[1], Peter J. Campbell[1,16], Thomas J. Mitchell[1,6], Iñigo Martincorena[1]*

Affiliations:

[1]Cancer, Ageing and Somatic Mutation Programme, Wellcome Sanger Institute, Hinxton CB10 1SA, UK.

[2]Department of Hematology, Erasmus University Medical Center, Rotterdam 3015 GD, The Netherlands.

[3]Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK.

[4]Department of Surgery, University of Cambridge, Cambridge CB2 0QQ, UK.

[5]NIHR Cambridge Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK.

[6]Department of Urology, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 0QQ, UK.

[7]The Royal Melbourne Hospital, 300 Grattan Street, Parkville, Victoria 3010, Australia.

[8]The University of Melbourne, Parkville, Victoria 3010, Australia.

[9]Academic Urology Group, Department of Surgery and Oncology, University of Cambridge, Cambridge CB2 0QQ, UK.

[10]Cambridge Urology Translational Research and Clinical Trials Office, University of Cambridge CB2 0QQ, UK.

[11]Department of Pathology, Royal Papworth Hospital NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0AY, UK.

[12]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, UK.

[13]Mobile Genomes and Disease, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain.

[14]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain.
[15]The Biomedical Research Centre (CINBIO), University of Vigo, Vigo 36310, Spain.
[16]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK.
*Correspondence to: *im3@sanger.ac.uk* (I.M.)

Abstract:

The extent of somatic mutation and clonal selection in the human bladder remains unknown. We sequenced 2,097 bladder microbiopsies from 20 individuals, using targeted (n=1,914), whole-exome (n=655) and whole-genome (n=88) sequencing. We found rampant positive selection in 17 genes. Chromatin remodeling genes were frequently mutated, whereas mutations were absent in several major bladder cancer genes. There was extensive inter-individual variation in selection, with different driver genes dominating the clonal landscape across individuals. Mutational signatures were heterogeneous across clones and individuals, suggestive of differential exposure to mutagens in the urine. Evidence of APOBEC mutagenesis was found in 22% of microbiopsies. Sequencing multiple microbiopsies from five patients with bladder cancer enabled comparisons to cancer-free individuals and across histological features. This study reveals a rich landscape of mutational processes and selection in normal urothelium, with large heterogeneity across clones and individuals.

One Sentence Summary:

Normal bladder urothelium is populated by mutant clones carrying cancer-driving mutations, with large heterogeneity in mutational signatures and selection across individuals.

Main Text:

Recent technological developments have started to enable the detection of somatic mutations in normal tissues (*1-15*). One observation derived from these studies is that, as we age, some tissues are colonized by mutant clones carrying driver mutations in cancer genes (*2, 3, 6-8, 11, 15*). These mutations confer a growth advantage driving clonal expansions, some of which are thought to represent the earliest steps towards cancer. However, the extent of this phenomenon remains unclear as driver mutations appear rare in other tissues (*4, 9, 10, 12*).

Bladder urothelium is an interesting tissue in this context. It is one of the slowest dividing epithelia in the human body, being largely quiescent in homeostasis, although able to regenerate quickly upon injury (*16*). Yet, bladder cancers arising from the urothelium have some of the highest mutation burdens of all major cancer types (*17*) and a rich landscape of driver mutations (*18, 19*). Bladder urothelium is also constantly bathed in urine, which can contain mutagenic and carcinogenic molecules known to increase risk of bladder cancer, such as aromatic amines from tobacco smoking, aristolochic acid from certain herbal medicines, and compounds present in dyes, solvents and fumes from occupational and environmental exposures (*20, 21*).

Somatic mutations in normal bladder

To characterize the mutational landscape of normal bladder urothelium both within and across individuals, we performed laser microdissection of small strips of urothelium. Microbiopsies had a median length of 855μm, typically containing a few hundred cells (Fig. 1A). In total, we studied 1,647 microbiopsies from 15 deceased transplant organ donors (ranging 25-78 years of age) and 450 microbiopsies from five patients with bladder cancer (49-75 years, table S1) (*22*). Formalin-free fixation and paraffin embedding were used to ensure high-quality morphology and genome sequencing (*22*).

To search for mutant clones, we performed targeted sequencing of 321 cancer-associated genes for 1,914 microbiopsies (median coverage of 89×) (*22*). To study mutation burden and signatures, copy number changes and selection outside of cancer genes, we performed whole-exome sequencing of 655 microbiopsies (median coverage of 72×) (*22*) and whole-genome resequencing of 88 microbiopsies dominated by large clones (median coverage of 33×, Fig. 1A) (*22*). By sequencing many biopsies per individual, we were able to study the heterogeneity in drivers, burden and signatures across clones and individuals.

In histologically-normal urothelium, we detected a median number of 40 mutations per exome and 1,879 mutations per genome, although the numbers varied considerably across microbiopsies (Fig. 1B and fig. S1) (*22*). Variant allele fractions (VAFs) were moderately low (median exome VAF = 0.13) and most mutations were detected in a single microbiopsy with few shared by adjacent microbiopsies (fig. S2), indicating that mutant clones are typically smaller than the microbiopsy sizes used in this study. Considering the allele fractions and the length of each microbiopsy, we estimate that most mutant clones are smaller than a few hundred micrometers in 1-dimensional sections of urothelium (Fig. 1C) (*22*), consistent with estimates derived from mitochondrial markers (*23*). This shows that histologically-normal bladder urothelium is a patchwork of small, typically microscopic, mutant clones.

Below, we first describe the mutational landscape of healthy bladder by focusing on data from the 15 transplant organ donors (Figs. 2 and 3), followed by an analysis of microbiopsies from the five patients with bladder cancer (Fig. 4).

Widespread positive selection in normal urothelium

To determine whether positive selection on certain genes drives these clonal expansions, we used the ratio of non-synonymous to synonymous mutation rates (dN/dS). Mutations driving clonal expansions become overrepresented among mutant clones reaching detectable sizes, which manifests as an excess of non-synonymous mutations in driver genes (*22*). We used *dNdScv*, an implementation of dN/dS that corrects for trinucleotide mutation rates, sequence composition and variable rates across genes (*19*). Applying it to the 321 cancer genes sequenced in 1,500 microbiopsies of normal urothelium from the transplant organ donors revealed significant positive selection on 12 genes (*22*): *KMT2D* (also known as *MLL2*), *KDM6A* (also known as *UTX*), *ARID1A*, *RBM10*, *EP300*, *STAG2*, *NOTCH2*, *CDKN1A*, *CREBBP*, *FOXQ1*, *RHOA* and *ERCC2* (Fig. 2A). Using restricted hypothesis testing on known bladder cancer genes and a dN/dS model at the level of single hotspots, we identified an additional five genes under selection: *KLF5*, *ZFP36L1*, *ELF3*, *GNA13* and *PTEN* (*22*). Overall, 17 genes were found to be under clear positive selection, conferring on the mutant cells a competitive advantage over neighboring cells.

The enrichment of non-synonymous mutations in positively-selected genes was large, with dN/dS ratios higher than 10 or even 100 (Fig. 2B). In most genes, selection on protein-truncating mutations (indels, nonsense and essential splice site substitutions) was stronger than on missense mutations, a pattern characteristic of tumor suppressor genes (*19*). In fact, while indels contributed just under 8% of all detected mutations across exomes and genomes, they accounted for 39% of all driver mutations. Clear exceptions were *RHOA*, *ERCC2* and *GNA13*, which displayed higher frequencies of missense mutations, typically at known oncogenic hotspots (Fig. 2B and fig. S3). Overall, based on the excess of non-synonymous mutations measured by dN/dS, we detected a total of 385 (CI95%: 357, 401) driver mutations across all microbiopsies (*22*).

We can integrate allele fractions to estimate the proportion of cells in bladder urothelium that carry a driver mutation, while accounting for the possibility of undetected copy number losses and mutations occurring in one or two alleles per cell (Fig. 2C) (*22*). This conservatively estimates that between 8 and 19% of cells carry a driver mutation in normal bladder in middle-age and elderly individuals.

Chromatin remodeling genes dominate the driver landscape

Of the 17 positively-selected genes, all but *NOTCH2* have been identified as bladder cancer genes from TCGA data (*18, 19*) (Fig. 2D). In contrast to the case of *NOTCH1* in normal esophagus (*7, 8*), the mutation frequency of these 17 genes is higher in bladder cancers than in normal urothelium from middle-age and elderly individuals in our cohort. This suggests that these mutations confer on the mutant cells an increased tumorigenic potential, even if the risk of progression of individual clones is extremely small. Most common bladder cancer genes can be classified into three functional groups: the RTK/Ras/PI3K pathway (such as *PIK3CA*, *FGFR3*, *ERBB2* and *ERBB3*), the p53/Rb pathway (such as *TP53*, *RB1* and *ATM*) and genes involved in chromatin remodeling (*18, 24*). Five of the top six most mutated driver genes in normal bladder are involved in chromatin remodeling, whereas mutations in RTK/Ras/PI3K or p53/Rb genes that are very common in bladder cancer are much rarer in normal urothelium (Fig. 2E).

The absence of mutations in some of the main bladder cancer genes was noteworthy. Across 1,500 microbiopsies, we only found three independent mutations in *TP53*, which is mutated in nearly 50% of muscle-invasive bladder cancer, and no mutations in *FGFR3*, which is mutated in 60-80% of non-muscle-invasive bladder cancers (*25*). We also did not detect any *TERT* promoter mutations across 55 whole-genomes of normal urothelium, despite it being mutated in around 70-80% of bladder cancers, including early stage bladder cancers (*26*). This suggests that these driver

mutations may not confer large clonal advantages in normal urothelium, but are key drivers of bladder cancer development. Detection of mutations in these genes in liquid biopsies may prove informative for early detection of bladder cancer (*26*).

The analyses above were restricted to the targeted panel of 321 known cancer genes. The extent of selection in normal tissues outside known cancer genes is less understood. It is conceivable that mutation of certain genes could drive benign clonal expansions in healthy tissues without contributing to tumorigenesis or even push cells down evolutionary paths away from cancer. Running *dNdScv* on all genes using 483 whole-exomes from normal urothelium, yielded seven genes under clear positive selection, all within the list of 17 genes above (*22*). This confirms that the main drivers of clonal expansions in normal urothelium are all known cancer genes. Somatic mutations could also lead to cellular death or differentiation, which would lead to a depletion of protein-altering mutations in surviving clones. While this dataset is not powered to detect negative selection at the level of individual genes, exome-wide dN/dS ratios excluding known cancer genes were close to, and not significantly lower than 1 (Fig. 2F). This is consistent with the vast majority of somatic coding point mutations being tolerated by normal cells and accumulating passively, in line with observations in cancer genomes (*19*). Similar non-significant results were obtained when focusing on putative antigenic regions of the exome, providing no clear evidence of immune editing against these mutant clones (fig. S4) (*22*).

Extreme variation in driver preference across individuals

Having identified many independent mutant clones per donor, we were able to study differences in selection across individuals. We used a dN/dS-based likelihood-ratio test that compares the relative enrichment of non-synonymous mutations in particular genes, while correcting for differences in mutation rates, mutation signatures, coverage and selection at other genes (*22*). This

analysis revealed striking differences in the landscape of clonal selection across donors (Fig. 2G and fig. S4). For example, one individual (T03_53F) had 35 different *KDM6A* mutations and two *ARID1A* mutations, whereas another (T06_59M) had four *KDM6A* mutations and 20 *ARID1A* mutations (Fig. 2, G-I). The four most frequently mutated genes in our dataset, *KMT2D*, *KDM6A*, *ARID1A* and *RBM10*, all showed highly-significant differences in selection across donors (Fig. 2G, q-values<0.05 from dN/dS likelihood-ratio tests) (*22*).

It is unclear whether these differences are driven by variability in environmental exposures or by the genetic background of each individual. No clear evidence of pathogenic germline mutations was found in these genes (*22*). *KDM6A* and *RBM10* are both located on the X-chromosome and *KDM6A* is known to escape X-chromosome inactivation, with some evidence suggesting that both *KDM6A* and *RBM10* are more frequently mutated in males across cancer types (*27*). However, in our limited cohort, *KDM6A* appears more frequently mutated in women than men, in line with previous observations in non-muscle-invasive bladder cancer (*28*). Larger cohorts would be required to establish robust associations between epidemiological factors and differences in somatic mutation rates and selection.

Large heterogeneity in burden and signatures across clones and donors

The whole-exome data showed an increase in the number of mutations detected with age, consistent with continual, irreversible accumulation of mutations during life (Fig. 3A). To estimate the mutation burden per cell despite the presence of multiple clones per microbiopsy, we used two alternative approaches to obtain lower-bounds from the whole-genome data: integration of allele frequencies and deconvolution of the major subclone (*22*). We estimate that, by middle age (50-65 years), cells in normal urothelium carry over 500-2,000 mutations per genome. This burden is

within the range observed for other normal tissues (*1, 4, 7*), but an order of magnitude lower than the typical burden of bladder cancers (Fig. 3B).

Analysis of the mutational spectra revealed striking differences across donors (Fig. 3C). To better understand this variation, we performed *de novo* mutational signature decomposition in 80 genomes of normal urothelium from all 20 individuals using a Bayesian hierarchical Dirichlet process, and matched these signatures to known signatures from cancer genomes (fig. S5-S6) (*22*). This identified four main signatures that contribute over 89% of all mutations in the dataset (Fig. 3, D-H). The same four signatures were found using non-negative matrix factorization (*SigProfiler*) (fig. S7A) (*22*).

One signature, the third most abundant, was clearly attributable to APOBEC mutagenesis (cosine similarity with SBS2+SBS13 = 0.995)(*29*). The high mutation burden in bladder cancers is largely driven by activation of APOBEC3 cytidine deaminases, which preferentially generate C>G and C>T changes in a TCN context (Fig. 3G) (*17*). APOBEC mutagenesis has been only rarely reported in normal tissues sequenced to date (*8, 9, 15*), but it occurs frequently in normal urothelium, contributing hundreds to thousands of mutations in the clones in which it is active (Fig. 3D).

The other three signatures did not match known signatures (fig. S6). Signatures A and B may contain a fraction of SBS5 mutations, which are common in bladder cancers (*17*), but they were stably extracted as separate from small amounts of SBS5 when using known signatures as priors or when adding cancer genomes to the signature extraction (fig. S7-S8) (*22*). Signature A is dominated by T>C changes, with a clear transcriptional strand bias suggestive of transcription-coupled damage or repair (Fig. 3E and fig. S9). Reanalysis of whole-genome data from the PCAWG consortium suggests a high contribution of signature A to some bladder cancer genomes (fig. S6C-E) (*22*). Signature B is dominated by C>T changes (Fig. 3F) and shares some

resemblance with SBS5 in combination with a C>T-rich signature with a modest transcriptional strand bias (fig. S6 and S9). Signature C has distinct peaks at T>A and T>G in an ATT context (Fig. 3H) and does not resemble any known signature or combination of signatures (fig. S6). It has a strong transcriptional strand asymmetry with lower mutation rates in transcribed regions (fig. S9), a pattern indicative of this signature being generated by DNA damage to thymines by adducts and subject to transcription-coupled repair (*9*). Signature C also has an extended sequence context dominated by adenines and thymines (fig. S10).

The relative contribution of different signatures within each individual was particularly interesting. APOBEC mutations are responsible for large differences in mutation burden and spectra between clones (Fig. 3D). This contrasts with signatures A-C, which show little variation across clones from the same individual but large differences between individuals (Fig. 3D). For example, signature A contributes ~70% of mutations in all clones from a 53 year-old woman (T03_53F), but is scarcely present (~5% of all mutations) in all clones from a 61 year-old woman (T08_61F). Similarly, signature C contributes over 25% of all mutations in six of the 15 donors, but is undetectable in others (Fig. 3D). The inter-individual differences in mutational signatures, together with the diverse etiology of bladder cancers, is suggestive of variable mutagenic exposures through the urine. This is exemplified by the presence of aristolochic acid mutagenesis in normal urothelium from Chinese patients (*30*). Smoking is a major risk factor of bladder cancer, increasing risk by 3-4 fold (*20*). No evidence of the smoking-associated signature (SBS4) was found in any of the individuals, including heavy smokers (table S1), a pattern consistent with the lack of SBS4 in bladder cancers from smokers (*31*). We used a linear mixed-effect regression model to test whether any of the four signatures found may be statistically associated with smoking or alcohol consumption. Despite the small cohort size, signature A was significantly associated with smoking

history (linear mixed-effect regression, *P*-value=9.4e-05, fig. S11) (*22*), raising the possibility that signature A may result from tobacco smoke mutagens excreted in the urine.

One additional source of heterogeneity across clones was exemplified by the microbiopsy with the highest mutation burden of the cohort, which contained ~6,500 mutations (Fig. 3D and fig. S12). This genome carried a hotspot mutation (N238T) in *ERCC2*, which is known to cause hypermutation in some bladder cancers through aberrant nucleotide excision repair (*32*). A total of 8 different *ERCC2* mutations were identified in the targeted and exome data, with clear positive selection acting on *ERCC2* (Fig. 2), suggesting that this mechanism is relatively common in normal urothelium.

Frequency and spatial distribution of APOBEC clones

APOBEC-induced mutations in normal urothelium displayed the characteristic replicational strand bias observed in human cancers and an extended sequence context suggestive of APOBEC3A being the main contributing enzyme (fig. S10) (*22, 33*). Analysis of APOBEC-positive genomes revealed extensive evidence of mutational clusters, known as kataegis (Fig. 3I) (*17*). These clusters were modest in size and displayed the typical strandedness observed in cancer genomes. While kataegis in cancers is often reported to occur near rearrangement breakpoints (*17*), this was not the case in normal urothelium. Overall, the patterns observed here are consistent with replication-associated APOBEC mutagenesis (*34*).

Analysis of the distribution of APOBEC-positive genomes in their tissue context revealed a suggestive example of spatial clustering of three APOBEC-positive clones (Fig. 3J). To study the frequency and spatial distribution of APOBEC-positive clones, we used signature fitting and a likelihood ratio test to annotate all exomes according to their evidence of APOBEC mutagenesis (*22*). Across donors, 22% of all microbiopsies of normal urothelium showed evidence of APOBEC

mutagenesis (likelihood-ratio test q-value<0.05, Fig. 3K). To determine whether APOBEC-positive clones tend to cluster in space, for each positive clone we calculated the fraction of positive clones surrounding it (Euclidean distance <1mm), both in the real data and in random permutations of the data (fig. S13) (*22*). This analysis suggests that APOBEC clones appear to be scattered uniformly in the tissue (permutation test *P*-value=0.92), without evidence of spatial clustering of unrelated clones, suggesting that APOBEC mutagenesis is typically triggered independently in individual cells across the urothelium.

Copy number and rearrangement analyses of normal urothelium revealed that the majority of clones carry no structural variants (*22*). Copy number alterations were detected in only 28% of urothelial exomes, with the most common changes involving whole or arm-level gains of chromosomes 13, 14, 15 and 16, and losses of chromosomes 9 and 21 (Fig. 3L). Across 55 genomes of normal urothelium, only 30 rearrangements and 3 retrotransposition events were detected (tables S7 and S11) (*22*). This is in stark contrast with bladder cancers, which display extensive aneuploidy, with an average of ~200 segmental alterations per exome and 1.7 retrotransposition events per genome (*35, 36*). This pattern is similar to that observed in other normal tissues (*3, 4, 7, 9, 37*), and it suggests that extensive structural changes are characteristic of later stages of carcinogenesis across a wide range of cancer types.

The mutational landscape in bladder cancer patients

Bladder cancer often presents with multiple synchronous tumors in different parts of the bladder. It remains unclear to what extent this is due to large premalignant clones colonizing distant parts of the bladder or to widespread changes in multiple independent clones across the bladder (*38*). To explore the mutational landscape of histologically-normal urothelium in bladder cancer patients

and to study the genomic changes underlying histologically abnormal areas, we performed laser microdissection of 450 microbiopsies from 19 distant biopsies from five bladder cancer patients.

Analysis of histologically-normal urothelium from bladder cancer patients revealed patterns similar to those observed in healthy bladders. As in transplant organ donors, mutant clones were small, typically constrained to single microbiopsies (fig. S2). There seems to be a modest increase in the number of mutations detected per exome (linear mixed-effect regression *P*-value=0.0068) and in their allele frequencies (*P*-value=0.00048) in some cystectomy samples (fig. S14) (*22*). However, differences should be interpreted with caution given the limited cohort size and the considerable inter-individual variation. The fraction of APOBEC-positive microbiopsies was similar in cystectomies and in age-matched transplant organ donors (25% vs 24%, Fisher's Exact Test *P*-value=0.91). Driver discovery in 223 microbiopsies of normal-urothelium from bladder cancer patients yielded a very similar driver landscape to that observed in the 15 transplant organ donors and the density of driver mutations detected per microbiopsy appeared comparable (*22*). Although a much larger number of patients would be required to accurately quantify differences between cohorts, these results suggest that the mutational landscape of histologically-normal urothelium from bladder cancer patients broadly resembles the patchwork of microscopic clones observed in healthy donors. They also suggest that widespread mutational changes in independent clones are unlikely to explain the emergence of multiple tumors in bladder cancer, consistent with the observation that synchronous tumors tend to be clonally related (*38-40*).

Areas of carcinoma *in situ* (CIS) were observed in three of the five cystectomies studied. CIS of the bladder is a flat, high-grade urothelial carcinoma restricted to the epithelial layer, which often appears concomitantly with more advanced tumors. 44 CIS microbiopsies were sequenced, including 11 whole-exomes and 5 whole-genomes. Phylogenetic analysis revealed that all CIS

areas sequenced within a patient were clonally related (Fig. 4, A to D, and fig. S15 and S16). In a 72-year-old patient (C04_72M), the same CIS clone was detected in two biopsies several centimeters away from the tumor and from one another, with most mutations being shared across distant biopsies (Fig. 4C). The phylogenetic tree provides a snapshot of the genome of the most recent common ancestor cell that gave rise to this clone. This cell had only a modestly increased burden, largely due to APOBEC, compared to other clones in normal urothelium, but had already acquired driver mutations in *ARID1A*, *RB1* and *TP53*, as well as a hotspot promoter mutation in *TERT* (Fig. 4C). In contrast to histologically-normal clones, the CIS showed extensive aneuploidy, including evidence of whole-genome duplication (Fig. 3M). Intriguingly, one of the terminal branches of the CIS clone showed an unusually-high number of CC>AA dinucleotide changes of uncertain origin (Fig. 4C and fig. S17). In a 67-year-old patient (C03_67M) we sequenced an area of CIS and an area of tumor from two separate biopsies. This revealed that the tumor and the CIS had originated from a common ancestor cell that had already acquired putative driver mutations in *NUP93*, *EPHA2* and *TERT*. The CIS and the tumor diverged early and each subsequently acquired an entirely different complement of driver mutations (Fig. 4D), providing a window into the early evolution of this tumor. This analysis corroborates that CIS clones are genetically highly aberrant and can colonize distant areas of the bladder, forming a hotbed from which invasive tumors can evolve (*40, 41*). A systematic analysis of tumor and non-invasive areas combining laser microdissection and genome sequencing could help shed light on the order of events in early bladder cancer evolution.

Laser microdissection also enabled us to study other histological changes observed in bladder cancer patients. Von Brunn's nests are groups of urothelial cells in the lamina propria, believed to arise from invagination of the surface urothelium (*42*). Although they are common in histological

sections from bladder cancer patients (Fig. 4B), they can also be seen in small numbers in healthy individuals. Sequencing of 98 microbiopsies revealed that most von Brunn's nests are single clones, with all cells within a nest derived from a single cell (Fig. 4, E and F). Phylogenetic reconstruction reveals that adjacent nests are clonally unrelated (Fig. 4D). The vast majority of von Brunn's nests sequenced did not carry a driver mutation; their driver landscape, mutation burden and largely diploid genomes resembled that of the adjacent histologically-normal urothelium. Overall, this is consistent with von Brunn's nests being benign ectopic growths not actively driven by specific mutations (*22*). Lymphoid aggregates are also common in cystectomy biopsies (Fig. 4A), reflecting adaptive immunity in the tumor microenvironment, and can also occur in healthy samples with evidence of inflammation (*43*). We microdissected 82 lymphoid aggregates for deep targeted sequencing, as the targeted gene panel contained probes for the B-cell and T-cell receptor loci (*22*). Unlike von Brunn's nests, lymphoid aggregates were highly polyclonal, with nearly all of the mutations detected at low allele fractions (Fig. 4G). The only exception was one clonal lymphoid aggregate, which also carried a lymphoid driver IgH/BCL2 translocation (fig. S18). This biopsy was from a donor who had been previously investigated for a possible lymphoma, although the relationship between the clonal lymphoid aggregate and the donor's clinical history is unclear. Across all lymphoid aggregates, 95% of mutations detected with the panel clustered in the *IGH* locus and had the characteristic signature of somatic hypermutation (SBS9) (Fig. 4H), confirming the presence of multiple clones of mature B lymphocytes in each aggregate sequenced. These examples showcase the power of laser microdissection and low-input sequencing to inform on the clonal composition and genetic changes underlying different histological structures.

Discussion

These data have revealed a rich mutational landscape in healthy and diseased bladder urothelium, with widespread positive selection, extensive APOBEC mutagenesis and large differences in mutation burden, signatures and selection across clones and across individuals.

The heterogeneity in mutational signatures and driver mutations across donors is particularly intriguing and appears larger than that reported in other tissues. Epidemiological studies have linked bladder cancer risk to a diversity of carcinogens, such as smoking, occupational or environmental exposures and recurrent infections (*20, 44*). Whether carcinogens are genotoxic (inducing mutations) or non-genotoxic (impacting cellular growth or the microenvironment), they are expected to leave distinct marks in the mutational landscape of normal tissues, altering mutation rates, mutation signatures, driver frequencies or clone sizes. Thus, the differences in the mutational landscape across individuals observed here may be expected to reflect the interplay between genetics and a lifetime of different exposures. The differences across donors might raise the possibility of developing personalized risk models (*45*). However, our results also suggest that differences in normal urothelium between healthy individuals and cancer patients may be subtle, consistent with theories predicting that modest differences in mutation and selection could have considerable impact on risk (*46, 47*). Systematic analyses of large cohorts of individuals will be needed to quantify the relationship between epidemiological factors, germline variants, changes in the mutational and selective landscape, and risk; enabling the development of mechanistic risk models of cancer development.

While somatic mutations have traditionally been studied in the context of cancer, the growing realization that some human tissues become colonized by mutant clones throughout life raises questions about their potential impact in ageing and other diseases. Laser microdissection and low-

input sequencing enable the study of somatic mutations associated with histological changes, and could shed new light on somatic evolution in cancer, ageing and non-malignant disease.

Acknowledgments:

We are grateful to the families of deceased transplant organ donors and to the patients with bladder cancer for their consent, and to the Cambridge Biorepository for Translational Medicine for access to human tissue. We thank P.H. Jones and J.C. Fowler for their early help with wholemounts; L. Alexandrov for advice on mutational signatures; K. Haase and P. van Loo for their advice on calling copy number changes in exome data using ASCAT; J.M.A. Lawson for artistic contribution to figures; D. Phillips for advice on carcinogen exposure in urine; P. Ellis, P. Nicola, M. Maddison, E. Anderson, S. Gamble, K. Roberts and A. Dooner for technical assistance; J. Hewinson and C. Hardy for their assistance with project management; J. Field-Rayner for consenting patients; and E. Cromwell for tissue processing. **Funding:** I.M. is funded by Cancer Research UK (C57387/A21777) and the Wellcome Trust. P.J.C. is a Wellcome Trust Senior Clinical Fellow. T.J.M. is funded by Cancer Research UK/Royal College of Surgeons Clinician Scientist Fellowship (C63474/A27176). L.M. is a recipient of a CRUK Clinical PhD fellowship (C20/A20917). Fresh cystectomy samples were acquired as part of the DIAMOND study "Evaluation of biomarkers in urological disease - NHS National Research Ethics Service reference 03/018", whose infrastructure is part-funded by the Cambridge NIHR BRC and CRUK Cambridge Cancer Centre Urological Malignancies programme. **Author contributions:** A.R.J.L. and I.M. conceptualized the project, with support from P.J.C., M.R.S. and T.J.M. A.R.J.L. and I.M. led the data analysis, with support from F.A., T.H.H.C., H.V. and S.Z. A.R.J.L led the experimental work, with support from Y.H., L.M.R.H. and A.Ca. T.M.B. and L.M contributed to method development. K.R., M.A.S., A.M., N.W., H.V., J.N., M.G. and I.M developed algorithms and software. L.O., C.L., and K.T.A.M helped with samples and project administration. A.Y.W., K.T.A.M., B.B., A.Co., W.T., B.T., V.G. and K.S.-P collected samples. J.N., J.M.C.T., M.G., K.S.-P., M.R.S., P.J.C., T.J.M. and I.M provided supervision. D.M.R provided histology support. I.M. and A.R.J.L wrote the manuscript, and all authors contributed to reviewing and editing it. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** Sequencing data is available in the European Genome-phenome Archive (EGA): EGAD00001006113, EGAD00001006114, EGAD00001006115, EGAD00001006116 and EGAD00001006117. Reproducible code is available in the supplementary material and in *https://doi.org/10.5281/zenodo.3966023*.

Supplementary Materials

Materials and Methods (including figs. S1 to S18, table S1 and captions for tables S2 to S11).
Tables S2 to S11.

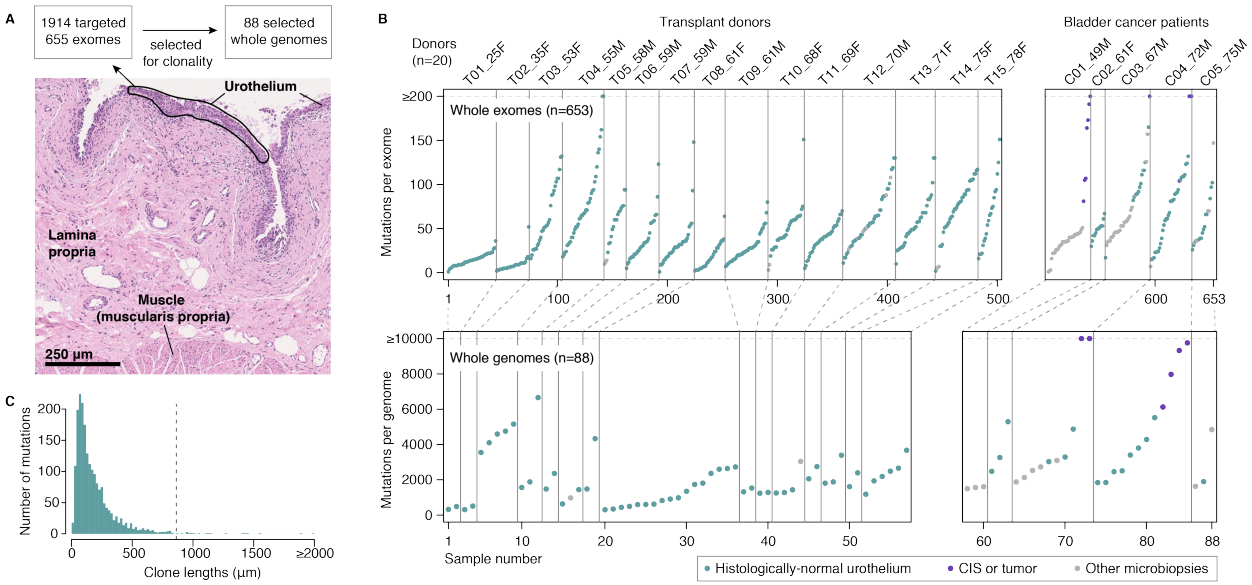


Fig. 1. Detection of somatic mutations in bladder by laser microdissection and low-input sequencing.

(**A**) Sequencing strategy and histology image of bladder mucosa (hematoxylin and eosin staining). (**B**) Combined number of substitutions and indels detected per exome (top) and whole-genome (bottom) across 15 transplant organ donors and 5 patients with bladder cancer. Donor identifiers contain age and gender information in suffix. (**C**) Distribution of estimated clone lengths for histologically normal urothelium (median indicated by a dashed line) (*22*).

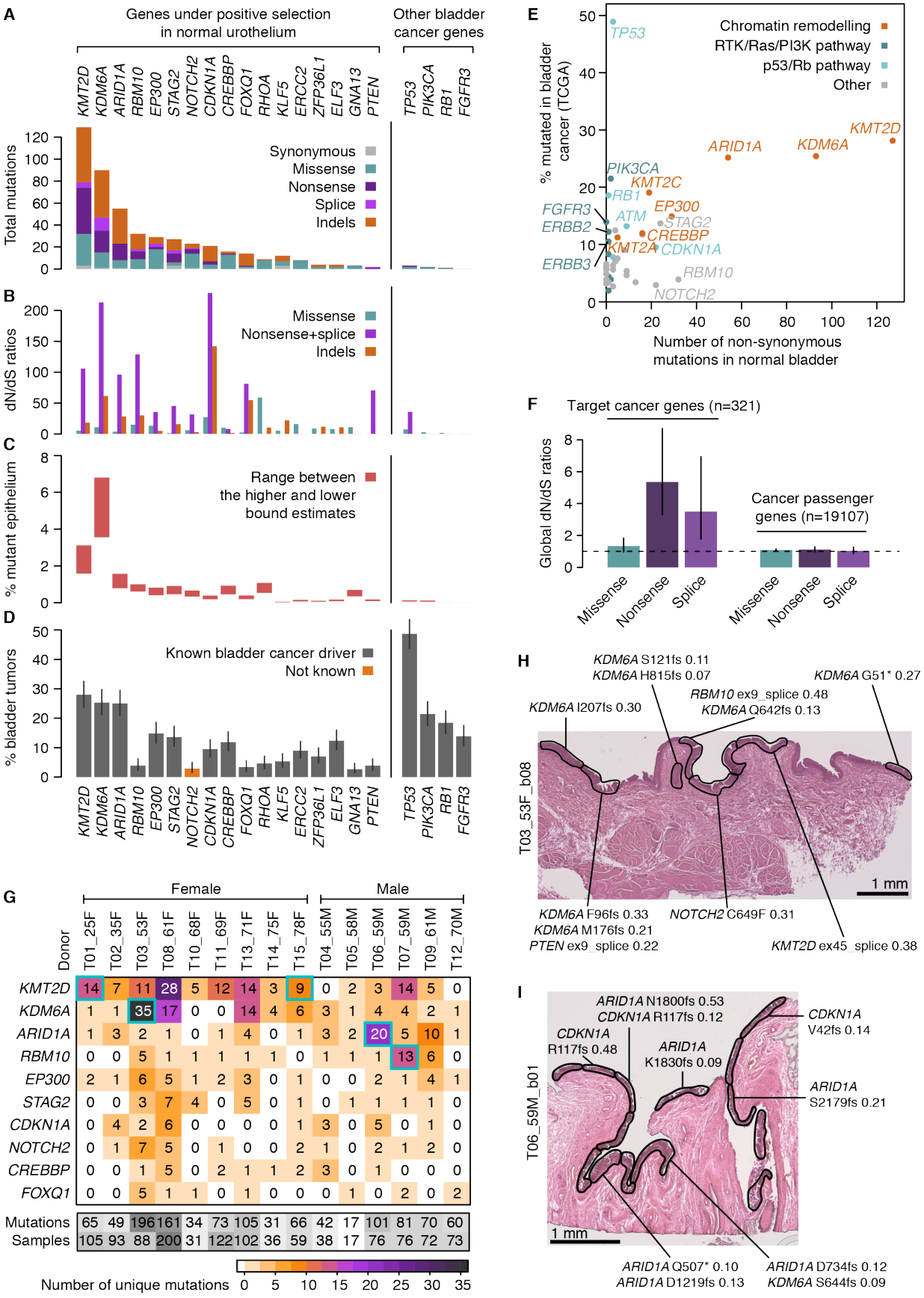

A
Genes under positive selection in normal urothelium
Other bladder cancer genes
KMT2D KDM6A ARID1A RBM10 EP300 STAG2 NOTCH2 CDKN1A CREBBP FOXQ1 RHOA KLF5 ERCC2 ZFP36L1 ELF3 GNA13 PTEN
TP53 PIK3CA RB1 FGFR3
Total mutations
Synonymous
Missense
Nonsense
Splice
Indels
B
dN/dS ratios
Missense
Nonsense+splice
Indels
C
% mutant epithelium
Range between the higher and lower bound estimates
D
% bladder tumors
Known bladder cancer driver
Not known
E
% mutated in bladder cancer (TCGA)
Chromatin remodelling
RTK/Ras/PI3K pathway
p53/Rb pathway
Other
TP53
KMT2D
KDM6A
ARID1A
PIK3CA
KMT2C
RB1
FGFR3
ATM
EP300
ERBB2
STAG2
CREBBP
KMT2A
CDKN1A
ERBB3
RBM10
NOTCH2
Number of non-synonymous mutations in normal bladder
F
Target cancer genes (n=321)
Global dN/dS ratios
Cancer passenger genes (n=19107)
Missense
Nonsense
Splice
G
Female
Male
Donor
T01_25F T02_35F T03_53F T08_61F T10_68F T11_69F T13_71F T14_75F T15_78F T04_55M T05_58M T06_59M T07_59M T09_61M T12_70M
KMT2D 14 7 11 28 5 12 14 3 9 0 2 3 14 5 0
KDM6A 1 1 35 17 0 0 14 4 6 3 1 4 4 2 1
ARID1A 1 3 2 1 0 1 4 0 1 3 2 20 5 10 1
RBM10 0 0 5 1 1 1 1 1 0 1 1 1 13 6 0
EP300 2 1 6 5 1 2 3 0 1 0 0 2 1 4 1
STAG2 0 0 3 7 4 0 5 0 1 0 1 1 1 1 0
CDKN1A 0 4 2 6 0 0 0 0 1 3 0 5 0 1 0
NOTCH2 0 1 7 5 0 1 0 0 2 1 0 2 1 2 0
CREBBP 0 0 1 5 0 2 1 1 2 3 0 1 0 0 0
FOXQ1 0 0 5 1 1 0 1 0 0 0 1 0 2 0 2
Mutations 65 49 196 161 34 73 105 31 66 42 17 101 81 70 60
Samples 105 93 88 200 31 122 102 36 59 38 17 76 76 72 73
Number of unique mutations 0 5 10 15 20 25 30 35
H
KDM6A S121fs 0.11
KDM6A H815fs 0.07
KDM6A G51* 0.27
KDM6A I207fs 0.30
RBM10 ex9_splice 0.48
KDM6A Q642fs 0.13
T03_53F_b08
1 mm
KDM6A F96fs 0.33
KDM6A M176fs 0.21
PTEN ex9_splice 0.22
NOTCH2 C649F 0.31
KMT2D ex45_splice 0.38
I
ARID1A N1800fs 0.53
CDKN1A R117fs 0.12
CDKN1A V42fs 0.14
CDKN1A R117fs 0.48
ARID1A K1830fs 0.09
ARID1A S2179fs 0.21
T06_59M_b01
1 mm
ARID1A Q507* 0.10
ARID1A D1219fs 0.13
ARID1A D734fs 0.12
KDM6A S644fs 0.09

Fig. 2. Positive selection of bladder cancer genes in normal urothelium from organ donors. In panels (A) to (D), analyses are shown for 17 genes under positive selection in normal urothelium and for four other genes frequently mutated in bladder cancer. (**A**) Number and consequence of mutations detected in histologically-normal urothelium. (**B**) Observed-to-expected ratios for missense substitutions, truncating (nonsense and essential splice site) substitutions, and indels. (**C**) Estimated percentage of urothelial cells bearing a mutation for donors aged ≥50 from samples with median on-target coverage ≥50×. (**D**) Percentage of urothelial carcinomas in The Cancer Genome Atlas (TCGA) with a non-synonymous substitution or indel. Error bars depict 95% Binomial confidence intervals. (**E**) Scatter plot comparing mutation frequency in bladder cancer (panel D) and the number of non-synonymous mutations in normal urothelium (panel A) for driver genes (colored by biological function) identified in this study and in (*18, 19*). (**F**) Comparison of dN/dS values for the 321 cancer genes in the targeted panel to 19,107 cancer passenger genes (defined in (*19*)). Dashed line indicates a dN/dS value of 1, indicating neutral expectation. (**G**) Heatmap showing the number of unique non-synonymous mutations in abundant (≥10 mutations) driver genes across transplant organ donors. Sample numbers refer to samples with at least one mutation. Blue boxes indicate statistically-significant combinations of gene and donor (*22*). (**H** and **I**) Histology images annotated with driver mutations and their cellular fractions in sequenced microbiopsies from two transplant organ donors exhibiting enrichment of drivers in *KDM6A* and *ARID1A* respectively.

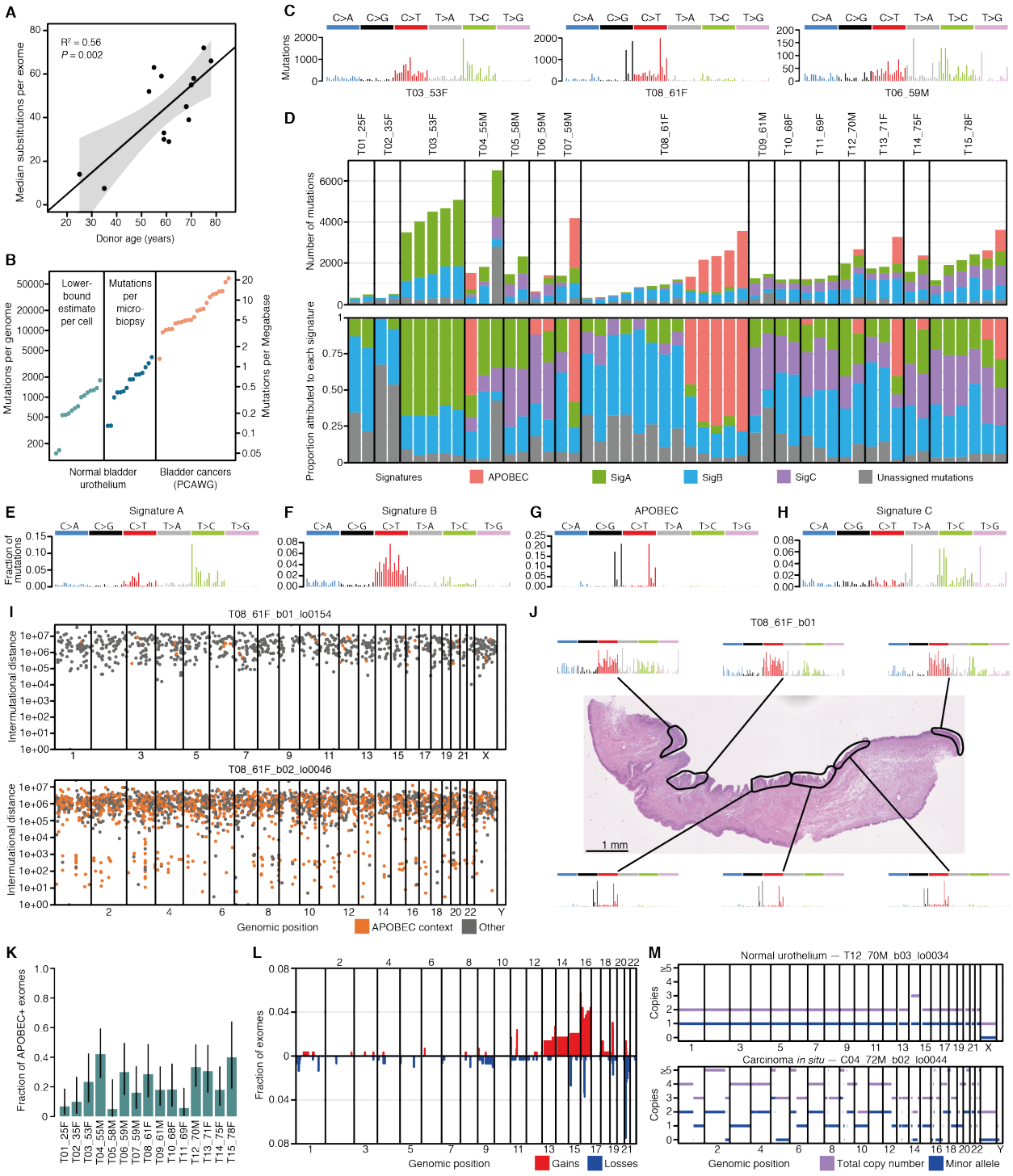

A
Median substitutions per exome
Donor age (years)
B
Mutations per genome
Mutations per Megabase
Lower-bound estimate per cell
Mutations per micro-biopsy
Normal bladder urothelium
Bladder cancers (PCAWG)
C
Mutations
T03_53F
T08_61F
T06_59M
D
Number of mutations
Proportion attributed to each signature
Signatures
APOBEC
SigA
SigB
SigC
Unassigned mutations
E
Signature A
F
Signature B
G
APOBEC
H
Signature C
Fraction of mutations
I
T08_61F_b01_lo0154
T08_61F_b02_lo0046
Intermutational distance
Genomic position
APOBEC context
Other
J
T08_61F_b01
1 mm
K
Fraction of APOBEC+ exomes
L
Fraction of exomes
Genomic position
Gains
Losses
M
Normal urothelium — T12_70M_b03_lo0034
Carcinoma in situ — C04_72M_b02_lo0044
Copies
Genomic position
Total copy number
Minor allele

Fig. 3. Mutation burden and signatures in normal urothelium.
(**A**) Scatter plot of donor age vs the median number of substitutions in high-coverage exomes (≥40× for ≥80% of the exome). The fitted line, R^2 value, and *P* value were obtained by linear regression. (**B**) Comparison of mutation burden between normal bladder urothelium and bladder cancers. In order to account for subclonality, both a mean lower-bound estimate per cell (*22*) and the mean number of mutations per microbiopsy are shown for whole-genomes from the 15 transplant organ donors. Bladder cancer data reflects total mutations per genome from Pan-Cancer Analysis of Whole Genomes (PCAWG)(*48*). (**C**) Raw mutational spectra for all urothelial genomes combined for three donors. (**D**) Number (top) and proportion (bottom) of mutations assigned to the four most abundant signatures extracted using a Bayesian hierarchical Dirichlet process (*22*) for urothelial genomes from transplant organ donors. The weak attribution of signature C to genomes from T08 may reflect overfitting to residual ATT>AAT alignment errors. (**E** to **H**) Bar plots depicting mutational spectra, split by type and trinucleotide context, of extracted signatures, as in (*17*). (**I**) Intermutational distance plots for urothelial clones free from and affected by APOBEC activity respectively, as in (*17*). (**J**) Histology image depicting variability in mutational processes between nearby urothelial microbiopsies. Mutational spectra are from independent clones. (**K**) Fraction of exomes with evidence of APOBEC mutagenesis (*22*). Error bars depict 95% Binomial confidence intervals. (**L**) Proportion of exomes from normal urothelium with large-scale copy number alterations in autosomes (*22*). Gains (red) and losses (blue) are shown above and below the x-axis respectively. (**M**) Copy number plots for representative whole-genomes of normal urothelium (top) and carcinoma *in situ* (bottom).

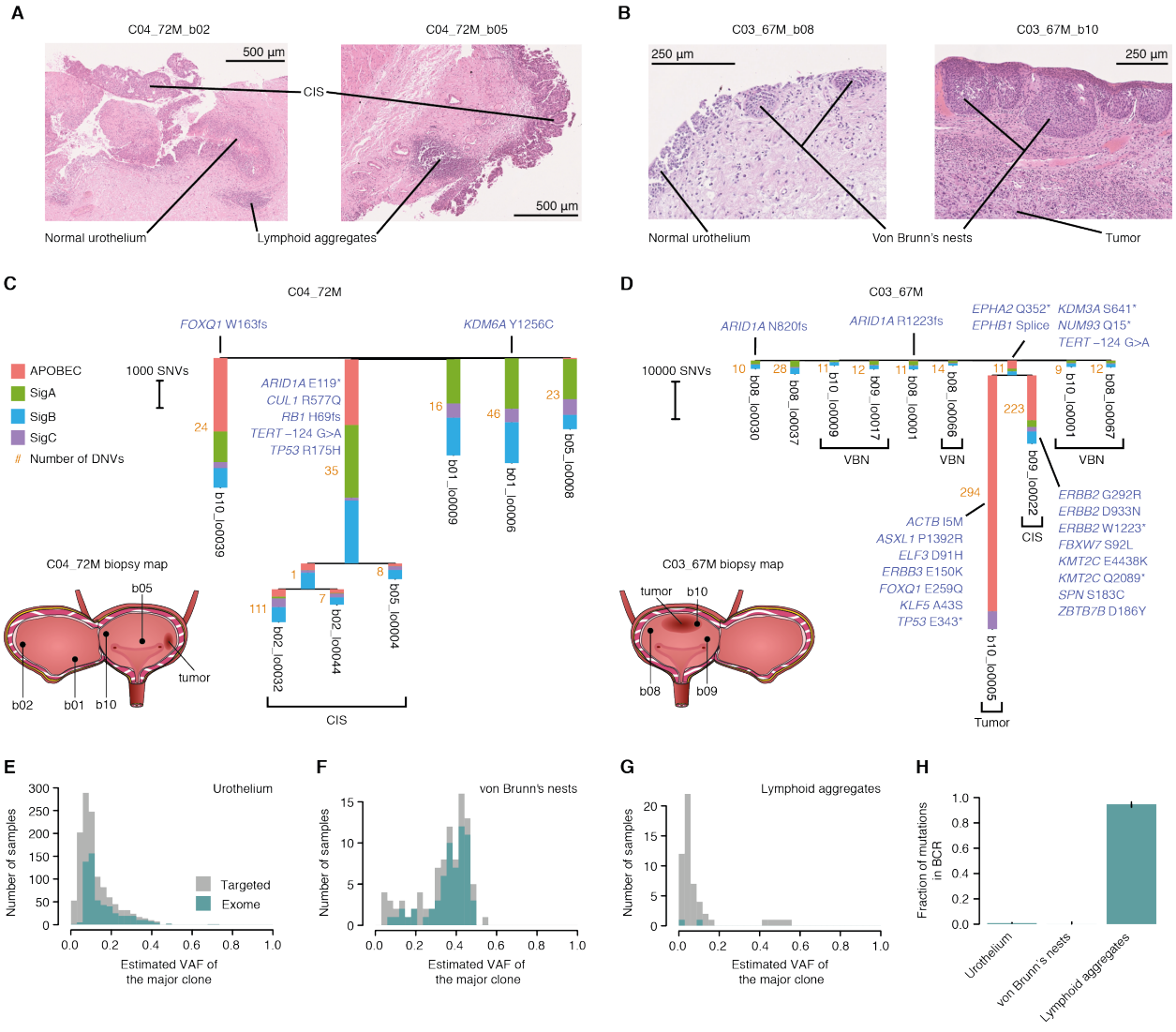


Fig. 4. The mutational landscape across histological features in patients with bladder cancer. (**A** and **B**) Histology images depicting features microdissected from cystectomy material for two patients with bladder cancer (C04_72M and C03_67M). (**C** and **D**) Phylogenetic reconstruction of the evolution of cancer and CIS clones (*22*). Only microbiopsies with a high degree of clonality (mean VAF ≥0.25) were included. Biopsy maps show the relative positions of macroscopic biopsies (b01-b10) within the bladder. Branches without a feature indicated are histologically-normal urothelium. Branch lengths depict single nucleotide variant (SNV) counts and the number next to each branch denotes assigned dinucleotide variants (DNVs). Driver genes identified in this study and in (*18, 19*) are annotated. Truncating mutations in *EPHB1* and *KDM3A* are annotated in the branch shared by the CIS and tumor for C03_67M. VBN, von Brunn's nest. (**E** to **G**) Histograms showing the estimated VAF of the major clone in targeted and exome sequencing data for three different histological features: urothelium, von Brunn's nests and lymphoid aggregates. (**H**) Proportion of mutations located within the B-cell receptor across histological features.

Science
AAAS

Supplementary Materials for

Extensive heterogeneity in somatic mutation and selection in the human bladder

Andrew R. J. Lawson, Federico Abascal, Tim H. H. Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, Mathijs A. Sanders, Anne Y. Warren, Krishnaa T. A. Mahbubani, Bethany Bareham, Timothy M. Butler, Luke M. R. Harvey, Alex Cagan, Andrew Menzies, Luiza Moore, Alexandra J. Colquhoun, William Turner, Benjamin Thomas, Vincent Gnanapragasam, Nicholas Williams, Doris M. Rassl, Harald Vöhringer, Sonia Zumalave, Jyoti Nangalia, José M. C. Tubío, Moritz Gerstung, Kourosh Saeb-Parsy, Michael R. Stratton, Peter J. Campbell, Thomas J. Mitchell, Iñigo Martincorena

Correspondence to: *im3@sanger.ac.uk*

This PDF file includes:

Materials and Methods
Figs. S1 to S18
Table S1
Captions for Tables S2 to S11

Other Supplementary Materials for this manuscript include the following:

Tables S2 to S11

Supplementary Methods

Contents

1. Sample collection and preparation
    1.1. Sample collection
    1.2. Sample preparation for laser-capture microdissection
    1.3. Library preparation of microbiopsy cell lysate
    1.4. Selection of libraries for different sequencing approaches
    1.5. p53 immunohistochemistry
2. DNA sequencing
    2.1. Targeted sequencing of 321 cancer-associated genes
    2.2. Exome sequencing
    2.3. Whole-genome sequencing
3. Mutation calling
    3.1. Substitution and indel calling from targeted data using ShearwaterML
    3.2. Substitution calling from exome and WGS data using CaVEMan
    3.3. Indel calling from exome and WGS data using cgpPindel
    3.4. Structural variant calling from exome data using ASCAT
    3.5. Structural variant calling from WGS data
4. Analyses of mutant clones and allele frequencies
    4.1. Estimation of mutation burden per cell
        4.1.1. Aggregating allele fractions
        4.1.2. Subclonal decomposition
    4.2. Approximate estimates of clone lengths
        4.2.1. Relationship between dN/dS and clone sizes
    4.3. Lower and upper bound estimates of the fraction of mutant epithelium
    4.4. Statistical pigeonhole principle
5. Selection and driver analyses
    5.1. Driver discovery in transplant donors
    5.2. Driver discovery outside of cancer genes
    5.3. Driver discovery in bladder cancer patients
    5.4. Estimation of the number of driver mutations
    5.5. Variation in selection between donors across genes
    5.6. Germline mutations in genes differentially selected across donors
    5.7. Mutation frequency and selection in bladder carcinomas from The Cancer Genome Atlas
    5.8. Selection at putative antigenic regions
6. Mutational signatures
    6.1. De novo signature extraction
    6.2. Alternative approaches for signature extraction
    6.3. Comparison to reference signatures
    6.4. Mutational signature analysis in bladder cancer genomes
    6.5. Transcriptional and replicational strand asymmetries
    6.6. Detection of APOBEC mutagenesis in exomes
    6.7. Spatial clustering of APOBEC-positive clones
    6.8. Statistical association between mutational signatures, smoking and alcohol consumption.

7. Comparison of normal urothelium between transplant donors and bladder cancer patients
8. Phylogenetic reconstruction

Supplementary Figures

Fig. S1. Comparison of mutations called across sequencing strategies.
Fig. S2. Clonal expansions.
Fig. S3. Distribution of mutations within selected genes.
Fig. S4. Additional selection analyses.
Fig. S5. Discovery of mutational signatures with HDP de novo.
Fig. S6. Matching extracted *de novo* mutational signatures to reference signatures.
Fig. S7. Alternative mutational signature extraction with SigProfiler and HDP with priors.
Fig. S8. Alternative mutational signature extraction with HDP and bladder cancer genomes.
Fig. S9. Transcriptional strand asymmetries of mutational signatures.
Fig. S10. Replicational strand asymmetries and extended contexts of mutational signatures.
Fig. S11. Statistical association between signature A and smoking history.
Fig. S12. Characterization of *ERCC2*-mutant clone
Fig. S13. Permutation test for the spatial clustering of APOBEC-positive clones.
Fig. S14. Mutation burden, VAFs and signatures in bladder cancer patients.
Fig. S15. Immunohistochemistry of p53.
Fig. S16. Heatmaps of early embryonic mutations from cystectomy phylogenies.
Fig. S17. Dinucleotide variants in cystectomy phylogenies.
Fig. S18. Clonal lymphoid aggregate.

Supplementary Tables

Table S1. Donor information.
Table S2. Microbiopsy information.
Table S3. Substitution and indel calls from targeted data.
Table S4. Substitution and indel calls from exome data.
Table S5. Substitution and indel calls from genome data.
Table S6. Copy number calls from exome data.
Table S7. Rearrangement calls from genome data.
Table S8. Driver discovery in transplant donors.
Table S9. Fraction of mutations attributed to each signature per sample.
Table S10. Mutational signatures extracted by HDP de novo.
Table S11. Retrotransposition calls from genome data.

Materials and Methods

1. Sample collection and preparation

1.1 Sample collection

Bladder specimens were obtained from two sources: (1) biopsies were taken from the dome of the bladder of deceased individuals from whom organs were being retrieved for transplantation; and (2) multi-region sampling was carried out on material removed by cystectomy as part of the treatment of bladder cancer patients. In the former instance, informed consent for the use of tissue in research was obtained from the donor's family as part of the Cambridge Biorepository for Translational Medicine program (REC reference: 15/EE/0152 NRES Committee East of England – Cambridge South). In the latter case, cystectomies were performed at the Cambridge University NHS Trust and informed consent was obtained from the patient prior to surgery (REC reference: 03/018 East of England - Cambridge Central Research Ethics Committee).

All samples were anonymized and were handled and processed in accordance with HTA guidelines. No sample size determination, randomization or blinding was carried out as this was a descriptive study. All samples were included for analysis, except those failing library preparation or sequencing.

1.2 Sample preparation for laser-capture microdissection

Immediately after collection, specimens underwent formalin-free fixation for 24 hours in PAXgene Tissue FIX containers (PreAnalytiX, Hombrechtikon, Switzerland) before being transferred to PAXgene STABILIZER solution (PreAnalytiX) for storage at -20 °C.

Prior to laser-capture microdissection, specimens were processed using a Tissue Tek VIP 6 AI tissue processor (Sakura Finetek, Leiden, Netherlands), embedded in paraffin and sectioned using an Accu-Cut SRM 200 microtome (Sakura Finetek). For each specimen, reference slides were prepared using 5 µm sections obtained at the beginning and end of each cutting session. These were mounted on Superfrost Plus glass microscope slides (VWR International, Lutterworth, UK), stained with hematoxylin and eosin (H&E; Leica Microsystems, Wetzlar, Germany), permanently coverslipped using CV Mount (Leica Microsystems) and imaged using a NanoZoomer 2.0-HT slide scanner (Hamamatsu Photonics, Hamamatsu, Japan). Sections for laser-capture microdissection were mostly cut at a thickness of 16 µm (see table S2 for exceptions) and mounted on polyethylene naphthalate (PEN) membrane glass slides (Leica Microsystems). These were stained with H&E, dipped in a xylene substitute, Neo-Clear (Merck, Darmstadt, Germany), and temporarily coverslipped before imaging on the slide scanner. Images obtained from the slide scanner were viewed using the NDP.view2 software.

Microbiopsies were dissected using an LMD7 microscope (Leica Microsystems). Detailed information on each microbiopsy is available in table S2. Images of the selected regions were captured immediately before and after microdissection. Proteolysis of isolated regions was performed using an Arcturus PicoPure DNA Extraction Kit (Thermo Fisher Scientific, Waltham, MA, USA). Cell lysate was stored at -20 °C prior to library preparation.

1.3 Library preparation of microbiopsy cell lysate

Library preparation was automated using an Agilent Bravo NGS workstation (Agilent, Santa Clara, CA, USA). Solid-phase reversible immobilization (SPRI) DNA purification was carried out using Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN, USA). In order to minimize gDNA losses caused by incomplete elution, subsequent library preparation steps used the entire post-elution sample (including beads) as input. Fragmentation, end repair, dA-tailing and adapter ligation steps were performed using an NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA). All libraries were prepared with a mean insert size of ~350 bp to ensure they were suitable for both sequence capture and whole-genome sequencing. Twelve cycles of PCR amplification were carried out using a KAPA HiFi PCR Kit (Roche, Wilmington, MA, USA). Libraries were eluted in 25 µl nuclease-free water (Thermo Fisher Scientific) and quantified using an AccuClear Ultra High Sensitivity dsDNA Quantification Kit (Biotium, Fremont, CA, USA).

1.4 Selection of libraries for different sequencing approaches

Having evaluated the relationship between library concentration and complexity, a minimum library concentration of 20 ng/µl was required for targeted, exome and whole-genome sequencing (see table S2 for exceptions). For most donors, all libraries with sufficient yield were initially sent for targeted sequencing of 321 cancer-associated genes. A small number of libraries from each donor were subsequently selected for whole-genome sequencing based on the presence of high variant allele fraction (VAF $\geq$0.2) mutations in the targeted sequencing data. By contrast, the library selection for exome sequencing was performed without referring to the VAFs in the targeted sequencing data in order to avoid biasing our mutation burden estimates. Instead, samples were selected for exome sequencing based on their library concentrations and the avoidance of duplicate regions. For four transplant donors (T04_55M, T05_58M, T10_68F and T14_75F) and two cystectomy patients (C01_49M and C05_75M), some or all of the exomes from these individuals do not have matched targeted sequencing data from the same library (tables S1 and S2).

1.5 p53 immunohistochemistry

Sections (5 µm thick) were mounted on Superfrost Plus glass microscope slides (VWR International) and endogenous peroxidase activity was blocked by incubation with hydrogen peroxide and methanol. Non-specific binding sites were blocked by incubating slides in 3% horse serum in Tris-buffered saline. Slides were incubated overnight at 4 °C with a mouse monoclonal p53 antibody (Santa Cruz Biotechnology, Dallas, TX, USA; Cat# sc-126, RRID:AB_628082) diluted 1 in 750 in 1.5% horse serum in Tris-buffered saline. Secondary antibody incubation and visualization were performed using a Vectastain Elite ABC HRP Kit (Vector Laboratories, Burlingame, CA, USA; Cat# PK-6100, RRID:AB_2336819). Slides were counterstained with hematoxylin (Vector Laboratories), coverslipped using CV Mount (Leica Microsystems) and imaged using a NanoZoomer 2.0-HT slide scanner (Hamamatsu Photonics).

2. DNA sequencing

2.1 Targeted sequencing of 321 cancer-associated genes

We designed a custom Agilent SureSelect bait set to capture the exonic regions of the following 321 cancer-associated genes:

ABL1, *ACVR1*, *ACVR1B*, *ACVR2A*, *AJUBA*, *AKT1*, *ALB*, *ALK*, *AMER1*, *APC*, *AR*, *ARHGAP35*, *ARID1A*, *ARID1B*, *ARID2*, *ARID5B*, *ASXL1*, *ATM*, *ATP1A1*, *ATP1B1*, *ATP2A2*, *ATP2B3*, *ATP7B*, *ATR*, *ATRX*, *AXIN1*, *AXIN2*, *B2M*, *BAP1*, *BCOR*, *BIRC3*, *BRAF*, *BRCA1*, *BRCA2*, *CACNA1D*, *CALR*, *CARD11*, *CASP8*, *CBFB*, *CBL*, *CBLB*, *CCND1*, *CCNE1*, *CD58*, *CD79A*, *CD79B*, *CDC73*, *CDH1*, *CDK12*, *CDK4*, *CDK6*, *CDKN1A*, *CDKN1B*, *CDKN2A*, *CDKN2B*, *CDKN2C*, *CEBPA*, *CFH*, *CIB3*, *CIC*, *CMTR2*, *CNOT3*, *COL2A1*, *CPA2*, *CREBBP*, *CRLF2*, *CSF1R*, *CSF3R*, *CTCF*, *CTNNA1*, *CTNNB1*, *CUL3*, *CUX1*, *CXCR4*, *CYLD*, *DAXX*, *DDR2*, *DDX3X*, *DICER1*, *DNM2*, *DNMT3A*, *EEF1A1*, *EGFR*, *EIF1AX*, *ELF3*, *EML4*, *EP300*, *EPHA2*, *EPS15*, *ERBB2*, *ERBB3*, *ERCC2*, *ERG*, *ERRFI1*, *ESR1*, *ETNK1*, *EZH2*, *FAM104A*, *FAM46C*, *FAM58A*, *FAT1*, *FAT2*, *FBXO11*, *FBXW7*, *FGFR1*, *FGFR2*, *FGFR3*, *FLT1*, *FLT3*, *FLT4*, *FOSL2*, *FOXA1*, *FOXA2*, *FOXL2*, *FOXP1*, *FOXQ1*, *FTH1*, *FTL*, *FUBP1*, *GAGE12J*, *GATA1*, *GATA2*, *GATA3*, *GATA4*, *GJA1*, *GNA11*, *GNA13*, *GNAQ*, *GNAS*, *GPS2*, *GRIN2A*, *H3F3A*, *H3F3B*, *HAMP*, *HFE*, *HFE2*, *HGF*, *HIST1H2BD*, *HIST1H3B*, *HLA-A*, *HLA-B*, *HLA-C*, *HNF1A*, *HOXB3*, *HRAS*, *IDH1*, *IDH2*, *IGF1R*, *IGSF3*, *IKBKB*, *IKZF1*, *IL6R*, *IL6ST*, *IL7R*, *IRF2*, *IRF4*, *JAK1*, *JAK2*, *JAK3*, *KCNJ5*, *KDM5C*, *KDM6A*, *KDR*, *KEAP1*, *KIT*, *KLF4*, *KLF5*, *KLF6*, *KMT2A*, *KMT2B*, *KMT2C*, *KMT2D*, *KRAS*, *LIPF*, *LRP1B*, *MAP2K1*, *MAP2K2*, *MAP2K4*, *MAP2K7*, *MAP3K1*, *MAX*, *MED12*, *MEN1*, *MET*, *MGA*, *MLH1*, *MPL*, *MSH2*, *MSH6*, *MTOR*, *MYC*, *MYCN*, *MYD88*, *MYOD1*, *NCOR1*, *NF1*, *NF2*, *NFE2L2*, *NFKBIE*, *NKX2-1*, *NOTCH1*, *NOTCH2*, *NOTCH3*, *NOTCH4*, *NPM1*, *NQO1*, *NRAS*, *NSD1*, *NT5C2*, *NTRK3*, *PALB2*, *PAX5*, *PBRM1*, *PCLO*, *PCMTD1*, *PDGFRA*, *PDYN*, *PHF6*, *PHOX2B*, *PIK3CA*, *PIK3R1*, *PIK3R3*, *PLCG1*, *POLE*, *POT1*, *POU2AF1*, *PPM1D*, *PPP2R1A*, *PPP6C*, *PRDM1*, *PREX2*, *PRKACA*, *PRKAR1A*, *PTCH1*, *PTEN*, *PTPN11*, *PTPN3*, *PTPRB*, *QKI*, *RAC1*, *RAC2*, *RAD21*, *RASA1*, *RB1*, *RBM10*, *RET*, *RHBDF2*, *RHOA*, *RHOB*, *RIT1*, *RNF43*, *ROBO2*, *RPL10*, *RPL22*, *RPL5*, *RPS6KA3*, *RREB1*, *RUNX1*, *SERPINA1*, *SETBP1*, *SETD2*, *SF3B1*, *SFTPA1*, *SFTPB*, *SFTPC*, *SH2B3*, *SLC10A1*, *SLC40A1*, *SMAD2*, *SMAD4*, *SMARCA4*, *SMARCB1*, *SMC3*, *SMO*, *SMTNL2*, *SOCS1*, *SOX2*, *SOX9*, *SPEN*, *SPOP*, *SRC*, *SRSF2*, *STAG2*, *STAT3*, *STAT5B*, *STK11*, *SUFU*, *TBL1XR1*, *TBX3*, *TCF7L2*, *TEK*, *TENM1*, *TERT*, *TET2*, *TFR2*, *TG*, *TGFBR2*, *TGIF1*, *TMEM170A*, *TMEM51*, *TNFAIP3*, *TNFRSF14*, *TP53*, *TP63*, *TRAF7*, *TSC1*, *TSC2*, *TSHR*, *TYRO3*, *U2AF1*, *UBR5*, *VEGFA*, *VHL*, *WT1*, *XBP1*, *XIRP2*, *XPO1*, *ZFHX3*, *ZFP36L1*, *ZNF750* and *ZRSR2*.

Highlighted in blue are 40 genes that were previously identified as being under significant positive selection in bladder cancer (*18, 19*). There are 22 additional genes that were found to be significantly mutated in urothelial carcinoma in these studies but were omitted from the targeted capture panel: *ACTB*, *ASXL2*, *C3orf70*, *CUL1*, *EPS8*, *HES1*, *KANSL1*, *MB21D2*, *MBD1*, *METTL3*, *NUP93*, *PARD3*, *PSIP1*, *RXRA*, *SF1*, *SPN*, *SPTAN1*, *SSH3*, *TAF11*, *TMCO4*, *USP28* and *ZBTB7B*. Another notable gene absent from the targeted panel is *UTY*, a member of the Histone H3 Lysine 27 (H3K27) demethylase gene family, which is present on the Y chromosome and is homologous to *KDM6A* on the X chromosome.

In addition to the cancer-associated genes listed above, we included the following regions in the targeted panel: two non-coding RNA genes (*MALAT1* and *NEAT1*); promoter regions for eight genes (*ARID1A*, *CDKN2A*, *EGFR*, *ERBB2*, *MYC*, *PLEKHS1*, *TERT* and *TP53*); three regions responsible for generating diversity in immune cells (*IGH*, *TRB* and *TRD*) and eighteen L1 retrotransposition hotspots. Common SNPs within or around the targeted genes were also included in the bait set to facilitate copy number analyses. The total size of the targeted panel was 1.99 Mb.

Samples were multiplexed to an average pool size of 32 and sequenced using 75 bp or 150 bp paired-end reads on Illumina HiSeq 4000 machines (table S2). Paired-end reads were aligned to human genome assembly GRCh37 using BWA-MEM (*49*). Duplicate reads were marked using biobambam (*50*). Library complexity and coverage statistics were calculated using Picard (http://broadinstitute.github.io/picard/). The median on-target coverage across all samples and genes was 89×. Across donors, median coverage ranged from 30× (C02_61F) to 152× (C03_67M). As the detection of low VAF mutations can be greatly impacted by index hopping or by sample contamination with DNA from other individuals, VerifyBamID (*51*) was run to ensure that the sequencing data included in this study were unaffected by these issues.

Due to the use of laser-captured microbiopsies instead of macroscopic biopsies, the coverage achieved in this study was lower than that of two previous studies from our team (*3, 7*). Coverage was limited by the library complexity achievable from small microbiopsies of 16µm-thick histology sections. However, the coverage per unit area of epithelium, and so the estimated sensitivity to detect microscopic mutant clones, is considerably higher in the current study compared to those previous studies.

2.2 Exome sequencing

Exome capture was performed using an Agilent SureSelect All Exon v5 bait set (S04380110). Samples were multiplexed and sequenced using 150 bp paired-end reads on either Illumina HiSeq 4000 (average pool size of 13) or Illumina NovaSeq (average pool size of 48) machines (table S2). Paired-end reads were aligned to human genome assembly GRCh37 using BWA-MEM (*49*). Duplicate reads were marked using biobambam (*50*) and sample contamination estimates were calculated using VerifyBamID (*51*). Library complexity and coverage statistics were calculated using Picard (http://broadinstitute.github.io/picard/). The median on-target coverage across all samples and genes was 72×. Across donors, median coverage ranged from 29× (T08_61F) to 99× (T10_68F).

2.3 Whole-genome sequencing

Whole-genome sequencing was performed on selected microbiopsies that were identified as likely having a high degree of clonality from the variant allele fractions observed in targeted or exome sequencing data of the same library (methods S1.4). Samples were sequenced using 150 bp paired-end reads on either Illumina HiSeq 4000 or Illumina NovaSeq machines (table S2). Paired-end reads were aligned to human genome assembly GRCh37 using BWA-MEM (*49*) and duplicate reads were marked using biobambam (*50*). The median coverage across all samples was 33×. Across samples, median coverage ranged from 17× (T06_59M_b01_lo0091) to 60× (T07_59M_b01_lo0038).

In order to test the reproducibility of variant calling from independent libraries for whole-genomes, two sets of duplicates (T08_61F_b01_lo0008 and T08_61F_b01_lo0048; T08_61F_b01_lo0035 and T08_61F_b01_lo0181) and one triplicate set (T08_61F_b01_lo0071, T08_61F_b01_lo0079 and T08_61F_b01_lo0091) were sequenced from microbiopsies of corresponding stretches of urothelium isolated from neighboring histology sections.

3. Mutation calling

3.1 Substitution and indel calling from targeted data using ShearwaterML

As in our previous work on sun-exposed skin and esophagus, ShearwaterML was used to detect substitutions and indels present at low VAFs in the targeted sequencing data (*3, 7, 52*). This algorithm is publicly available as part of the *deepSNV* R package (https://github.com/gerstung-lab/deepSNV). In order to generate the base-specific error model, we used a collection of 75 smooth muscle, lamina propria and blood vessel microbiopsies (table S2) that were processed using the same library preparation protocol as the urothelium samples (methods S1.3), resulting in an average background coverage of 6,210×.

Unmapped reads, duplicate reads, failed reads, secondary and supplementary alignments, reads with mapping quality scores <55 and bases with Phred quality scores <30 were excluded from coverage calculations. Overdispersion values were estimated within the interval [10^{-6}, 0.32]. *P*-values were subject to multiple testing correction using Benjamini & Hochberg's False Discovery Rate (*53*) and a q-value cut-off of 0.01 was used to call somatic mutations. Consecutive substitutions were merged into a single event, as were consecutive indels providing their VAFs were found to be compatible using a Fisher's exact test. Additional filtering against artefacts introduced at cruciform DNA sites was applied as described in the following section (methods S3.2).

Many common germline SNPs were already absent from the ShearwaterML calls due to their presence in the matched normal panel. Mutations present at VAF ≥0.25 in any of the non-urothelium microbiopsies were excluded from all other samples from the same donor as putative rare germline SNPs.

Substitution and indel calls for targeted data are available in table S3.

3.2 Substitution calling from exome and WGS data using CaVEMan

For exome and whole-genome sequencing data, substitutions were called using CaVEMan (Cancer Variants through Expectation Maximization) (https://cancerit.github.io/CaVEMan/) (*54*). In order to increase the sensitivity of CaVEMan for calling subclonal variants, the following parameters were used: mutant copy number = 5; wild type copy number = 2; and normal contamination = 0.1.

The following smooth muscle, lamina propria or blood vessel microbiopsies were used as matched normals for variant calling in the exomes:

T01_25F_b01_lo0056, T02_35F_b06_lo0049, T03_53F_b08_lo0030, T04_55M_b01_lo0011, T05_58M_b01_lo0028, T06_59M_b01_lo0220, T07_59M_b01_lo0047, T08_61F_b01_lo0207, T09_61M_b02_lo0016, T10_68F_b01_lo0047, T11_69F_b03_lo0024, T12_70M_b04_lo0023, T13_71F_b06_lo0070, T14_75F_b02_lo0019, T15_78F_b09_lo0008, C01_49M_b01_lo0023, C02_61F_b01_lo0036, C03_67M_b09_lo0035, C04_72M_b05_lo001 and C05_75M_b10_lo0006.

Similarly, the following microbiopsies were used as matched normals for variant calling in the whole-genomes:

T01_25F_b04_lo0040, T02_35F_b05_lo0008, T03_53F_b08_lo0029, T04_55M_b01_lo0011, T05_58M_b01_lo0028, T06_59M_b01_lo0220, T07_59M_b01_lo0048, T08_61F_b01_lo0084, T09_61M_b01_lo0012, T10_68F_b01_lo0047, T11_69F_b05_lo0089, T12_70M_b02_lo0018, T13_71F_b06_lo0051, T14_75F_b02_lo0019, T15_78F_b09_lo0028, C01_49M_b01_lo0023, C02_61F_b01_lo0012, C03_67M_b08_lo0055, C04_72M_b06_lo0055 and C05_75M_b10_lo0021.

As reported previously (*10, 37*), the enzymatic fragmentation step used in our low-input library preparation protocol can introduce artefactual calls from the incorrect processing of cruciform DNA that are not excluded by the standard filters in the CaVEMan algorithm. Therefore, the following post-processing steps were carried out:

(1) Variants where the median alignment score of reads supporting the variant is <120 were excluded.
(2) Variants where the median number of bases clipped from the supporting reads was >0 were excluded.
(3) Variants that were supported by <3 supporting read pairs (fragments) were excluded.
(4) For variants that were supported by a low number of reads (0-1) on a particular strand, it was required of the other strand that either ≤90% of supporting reads had the variant located within the first 15% of the read or that the median absolute deviation of the variant position was >0 and the standard deviation of the variant position was ≥4.
(5) For variants with sufficient support from both strands (≥2 reads), it was required for both strands separately that ≤90% supporting reads had the variant located within the first 15% of the read or that the median absolution deviation and standard deviation were both greater than 2 or that the standard deviation for one strand was >10.
(6) For exomes, if the same variant was called in ≥⅓ of samples from a donor, it was deemed likely to be a germline variant and was excluded. The 96 variants removed by this filter were manually reviewed and all deemed unlikely to be somatic variants. This filter removed a median of 2 variants from the cystectomy patients and a median of 3 variants from the transplant donors.

For samples with matched targeted sequencing data, there was good agreement between the variants called by ShearwaterML and CaVEMan (fig. S1, A to C and G to I). Most discordant calls were due to drop-out of low VAF mutations in the exome or whole-genome call sets, which had lower coverage than the targeted data.

Substitution calls for the exomes and whole-genomes are available in tables S4 and S5 respectively.

3.3 Indel calling from exome and WGS data using cgpPindel

For exome and whole-genome sequencing data, small insertions and deletions (indels) were called using cgpPindel (https://github.com/cancerit/cgpPindel) (*55*). The same matched normal samples were used for each donor as for substitution calling with CaVEMan (methods S3.2). In addition, the following post-processing steps were carried out:

(1) Variants were required to pass the simple repeat filter (F017).
(2) Variants with a REP score >3 were excluded.
(3) Variants with a reported VAF of 0 were excluded.
(4) For exomes, variants that were called in multiple samples were manually reviewed. Of the 169 variants called in more than one sample, 55 were excluded as putative germline or artefactual calls.

As with substitutions, there was a high degree of concordance between Shearwater and cgpPindel calls for microbiopsies with both targeted and exome sequencing data (fig. S1, D to F). However, the recovery of high VAF calls by cgpPindel from within the targeted bait region was considerably worse for whole-genome samples than for exome samples (fig. S1, J to L). This difference in behavior was largely due to the performance of F016, one of the cgpPindel filters that is applied solely to whole-genome sequencing data. Removing this filter resulted in a large number of calls that were deemed likely to be artefacts and so the final call set for the whole-genome data has had the F016 filter applied.

Indel calls for the exomes and whole-genomes are available in tables S4 and S5 respectively.

3.4 Structural variant calling from exome data using ASCAT

For the exome data, structural variants were called using ASCAT (allele-specific copy number analysis of tumors) (*56, 57*). The same matched normal samples were used for each donor as for substitution calling with CaVEMan (methods S3.2). Alleles were counted at SNP sites identified outside of the HLA locus in phase 3 of the 1000 Genomes Project (*58*). B-allele fractions and logR values were calculated for SNPs with at least 20× coverage in the matched normal sample. A penalty score of 150 was used for ASPCF segmentation.

Of the 655 non-reference exomes sequenced, structural variant calls are provided for 370 of them. The list of samples that were used for structural variant calling are given in the second sheet of table S6. Only calls from 295 urothelial samples are plotted in Fig. 3L. The following exclusion criteria were applied for selecting exomes for structural variant calling:

(1) As ASCAT is not well-suited to the detection of subclonal copy number variants, we excluded 246 samples that did not exhibit strong evidence of containing a major subclone from their substitution and indel calls. In order for a sample to be included, we required it to have at least two substitution or indel calls with VAF ≥0.2 and total depth ≥30.
(2) Two samples were excluded as ASCAT was unable to find an optimal ploidy solution.

(3) 37 samples were excluded as the goodness of fit of the ASCAT solution was <95.

We required structural variants to span at least 2 Mb in order to be called. Regions smaller than this were iteratively collapsed providing they were surrounded by segments with identical copy number states.

Despite these attempts to remove all artefactual copy number calls, it is worth noting that some are likely still present within table S6. Of the 37 samples excluded for having poor goodness of fit, 17 were from T12_70M. The two urothelial samples with the highest number of copy number calls (T12_70M_b08_lo0015 has 12 calls and T12_70M_b01_lo0009 has 8 calls) both come from this transplant organ donor and they both have goodness of fit values that are close to the threshold (95.03 and 96.64). Additionally, there are small regions close to the 2 Mb cut-off within Fig. 3L that show evidence of both gains and losses in different samples. Manual inspection of these regions indicate that most, if not all, of these are likely to be artefactual calls caused by variable coverage at these sites between samples as they show aberrant logR values with no deviation in B-allele fraction.

Exome copy number variants and the list of samples for which variants were called are available in table S6.

3.5 Structural variant calling from WGS data

Copy number changes in whole-genomes were called using an implementation of the Battenberg algorithm (https://github.com/cancerit/cgpBattenberg) (*59*). The same matched normal samples were used for each donor as for substitution calling with CaVEMan (methods S3.2).

In order to identify rearrangement breakpoints associated with structural variants, we ran the BRASS (breakpoints via assembly) algorithm (https://github.com/cancerit/BRASS) (*60*), which identifies discordant read pairs, groups them based on their mapping locations and attempts to assemble these groups of reads to reveal the sequences spanning breakpoints. The same matched normal samples were used for BRASS as for substitution calling (methods S3.2).

As has been observed previously (*10*), post-process filtering of BRASS output is necessary for low-input sequencing to account for an increased proportion of artefactual calls compared to bulk sequencing, which is due to elevated duplicate rates for low complexity libraries leading to incomplete duplicate removal as a result of frameshifts at repetitive sites. Additional statistics were calculated by running AnnotateBRASS (https://github.com/MathijsSanders/AnnotateBRASS). Recurrent artefacts were removed using an unmatched panel of 136 whole-genomes derived from microdissected pancreas samples that had undergone the same library preparation process (methods S1.3).

BRASS calls were retained if they met the following criteria (previously described in greater detail in (*37*)):

(1) ≥4 unique reads supporting each breakpoint.
(2) Non-zero variance in the start positions of the reads supporting each breakpoint.

(3) Reads not supporting the structural variant should not have their mates mapped to a large number of different chromosomes (i.e. different chromosome score <25 for each side of the breakpoint with a sum of the two different chromosome scores ≤40).
(4) ≤50% of supporting reads on at least one side of the breakpoint should have an alternative alignment (XA-tag) or an alternative alignment score similar to the current alignment score (XS-tag).
(5) Of reads not supporting the structural variant, <5% should have a discordant insert size (i.e. ≥1000 bp) for both breakpoints.
(6) No read pairs supporting the variant were present in the matched control sample.
(7) The structural variant was not detected in the unmatched panel of microdissected pancreas samples.
(8) Read pairs supporting the structural variant should not have widely divergent clipping positions for both breakpoints.

The rearrangement breakpoints that pass these filters are available in table S7.

Somatic mobile element insertions were called in whole-genomes using an implementation of the TraFiC algorithm (https://gitlab.com/mobilegenomes/TraFiC). A detailed description of the method can be found in (*36*). To increase sensitivity to candidate mobile element insertions, independent clusters were considered in addition to reciprocal clusters. Calls from independent clusters were retained if they met the following criteria:
(1) ≥6 discordant reads supporting the cluster.
(2) ≥1 discordant reads supporting the reciprocal cluster.
(3) ≥1 clipped reads supporting a polyA tail longer than 10 bp.

Mobile element insertions identified in a given donor were clustered using a breakpoint window of ±30 bp to generate a list of non-redundant insertions. This unified list was re-genotyped in all the microbiopsies. Only re-genotyped insertions supported by more than 4 discordant reads were retained. All calls were confirmed by visual inspection using IGV.

Microbiopsies were selected for retrotransposition calling based on the product of sequencing depth and clonality. Only those whose product of mean coverage and median VAF was greater than 6 were included (table S11, tab 2). The same matched control samples were used for each donor as for substitution calling with CaVEMan (methods S3.2).

The list of the retrotransposition events detected is given in table S11 (tab 1). Across 55 genomes of normal urothelium and 10 genomes of von Brunn's nests, we detected 3 and 2 retrotranspostion events, respectively. All carcinoma *in situ* (n = 5) and tumor (n = 1) microbiopsies analyzed harbored at least one retrotransposition event. Several retrotransposition events were found to be shared across different carcinoma *in situ* microbiopsies.

4. Analyses of mutant clones and allele frequencies

4.1 Estimation of mutation burden per cell

Mutation burden in cancer genomes is a term used to refer to the number of mutations in a cancer. This is typically calculated by dividing the number of mutations detected in a cancer by the size of the genome sequenced. Although this is a common practice, cancer genomes can contain multiple subclones and the approach above is only an approximation to the number of mutations per cell in a cancer.

This problem is exacerbated when sequencing samples of normal tissues with multiple clones. Each microbiopsy can contain multiple independent clones and estimation of the mutation burden per cell requires accounting for the fraction of cells that carry each mutation. In this study, we used two alternative approaches to estimate the mutation burden per cell (Fig. 3B), both of which have been used before in normal tissues: aggregating allele fractions (*3, 7*) and subclonal decomposition (*3*).

4.1.1 Aggregating allele fractions

As described in previous studies (*3, 7*), the average number of detectable mutations per cell per megabase (Mb) in a given microbiopsy can be estimated as follows:

$$\beta_s = \sum_j \rho_j / L_{Mb} \approx 2 \sum_j v_j / L_{Mb} \qquad \text{(Eq. 1)}$$

Where L_{Mb} is the size of the region sequenced (in megabases), p_j is the fraction of cells of a sample carrying a mutation (j) and v_j is the variant allele fraction (VAF) of the mutation. In the absence of copy number changes, mutations are heterozygous and p_j can be approximated by $2v_j$. For a more general expression valid for other copy number configurations, please refer to (*3, 7*). Summing across all of the microbiopsies from a donor, we can estimate the average burden of detectable mutations per cell in each donor.

$$\beta \approx \frac{2}{L_{Mb} S} \sum_s \sum_j v_j \qquad \text{(Eq. 2)}$$

Where S is the number of microbiopsies from a donor.

This is a lower-bound estimate of the true mutation burden per cell as it is restricted to detectable mutations. It is particularly suitable when dealing with large numbers of mutations with low allele fractions, such as those found in ultra-deep sequencing of macroscopic biopsies of normal tissues (*3, 7*).

4.1.2 Subclonal decomposition

If a single cell expands into a large clone, clustering of VAFs can be used to identify the mutations present in the clone and, in doing so, to measure the number of mutations in that cell, providing an alternative estimate of the mutation burden per cell. This is particularly suitable when dealing with samples dominated by a major clone, such as when sequencing colonic crypts (*9*).

We can model allele fractions in a sample using a binomial or beta-binomial mixture model, where each component represents a subclone (*3*). Here, we used a mixture of beta-binomial distributions with modest overdispersion (ρ=1e-4), fitted with an Expectation-Maximization (EM) algorithm. We first fitted a model with one component (one subclone) and then with increasing numbers of subclones, using a Likelihood-Ratio Test to determine whether each additional subclone significantly improved the fit of previous model, stopping when the test became non-significant.

This mixture model was used for subclonal deconvolution in two sections in this manuscript: as an alternative approach for estimating the mutation burden per cell (Fig. 3B), and to identify the major subclone in the analysis of different histological features from cystectomy biopsies (Fig. 4, E to G). For each microbiopsy, 50 random initializations of the EM algorithm were run to minimize the risk of falling into local minima.

For microbiopsies dominated by one major clone, aggregating allele fractions can underestimate the mutation burden per cell more than subclonal deconvolution. Since both approaches provide lower-bound estimates, Fig. 3B shows the maximum value of these two approaches for each microbiopsy.

4.2 Approximate estimates of clone lengths

Clone sizes cannot be directly measured using sequencing data, but approximate estimates can be obtained considering the fraction of mutant cells within a biopsy. In previous studies of skin and esophagus, which used deep sequencing of small areas of epithelium, clone sizes were estimated by multiplying the fraction of mutant cells within a biopsy by the area of the biopsy (*3, 7*).

The proportion of cells in a microbiopsy that carry a mutation can be estimated using the variant allele fraction. Equations to do so in the presence of copy number changes are available in (*3, 7*). However, as we have shown here, copy number changes are rare in normal urothelium. In the absence of copy number changes, as described in methods S4.1, this fraction can be estimated as $2v_j$ (two times the allele fraction of a mutation).

Microbiopsies pose the additional challenge that histological sections are only 16 µm thick (1 or 2 cells thick on average), cutting across clones. While we cannot estimate clone areas from such sections, if we assume that clones are compact in space, we can obtain lower-bound estimates of the length of a microbiopsy or strip of urothelium occupied by a clone by multiplying the length of the microbiopsy by the fraction of mutant cells in it (clone length, Fig. 1C). If a clone extends to more than one microbiopsy within a section, we summed the clone lengths estimated in each microbiopsy.

The relationship between the length of a clone (as defined above) and the area occupied by the clone is not straightforward, as it depends on the shape of the clone in 2-dimensions. For example, if clones were perfectly circular in 2-dimensions, random 1-dimensional cuts across them would give clone lengths that follow the distribution of chord lengths for a circle. The relationship between the mean chord length (d) and the radius of a circle (R) is: $R = \pi d/4 \sim 0.79d$. However,

given the unavoidable uncertainties in the estimation of clone sizes, we only provide clone lengths as approximate estimates.

4.2.1 *Relationship between dN/dS and clone sizes*

Using deep sequencing of microbiopsies, only mutations that reach a minimum clone size will be detectable by *ShearwaterML* or *CaVEMan*. This gives dN/dS ratios a simple interpretation in terms of clonal expansions. For example, we see a dN/dS of ~200 for truncating (nonsense or essential splice site) mutations in *KDM6A* in normal urothelium. This means that, a cell carrying a *KDM6A* truncating mutation is ~200 times more likely to reach a detectable clone size in our dataset than a cell carrying a *KDM6A* synonymous mutation. There can be multiple mechanisms behind signals of positive selection, such as increased proliferation, reduced differentiation, increased survival, etc. However, independently of the underlying mechanisms, dN/dS ratios >>1 typically imply statistically increased clonal growth.

For the reasons above, we may expect driver mutations under positive selection in normal tissues to have larger clone sizes on average. This trend has been shown before (*3, 7*). However, these studies also suggested that direct comparisons of estimated clone sizes between driver and passenger mutations tend to be less sensitive to selection than dN/dS ratios. There are several reasons for this. First, many synonymous mutations reach larger clone sizes by co-occurring with driver mutations (hitchhiking). Second, because of the life-long generation of new mutations, the frequency distributions of clone sizes tend to be wide, dominated by recently generated and small clones, which can reduce or confound the differences. Further, spatial constraints in a solid tissue can impose limits to clone sizes (colonic crypts being an extreme example). In that situation, drivers can still favour clonal growth leading to dN/dS signals of selection, and yet detectable clone sizes of drivers and passengers could be limited by these constraints reducing differences between detectable driver and passenger mutations.

Despite the limitations above, we can compare the estimated clone sizes (lengths) of driver and putative passenger mutations in normal urothelium. We annotated all non-synonymous mutations in any of the 17 positively-selected genes as putative driver mutations, and all mutations (non-synonymous or synonymous) in non-significant genes as putative drivers. To minimize the risk of hitchhiking, we avoided annotating putative passenger mutations in microbiopsies with one or more driver mutations. In line with previous observations in skin and esophagus (*3, 7*), the estimated clone sizes of driver mutations tend to be statistically larger than those of passenger mutations, although this difference only reaches statistical significance at individual gene level for some of the positively-selected genes (fig. S2B).

4.3 Lower and upper bound estimates of the fraction of mutant epithelium

Estimates of the fraction of cells in the urothelium that carry mutations in a given gene were obtained as described in (*7*).

Briefly, when a single mutation is observed in a given gene in a microbiopsy in the absence of a copy number change at the locus, the fraction of mutant cells is $\rho = 2v_j$. If we conservatively assume that there could be undetected copy number changes affecting the locus in cells carrying

this mutation, this estimate of the fraction of mutant cells could be inflated. For example, in the presence of copy-neutral loss of heterozygosity (LOH) with a mutation present in both alleles, the correct estimate would be $\rho = v_j$, and in the presence of a copy number loss of the wild-type copy, the estimate would be $\rho = 2v_j/(1+v_j) < 2v_j$.

When more than one driver mutation is found in the same gene in the same biopsy, these mutations can occur in different cells, different alleles of the same cells or even affect the same allele. As it was shown in (*7*), assuming a maximum of two non-synonymous mutations per cell per gene, and considering different possibilities of compound heterozygosity, copy-neutral LOH and hemizygous losses, the total fraction of mutant cells for a given gene in a sample ranges from the lower bound $\rho_{low} = \Sigma_j v_j$ (which corresponds to bi-allelic mutation of the gene in all mutant cells of the sample by either copy-neutral LOH or compound heterozygosity) to the higher bound $\rho_{high} = 2\Sigma_j v_j$ (which corresponds to one mutant allele per gene at a diploid locus). Mutations in sex chromosomes in males have estimated mutant cell fractions of $\Sigma_j v_j$.

The approach above yields conservative estimates for the fraction of cells in the urothelium that carry a non-synonymous mutation in a given driver gene (Fig. 2C). To aggregate these estimates across genes and estimate the fraction of the urothelium that carries a non-synonymous mutation in any of the 17 drivers, we need to know how driver mutations are clonally related when more than one driver gene is seen mutated in the same biopsy. When allele frequencies are high enough, we can apply the pigeonhole principle (*7, 59*) to determine whether mutations are co-occurring in the same clone, however, allele frequencies are typically too low. However, we can obtain lower bound estimates by assuming that all driver mutations in a sample are clonally nested ($\max(\rho_{low})$) and upper bounds by assuming that they occur in different cells ($\min(1,\text{sum}(\rho_{high})$).

Across all microbiopsies from donors of age ≥50 years with median coverage ≥50x, the fraction of the urothelium that carries a driver mutation was estimated to range between 8% and 19%.

4.4 Statistical pigeonhole principle

As described above, allele frequencies together with copy number data can be used to estimate the fraction of cells that carry a mutation in a sample. The pigeonhole principle can be used to infer whether two or more mutations co-occur in the same cells within a microbiopsy (*59*). Let ρ_A and ρ_B be the fraction of cells carrying mutations A and B in a given sample. The pigeonhole principle states that if $\rho_A+\rho_B>1$, then both mutations cannot be in unrelated clones but must be nested or co-clonal.

To screen for examples of co-occurring pairs or groups of mutations in this dataset, we applied a statistical version of the pigeonhole principle that determines whether the sum of the mutant cell fraction of two or more mutations within a microbiopsy are significantly higher than 1. To do so we used the approach described in (*7*). Only 10 microbiopsies had allele frequencies high enough to violate the pigeonhole principle (fig. S4A), identifying at least three clones carrying two driver mutations in canonical driver genes: two clones in two different donors with *KDM6A* and *RHOA* and one clone with *KDM6A* and *ATM*.

5. Selection and driver analyses

To quantify the extent of selection and to identify positively-selected genes, we used the *dNdScv* algorithm, a maximum-likelihood implementation of dN/dS specifically developed for somatic mutation data (https://github.com/im3sanger/dndscv).

The method is described in detail in the original publication (*19*). Briefly, in the *dNdScv* algorithm, dN/dS ratios are calculated using a context-dependent substitution model with 192 rate parameters to model each substitution type in each trinucleotide context, including transcriptional strand biases. Since somatic mutation datasets are sparse, the substitution model is fitted on all genes of interest. The density of mutations has been shown to vary considerably across genes, often depending on the chromatin state, expression level and replication time (*61*). This is modeled by *dNdScv* using a negative binomial regression with epigenomic covariates, in effect modeling the variation in mutation density across genes (beyond that expected from the substitution model and sequence composition) as being Gamma-distributed. Likelihood ratio tests are used to detect selection on missense and truncating (nonsense and splice site) substitutions, and a separate negative binomial regression model is used to detect positive selection on indels.

5.1 Driver discovery in transplant donors

In order to identify genes under positive selection in normal bladder we first combined the mutation calls from targeted and exome data from the 15 transplant donors. To do so, redundant calls from microbiopsies sequenced by both targeted and exome sequencing were collapsed, substitutions within 10 bp of an indel were filtered out and indels within 10 bp of one another were collapsed into a single event.

To avoid counting single mutational events multiple times, mutations shared across microbiopsies within an individual were conservatively collapsed into single entries. *dNdScv* was then run on the list of 321 genes in the targeted gene panel (methods S2.1) on the combined targeted and exome data. As in (*7*), we used default settings for *dNdScv* (version 0.0.0.9), with the exception of excluding from the fitting of the indel background model those genes found as significant using *dNdSloc*, and using the arguments *max_muts_per_gene_per_sample=Inf*, *max_coding_muts_per_sample=Inf*.

As shown in table S8, this analysis identified 12 genes under significant positive selection based on q-values <0.01 for all mutations combined (*qglobal_cv*<0.01) or for missense mutations separately (*qmis_cv*<0.01). Of these, 11 are known bladder cancer genes based on analysis of TCGA data by (*18*) or by (*19*). To increase our sensitivity to detect bladder cancer genes under positive selection, we also used restricted hypothesis testing (a separate FDR adjustment for 62 known bladder cancer genes from the two sources above) (*62*) with q-value<0.10. This identified an additional 4 bladder cancer genes as being under positive selection in normal urothelium from the 15 transplant donors.

Oncogenes are typically mutated at specific hotspot sites. The *dNdScv* algorithm does not exploit this information but two functions in the *dNdScv* R package have been developed to detect significant positive selection at hotspot sites: *sitednds* (at the level of single sites) and *codondnds* (at the level of single codons). Applying both functions (using the LNP background model (*63*))

identified significance recurrence at codon R200 of *GNA13* and codon F106 of *RHOA*, each with 3 mutations and q-value=0.001. RHOA had already been detected as significant in the gene-level analysis (table S8), but the hotspot analysis added *GNA13* as a positively selected gene in normal urothelium.

Outside of the significant thresholds above, we also found 4 missense mutations in *SPOP*, all at or near known hotspots. Despite their obvious clustering, they did not affect exactly the same site or codon and escaped detection with *sitednds* or *codondnds*. Interestingly, there is evidence of these hotspots in the TCGA bladder cancer dataset, with 3 mutations affecting codon R130 in *SPOP*. It is likely that *SPOP* is positively selected in normal urothelium and in bladder cancer.

5.2 Driver discovery outside of cancer genes

The availability of 483 whole-exomes from normal urothelium from the 15 transplant donors, allowed us to search for evidence of selection outside of known cancer genes. To do so, we ran *dNdScv* on all protein-coding genes in the genome using default arguments. Using q-values<0.01 for all mutations (*qglobal_cv*<0.01) or for missense mutations separately (*qmis_cv*<0.01) identified eight significant genes. Only one gene, *KRTAP5-3*, had not been found in the analysis of the targeted genes and was not a known bladder cancer gene. The only mutations in the gene were 4 indels and close evaluation of the calls and the locus revealed that the *KRTAP5-3* gene contains many small repeats and is prone to indel artefacts caused by misalignment, as suggested by a high density of indel calls at the locus in polymorphism datasets. This strongly suggested that *KRTAP5-3* was not a genuine hit and this gene was filtered out.

5.3 Driver discovery in bladder cancer patients

To study the driver landscape in histologically-normal urothelium from the patients with bladder cancer, we first ran *dNdScv* as described above (methods S5.1) on targeted and exome data from these patients alone. This identified four genes under significant positive selection based on *qglobal_cv*<0.01 or *qmis_cv*<0.01: *RBM10*, *ARID1A*, *RHOA* and *EEF1A1*. Only *EEF1A1* was not in the list of 17 genes under clear positive selection in normal urothelium from the transplant donors. However, its q-value in the transplant donors was just outside the conservative significance cut-off chosen in this study (qglobal_cv=0.0147, table S8).

We also combined all microbiopsies of histologically-normal urothelium across the 15 transplant donors and the 5 patients with bladder cancer to increase the power for driver discovery. Using the cut-offs described in methods S5.1 yielded a very similar list of significant genes with stronger support for *EEF1A1* and the addition of *KMT2C*, which was borderline significant but only under restricted hypothesis testing (q-value RHT = 0.0237).

Altogether, the driver landscape in normal urothelium from the five patients with bladder cancer appeared similar to that of the 15 transplant donors. *EEF1A1* and *KMT2C* emerged as likely drivers of clonal expansions in normal urothelium.

Finally, to study the driver landscape in von Brunn's nests and to determine whether they were driven by mutations in cancer genes, we ran *dNdScv* separately on 95 microbiopsies of von Brunn's nests. There were only 91 unique mutations in the list of 321 targeted genes in these microbiopsies

and running *dNdScv* only identified *ARID1A* as near-significant (*qglobal_cv*=0.0303), with only 4 mutations in the gene across the 95 microbiopsies. Global dN/dS ratios for the targeted genes were also modest in von Brunn's nests, in line with those of normal urothelium.

5.4 Estimation of the number of driver mutations

dN/dS ratios can be used to estimate the excess of non-synonymous mutations compared to the neutral expectation, providing an estimate of the number of positively-selected mutations in a set of mutations (*19*). In this study, a conservative estimate of 385 (CI95%: 357, 401) driver mutations in the 17 genes detected under positive selection was obtained by calculating the excess of non-synonymous substitutions and indels in them, using the approach described in (*7*). Confidence intervals for the excess of non-synonymous substitutions were calculated using the global dN/dS ratios obtained by Poisson regression in *dNdScv*. For consistency with the background model used in *dNdScv*, the predicted excess for indels was calculated by estimating the number of indels per coding base-pair in the 315 genes not detected as significant by *dNdSloc*. The six genes that were identified as significant by *dNdSloc* were: *KMT2D*, *KDM6A*, *ARID1A*, *STAG2*, *RBM10* and *EP300*. Confidence intervals for the global observed/expected ratio for indels were calculated using the confidence interval for the ratio of two Poisson observations.

5.5 Variation in selection between donors across genes

This analysis was done as described in (*7*). To evaluate whether certain genes were under stronger positive selection in one individual than in others, we compared dN/dS ratios between donors using a Likelihood-Ratio Test for each gene in each donor. By using dN/dS ratios, this approach corrected for differences in the mutation density across genes caused by differences in sequencing coverage or mutational signatures, rather than selection.

Let $\omega_{g,1}$ and $\omega_{g,2}$ be the maximum-likelihood estimates (MLEs) for the dN/dS ratios for gene *g* in datasets 1 (one particular individual) and 2 (all other individuals), respectively. Here we used MLEs from the *dNdScv* model as they combine local and global information to estimate the background mutation rate in a gene. We can test for higher dN/dS ratios in a gene in a given individual using a one-sided test with the following null and alternative hypotheses:

H_0: $\omega_{g,1} \leq \omega_{g,2}$
H_1: unconstrained $\omega_{g,1}$ and $\omega_{g,2}$

However, higher dN/dS ratios in a given gene in an individual could also be due to stronger selection across all driver genes in the donor, rather than reflecting differences in relative selection between genes. Instead, we used a more conservative test by removing the effect of global differences in dN/dS ratios across the 17 driver genes identified in this study. Let $\boldsymbol{\omega_1}$ be the MLE of the global dN/dS ratio from all genes other than the gene being tested. We can then estimate a gene-specific relative enrichment of non-synonymous mutations over $\boldsymbol{\omega_1}$ as a multiplicative factor ($\omega'_{g,1}$): $\omega_{g,1}=\boldsymbol{\omega_1}\omega'_{g,1}$. Based on this, a more conservative Likelihood-Ratio Test can be used to test for an enrichment of non-synonymous mutations in a gene in a donor compared to all other donors, while removing the effect of differences in mutational signatures and global differences in the signal of selection:

H_0: $\omega'_{g,1} \leq \omega'_{g,2}$
H_1: unconstrained $\omega'_{g,1}$ and $\omega'_{g,2}$

This test was used in Fig. 2G, running it on microbiopsies of normal urothelium (targeted and exome data) from the 15 transplant donors. The analysis was restricted to the 10 of the 17 driver genes that had ≥10 non-synonymous mutations in the dataset (list shown in Fig. 2G). After multiple testing correction using the Benjamini-Hochberg procedure, the following five gene-donor pairs were found to be significant (q-value<0.05):

ARID1A in T06_59M (q-value=0.0042).
KMT2D in T01_25F (q-value=0.0042).
KMT2D in T15_78F (q-value=0.024).
RBM10 in T07_59M (q-value=0.024).
KDM6A in T03_53F (q-value=0.043).

Fig. 2G also suggests that the density of driver mutations varies across individuals, with individuals such as T12_70M having a seemingly lower fraction of coding mutations annotated as possible drivers. To test this formally, while accounting for the discrete number of mutations detected, we can use an overdispersion test. Specifically, we can use a likelihood ratio test with a null hypothesis stating that the fraction of all coding mutations that are annotated as drivers is the same across individuals (with binomial variation), and an alternative hypothesis where the relative density of driver mutations varies across individuals (beta-binomial model, with a beta distribution reflecting variation in driver density across individuals) (1 degree of freedom). The implementation of this test is available in the supplementary code. This analysis confirmed that the density of driver mutations varied significantly, independently of mutation burden, across the 15 transplant donors (*P*-value=2.9e-5, likelihood ratio test).

5.6 Germline mutations in genes differentially selected across donors

Germline variants, including single-nucleotide polymorphisms (SNPs) and indels, were called with gatk-4.0.1.2 *HaplotypeCaller*. The functional impact of SNPs was annotated with Variant Effect Predictor (VEP, v97.3) using options *--everything* and *--canonical*. We kept variants with functional impacts classified as MODERATE or HIGH by VEP on canonical transcripts.

To explore whether the high frequency of *KDM6A* mutations in T03_53F (Fig. 2, G and H) could be explained by the presence of some germline deleterious variation, we curated a collection of genes related to histone modifications, including histone demethylases (*KDM2A, KDM4D, KDM4E, KDM5A, KDM2B, KDM8, KDM6B, KDM4B, KDM1A, KDM4A, KDM5B, KDM3A, KDM3B, KDM1B, KDM7A, KDM4C, KDM5C, KDM6A, KDM5D, RSBN1, HR, PHF2*), histone binding (*RBBP4, RBBP7*), histone-lysine N-methyltransferases (*KMT2A, KMT2D, KMT2B, KMT2E, KMT2C, DOT1L, SETD1B, SETD8, SETDB2, SETD3, SETD1A, SETD6, SETDB1, SETD4, SETD5, SETD2, SETD7, SETD9, EZH2, ASH2L*), histone deacetylases (*HDAC7, HDAC5, HDAC1, HDAC10, HDAC4, HDAC11, HDAC3, HDAC2, HDAC9, HDAC6, HDAC8),* histone acetyl transferases *(NCOA1, NCOA3, NCOA2, NCOA6, NCOA7, NCOA4, CLOCK, KAT5, KAT6A, CREBBP, KAT6B, KAT7, KAT2B, KAT8, KAT2A, ATF2*), and other functionally related

genes (*HCFC1, RBBP5, WDR82, WDR5, CXXC1, DPY30, PAXIP1, PAGR1*). We searched for potentially deleterious variants in these genes, using *VEP*, *SiFT* and *PolyPhen* functional impact predictions.

Similarly, we explored whether there were deleterious germline variants in the donor T06_59M that may explain the high frequency of *ARID1A* mutations in this individual (Fig. 2, G and I). We scanned the following functionally related genes: *ARID1A, ARID1B, ARID2, ARID3A, ARID3B, ARID3C, ARID4A, ARID4B, ARID5A, ARID5B, SMARCA1, SMARCA2, SMARCA4, SMARCA5, SMARCAD1, SMARCAL1, SMARCB1, SMARCC1, SMARCC2, SMARCD1, SMARCD2, SMARCD3, SMARCE1, PHF10, PBRM1*, and *TERT*.

None of these analyses revealed a clear link between *KDM6A* and *ARID1A* mutation recurrence in T03_53F and T06_59M and germline deleterious variation.

5.7 Mutation frequency and selection in bladder carcinomas from The Cancer Genome Atlas

To determine the frequency of non-synonymous substitutions and indels in bladder cancer for the genes shown in Fig. 2D and Fig. 2E, we used the public MC3 mutation calls from TCGA for 411 muscle-invasive bladder cancers. *dNdScv* was used to annotate coding variants (*annotmuts* output object) for these figures.

Driver discovery on the list of 321 target genes was performed using *dNdScv* with default parameters, excluding genes found as significant by *dNdSloc* (q-value<0.05) from the indel background model of *dNdScv* (*kc* argument). Instead of running *dNdScv* on the 321 genes using the *gene_list* argument, *dNdScv* was run on the entire exome to use all available information and restricted hypothesis testing on the list of 321 target genes was used to obtain q-values restricted to the list of target genes. Genes with *qglobal_cv*<0.01 or *qmis_cv*<0.01 were considered as significant in bladder cancer for Fig. 2D, which yielded a list of 47 significant genes from the list of 321 targeted genes. Of these, 36 were already found as significant in (*18*) or (*19*). The full list of 47 significant genes (sorted by *qglobal_cv* and with genes not previously found to be significant in red) was the following: *TP53, PIK3CA, FGFR3, CDKN2A, ARID1A, KDM6A, KMT2D, RB1, STAG2, ELF3, CDKN1A, ZFP36L1, TSC1, EP300, RHOB, CREBBP, FBXW7, KMT2C, RHOA, ERCC2, HRAS, ERBB2, FOXQ1, KRAS, ERBB3, PTEN, ARID2, KLF5, KMT2A, FAT1, FOXA1, EPHA2, RBM10, ATM, NFE2L2, ASXL1, HIST1H3B, GPS2, RUNX1, ARHGAP35, GATA3, NF1, ARID1B, FOXP1, ZNF750, BAP1* and *NCOR1*.

5.8 Selection at putative antigenic regions

To look for evidence of possible immune editing against mutant clones in normal urothelium, we searched for negative selection in putative antigenic regions. We used the dNdScv package to calculate dN/dS ratios in exonic regions overlapping putative antigenic sites, using a full 192-rate trinucleotide substitution model. We used three different sources of antigenic regions, following Eynden et al. and Zapata et al. (*64, 65*):

(1) A large set of predicted HLA-binding regions from Eynden et al. (*64*), based on the harmonic mean of Kd values for 6 common HLAs (using a cut-off of 500nM on the harmonic mean). These regions extend ~22% of the exome.

(2) All epitopes in the Immune Epitope Database and Analysis Resource (IEDB) mapped to the hg19 assembly of the human genome by Eynden et al. (*64*).
(3) The intersection of the two sets of regions above, similar to the approach used by Zapata et al. (*65*).

All dN/dS ratios for missense and truncating mutations in the three sets of regions above were close to and not significantly different from 1, in line with the observation in cancer genomes from (*64*). This analysis revealed no clear evidence of immune editing acting on mutant clones in normal urothelium. However, we note that lack of evidence does not demonstrate lack of immune editing. Larger datasets and improved prediction of antigenic sites will be required for a more detailed analysis.

6. Mutational signatures

6.1 *De novo* signature extraction

De novo extraction of mutational signatures was performed using a Bayesian hierarchical Dirichlet process (*66*) implemented in the *HDP* R package (https://github.com/nicolaroberts/hdp). The units of signature extraction were substitutions called in whole-genomes derived from microbiopsies of urothelium and von Brunn's nests for both transplant donors and cystectomy patients. Substitutions were classified by mutation type (annotated from the pyrimidine base) and trinucleotide context. For the duplicate and triplicate whole-genomes (methods S2.3), unique mutations from each set of samples were treated as a single sample to avoid double counting of mutations. No prior signatures were assigned to frozen nodes. The hyperparameters for the α clustering parameter (α and β) were both set to 1. Extraction was started with 10 data clusters (parameter 'initcc'). The first 20,000 iterations of the Gibbs sampler were not collected (parameter 'burnin') after which 100 posterior samples were collected (parameter 'n') at intervals of 1000 iterations (parameter 'space'). After each Gibbs sampling iteration, three iterations of concentration parameter sampling were performed (parameter 'cpiter'). The results from 10 chains with different seeds were combined for signature extraction. Clusters with cosine similarity >0.9 were merged. Clusters with no significant data categories were combined into a null signature.

In addition to the null signature, nine signatures were extracted (fig. S5). The four most abundant signatures accounted for >89% of mutations. As the remaining five signatures only accounted for a small proportion of mutations (each <3%), these were excluded from further analyses and any mutations assigned to these low abundance signatures were binned alongside the null signature mutations into the "Unassigned mutations" category in Fig. 3D and fig. S14C.

6.2 Alternative approaches for signature extraction

In addition to de novo signature extraction with HDP we performed signature extraction in three complementary ways:
(1) Non-negative matrix factorization (NMF) using SigProfiler.
(2) HDP conditioning the algorithm on known bladder cancer signatures as priors.
(3) HDP de novo combining all genomes of normal urothelium with 23 published bladder cancer genomes from the PCAWG consortium.

Non-negative matrix factorization (NMF) using SigProfiler. In order to determine whether the signatures identified by HDP were robust, we also performed *de novo* signature extraction using SigProfiler (*29*) (https://github.com/cancerit/docker-sigprofiler). Unlike HDP, signature extraction with SigProfiler is based on non-negative matrix factorization. The same samples were used for *de novo* extraction with SigProfiler as for HDP, with duplicate and triplicate samples collapsed as described in methods 6.1. The upper bound for the number of signatures extracted by SigProfiler was set to 10. Assessment of the stability and accuracy of the possible solutions identified the most likely solution to be four extracted signatures.

The four signatures extracted with SigProfiler are nearly identical to those extracted by HDP. Cosine similarities between SigProfiler's and HDP's signatures were 0.99 for signature A, 0.997 for signature B, 0.98 for signature C and 0.997 for APOBEC (fig. S7A). This confirms that the results presented in the main text are not contingent on the use of HDP.

HDP conditioning the algorithm on known bladder cancer signatures as priors. In addition to *de novo* signature extraction, we also ran HDP after conditioning the algorithm with signatures previously identified in bladder cancer. An analogous approach was used for signature extraction in normal colorectal epithelial cells by Lee-Six *et al*. (*9*). We conditioned HDP with the four single-base substitution signatures identified in ≥10% bladder cancers (COSMIC signatures version 3; https://cancer.sanger.ac.uk/cosmic/signatures)(*29*): SBS1, SBS2, SBS5 and SBS13. A weighting of 10,000 pseudocounts was used for each of these signatures to provide a strong prior. The pseudocount nodes are frozen, such that the pseudocounts are unable to leave their initial clusters during the Dirichlet process, but counts from the dataset may join these clusters. Sampling parameters for HDP are as described in methods 6.1.

The use of a strong prior was designed to facilitate the detection of signatures previously identified in bladder cancer genomes, as well as to help separate putative novel signatures from already known signatures. This analysis yielded the same four dominant signatures in normal urothelium identified *de novo* by HDP and SigProfiler, with the exception of splitting the APOBEC signature into its two components (SBS2 and SBS13). Cosine similarities between HDP using priors and HDP *de novo* were 0.999 for signature A, 0.998 for signature B, and 0.999 for signature C (fig. S7D). Importantly, signatures A, B and C, were stably extracted as separate from SBS5 and SBS1. This analysis also suggests that SBS5 is likely present in modest quantities in normal urothelium. If we use the stringent criterion used in the main text, where a signature is only reported in a sample if its 95% credibility interval does not extend to zero, SBS5 is only detected in four genomes of normal urothelium (only two genomes excluding samples from patients with bladder cancer) (fig. S7B). If we use a more liberal approach, accepting mutational signature attributions independently of their credibility intervals, a small minority of mutations in most samples of normal urothelium may be attributable to SBS5 (fig. S7C). Overall, this analysis confirms that the four main signatures described in the main text are stably extracted despite the use of strong priors and appear largely separate from SBS5.

HDP de novo combining all genomes of normal urothelium with 23 published bladder cancer genomes from the PCAWG consortium. An alternative approach of utilizing pre-existing information to identify signatures in normal tissues is to include previously sequenced cancer genomes as samples in *de novo* signature extraction, as in Brunner *et al*. (*10*). Therefore, we also

ran HDP including the 23 PCAWG bladder cancer samples. In order to maximize their weighting, cancer samples were assigned to the same parent node as the normal urothelium samples. Sampling parameters for HDP are as described in methods 6.1.

This analysis yielded very similar results to those presented in the main text. The four commonest mutational signatures in samples of normal urothelium were analogous to the four signatures in the main text. APOBEC was partially split into two components. Cosine similarities between HDP using PCAWG bladder cancer genomes and HDP on normal urothelium alone (main text) were 0.98 for signature A, 0.99 for signature B, and 0.997 for signature C (fig. S8). Two signatures with partial resemblance to SBS5 (cosine similarity 0.86) and SBS1 (cosine similarity 0.91) together with a C>A-rich signature, contributed a modest number of mutations across normal urothelium genomes.

Overall, the four alternative approaches used to extract mutational signatures in this manuscript yielded similar results, identifying the four signatures in the main text as the most dominant signatures in normal urothelium. In additional, smaller contributions from other signatures, particularly SBS5, appear likely. Signature C has an interesting profile with distinct peaks of T>A and T>G at ATT sites. As noted in methods S6.5, the transcription strand asymmetry of these peaks appears opposite to that of T>C mutations in this signature, raising the possibility of incomplete separation from signature A, although an unambiguous assessment was not possible due to the overlap with signature A. Ultimately, larger datasets of bladder cancer genomes and of normal urothelium would be expected to expand and refine the signatures extracted above.

6.3 Comparison to reference signatures

In addition to *de novo* signature extraction, the closest fit from linear combinations of signatures previously described in cancer genomes (*67*) was identified using the *DeconstructSigs* R package (*68*) for the urothelium and von Brunn's nests whole-genomes from transplant donors and cystectomy patients. When all previously described signatures were used (fig. S5A), a substantial proportion of mutations were attributed to signatures rarely observed in bladder cancer (such as SBS12, SBS16 and SBS19). Alternatively, when signature fitting was carried out solely using signatures frequently observed in bladder cancer (SBS1, SBS2, SBS3, SBS5 and SBS13), the best fit often had a low cosine similarity with the observed mutational spectra. In conjunction, these two results suggest that signatures not previously described in bladder cancer were present in normal urothelium.

The four main extracted signatures (methods S6.1) were fitted to previously described cancer signatures using the *DeconstructSigs* R package (*68*). For each of the four extracted signatures, mutation spectra were simulated using 1,000,000 mutations and the closest fits were identified for linear combinations of individual, pairs, triplets or an unlimited number of signatures previously extracted from the PCAWG data using SigProfiler (*67*) (fig. S6A). This process was repeated for the signatures extracted from the PCAWG data using SigAnalyzer (*67*) (fig. S6B).

Cosine similarities were calculated across the 96 trinucleotide mutation contexts (i) between the extracted signatures (A) and the fitted linear combinations (B) as follows:

$$\text{Similarity} = \frac{\sum_{i=1}^{96} A_i B_i}{\sqrt{\sum_{i=1}^{96} A_i^2} \sqrt{\sum_{i=1}^{96} B_i^2}} \quad \text{(Eq. 3)}$$

6.4 Mutational signature analysis in bladder cancer genomes

In order to determine whether there was any evidence for the presence of the novel extracted signatures in bladder cancers, we used the *DeconstructSigs* R package (*68*) to fit linear combinations of signatures previously identified in bladder cancer (SBS1, SBS2, SBS3, SBS5 and SBS13) and the three novel extracted signatures (Signatures A, B and C) to 23 bladder cancers from the PCAWG dataset (fig. S6C).

Comparable results were also obtained by re-running *HDP* with the PCAWG bladder cancer genomes as additional samples (methods S6.1). This alternative approach identified Signature A as contributing >20% mutations in PCAWG samples 1, 4, 7 and 20, as well as >50% mutations in PCAWG sample 9. By contrast, none of the PCAWG samples had >20% of their mutations attributable to either Signature B or Signature C.

6.5 Transcriptional and replicational strand asymmetries

Mutations within gene bodies were annotated as to whether the pyrimidine base was located on the template or coding strand. Gene co-ordinates were obtained from the RefSeq database (*69*). Mutations occurring at loci where transcription occurs from both strands were excluded. Additionally, as the trinucleotide composition of exonic regions differs from the genome-wide trinucleotide composition, only mutations within introns were used for identifying transcriptional strand asymmetries.

The probability, p, that a particular mutation, i, could be assigned to a given signature, j, in genome k was calculated as follows:

$$p_{i,j,k} = \frac{w_{j,k} \cdot f_{i,j}}{\sum_j w_{j,k} \cdot f_{i,j}} \quad \text{(Eq. 4)}$$

where $w_{j,k}$ is the proportion of mutations assigned to signature j in genome k by *HDP* (methods S6.1) and $f_{i,j}$ is the fraction of mutations in signature j that are the same substitution type and occur at the same trinucleotide context as mutation i.

Mutation assignment probabilities were used in two different ways for analyzing transcriptional strand asymmetries:

(1) For some analyses, only mutations with an assignment probability >0.7 were included (fig. S9, A-D and I-L). Advantages of this approach are that Poisson tests can readily be applied to mutation counts and it reduces issues caused by bleed-through of overlapping signatures. However, a disadvantage of this approach is that the exclusion of ambiguous mutational contexts means that the catalogue of analysed mutations can differ substantially from the complete extracted signature, as exemplified by the lack of non-C>T mutations in fig. S9B.

(2) Alternatively, mutation assignment probabilities can be summed together (fig. S9, E to H). The resultant profile closely matches the extracted signature but bleed-through of overlapping signatures can give seemingly contradictory results, as exemplified by the T>C mutations appearing to have the opposite asymmetry to T>A and T>G mutations for Signature C in fig. S9H.

For the analysis of transcriptional asymmetry in highly vs. lowly expressed genes (fig. S9, I to L), genes in the top and bottom expression quartiles of the PCAWG bladder cancer expression data were compared. Transcribed regions were identified according to Ensembl gene annotations, with regions 10 kb upstream and downstream of genes used as controls. Mutation rates were calculated by dividing mutation counts from the template strand by the number of bases of that type in the region. Significant differences between template and coding strands were assessed using Poisson tests. Confidence intervals were calculated using binomial tests.

Analyses of replicational strand asymmetries and extended nucleotide contexts both used the per mutation assignment probabilities described above. The threshold approach was used in fig. S10, A-D and I-L, and the sum of assignment probabilities approach was used in fig. S10, E to H. The assignment of mutations to left- and right-replicating regions was performed using previously defined classifications of genomic regions (*70*).

6.6 Detection of APOBEC mutagenesis in exomes

To determine whether there was statistical evidence of APOBEC mutagenesis in individual exomes, we modeled the mutations observed in each exome as being multinomial draws from a mixture of the four mutational signatures found in this study. Since these mutational signatures were identified from whole-genome data, we used the genome-to-exome normalization from the *DeconstructSigs* R package (*68*). An Expectation-Maximization algorithm was then used to estimate the maximum-likelihood contribution of each signature, as described in (*19*). A Likelihood-Ratio Test was constructed by calculating the likelihood of explaining the mutations in each exome using either three signatures (Signatures A-C) or four signatures (Signatures A-C and APOBEC), using the multinomial density function (*dmultinom* in R). Multiple testing correction was performed using the Benjamini-Hochberg procedure and exomes with q-values<0.05 for the presence of APOBEC were considered APOBEC-positive for the analyses shown in the main text.

6.7 Spatial clustering of APOBEC-positive clones

Analysis of the spatial clustering of APOBEC-positive clones was initially restricted to the 437 exomes of histologically-normal urothelium from transplant donors that were derived from microbiopsies containing a single cut and had had their position annotated on a SlideScanner image (table S2). In order to avoid double-counting clones that spanned multiple microbiopsies, samples which shared ≥5 mutations were identified. Exomes were iteratively excluded from networks of shared samples until none of the remaining samples shared ≥5 mutations by applying the following hierarchical filters:

(1) Exclude the most connected sample in the network

(2) Exclude the sample that has the fewest other microbiopsies from the same section remaining
(3) Exclude the first sample alphabetically

After excluding duplicate samples, 362 exomes remained. Exomes were classified as APOBEC-positive or APOBEC-negative as described above (methods S6.6). Microbiopsies from the same section and with a Euclidean distance ≤1 mm between the centroids of their SlideScanner annotations were classified as neighbors. Of the 362 non-duplicate samples, 274 had at least one neighbor. For the 51 APOBEC-positive clones with neighbors, the proportion of neighbors that were also APOBEC-positive was calculated. The observed average proportion of neighbors of APOBEC-positive clones that were also APOBEC-positive was 0.13.

Labels indicating whether a microbiopsy was APOBEC-positive or APOBEC-negative were randomly permuted between samples from the same transplant donor (including samples with no neighbors). The average proportion of APOBEC-positive neighbors for APOBEC-positive clones with neighbors was calculated for 10,000 iterations of this permutation test. No evidence was found for a higher degree of spatial clustering of APOBEC-positive samples than expected by chance (fig. S13).

6.8 Statistical association between mutational signatures, smoking and alcohol consumption

To test for possible associations between the four mutational signatures and smoking or alcohol consumption history, we used linear mixed-effect regression models (*lme4* package in R). The relative contribution of each signature was used as the response variable, smoking or alcohol consumption history were used as fixed effects, and a random effect was used for the intercept across individuals. Unlike simpler regression models, the use of a random intercept model controls for the non-independence of multiple genomes per individual.

Smoking history was available for all 20 individuals in the study (Table S1). We used two alternative regression models on smoking history: (1) using a binary classification (*heavy smoker*, defined as >10 pack-years, and *never/low-smokers*, defined as ≤5 pack-years), and (2) using pack years. Alcohol consumption history was only available for the 15 transplant donors and was encoded into three levels for the regression analysis (1: *none* or *rare*, 2: *light* or *moderate*, 3: *heavy*) (Table S1).

The R code used for this analysis is available within the supplementary code provided, but for illustrative purposes the pseudocode for the regression models on: (1) binary smoking history, (2) pack years, and (3) alcohol consumption, is shown below:

lmer(signaturefraction ~ donortype + gender + smokingbinary + (1|donorID), REML=F)
lmer(signaturefraction ~ donortype + gender + packyears + (1|donorID), REML=F)
lmer(signaturefraction ~ smokingbinary + gender + alcohol + (1|donorID), REML=F)

P-values for the association with smoking and alcohol history were obtained using likelihood-ratio tests (*anova* function). These analyses revealed a significant positive association between signature A and smoking history, both using the binary classification (*P*-value=9.4e-5) and pack years (*P*-

value=0.0033) (fig. S11). No other signature showed a significant association with either smoking or alcohol history (all *P*-values>0.05).

7. Comparison of normal urothelium between transplant donors and bladder cancer patients

It should be noted that this dataset was not adequately powered to identify modest epidemiological associations, as our cohort included only 20 patients and there was extensive inter-individual variation in mutation burden and selection. We include these analyses here for completeness, but they should be considered preliminary analyses with limited statistical power. Much larger cohorts will be needed to study differences in the mutational landscape as a function of epidemiological factors.

To explore whether there were differences in mutation burden and allele fractions in histologically-normal urothelium between transplant donors and patients with bladder cancer, or as a function of smoking history or gender, we used linear mixed-effect models. We used the *lme4* package in R to fit different mixed-effects regression models with different fixed effects and a random effect for the intercept across donors. Unlike simpler regression models, the use of this random intercept model controls for the non-independence of multiple microbiopsies per donor.

Before applying the regression models, we removed the confounding effect of variable sequencing coverage across microbiopsies and across donors using *in silico* subsampling of the mutation calls. Microbiopsies with a median on-target coverage after PCR duplicate removal ≥30x in ≥80% of the target regions were included in this analysis. For each mutation call, 30 reads were selected at random and mutation calls supported by <4 reads were removed. This largely removes differences in coverage across samples, ensuring a uniform coverage of ~30x across microbiopsies and donors. After *in silico* subsampling, the number of mutations per exome and their mean VAF were used as outcome variables for the regression models below. Although age could be included as a confounding factor in the regression model, this would rely on certain assumptions about the shape of the relationship between age and the outcome variables. Instead, we excluded from the analysis the two youngest transplant organ donors, which resulted in similar ages for transplant donors (mean 64.4 years) and patients with bladder cancer (mean 64.8 years). In total, we used data from 421 exomes across 16 individuals (two of the 18 middle-age or elderly individuals did not have any exomes with sufficient coverage after *in silico* subsampling).

To test for differences in mutation burden or mean VAFs between transplant donors and patients with bladder cancer, we used the following regression models (R code):

model = lmer(mutperexome ~ patient_type + gender + smokingclass + (1|patient), REML=F)
model = lmer(meanvafperexome ~ patient_type + gender + smokingclass + (1|patient), REML=F)

patient_type indicated whether the patient was a transplant donor or a patient with bladder cancer. s*mokingclass* was a binary variable indicating whether the patient was a heavy smoker (≥10 pack years) or a moderate/never smoker (<10 pack years). A simple binary classification for the

smoking history was chosen given the modest size of the cohort, but this analysis should be considered preliminary, as mentioned above.

The likelihood of the regression model above was then compared to the likelihood of a null model without *patient_type* as predictor using an ANOVA test. This yielded a *P*-value=0.00682 for the difference in the number of mutations per exome in patients with bladder cancer vs transplant donors (fig. S14), and *P*-value=0.000481 for differences in mean VAFs (suggesting significantly larger clone sizes in patients with bladder cancer).

Using analogous tests, no significant effects were found for *smokingclass* (*P*-value=0.166 and *P*-value=0.135, respectively) or *gender* (*P*-value=0.235 and *P*-value=0.339, respectively). However, as discussed above, larger datasets may be expected to detect significant effects for these variables, given their association with bladder cancer risk.

To test for differences in driver density, we used a regression model using the number of driver mutations per exome after subsampling as the outcome variable. In particular, we used the number of non-synonymous mutations in any of the 17 driver genes as the outcome variable in a negative binomial regression (*MASS* R package) using the number of samples as the offset and *age* and *patient_type* as predictors. This analysis yielded a *P*-value=0.069 (nominally lower density of driver mutations in normal urothelium from patients with bladder cancer, although this difference was not significant). Again, much larger cohorts of patients will be required to test this accurately, as the subsampled dataset only contained four patients with bladder cancer.

8. Phylogenetic reconstruction

In order to identify early embryonic mutations, we ran CaVEMan in "unmatched" mode. Instead of using a matched normal, we used a synthetic sample (PDv37is), thus avoiding filtering out early embryonic mutations that may also be present in a matched normal sample and so would be incorrectly identified as germline variants. Calls were retained if they passed the default CaVEMan filters, as well as if the median number of clipped bases for reads supporting the variant was <5 and the median alignment score for supporting reads was >125. Identification of early embryonic mutations was restricted to samples that did not carry large copy number alterations, as these would impact the performance of subsequent filters (such as the overdispersion filter).

In order to distinguish *bona fide* early embryonic variants from germline variants and artefacts, we ran cgpVAF (https://github.com/cancerit/vafCorrect) to obtain read counts for each of the mutations called using unmatched CaVEMan across all samples from the same donor. Mapping and base qualities >30 were required for reads to be counted by cgpVAF. Sites with a global coverage lying outside of the 5th to 95th percentile were excluded as likely problematic regions. Variants with a global VAF >0.4 (or >0.8 for variants on the sex chromosomes in males) were filtered as germline SNPs. We required each putative early embryonic variant to be supported by ≥3 reads in ≥2 samples. We fitted a beta-binomial distribution to each call to estimate the overdispersion across samples, as germline variants and artefacts are likely to show low overdispersion, whereas true embryonic mutations will show higher overdispersion. We empirically found that the overdispersion threshold $\log(\rho) > -2$ properly separated noise from true

variation. This processing resulted in 20 and 35 early embryonic calls in C03_67M and C04_72M, respectively (fig. S16).

To call non-embryonic somatic mutations, we followed the approach described above (methods S3.2). We used cgpVAF to obtain the counts of reads supporting the mutant and reference bases across all samples and mutations. For each sample, we calculated the mean VAF across all sites with ≥2 mutant reads. To build the tree we selected clonal or nearly clonal samples, requiring that the mean VAF was ≥0.25. For carcinoma *in situ* and tumor samples, we instead required that the aberrant cell fraction calculated by ASCAT was ≥50%, as the high level of copy-number alterations meant that mean VAF was not a reliable estimator of clonality.

Early embryonic and non-embryonic somatic calls were combined together for the clonal samples. To binarize the VAF matrix into an absent/present mutation matrix, we required the VAF to be >0.15 for variants to be considered present. VAFs between 0.05 and 0.15 were assigned an unknown status. Otherwise, the mutation was flagged as absent. This was done to reduce noise due to residual artefacts or low frequency embryonic mutations. We used the resulting binarized matrix to estimate a maximum parsimony tree using the *phangorn* R package (*71*). We applied the *pratchet* method to find the best topology (*72*) and the *acctran* method to estimate branch lengths.

We assigned SNVs and DNVs to branches using a maximum likelihood approach. To estimate the contribution of SNV signatures A, B, C, and APOBEC to each branch of the resulting tree, we used R package *DeconstructSigs (68)*. The resultant trees were annotated for non-synonymous mutations occurring in a set of 75 genes of interest. This set of genes consisted of bladder cancer drivers found in (*19*), TCGA-bladder (*18*), and this study. For the branch shared by the tumor and the CIS in C03_67M, we also annotated high impact mutations in *KDM3A* and *EPHB1* (Fig. 4D).

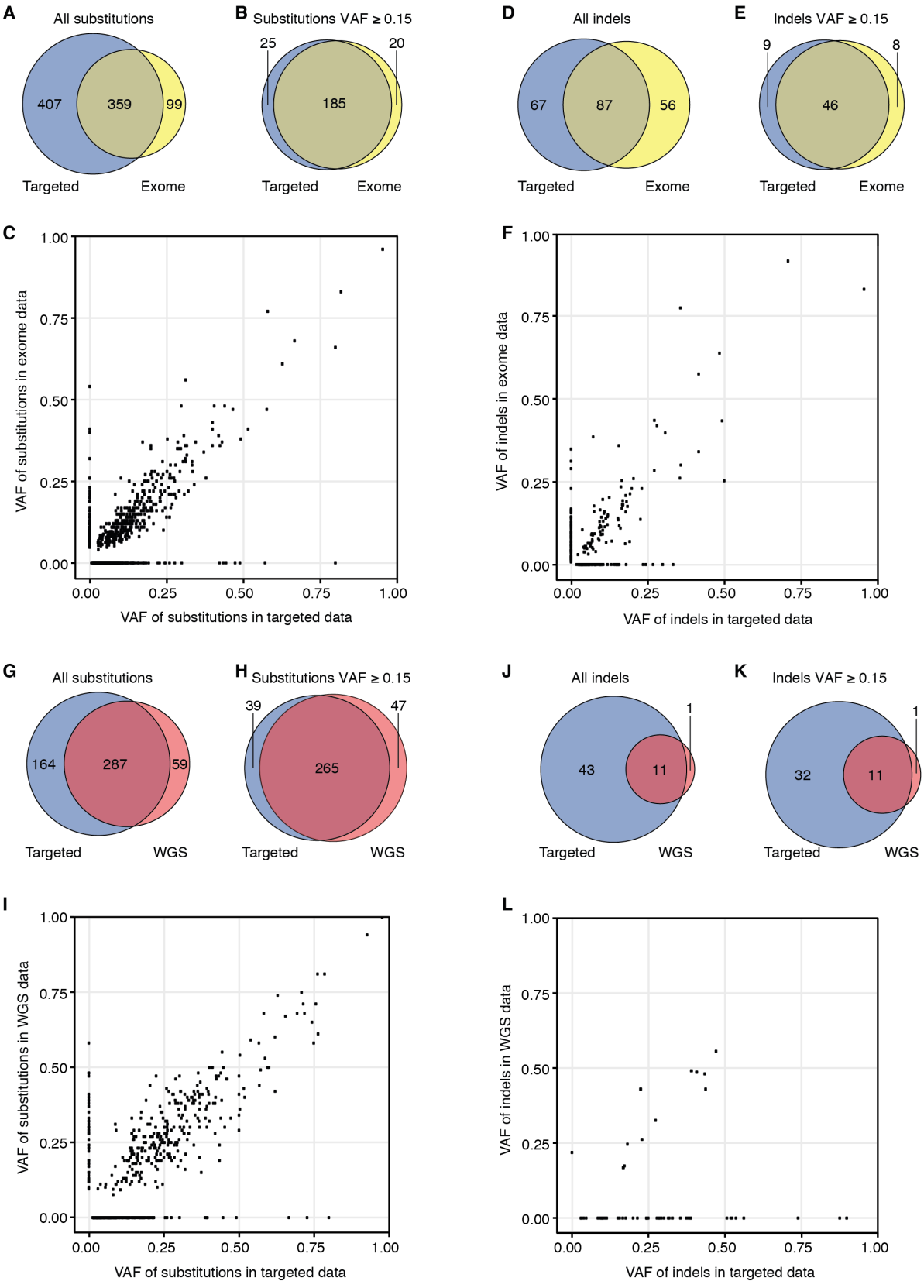

A
All substitutions
407
359
99
Targeted
Exome
B
Substitutions VAF ≥ 0.15
25
185
20
Targeted
Exome
D
All indels
67
87
56
Targeted
Exome
E
Indels VAF ≥ 0.15
9
46
8
Targeted
Exome
C
VAF of substitutions in exome data
VAF of substitutions in targeted data
0.00
0.25
0.50
0.75
1.00
F
VAF of indels in exome data
VAF of indels in targeted data
G
All substitutions
164
287
59
Targeted
WGS
H
Substitutions VAF ≥ 0.15
39
265
47
Targeted
WGS
J
All indels
43
11
1
Targeted
WGS
K
Indels VAF ≥ 0.15
32
11
1
Targeted
WGS
I
VAF of substitutions in WGS data
VAF of substitutions in targeted data
L
VAF of indels in WGS data
VAF of indels in targeted data

Fig. S1. Comparison of mutations called across sequencing strategies.

(**A** to **C**) Comparison of substitutions called in exome and targeted sequencing data. Substitutions were called using ShearwaterML for targeted sequencing and CaVEMan for exome sequencing. (**A**) Venn diagram for all substitutions called within the overlap of the exome and targeted capture regions from libraries that underwent both exome and targeted sequencing. (**B**) Venn diagram for substitutions with VAF $\geq$ 0.15 in either exome or targeted sequencing data. (**C**) Scatter plot comparing VAFs for exome and targeted substitutions. (**D** to **F**) Comparison of indels called in exome and targeted sequencing data. Indels were called using ShearwaterML for targeted sequencing and cgpPindel for exome sequencing. Panel descriptions are the same as A to C for indels rather than substitutions. To account for inconsistencies in the assignment of indel positions in repetitive sequences, a 20 bp window around the reported position was used when determining whether an indel was present in both exome and targeted calls. (**G** to **I**) Comparison of substitutions called in whole-genome and targeted sequencing data. Substitutions were called using CaVEMan for whole-genome sequencing. Panel descriptions are the same as A to C for whole-genome sequencing rather than exome sequencing. (**J** to **L**) Comparison of indels called in whole-genome and targeted sequencing data. Indels were called using cgpPindel for whole-genome sequencing. Panel descriptions are the same as A to C for indels rather than substitutions and for whole-genome rather than exome sequencing.

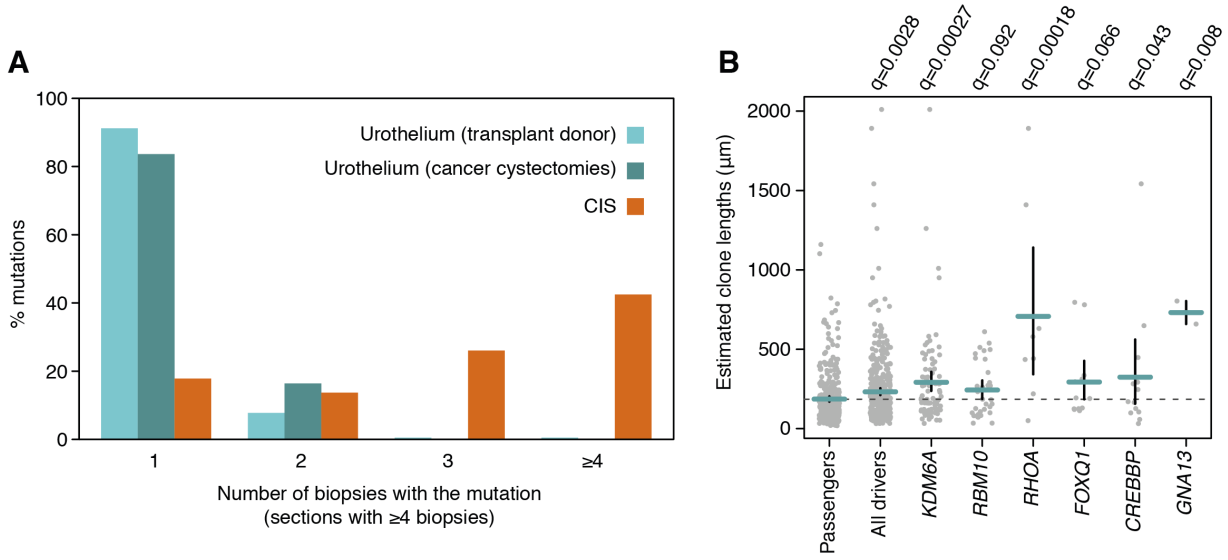


Fig. S2. Clonal expansions.

(**A**) Sharing of mutations across microbiopsies. Histogram showing the proportion of mutations that were detected in 1, 2, 3 or ≥4 microbiopsies within a given section for histologically-normal urothelium from transplant donors and bladder cancer patients, as well as for carcinoma *in situ* microbiopsies. Only sections containing ≥4 microbiopsies were included in this analysis. (**B**) Estimated clone lengths (µm) for putative drivers and putative passenger mutations. Putative driver mutations liberally include all non-synonymous mutations in the 17 positively-selected driver genes. Putative passenger mutations include any mutation in non-significant genes from samples without a putative driver mutation (to minimize the risk of hitchhiking with drivers). Horizontal green bars depict the means and error bars the confidence intervals for the means (obtained by bootstrapping). Dashed line represents the mean for passenger mutations. q-values were calculated with a permutation test using the mean as the statistic (to be sensitive to large clones) with 100,000 permutations. Only significant genes (q-value<0.1) are shown as separate genes in the figure.

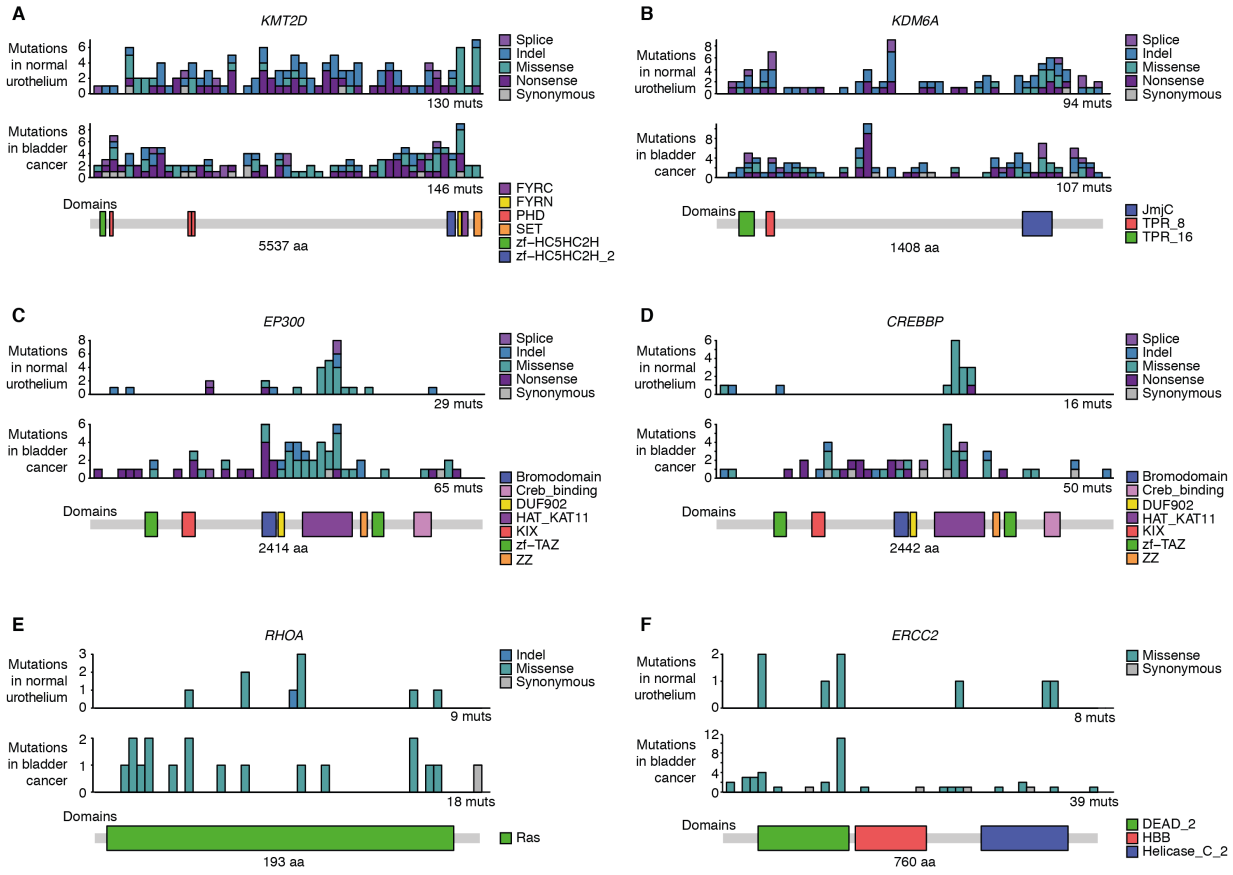


Fig. S3. Distribution of mutations within selected genes.

In panels A-F, the distribution of independent mutations within genes are shown for histologically-normal urothelium (top) and bladder cancer data from The Cancer Genome Atlas (middle) above a gene domain diagram (bottom). The number of bins (50) is constant for each gene and so the bin width is variable between genes. (**A** and **B**) Mutations are distributed throughout the gene bodies of *KMT2D* and *KDM6A* and there is a large number of nonsense mutations, consistent with these genes acting as tumor suppressor genes in bladder cancer. (**C** and **D**) Similar mutation distributions are seen for the homologues *EP300* and *CREBBP*, with enrichment of mutations in the histone acetyltransferase domain apparent in both for histologically-normal urothelium. (**E** and **F**) Enrichment of mutations at oncogenic hotspots and a lack of truncating mutations in *RHOA* and *ERCC2*.

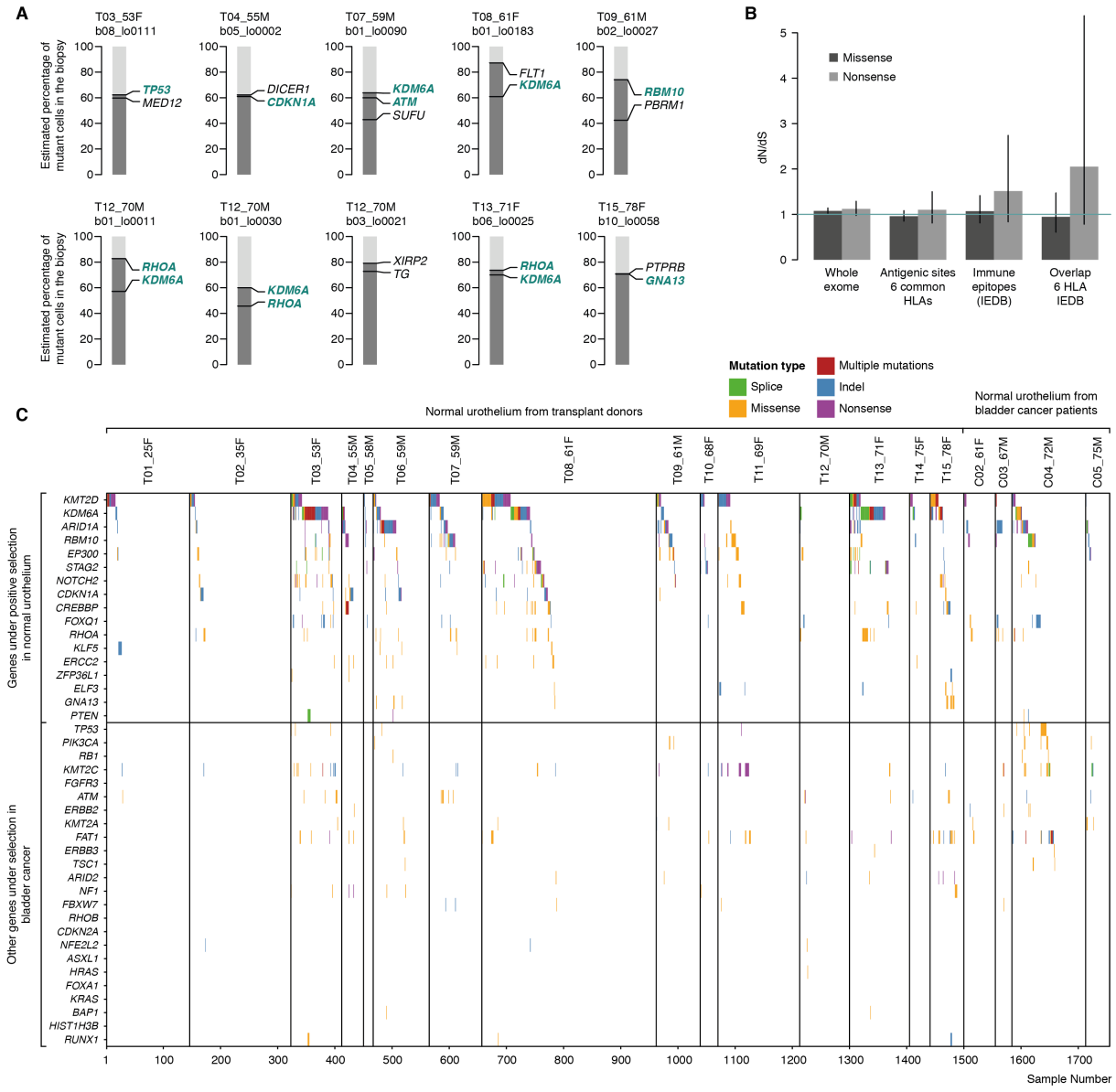


Fig. S4. Additional selection analyses.

(**A**) Results from the statistical pigeonhole principle analysis (methods S4.4). Each plot corresponds to a different microbiopsy and so only nine unique clones are shown as the same clone is identified for T12_70M_b01_lo0011 and T12M_70M_b01_lo0030. Bladder cancer driver genes are highlighted in green. (**B**). dN/dS ratios on all non-driver genes (whole-exome) and in putative antigenic regions using three different definitions: antigenic sites using the harmonic mean across 6 common HLAs (from (*64*)), immune epitopes from the Immune Epitope Database and Analysis Resource (IEDB) (from (*64*)), and the intersection of both of the above. All confidence intervals include 1, revealing not clear evidence of immune editing of mutant clones. (**C**) The presence of non-synonymous mutations from the combined targeted and exome calls are shown across microbiopsies for genes identified as drivers in this study and for other genes known to be drivers in bladder cancer. Redundant mutations caused by sequencing the same library by both exome and

targeted sequencing are excluded. However, redundancies caused by sequencing the same stretch of urothelium from adjacent sections are not excluded in this representation. Driver genes are sorted by the number of mutations detected in histologically-normal urothelium (above line) or by the frequency with which they are mutated in The Cancer Genome Atlas data (below line). Microbiopsies are sorted by donor type, donor age and then by the presence of mutations in successive genes in the driver gene list.

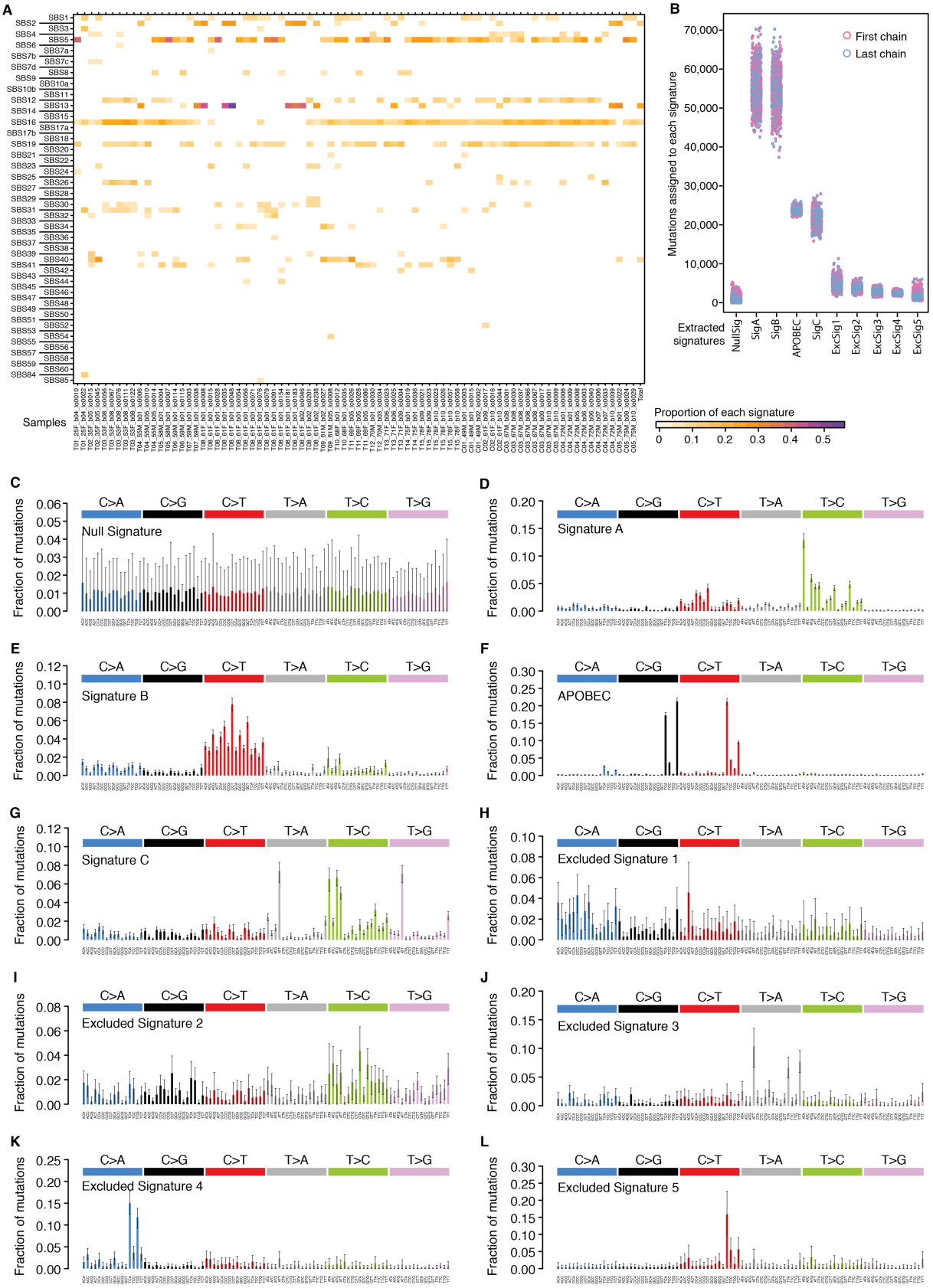
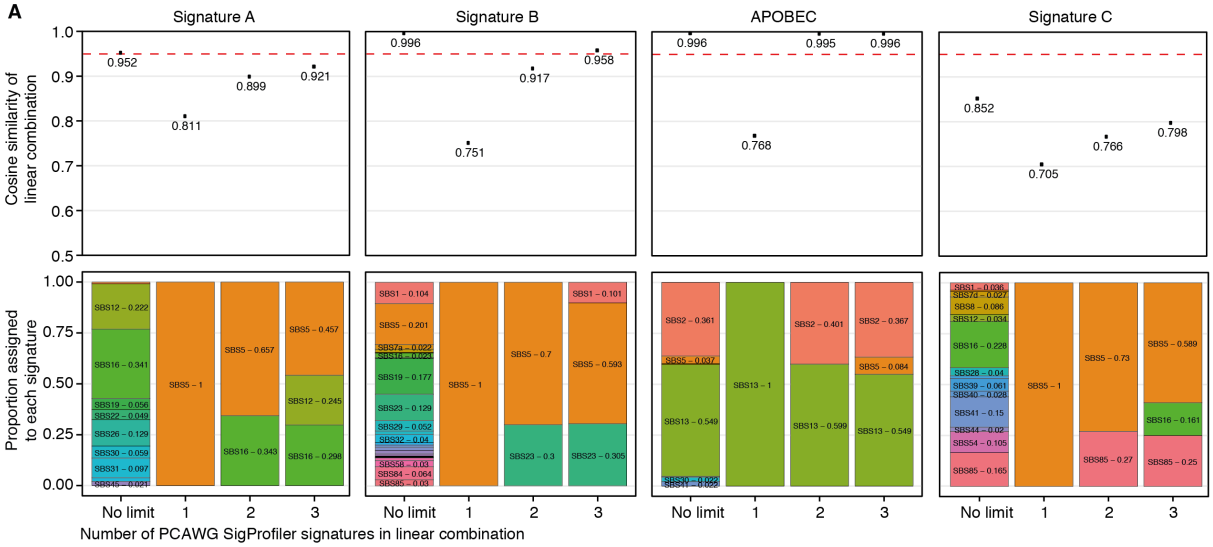
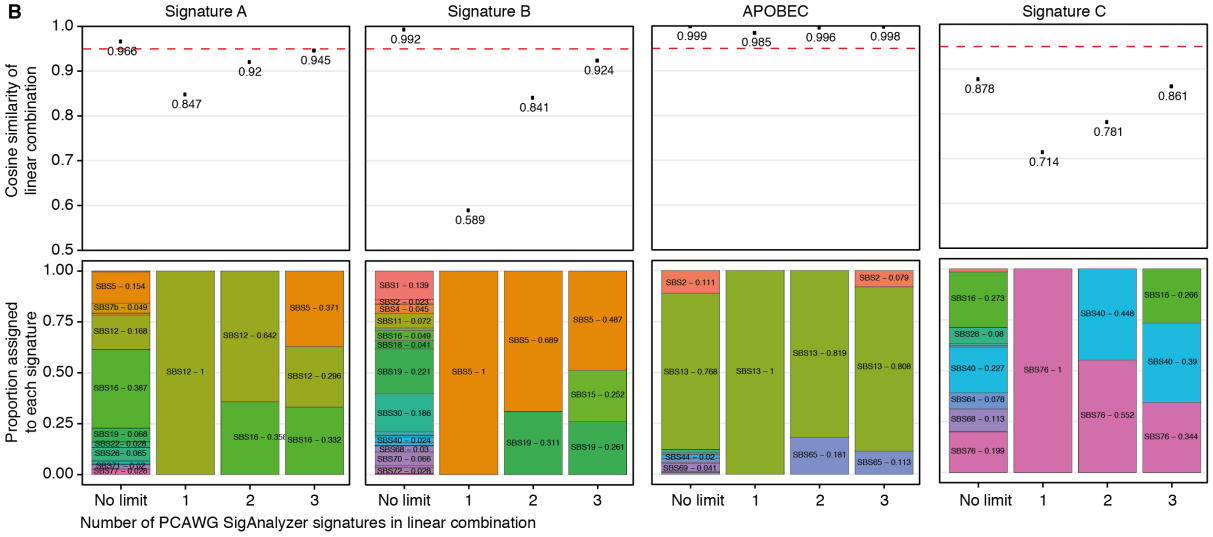
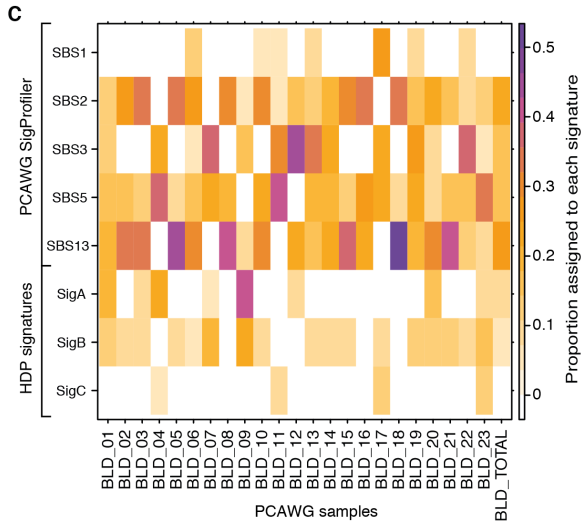

A
SBS1
SBS2
SBS3
SBS4
SBS5
SBS6
SBS7a
SBS7b
SBS7c
SBS7d
SBS8
SBS9
SBS10a
SBS10b
SBS11
SBS12
SBS13
SBS14
SBS15
SBS16
SBS17a
SBS17b
SBS18
SBS19
SBS20
SBS21
SBS22
SBS23
SBS24
SBS25
SBS26
SBS27
SBS28
SBS29
SBS30
SBS31
SBS32
SBS33
SBS34
SBS35
SBS36
SBS37
SBS38
SBS39
SBS40
SBS41
SBS42
SBS43
SBS44
SBS45
SBS46
SBS47
SBS48
SBS49
SBS50
SBS51
SBS52
SBS53
SBS54
SBS55
SBS56
SBS57
SBS58
SBS59
SBS60
SBS84
SBS85
Samples
B
First chain
Last chain
Mutations assigned to each signature
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Extracted signatures
NullSig
SigA
SigB
APOBEC
SigC
ExcSig1
ExcSig2
ExcSig3
ExcSig4
ExcSig5
Proportion of each signature
0
0.1
0.2
0.3
0.4
0.5
C
Null Signature
D
Signature A
E
Signature B
F
APOBEC
G
Signature C
H
Excluded Signature 1
I
Excluded Signature 2
J
Excluded Signature 3
K
Excluded Signature 4
L
Excluded Signature 5
C>A
C>G
C>T
T>A
T>C
T>G
Fraction of mutations

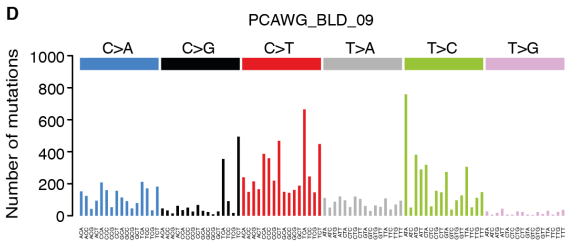
Fig. S5. Discovery of mutational signatures with HDP *de novo*.

(**A**) Heatmap depicting the linear combination of previously described cancer mutational signatures (*67*) identified as the best fit by *DeconstructSigs* (*68*) for 80 urothelium and von Brunn's nests whole-genomes (methods S6.3). (**B**) The number of mutations assigned to the null signature and the nine signatures extracted by *HDP* (methods S6.1). Each circle corresponds to a different collected iteration, colored by which of the ten chains initialized with a different seed they are derived from. (**C** to **L**) Trinucleotide mutational profiles for the null signature and the nine signatures extracted by *HDP*. Error bars depict 95% credibility intervals.


A
Signature A
Signature B
APOBEC
Signature C
Cosine similarity of linear combination
Proportion assigned to each signature
No limit
1
2
3
Number of PCAWG SigProfiler signatures in linear combination


B
Signature A
Signature B
APOBEC
Signature C
Cosine similarity of linear combination
Proportion assigned to each signature
No limit
1
2
3
Number of PCAWG SigAnalyzer signatures in linear combination


C
PCAWG SigProfiler
HDP signatures
SBS1
SBS2
SBS3
SBS5
SBS13
SigA
SigB
SigC
Proportion assigned to each signature
PCAWG samples
BLD_TOTAL


D
PCAWG_BLD_09
C>A
C>G
C>T
T>A
T>C
T>G
Number of mutations

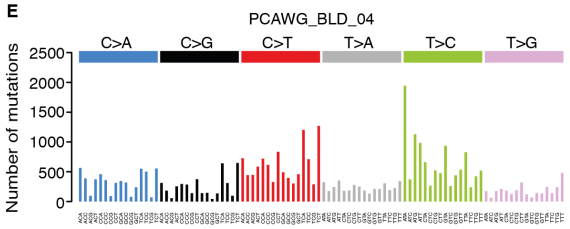

E
PCAWG_BLD_04
C>A
C>G
C>T
T>A
T>C
T>G
Number of mutations

Fig. S6. Matching extracted *de novo* mutational signatures to reference signatures.

(**A** and **B**) Results of fitting the four most abundant extracted signatures from this study to the two catalogues of mutational signatures (extracted using SigProfiler and SigAnalyzer respectively) previously identified from the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset (*67*). Signature fitting was performed using *DeconstructSigs (68)* (methods S6.3). Cosine similarity (top) and linear combination of signatures (bottom) is shown for fitting to an unlimited number, 1, 2 or 3 previously described signatures. The dashed red line corresponds to a cosine similarity of 0.95. (**C**) Heatmap depicting the linear combinations of mutational signatures observed in bladder cancer (*67*) or described here that are the best fits for 23 bladder cancer genomes from the PCAWG dataset (*67*) (methods S6.4). (**D** and **E**) Mutational spectra for the two PCAWG bladder cancer genomes identified as having the highest proportion of Signature A mutations.

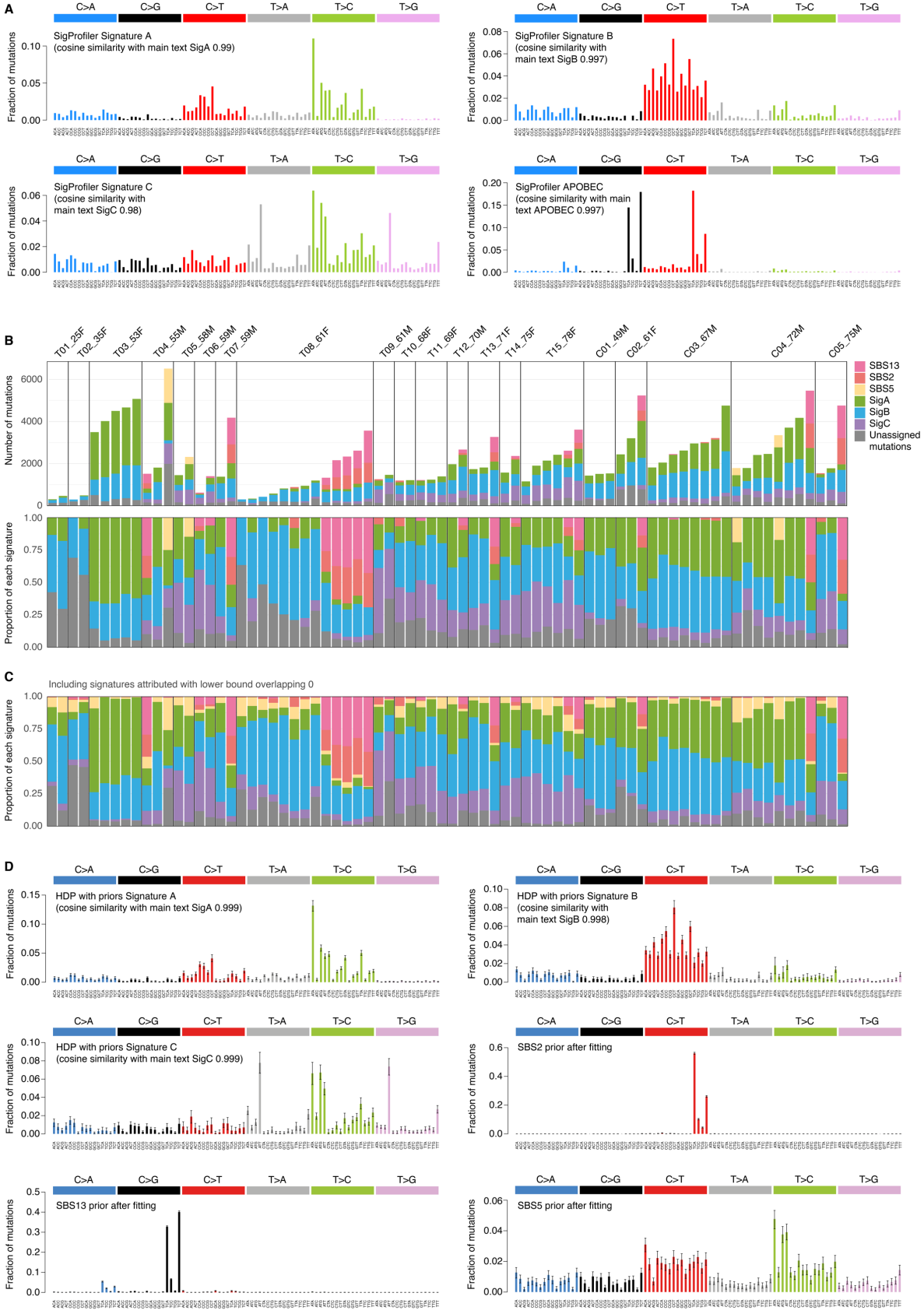
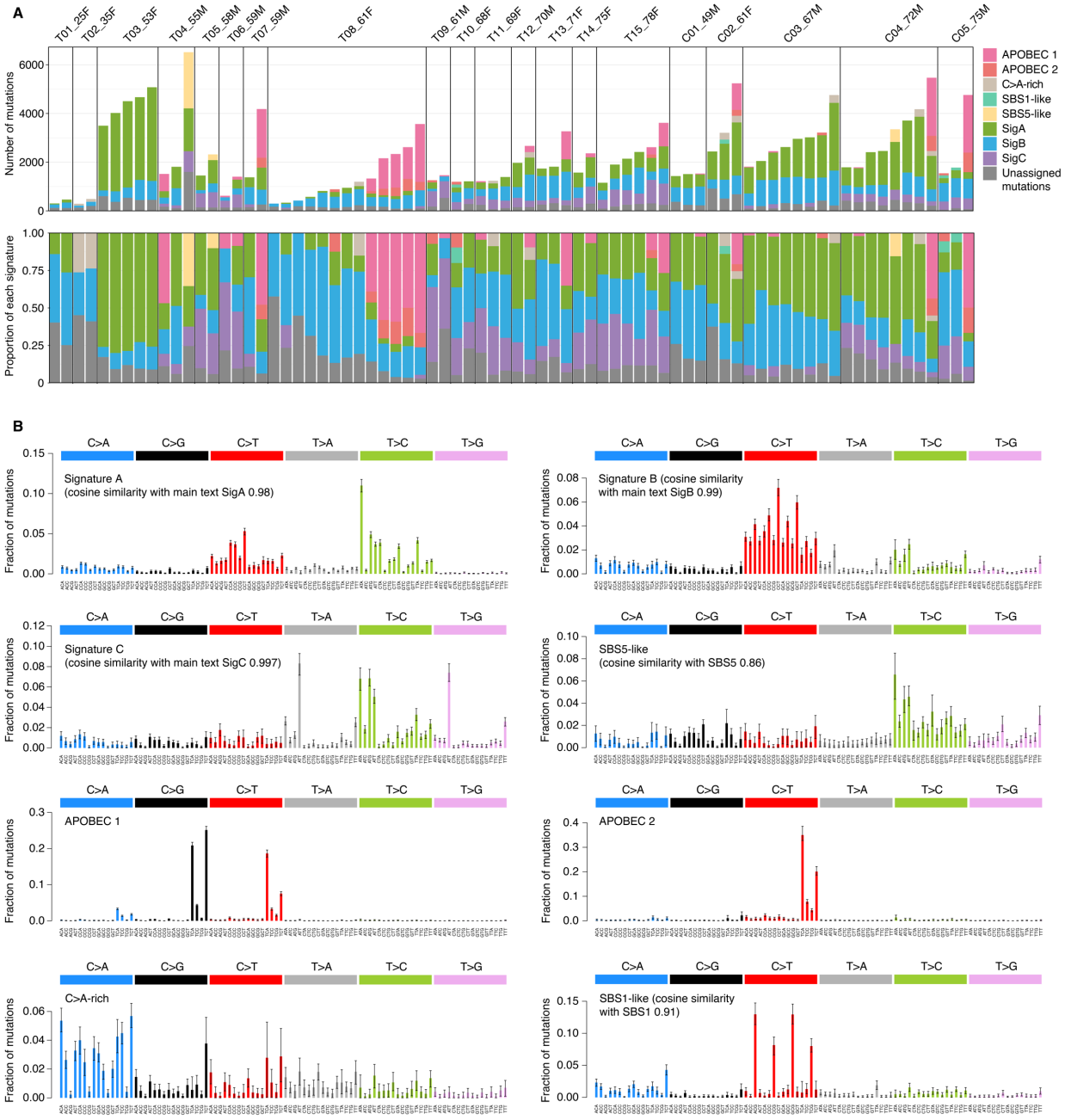

A
C>A
C>G
C>T
T>A
T>C
T>G
Fraction of mutations
SigProfiler Signature A
(cosine similarity with main text SigA 0.99)
SigProfiler Signature B
(cosine similarity with
main text SigB 0.997)
SigProfiler Signature C
(cosine similarity with
main text SigC 0.98)
SigProfiler APOBEC
(cosine similarity with main
text APOBEC 0.997)
B
T01_25F
T02_35F
T03_53F
T04_55M
T05_58M
T06_59M
T07_59M
T08_61F
T09_61M
T10_66F
T11_69F
T12_70M
T13_71F
T14_75F
T15_76F
C01_49M
C02_61F
C03_67M
C04_72M
C05_75M
Number of mutations
Proportion of each signature
SBS13
SBS2
SBS5
SigA
SigB
SigC
Unassigned
mutations
C
Including signatures attributed with lower bound overlapping 0
D
HDP with priors Signature A
(cosine similarity with main text SigA 0.999)
HDP with priors Signature B
(cosine similarity with
main text SigB 0.998)
HDP with priors Signature C
(cosine similarity with main text SigC 0.999)
SBS2 prior after fitting
SBS13 prior after fitting
SBS5 prior after fitting

Fig. S7. Alternative mutational signature extraction with SigProfiler and HDP with priors.

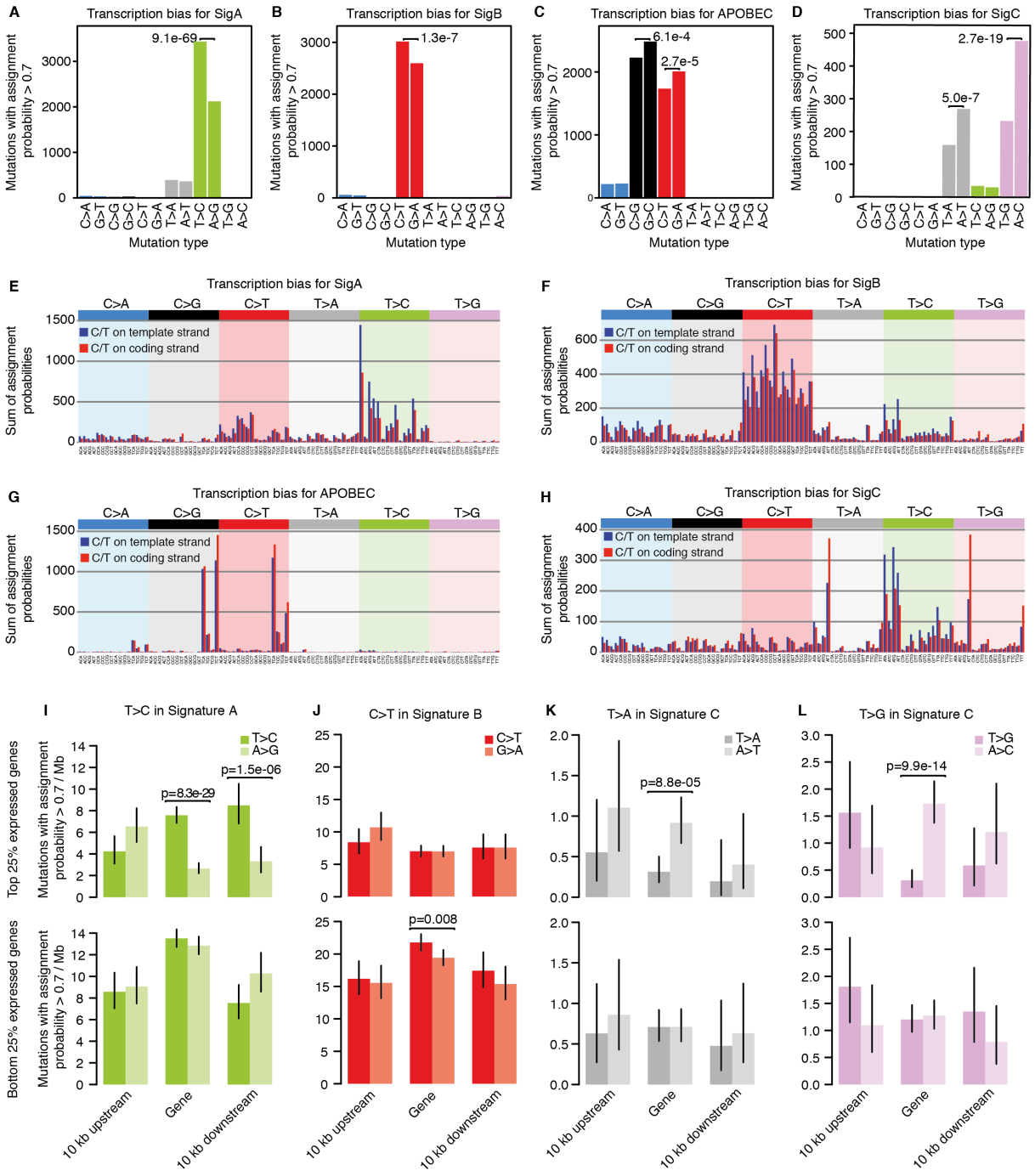
(**A**) Mutational spectra of the four signatures extracted de novo using non-negative matrix factorization (SigProfiler) (methods S6.2). (**B**) Number (top) and proportion (bottom) of mutations assigned to the four most abundant signatures extracted using a HDP with priors (methods S6.2) for urothelium and von Brunn's nest genomes from transplant donors and bladder cancer patients. Only signatures whose 95% credibility interval does not extend to zero are represented here, as in Fig. 3D. (**C**) Proportion of mutations assigned to each signature including signatures with credibility intervals extending to zero. This reveals a possible low level contribution of SBS5 across most samples. (**D**) Mutational spectra of the six signatures extracted using HDP with priors (methods S6.2).


A
T01_25F
T02_35F
T03_53F
T04_55M
T05_58M
T06_59M
T07_59M
T08_61F
T09_61M
T10_68F
T11_69F
T12_70M
T13_71F
T14_75F
T15_78F
C01_49M
C02_61F
C03_67M
C04_72M
C05_75M
Number of mutations
Proportion of each signature
APOBEC 1
APOBEC 2
C>A-rich
SBS1-like
SBS5-like
SigA
SigB
SigC
Unassigned mutations
B
C>A
C>G
C>T
T>A
T>C
T>G
Fraction of mutations
Signature A
(cosine similarity with main text SigA 0.98)
Signature B (cosine similarity
with main text SigB 0.99)
Signature C
(cosine similarity with main text SigC 0.997)
SBS5-like
(cosine similarity with SBS5 0.86)
APOBEC 1
APOBEC 2
C>A-rich
SBS1-like (cosine similarity
with SBS1 0.91)

5

Fig. S8. Alternative mutational signature extraction with HDP and bladder cancer genomes.

(**A**) Number (top) and proportion (bottom) of mutations assigned to the eight most abundant signatures extracted using a HDP *de novo* on normal urothelium and 23 bladder cancer genomes from the PCAWG consortium (methods S6.2). Only signatures whose 95% credibility interval does not extend to zero are represented here, as in Fig. 3D. (**B**) Mutational spectra of the eight signatures extracted using HDP de novo combining normal urothelium and PCAWG bladder cancer genomes (methods S6.2).


A
Transcription bias for SigA
9.1e-69
Mutations with assignment probability > 0.7
Mutation type
B
Transcription bias for SigB
1.3e-7
C
Transcription bias for APOBEC
6.1e-4
2.7e-5
D
Transcription bias for SigC
2.7e-19
5.0e-7
C>A G>T C>G G>C C>T G>A T>A A>T T>C A>G T>G A>C
E
Transcription bias for SigA
F
Transcription bias for SigB
G
Transcription bias for APOBEC
H
Transcription bias for SigC
C>A C>G C>T T>A T>C T>G
C/T on template strand
C/T on coding strand
Sum of assignment probabilities
I
T>C in Signature A
p=8.3e-29
p=1.5e-06
J
C>T in Signature B
p=0.008
K
T>A in Signature C
p=8.8e-05
L
T>G in Signature C
p=9.9e-14
Top 25% expressed genes
Bottom 25% expressed genes
Mutations with assignment probability > 0.7 / Mb
10 kb upstream
Gene
10 kb downstream

5

Fig. S9. Transcriptional strand asymmetries of mutational signatures.

(**A** to **D**) Transcriptional strand asymmetries across mutation types for the four main mutational signatures. Only mutations with an assignment probability >0.7 are included (methods S6.5). Mutation types are annotated by the template strand base. Significant results from Poisson tests adjusted for multiple testing correction using Benjamini & Hochberg's False Discovery Rate are indicated. (**E** to **H**) Sum of assignment probabilities across trinucleotide contexts, split by whether the pyrimidine is on the template or coding strand, for the four main mutational signatures. (**I** to **L**) Characterization of transcriptional strand asymmetries from the three novel signatures in highly- (above) and lowly-expressed (below) genes and their flanking sequence (methods S6.5).

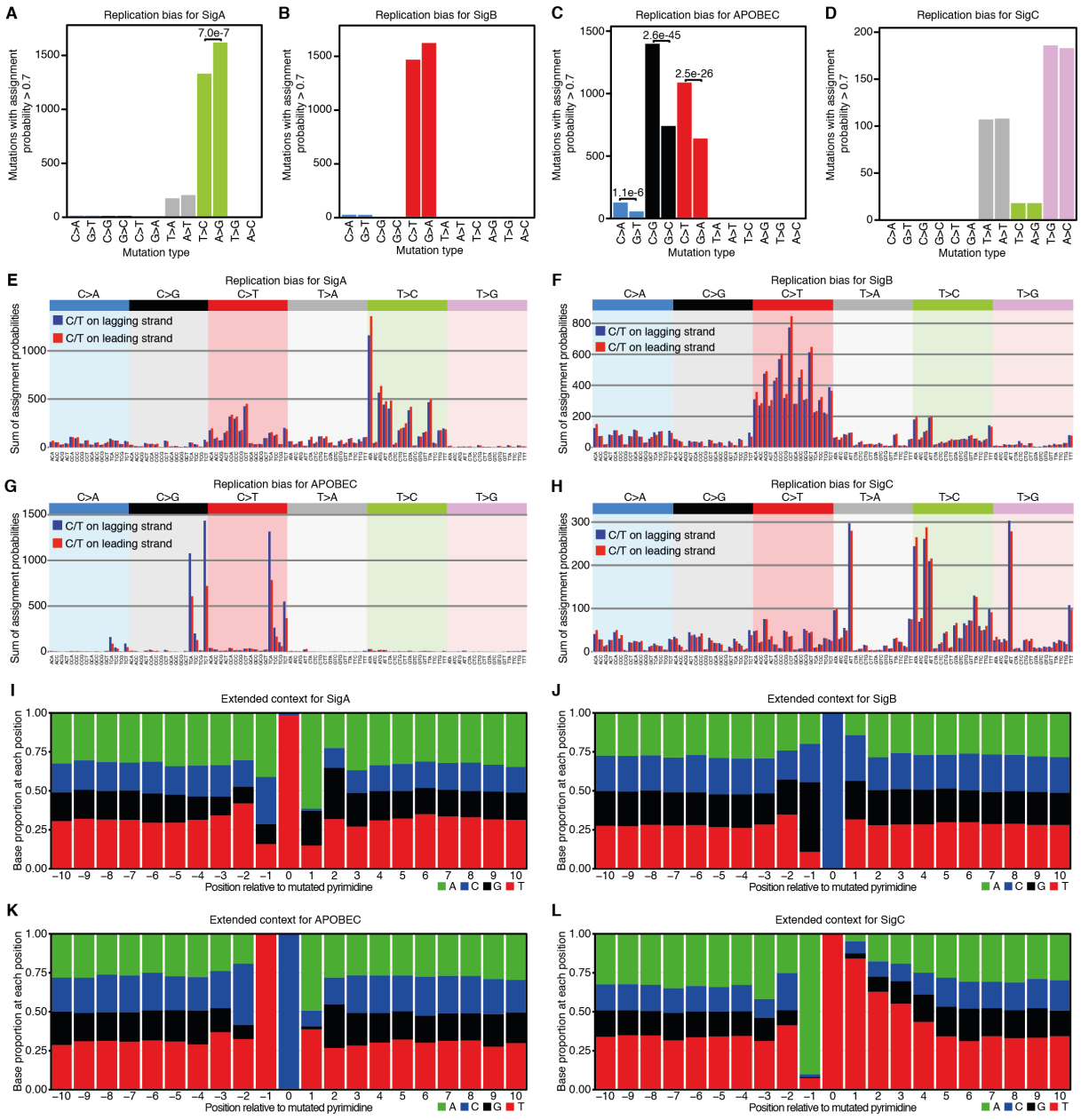


Fig. S10. Replicational strand asymmetries and extended contexts of mutational signatures. (**A** to **D**) Replicational strand asymmetries across mutation types for the four main mutational signatures. Only mutations with an assignment probability >0.7 are included (methods S6.5). Mutation types are annotated by the lagging strand base. Significant results from Poisson tests adjusted for multiple testing correction using Benjamini & Hochberg's False Discovery Rate are indicated. (**E** to **H**) Sum of assignment probabilities across trinucleotide contexts, split by whether the pyrimidine is on the lagging or leading strand, for the four main mutational signatures. (**I** to **L**) Extended nucleotide context for four main mutational signatures. The proportion of bases for the 10 nt either side of the mutated pyrimidine are shown. Only mutations with an assignment probability >0.7 are included.

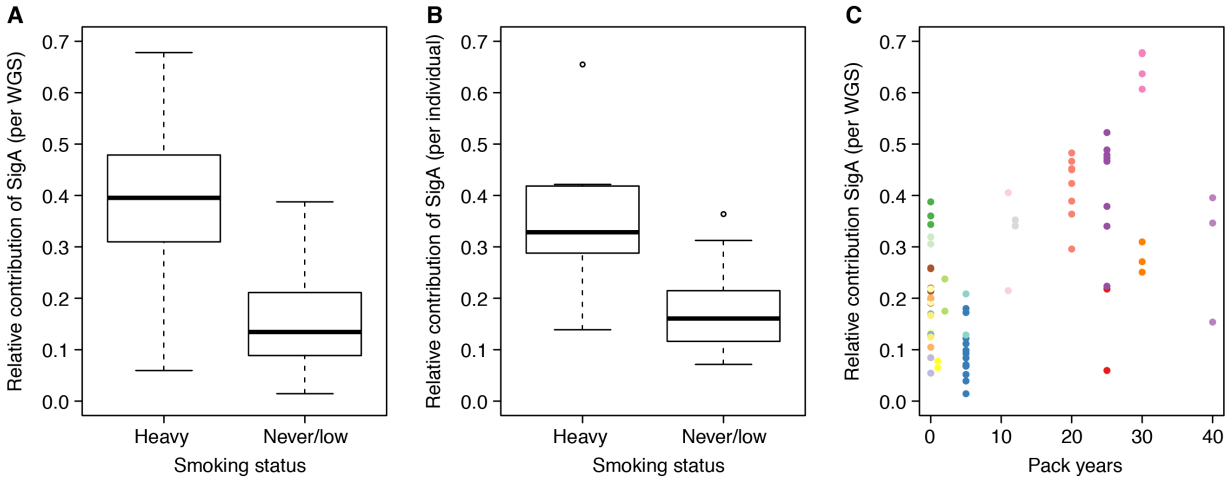


Fig. S11. Statistical association between signature A and smoking history.
(**A**) Boxplot representation of the fraction of mutations assigned to signature A per genome (with multiple genomes per individual represented as separate data points), classified according to smoking status (heavy smoker is defined as >10 pack-years, and never/low is defined as ≤5 pack-years). (**B**) Analogous boxplot representation of the mean relative contributions of signature A per individual. (**C**) Scatter plot of the relative contribution of signature A per genome as a function of pack years. Every dot represents a genome and each individual is represented in a different color. Mixed-effect regression revealed a significant association between binary smoking status and signature A (linear mixed-effect regression *P*-value=9.4e-05) and between pack years and signature A (linear mixed-effect regression *P*-value=0.0033) (methods S6.8).

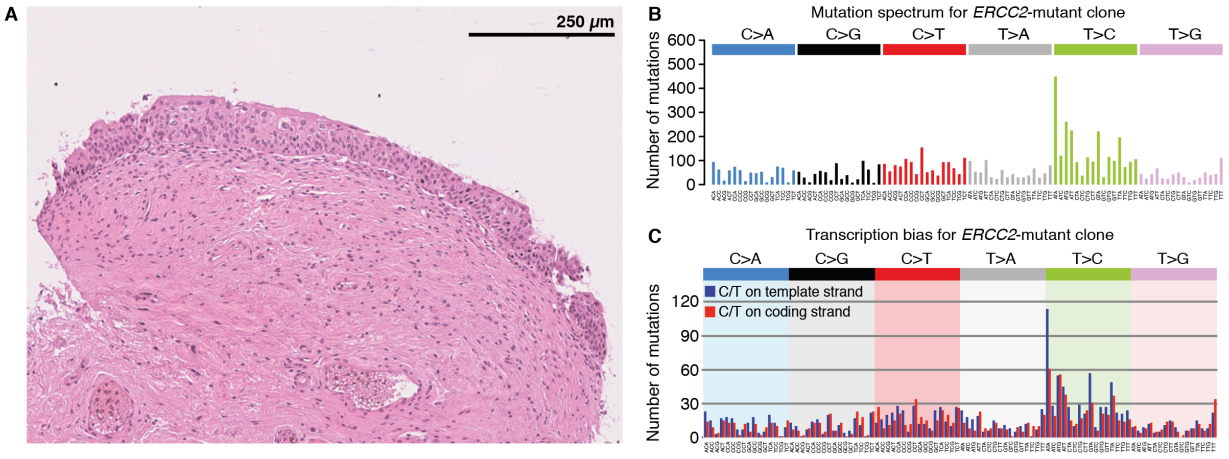


Fig. S12. Characterization of *ERCC2*-mutant clone.

(**A**) Histology image depicting stretch of urothelium containing the *ERCC2*-mutant clone identified in transplant donor T04_55M. (**B**) Mutational spectrum for the *ERCC2*-mutant clone. (**C**) Transcriptional asymmetry plot for the *ERCC2*-mutant clone.

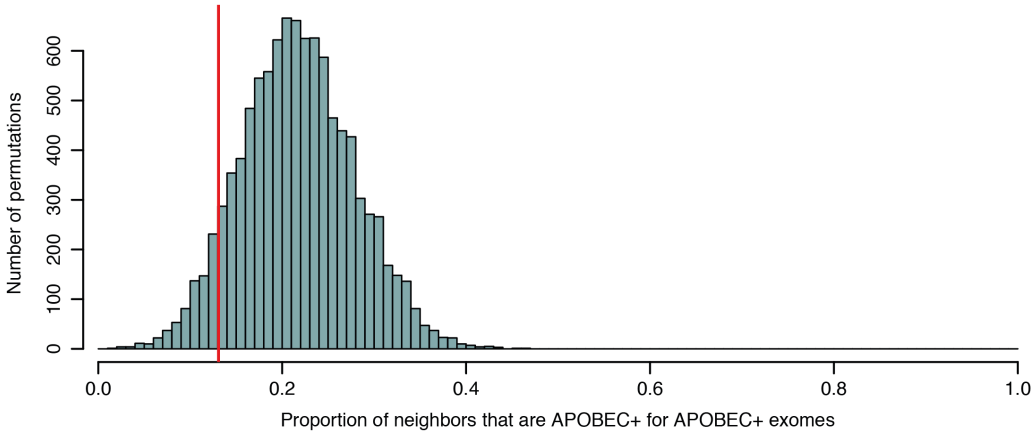


Fig. S13. Permutation test for the spatial clustering of APOBEC-positive clones.
Histogram depicting the distribution obtained from the spatial clustering of APOBEC permutation test (methods S6.7). Red line indicates the observed proportion of neighbors of APOBEC-positive exomes that were also APOBEC-positive.

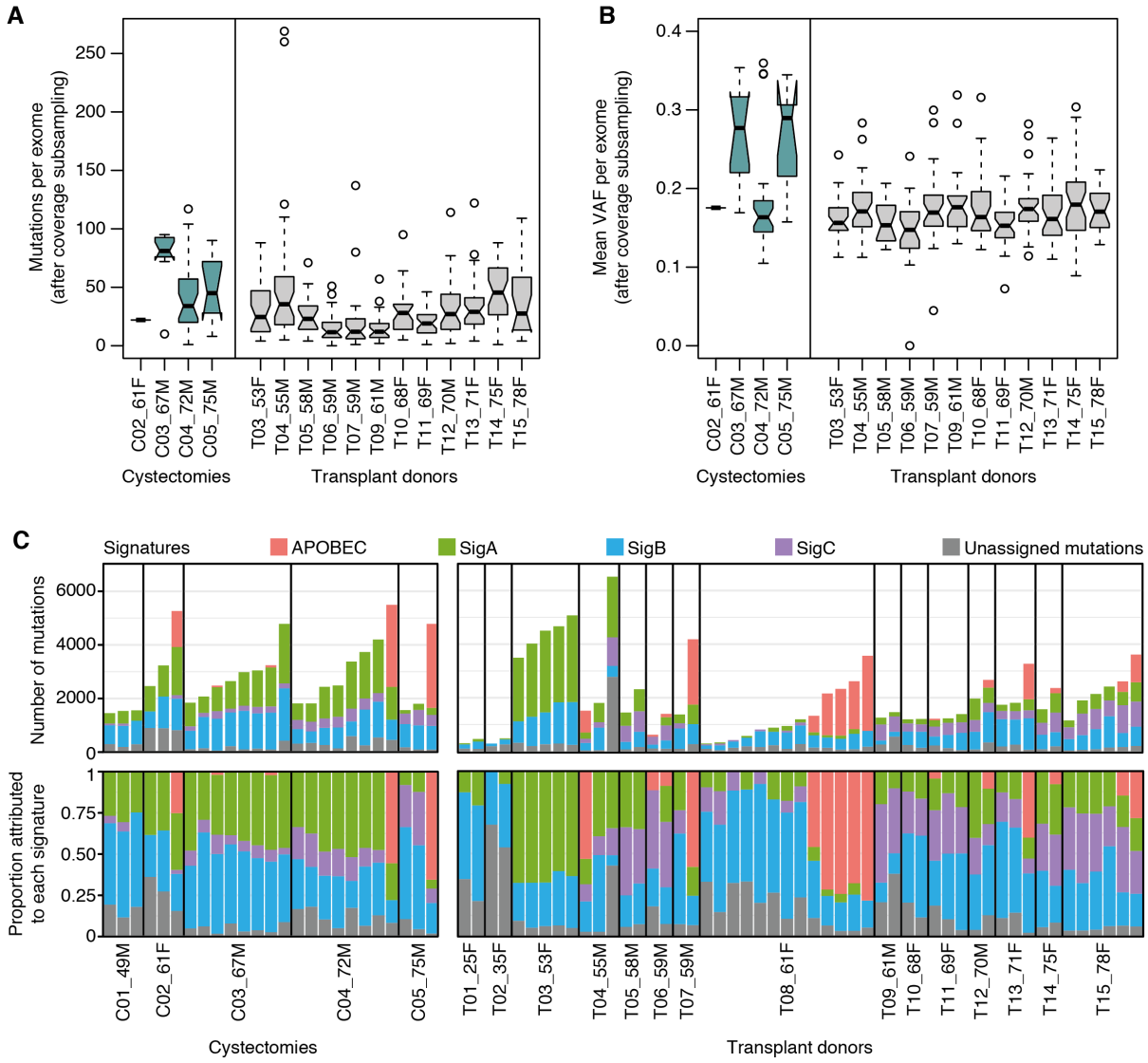


Fig. S14. Mutation burden, VAFs and signatures in bladder cancer patients.

(**A** and **B**) Box and whisker plots showing the number of mutations and the mean VAFs per exome in histologically-normal urothelium for bladder cancer patients and transplant donors. Subsampling was performed to account for differences in coverage (methods S7). (**C**) Extended version of Fig. 3D showing the number (top) and proportion (bottom) of mutations assigned to the four most abundant signatures extracted using a Bayesian hierarchical Dirichlet process (methods S6.1) for urothelium and von Brunn's nest genomes from transplant donors and bladder cancer patients.

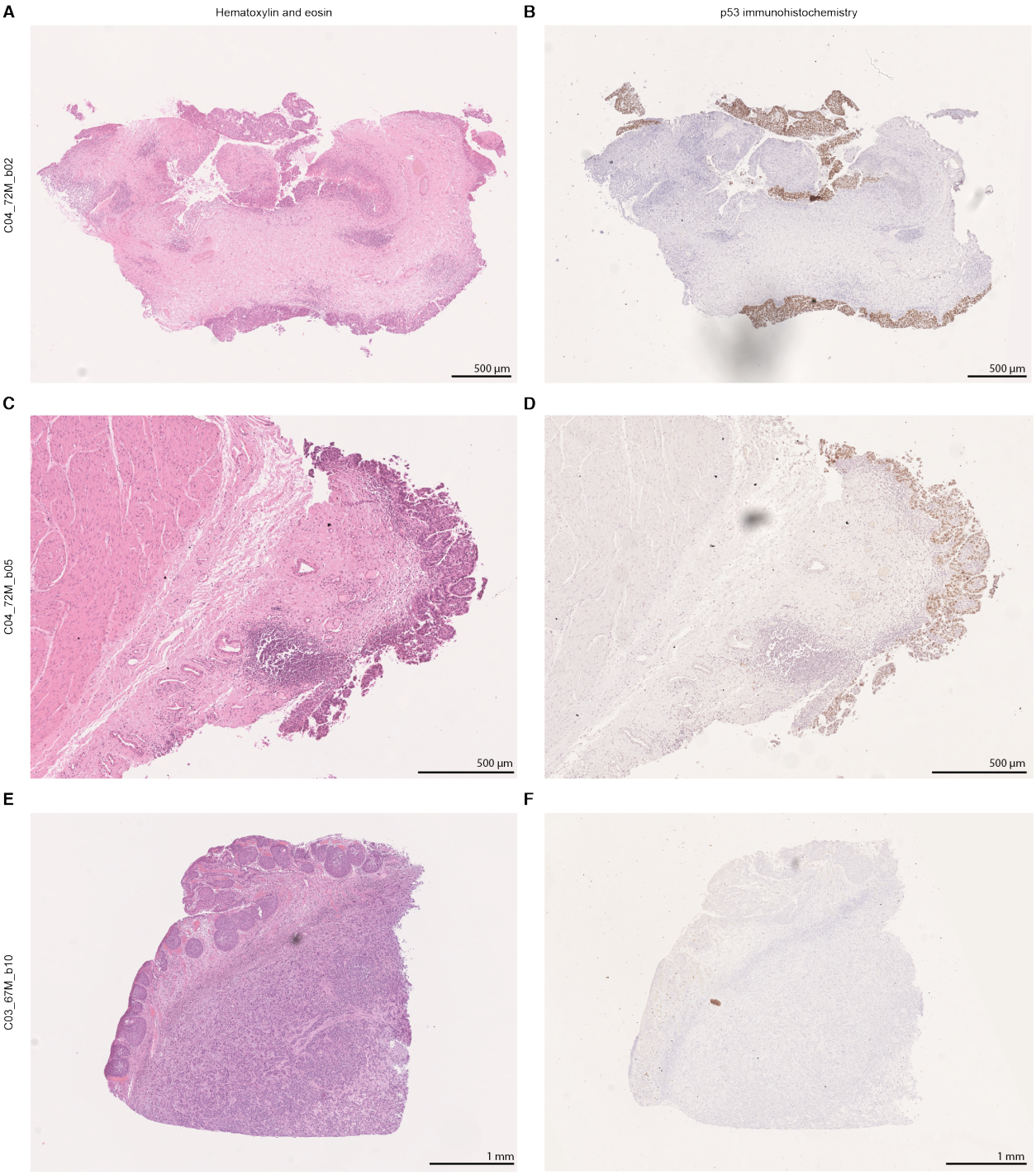


Fig. S15. Immunohistochemistry of p53.

(**A**, **C** and **E**) H&E-stained sections containing regions of carcinoma *in situ* and tumor for three biopsies taken from two bladder cancer patients (C04_72M and C03_67M). (**B**, **D** and **F**) Immunohistochemistry of p53 for matched sections. Staining is observed in the carcinoma *in situ* regions for C04_72M, due to the presence of a stabilizing *TP53* R175H mutation that is shared across the biopsies. No staining is seen in the tumor region for C03_67M, as in this case *TP53* is instead inactivated by a nonsense mutation (E343*).

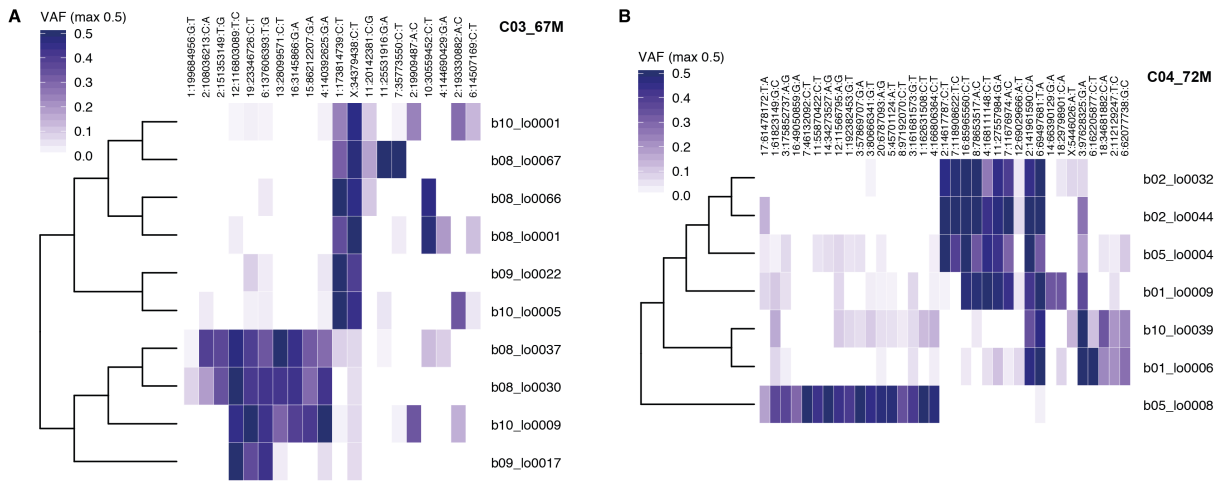


Fig. S16. Heatmaps of early embryonic mutations from cystectomy phylogenies.

(**A** and **B**) Heatmaps depicting the VAFs of early embryonic mutations identified in specimens used for phylogenetic reconstruction from two bladder cancer patients (methods S8).

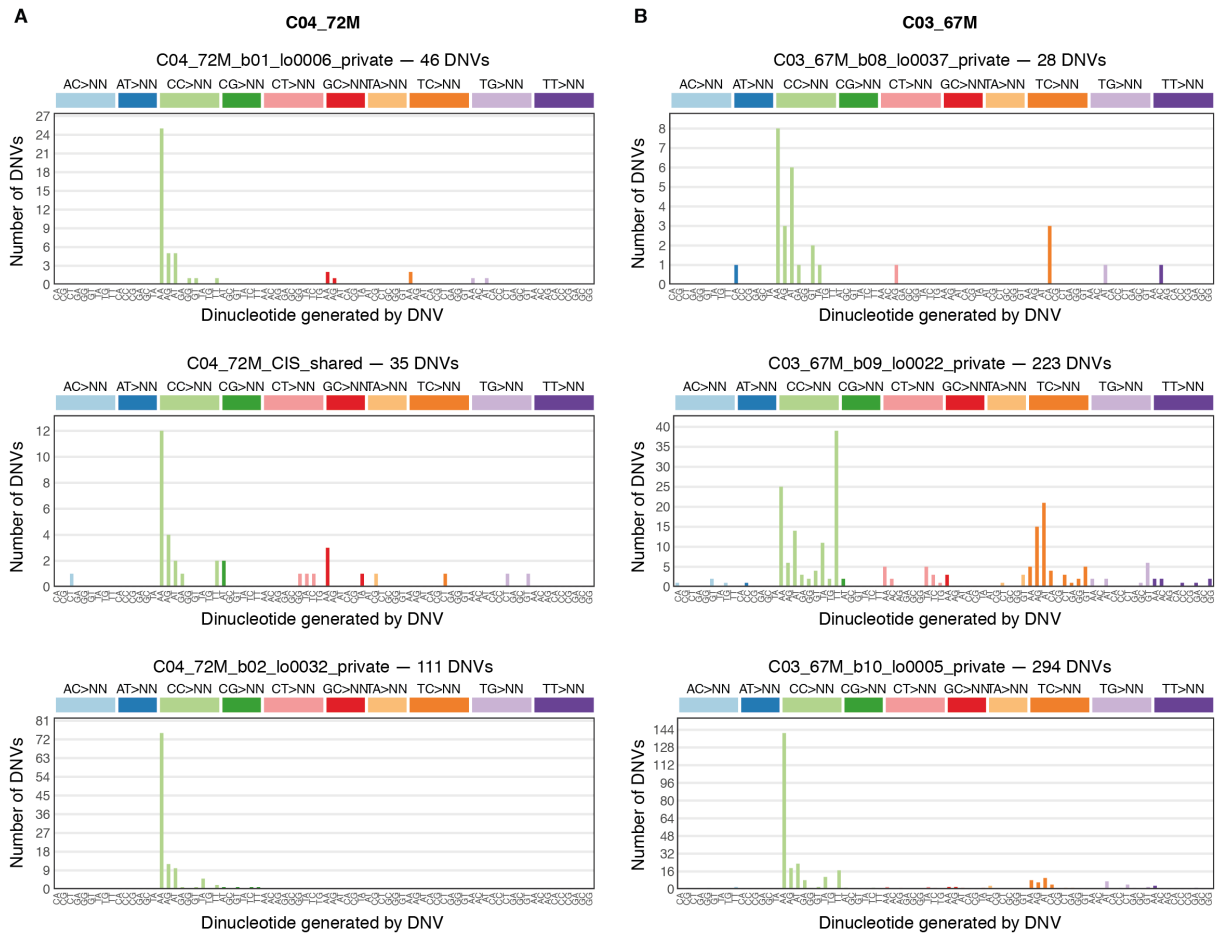


Fig. S17. Dinucleotide variants in cystectomy phylogenies.

(**A** and **B**) Profiles of dinucleotide variants (DNVs), as in (*67*), for branches of the phylogenies reconstructed from two bladder cancer patients. Of note, the C04_72M_b02_lo0032_private branch contains many more DNVs than expected compared to the number of private single nucleotide variants (SNVs). The DNV profile in this branch closely matches the previously identified DBS2 (*67*).

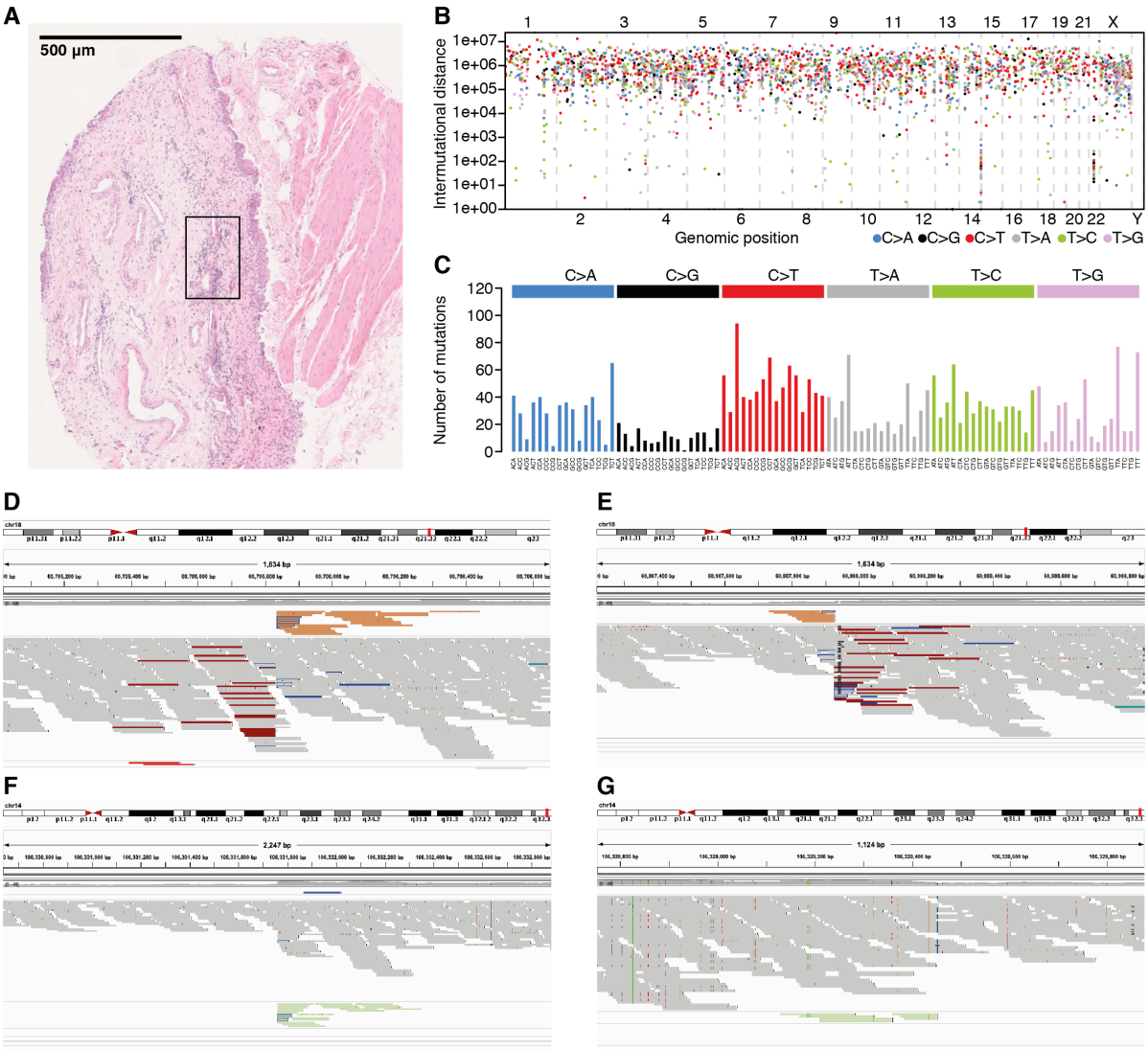


Fig. S18. Clonal lymphoid aggregate.

(**A**) Histology image depicting clonal lymphoid aggregate identified in transplant donor T11_69F. (**B**) Intermutational distance plot shows clustering of mutations at the immunoglobulin heavy locus on chromosome 14 and the immunoglobulin lambda locus on chromosome 22. (**C**) Mutation spectrum for the clonal lymphoid aggregate. The spectrum is highly similar to SBS9, the signature associated with mutations arising from replication by polymerase η as part of somatic hypermutation in lymphoid cells (*67*). (**D** to **G**) The four breakpoints associated with the *IgH*/*BCL2* translocation. Panels D and E show either side of the *BCL2* gene. The read pairs highlighted in dark red indicate that the locus has been excised from chromosome 18 and the reads in orange have their mates on chromosome 14 within the *IgH* locus. Panels F and G show reads from the *IgH* locus supporting the transposition of *BCL2*. The reads in green have their mates on chromosome 18. Transposition of *BCL2* into the *IgH* locus mediated by V(D)J recombinase has previously been described for follicular lymphoma (*73*). Images are from Integrative Genomics Viewer (IGV) (*74, 75*).

Donor ID	Gender	Age	Tobacco intake*	Alcohol intake†	BMI (kg/m^2)	Additional Information	Libraries sequenced‡		
							Targeted	Exome	WGS
T01_25F	Female	25	Smoker (5)	Rare	24	-	190 (146)	45 (44)	3 (2)
T02_35F	Female	35	Ex-smoker (1)	Rare	24	Hemangiopericytoma	187 (177)	31 (30)	3 (2)
T03_53F	Female	53	Smoker (30)	Rare	39	-	93 (89)	31 (30)	6 (5)
T04_55M	Male	55	Smoker (40)	Heavy	28	Benign testicular lump	-	39 (38)	4 (3)
T05_58M	Male	58	Smoker (12)	Heavy	22	-	-	21 (17)	3 (2)
T06_59M	Male	59	Ex-smoker (25)	Moderate	26	-	103 (98)	31 (30)	4 (2)
T07_59M	Male	59	Smoker (2)	Rare	25	Heroin user	94 (92)	33 (32)	3 (2)
T08_61F	Female	61	Smoker (5)	None	28	History of urinary tract infections (UTIs)	312 (305)	30 (28)	18 (17)
T09_61M	Male	61	Non-smoker	Rare	29	-	84 (77)	40 (39)	3 (2)
T10_68F	Female	68	Non-smoker	Rare	34	Thyroid adenoma / nodal hyperplasia	-	34 (31)	3 (2)
T11_69F	Female	69	Non-smoker	Rare	24	UTI; investigated for lymphoma	155 (143)	36 (35)	5 (3)
T12_70M	Male	70	Ex-smoker (11)	Rare	25	-	111 (87)	49 (38)	3 (2)
T13_71F	Female	71	Non-smoker	Light	19	-	123 (105)	37 (36)	4 (3)
T14_75F	Female	75	Non-smoker	Light	27	Type 2 diabetes; benign breast lump	-	40 (36)	3 (2)
T15_78F	Female	78	Non-smoker	Rare	22	-	61 (59)	21 (20)	7 (6)
C01_49M	Male	49	Ex-smoker (30)	NR	34	pT1G3 & CIS	24 (0)	40 (0)	4 (0)
C02_61F	Female	61	Non-smoker	NR	30	pT2G3. 4 cycles of chemotherapy	65 (55)	14 (13)	4 (3)
C03_67M	Male	67	Ex-smoker (20)	NR	25	pT2G3	98 (29)	42 (12)	11 (3)
C04_72M	Male	72	Ex-smoker (25)	NR	23	pT1G3 & CIS	169 (129)	40 (33)	13 (8)
C05_75M	Male	75	Non-smoker	NR	30	pT1G3 & CIS	45 (37)	21 (14)	4 (1)

Table S1. Donor information

* The number of pack-years are shown in parentheses for current and ex-smokers. † Alcohol intake classifcations are: Rare (<1 U/day); Light (1-2 U/day); Moderate (3-6 U/day); Heavy (7-9 U/day); NR – Not recorded. ‡ The number of libraries sequenced from regions of histologically normal urothelium are shown in parentheses.

Captions for Supplementary Tables

Table S2. Microbiopsy information.
TableS2_MicrobiopsyInformation.xlsx

Table S3. Substitution and indel calls from targeted data.
TableS3_TargetedSubsAndIndels.xlsx

Table S4. Substitution and indel calls from exome data.
TableS4_ExomeSubsAndIndels.xlsx

Table S5. Substitution and indel calls from genome data.
TableS5_GenomeSubsAndIndels.xlsx

Table S6. Copy number calls from exome data.
TableS6_ExomeCNVs.xlsx

Table S7. Rearrangement calls from genome data.
TableS7_RearrangementCalls.xlsx

Table S8. Driver discovery in transplant donors.
TableS8_DriverDiscoveryInTransplantDonors.xlsx

Table S9. Fraction of mutations attributed to each signature per sample.
TableS9_MutationSignatureAttribution.xlsx

Table S10. Mutational signatures extracted by HDP *de novo*.
TableS10_SignatureTrinucleotideFrequencies.xlsx

Table S11. Retrotransposition calls from genome data.
TableS11_RetrotranspositionEvents.xlsx

References and Notes:

1. J. S. Welch *et al.*, The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278 (2012).
2. S. Jaiswal *et al.*, Age-related clonal hematopoiesis associated with adverse outcomes. *The New England journal of medicine* **371**, 2488-2498 (2014).
3. I. Martincorena *et al.*, Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (New York, N.Y.)* **348**, 880-886 (2015).
4. F. Blokzijl *et al.*, Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264 (2016).
5. M. A. Lodato *et al.*, Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science (New York, N.Y.)* **359**, 555-559 (2018).
6. K. Suda *et al.*, Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell reports* **24**, 1777-1789 (2018).
7. I. Martincorena *et al.*, Somatic mutant clones colonize the human esophagus with age. *Science (New York, N.Y.)* **362**, 911-917 (2018).
8. A. Yokoyama *et al.*, Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312-317 (2019).
9. H. Lee-Six *et al.*, The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).
10. S. F. Brunner *et al.*, Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538-542 (2019).
11. L. Moore *et al.*, The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640-646 (2020).
12. K. Yizhak *et al.*, RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science (New York, N.Y.)* **364**, (2019).
13. I. Franco *et al.*, Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. *Nat Commun* **9**, 800 (2018).
14. J. D. Krimmel *et al.*, Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc Natl Acad Sci U S A* **113**, 6005-6010 (2016).
15. K. Yoshida *et al.*, Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266-272 (2020).
16. C. Wang, W. T. Ross, I. U. Mysorekar, Urothelial generation and regeneration in development, injury, and cancer. *Developmental dynamics : an official publication of the American Association of Anatomists* **246**, 336-343 (2017).
17. L. B. Alexandrov *et al.*, Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
18. A. G. Robertson *et al.*, Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171**, 540-556.e525 (2017).
19. I. Martincorena *et al.*, Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e1021 (2017).
20. S. Letasiova *et al.*, Bladder cancer, a review of the environmental risk factors. *Environmental health : a global access science source* **11 Suppl 1**, S11 (2012).
21. S. L. Poon *et al.*, Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome medicine* **7**, 38 (2015).
22. Supplementary materials and methods.

23. N. T. Gaisa *et al.*, The human urothelium consists of multiple clonal units, each maintained by a stem cell. *The Journal of pathology* **225**, 163-171 (2011).
24. Y. Gui *et al.*, Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nature genetics* **43**, 875-878 (2011).
25. S. Hernandez *et al.*, Prospective study of FGFR3 mutations as a prognostic factor in nonmuscle invasive urothelial bladder carcinomas. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **24**, 3664-3671 (2006).
26. Y. Allory *et al.*, Telomerase reverse transcriptase promoter mutations in bladder cancer: high frequency across stages, detection in urine, and lack of association with outcome. *European urology* **65**, 360-366 (2014).
27. A. Dunford *et al.*, Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nature genetics* **49**, 10-16 (2017).
28. C. D. Hurst *et al.*, Genomic Subtypes of Non-invasive Bladder Cancer with Distinct Metabolic Profile and Female Gender Bias in KDM6A Mutation Frequency. *Cancer cell* **32**, 701-715.e707 (2017).
29. L. B. Alexandrov *et al.*, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
30. Li et al. Co-submitted manuscript.
31. L. B. Alexandrov *et al.*, Mutational signatures associated with tobacco smoking in human cancer. *Science (New York, N.Y.)* **354**, 618-622 (2016).
32. J. Kim *et al.*, Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature genetics* **48**, 600-606 (2016).
33. K. Chan *et al.*, An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nature genetics* **47**, 1067-1072 (2015).
34. H. Vöhringer, M. Gerstung, Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *bioRxiv*, 850453 (2019).
35. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322 (2014).
36. B. Rodriguez-Martin *et al.*, Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nature genetics* **52**, 306-319 (2020).
37. L. Moore *et al.*, The mutational landscape of normal human endometrial epithelium. *bioRxiv*, 505685 (2018).
38. O. Acar *et al.*, Determining the origin of synchronous multifocal bladder cancer by exome sequencing. *BMC cancer* **15**, 871 (2015).
39. M. B. H. Thomsen *et al.*, Comprehensive multiregional analysis of molecular heterogeneity in bladder cancer. *Sci Rep* **7**, 11702 (2017).
40. T. Majewski *et al.*, Whole-Organ Genomic Characterization of Mucosal Field Effects Initiating Bladder Carcinogenesis. *Cell reports* **26**, 2241-2256.e2244 (2019).
41. K. Shin *et al.*, Cellular origin of bladder neoplasia and tissue dynamics of its progression to invasive carcinoma. *Nat Cell Biol* **16**, 469-478 (2014).
42. K. E. Volmar, T. Y. Chan, A. M. De Marzo, J. I. Epstein, Florid von Brunn nests mimicking urothelial carcinoma: a morphologic and immunohistochemical comparison to the nested variant of urothelial carcinoma. *The American journal of surgical pathology* **27**, 1243-1252 (2003).

43. M. Koti *et al.*, Tertiary Lymphoid Structures Associate with Tumour Stage in Urothelial Bladder Cancer. *Bladder cancer (Amsterdam, Netherlands)* **3**, 259-267 (2017).
44. S. H. Vermeulen *et al.*, Recurrent urinary tract infection and risk of bladder cancer in the Nijmegen bladder cancer study. *Br J Cancer* **112**, 594-600 (2015).
45. S. Abelson *et al.*, Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400-404 (2018).
46. P. Armitage, R. Doll, The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer* **8**, 1-12 (1954).
47. C. Tomasetti, L. Marchionni, M. A. Nowak, G. Parmigiani, B. Vogelstein, Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A* **112**, 118-123 (2015).
48. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
49. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
50. G. Tischler, S. Leonard, biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine* **9**, 13 (2014).
51. G. Jun *et al.*, Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848 (2012).
52. M. Gerstung, E. Papaemmanuil, P. J. Campbell, Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198-1204 (2014).
53. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
54. D. Jones *et al.*, cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15.10.11-15.10.18 (2016).
55. K. M. Raine *et al.*, cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15.17.11-15.17.12 (2015).
56. P. Van Loo *et al.*, Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915 (2010).
57. K. M. Raine *et al.*, ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinformatics* **56**, 15.19.11-15.19.17 (2016).
58. A. Auton *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
59. S. Nik-Zainal *et al.*, The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
60. P. J. Campbell *et al.*, Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics* **40**, 722-729 (2008).
61. M. S. Lawrence *et al.*, Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
62. M. S. Lawrence *et al.*, Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
63. J. M. Hess *et al.*, Passenger Hotspot Mutations in Cancer. *Cancer cell* **36**, 288-301.e214 (2019).

64. J. Van den Eynden, A. Jimenez-Sanchez, M. L. Miller, E. Larsson, Lack of detectable neoantigen depletion signals in the untreated cancer genome. *Nature genetics* **51**, 1741-1748 (2019).
65. L. Zapata *et al.*, Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol* **19**, 67 (2018).
66. Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei, Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* **101**, 1566--1581 (2006).
67. L. B. Alexandrov *et al.*, The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*, 322859 (2019).
68. R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, C. Swanton, DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).
69. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).
70. N. J. Haradhvala *et al.*, Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-549 (2016).
71. K. P. Schliep, phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593 (2011).
72. K. C. Nixon, The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis. *Cladistics* **15**, 407-414 (1999).
73. J. W. Vaandrager, E. Schuuring, K. Philippo, P. M. Kluin, V(D)J recombinase-mediated transposition of the BCL2 gene to the IGH locus in follicular lymphoma. *Blood* **96**, 1947-1952 (2000).
74. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).
75. H. Thorvaldsdottir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).